Mimicking Radiologists: A Coarse-to-Fine Framework with Structural Sparse Tokens for Dual-LLM Computed Tomography Report Generation

Hong Liu[2,3], Dong Wei[3], Yefeng Zheng[3,4], Xian Wu[3⋆], and Liansheng Wang[1,2⋆⋆]

[1] School of Informatics, Xiamen University, Xiamen, China
lswang@xmu.edu.cn, liuhong@stu.xmu.edu.cn
[2] National Institute for Data Science in Health and Medicine, Xiamen University
[3] Tencent Jarvis Lab, Tencent Healthcare (Shenzhen) Co., Ltd., Shenzhen, China
{donwei,yefengzheng,kevinxwu}@tencent.com
[4] Medical Artificial Intelligence Laboratory, Westlake University, Hangzhou, China
{zhengyefeng}@westlake.edu.cn

Abstract. Computed tomography report generation (CTRG) automates radiology reporting to reduce clinical workload and facilitate patient care. Recent efforts in applying the rapidly developing large language models (LLMs) have advanced the field; yet, they still face a fundamental challenge due to the large volume of 3D data: effectively reducing high feature redundancy and computational burden while simultaneously extracting information-rich representations. To address this challenge, this work presents a novel CTRG framework that fully imitates the coarse-to-fine visual search pipeline practiced by radiologists. Our framework first employs a ViT-based image-text alignment architecture to extract a global token and local patch tokens for each anatomical structure, enhanced by a mask-guided sparse negative-entropy loss. An abnormality-proposal LLM then processes the global token and clinical metadata to propose a shortlist of candidate abnormalities. This shortlist prompts an abnormality-prompted local token filter (AP-LTF) to select the most informative patch tokens, effectively reducing redundancy while preserving critical information. Finally, a report-generation LLM takes in the global token, selected local tokens, and clinical metadata to compose a full report. During training, ground-truth abnormalities are prefixed to the reference report to enhance awareness of clinic-relevant findings, and group relative policy optimization (GRPO) aligns the abnormality-proposal and report-generation LLMs for collaborative efficacy. Experimental results on two public CTRG datasets demonstrate the superior performance of our framework compared to existing state-of-the-art

⋆ Corresponding author.
⋆⋆ Corresponding author.
H. Liu and D. Wei——Contributed equally; H. Liu contributed during an internship at Tencent.

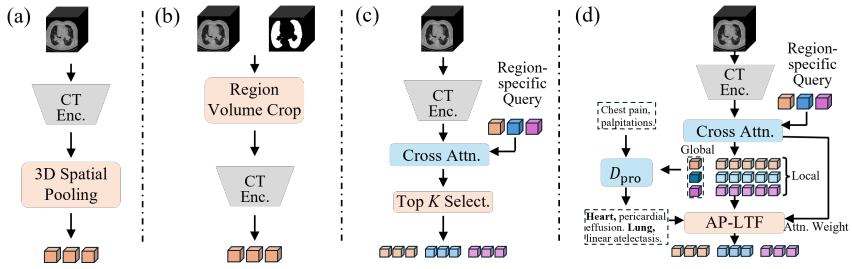


Fig. 1: Comparison of token compression strategies for CTRG. (a) 3D spatial pooling [2, 8] achieves high compression rate but incurs significant information loss. (b) Mask-guided volume cropping [12, 22] yields region-specific tokens but omits intra-region redundancy. (c) [26] selects the top K tokens with the largest attention weights to region-specific queries, yet may be subject to dominant "sink tokens" [14, 29]. (d) Our method first proposes a shortlist of candidate abnormalities using only a single global token per structure and clinical metadata. Then, an abnormality-prompted local token filter (AP-LTF) combines the shortlist, a learned score, and the attention weight for a comprehensive token selection, thereby preventing sink tokens from dominating.

methods in terms of clinical efficacy, RaTEScore, and GREEN scores. Ablation studies further validate the effectiveness of its novel design.

Keywords: CT report generation · Abnormality-prompted token filtering · Dual-LLM framework · Group relative policy optimization

1 Introduction

Radiology reports of radiographs, such as computed tomography (CT) and X-rays, are essential for clinical diagnosis and timely treatment. However, reading volumetric CT scans requires concentrated visual search across hundreds of slices and careful inspection of subtle, localized abnormalities. In practice, this process is time-consuming and error-prone—especially under staff shortages and heavy workloads, causing nontrivial rates of missed findings and delayed reporting [6].

Various works have attempted automatic 3D CT report generation (CTRG). Exploratory methods approached CTRG by connecting a volumetric encoder to a large language model (LLM) [2, 8], in which the encoder extracted CT image features, and the LLM received projected image features, generating reports using its strong language capabilities. To improve the quality of extracted CT image features, various works explored effective CT representation learning, including contrastive learning [7, 26], vital slice selection [38], and attention mechanisms [13, 18]. However, CTRG methods face a fundamental challenge: the large volumes of CT scans present significant information redundancy and detrimental lesion sparsity, while also incurring substantial computational costs.

A straightforward approach to reducing the computational costs is spatial pooling [2, 8] (Fig. 1(a)), however, with significant information loss. Recent studies extracted region-specific representations from cropped volumes based on segmentation masks [12, 22] (Fig. 1(b)). Compared with global feature-based methods, these region-centric methods are less prone to overlooking abnormalities due to failure to explore local details. Nonetheless, they do not address the redundancy in each region, e.g., the lung. Liu et al. [26] proposed selecting the top 10 patch tokens with the largest attention weights to a region-specific query (prototype) for each region, thereby substantially reducing intra-region redundancy and computational costs (Fig. 1(c)). However, recent studies on "sink tokens" [14, 29], i.e., low-semantic tokens that attract disproportionately large attention, suggest that token selection solely based on attention weights may be adverse for tasks requiring fine-grained, localized evidence, such as detailed radiology reporting.

These limitations contrast with how radiologists work. They perform a coarse-to-fine visual search: first scanning through the entire stack to flag suspicious areas—guided by clinical metadata (e.g., chief complaint, medical history) if available—and then inspecting those areas in detail, focusing on reporting abnormal findings therein. This efficient information retrieval capability is attributed to a mechanism known as visual search, which has been extensively studied in cognitive and vision sciences [34, 40].

Following the radiologists' clinical practice, we propose a novel CTRG framework that fully imitates the coarse-to-fine visual search pipeline. As the prerequisite, we employ the vision Transformer (ViT)-based [15] image-text alignment architecture in [26] for structure-wise visual token extraction, obtaining a global visual token and a pool of local patch tokens for each anatomical structure of interest. For better extraction of structure-wise tokens, we propose a mask-guided sparse negative-entropy loss that encourages each patch token to concentrate on the structure it resides in. Then, analogous to the quick but coarse abnormality-scanning process of radiologists, we feed only the global visual token, plus clinical metadata if available, into an abnormality-proposal LLM, yielding a concise, patient-specific shortlist of region-wise candidate abnormalities. Next, we input these candidate abnormalities to an abnormality-prompted local token filter (AP-LTF) to select the most informative local patch tokens from each structure's token pool (Fig. 1(d)). For each token, AP-LTF considers its similarity to the candidate abnormalities and a learned score, in addition to the original attention weight, thereby effectively preventing sink tokens from dominating. The selected patch tokens are more likely to contain potential abnormalities than those discarded, similar to the suspicious areas flagged by radiologists for further scrutiny. Meanwhile, the selection process effectively reduces intra-region redundancy and computational burden. Finally, we feed the global visual token and selected local patch tokens of all structures, along with clinical metadata, to a report-generation LLM that composes a full radiology report. To enhance the LLM's awareness of clinic-relevant abnormalities and negations, we prefix a shortlist of ground-truth abnormalities parsed from the reference report for training. In addition, we propose using group relative policy optimization (GRPO) [35] to

align the abnormality-proposal LLM with the report-generation LLM and to close their distribution gap, thereby enabling effective collaboration. The reward considers both the clinical efficacy and linguistic fidelity of the final report.

In summary, our main contribution is a novel CTRG framework that fully imitates the coarse-to-fine visual search pipeline practiced by radiologists, with the following novelties:

- We propose a mask-guided sparse negative entropy loss for more accurate structure-specific token extraction.
- We implement the coarse-to-fine paradigm with a dedicated abnormality-proposal LLM and a dedicated report-generation LLM.
- We propose AP-LTF, which considers the candidate abnormalities proposed by the first LLM, a learned score, and attention weight for comprehensive token filtering.
- We introduce a GRPO-based collaborative refinement to close the gap between the two LLMs for optimal candidate abnormality proposal and final report generation.
- We conduct extensive experiments on two public CTRG datasets, to demonstrate the improvements over state-of-the-art (SOTA) methods and validate the effectiveness of the novel components of our framework.

2 Related Work

In this section, we review the literature related to CT report generation (CTRG) from two perspectives: effective CT representation learning and efficient redundancy reduction, both of which are crucial to high-quality CTRG.

CT Representation Learning. Tang et al. [38] proposed a scan localizer network trained with disease classification supervision (using manually labeled medical terms) to select critical slices in a CT sequence for report generation. CT2Rep [18] proposed a cross-attention multimodal fusion module and hierarchical memory to incorporate longitudinal multimodal data. Chen et al. [13] introduced a disease prototype memory bank and a disease-aware attention module to better capture diagnostic cues. Li et al. [24] proposed μ^2Tokenizer, an intermediate layer that integrates multiscale visual tokens with text, and applies direct preference optimization (DPO) to improve generation quality. These approaches improved CT image representation quality, but did not differentiate the features' originating regions for more localized representation.

To capture region-specific details, [12] and [22] utilized organ segmentation masks as guidance to crop region-specific areas at the input and token levels, respectively. However, both required running a segmentation network during inference, which adds extra complexity and latency to deployment. Liu et al. [26] proposed structure-wise image-text alignment by parsing sentences describing different anatomical regions from the ground truth reports as supervision. However, the textual supervision is not as precise as the segmentation masks. Our

method introduces a mask-guided sparse negative entropy loss to the structure-wise image-text alignment architecture. This loss utilizes segmentation masks to learn more localized, region-specific representations while avoiding the additional burden of the segmentation network during deployment.

CT Volume Redundancy Reduction. In multimodal large language models (MLLMs), visual tokens are often highly redundant, inflating computational cost. Token compression methods typically fall into two classes: training-based and training-free. Training-based methods learn compact, model-specific visual encodings via retraining to realize token reduction. QueCC [23] injected query semantics into the visual grid and applied a learnable convolutional cross-attention compressor, and LLaVA-Mini [44] pre-fuses visual features into text tokens, both yielding as few as one token fed to the LLM. Training-free methods exploit heuristics and token-selection/merging strategies without retraining [37], enabling plug-and-play compression but at the risk of larger accuracy loss at high sparsity. FastV [9] noticed inefficient attention and proposed reducing image tokens after the early layers of LLMs. SparseVLM [46] relied on text-guided, training-free scoring to pick visual tokens most relevant to the prompt. VisionZip [42] proposed text-independent reduction by selecting dominant tokens and merging similar ones. MustDrop [27] inserted dedicated compression modules across encoding, prefilling, and decoding stages of MLLMs.

With the development of LLMs, researchers have sought to improve CTRG by leveraging their strong reasoning and text-generation capabilities. 3D-CT-GPT [8] and M3D [2] generated reports from CT volumes by bridging CT-specific ViTs and LLMs via a linear projection layer and multi-layer perceptrons, respectively. Three-dimensional spatial pooling was implemented to reduce the number of visual tokens and computational cost, but at the cost of significant information loss. [26] proposed selecting the top K tokens with the largest attention weights to a learned structure-specific prototype for each anatomical structure of interest. However, selecting a few tokens purely based on attention weights may cause the selected tokens to be overwhelmed by low-semantic tokens that attract disproportionately large attention, also known as "sink tokens" [14,29], and overlook detail-oriented tokens crucial for fine-grained radiology reporting. Our work first proposes a shortlist of candidate abnormalities using a single global token per structure as input. Then, it uses this list as a prompt to an abnormality-prompted local token filter (AP-LTF), which combines each token's similarity to the candidate abnormalities, a learned score, and the attention weight for comprehensive token filtering.

3 Method

Fig. 2 shows an overview of our framework. As a prerequisite, we employ the image-text alignment strategy in [26] for structure-wise visual token extraction, yielding a global, summary token $\boldsymbol{s}^v$ and a set of local patch tokens $\mathcal{G}$ for each anatomical structure of interest (Section 3.1). We enhance the extraction with

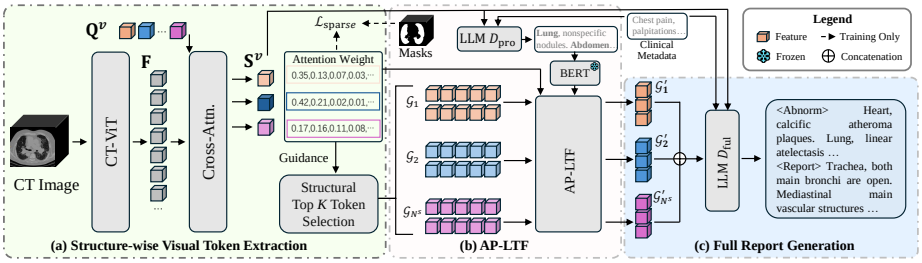


Fig. 2: Overview of our proposed framework.

a new mask-guided sparse negative entropy loss $\mathcal{L}_{\text{sparse}}$ (Section 3.2). Then, an abnormality-proposal LLM D_{pro} proposes a shortlist of candidate abnormalities, combining global visual tokens with clinical metadata as input. D_{pro} is refined with the group relative policy optimization (GRPO; Section 3.5). The shortlist serves as the prompt for the abnormality-prompted local token filter (AP-LTF; Section 3.3), which selects the most informative patch tokens for fine-grained report generation (Section 3.4). This process mimics clinical practice, in which radiologists first perform a quick initial scan of coarse, contextual information to identify relevant local image regions for detailed, complete reporting.

3.1 Preliminaries: Structure-wise Visual Token Extraction

Our framework employs the image-text alignment architecture in [26] for structure-wise visual token extraction (Fig. 2(a); detailed in Supplementary Section 7; here we provide a high-level introduction as preliminaries). Given a CT volume $\mathbf{X} \in \mathbb{R}^{D\times H\times W}$, a CT-ViT [19] is used to extract patch tokens $\mathbf{F} \in \mathbb{R}^{N^v\times L}$, where N^v is the number of raw visual tokens and L is the token dimension. Then, a set of learnable structure-specific visual queries $\mathbf{Q}^v = [\boldsymbol{q}_i^v]_{i=1}^{N^s} \in \mathbb{R}^{N^s\times L}$—with each $\boldsymbol{q}_i^v$ for an anatomical structure (e.g., mediastinum, lung) and N^s indicating the total number of structures of interest—extracts a global, summarizing token $\boldsymbol{s}_i^v$ for each structure via cross attention:

$$\mathbf{A}^v = \text{softmax}\left(\mathbf{Q}^v\mathbf{W}_0(\mathbf{F}\mathbf{W}_1)^\top\right), \quad \mathbf{S}^v = [\boldsymbol{s}_i^v]_{i=1}^{N^s} = \mathbf{A}^v(\mathbf{F}\mathbf{W}_2), \tag{1}$$

where $\mathbf{W}_0$, $\mathbf{W}_1$, and $\mathbf{W}_2$ are linear projection matrices, and $\mathbf{A}^v \in \mathbb{R}^{N^s\times N^v}$ is the pair-wise similarity matrix/attention weights between the structural queries and patch tokens. In addition to the global token $\boldsymbol{s}_i^v$, a group of K localized patch tokens $\mathcal{G}_i = \{\boldsymbol{g}_{i,j}\}_{j=1}^K \in \mathbb{R}^{K\times L}$ is also selected from $\mathbf{F}$ for the i^{th} structure. Concretely, the top K patch tokens with the largest attention weights to a specific structure visual query are selected for that structure according to $\mathbf{A}^v$. A proper K value is empirically determined to strike a balance between efficiency and performance. Finally, all extracted global and local tokens are jointly input to a language decoder (e.g., LLaMA-3.2-3B [16]) for report generation.

To learn the structure-specific visual queries $\mathbf{Q}^v$, CT reports are used for supervision. In brief, the reports are parsed into sections describing different

anatomical structures, and subsequently encoded into a set of structure-specific textual tokens $\mathbf{S}^t = [\boldsymbol{s}_i^t]_{i=1}^{N^s}$. Concretely, each sentence in the report is assigned to a structure according to a keyword-matching rule. A pretrained, frozen text encoder [17] then encodes the sentences for each structure, and the class (CLS) token of each structure's encoding is taken as $\boldsymbol{s}_i^t$. The global visual tokens $\mathbf{S}^v$ are then aligned with $\mathbf{S}^t$ by a cross-modal image-to-text contrastive loss $\mathcal{L}_{\text{so-itc}}$ over a per-structure textual memory bank. As a single one-hot target over-penalizes semantically similar negatives, we additionally align the image-to-text similarity with a text-text-similarity soft target via a KL divergence loss $\mathcal{L}_{\text{so-kl}}$. The overall pretraining objective is $\mathcal{L}_{\text{so-pre}} = 0.5\,\mathcal{L}_{\text{so-itc}} + 0.5\,\mathcal{L}_{\text{so-kl}}$, with full formulations detailed in the supplementary material.

3.2 Mask-guided Sparse Negative Entropy Loss

Although requiring no manual annotation, the image-text alignment above provides only indirect supervision about visual structures. Recent advances in prompted segmentation have enabled rapid and convenient segmentation of medical images. In this work, we propose employing the SAT model [48] to generate segmentation masks of anatomical structures, and utilizing these masks to provide direct visual guidance. SAT segments 3D medical volumes given natural-language prompts (e.g., "Kidney"); we prompt it for the 197 chest-CT classes and group them into the 10 structures of interest following [45]. Specifically, we propose a mask-guided sparse negative entropy loss:

$$\mathcal{L}_{\text{sparse}} = \frac{1}{\|\mathcal{H}\|}\sum\nolimits_{(i,j)\in\mathcal{H}} -\log\bigl(1-\mathbf{A}_{i,j}^v\bigr),\ \mathcal{H} = \{(i,j) \mid \mathbf{M}_{i,j} = 0 \wedge \mathbf{A}_{i,j}^v > \eta\}, \quad (2)$$

where $\mathbf{M} \in \{0,1\}^{N^s \times N^v}$ is the token-level structural mask aligned with the patchified tokens and η is a threshold. $\mathbf{M}_{i,j} = 1$ indicates the j^{th} token (patch) is within the i^{th} structure and 0 outside. Practically, considering the imperfect segmentation accuracy and patch-induced quantification effects, we consider a patch is in a structure if $\geq$30% of the patch overlaps with the structure's mask. In effect, $\mathcal{L}_{\text{sparse}}$ encourages each visual token to concentrate its attention on the structure it belongs to, by penalizing excessive attention elsewhere (i.e., $\mathbf{M}_{i,j} = 0$), thus facilitating learning of improved structure-specific features with fine granularity. Although we are aware that the prompt segmentation by SAT may not be highly precise, the ablation study in Table 2 confirms the effectiveness of $\mathcal{L}_{\text{sparse}}$. Note that the structure masks are only needed for training, not inference.

3.3 Abnormality-Prompted Local Token Filter (AP-LTF)

In [26], K patch tokens with the largest attention weights to the structure's visual query are selected for each structure of interest. Studies [14,29] have shown that in cross-attention and self-attention architectures, some low-semantic tokens can attract disproportionately large attention—a phenomenon termed "attention sinks". These sink tokens typically provide coarse or summary information that

may be helpful for tasks dominated by global semantics, such as classification, but can be detrimental for tasks requiring fine-grained, localized evidence, e.g., detailed radiology reporting. Therefore, merely selecting a few tokens ($K = 10$ in [26]) with the largest attention weights risks favoring attention-sink tokens and overlooking informative local patches for fine-grained CTRG.

To overcome this issue, we propose an abnormality-prompted local token filter (AP-LTF; Fig. 2(b)). Concretely, we train an LLM model with the common, chained next-token prediction objective in prior, denoted by D_{pro}, to propose a shortlist of candidate abnormalities using the global visual tokens $\mathbf{S}^v$ and clinical metadata, if available (e.g., symptoms, chief complaint, medical history). Specifically, the metadata is incorporated as part of the prompt to D_{pro}, e.g., "Hepatocellular carcinoma (HCC), post-operative control", or omitted when unavailable. The training target is parsed from the ground truth report in the format "struct1, abnormality. struct2, none. struct3, abnormality1, abnormality2. ..." On average, the list of candidate abnormalities contains 45 tokens per CT volume. Then, a BERT-based text encoder pretrained on CT volumes and reports [17] encodes the candidate abnormalities into a group of structure-specific textual query tokens $\mathbf{Q}^t = \{\boldsymbol{q}_i^t\}_{i=1}^{N^s}$, with each $\boldsymbol{q}_i^t$ for a structure. Next, each $\boldsymbol{q}_i^t$ is used by a lightweight AP-LTF (shared by all structures) to query the most informative local visual tokens from $\mathcal{G}_i$. Specifically, AP-LTF computes a combined score for each candidate token $\boldsymbol{g}_{i,j}$ by fusing three evidence sources: similarity to the query $\boldsymbol{q}_i^t$, a scalar score output by a learned linear layer, and the attention weight in $\mathbf{A}^v$:

$$s_{i,j} \;=\; \alpha_0\,\mathrm{sim}(\boldsymbol{q}_i^t, \boldsymbol{g}_{i,j}) \;+\; \alpha_1\,\mathrm{lin}\left(\,\mathrm{cat}(\boldsymbol{g}_{i,j}, \boldsymbol{q}_i^t)\right) \;+\; \alpha_2 \mathbf{A}_{i,j}^v, \tag{3}$$

where $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity function, $\mathrm{lin}(\cdot)$ is a learned linear layer that outputs a scalar, cat is concatenation operation, $\boldsymbol{g}_{i,j}$ is the j^{th} token in $\mathcal{G}_i$, and α_0, α_1, and α_2 are learnable weights.

These scores are converted to a probability distribution over the K candidates via softmax:

$$p_{i,j} = \exp(s_{i,j}) \big/ \sum\nolimits_{k=1}^{K} \exp(s_{i,k}). \tag{4}$$

Based on p, the top K' tokens are selected for each structure to form $\mathcal{G}'$, the refined collection of local visual tokens, which will be used to generate the full report. Since a discrete top K' selection is preferred, we use a straight-through estimator [20] to allow gradient flow while producing sparse selections. Specifically, the straight-through estimator performs the hard (discrete) top K' selection in the forward pass but uses a continuous surrogate for the backward pass, allowing gradients to flow through the non-differentiable operation. This simple approximation enables end-to-end training with sparse, hard selections while retaining practical gradient-based optimization.

It is worth mentioning that the AP-LTF requires the initial local token pool $\mathcal{G}_i$ to include sufficient "non-sink" tokens for effective filtering. Therefore, we increase the K value from $K = 10$ used in [26], and empirically determine the optimal values for K and K' in Fig. 4.

The AP-LTF is trained with a structural contrastive objective to align the averaged filter-through token $\overline{\boldsymbol{g}}_i = \text{avg}(\boldsymbol{g}_{i,j})$ with the corresponding ground-truth-parsed structure-specific textual token $\boldsymbol{s}_i^t$ (described in Section 3.1):

$$\mathcal{L}_{\text{filt}} = -\sum_{i=1}^{N^s} \log \frac{\exp\big(\text{sim}(\overline{\boldsymbol{g}}_i, \boldsymbol{s}_i^t)/\tau\big)}{\sum_{n=1}^{N^q} \exp\big(\text{sim}(\overline{\boldsymbol{g}}_i, \boldsymbol{s}_n^t)/\tau\big)}, \tag{5}$$

where τ is a learnable temperature parameter and $N^q = 1000$ is the size of a textual token memory bank [26].

3.4 Abnormality-Prefixed Supervised Fine-Tuning

To generate the full reports, we train another LLM, denoted by D_{ful}, supervised by the ground truth CT reports (Fig. 2(c)). The input to D_{ful} includes the global visual tokens $\mathbf{S}^v$, clinical metadata $\mathbf{T}^{\text{cli}}$ (if available), and the AP-LTF filtered local visual tokens $\{\mathcal{G}_i'\}$. During supervised fine-tuning, we use the abnormalities parsed from ground truth reports as prompts to AP-LTF, ensuring that the training of D_{ful} is not affected by the performance of D_{pro}. To boost D_{ful}'s awareness of clinically relevant findings, we also prefix each ground truth report with the ground-truth-parsed abnormalities to form the generation target (cf. Fig. 2(c)). Note that the prefix is omitted for performance evaluation. Thus, D_{ful} is trained with the loss:

$$\mathcal{L}_{\text{rg}} = -\sum \log P(\boldsymbol{t}_m | \boldsymbol{t}_{1:m-1}, \mathbf{S}^v, \{\mathcal{G}'\}, \mathbf{T}^{\text{cli}}), \tag{6}$$

where P is the conditioning probability of predicting the next token $\boldsymbol{t}_m$ in the abnormality-prefixed target report.

3.5 GRPO-based Collaborative Refinement

For the full report generation, a distributional gap exists between training and inference (abnormalities parsed from ground truth reports versus those proposed by D_{pro} based on global visual tokens). To close this gap, we freeze D_{ful} and AP-LTF and treat D_{pro} as a proposal policy π_θ to be optimized so that its proposals, when collaborating with the trained D_{ful} and AP-LTF, yield high-quality reports. The generated report receives a composite reward:

$$r = \lambda \cdot \text{CE-F1} + (1 - \lambda) \cdot \text{BLEU-4}, \tag{7}$$

where $\lambda \in [0, 1]$ is a hyperparameter, CE-F1 is the clinical-efficacy F1 score measuring clinical correctness of predicted findings, and BLEU-4 measures linguistic quality. We optimize D_{pro} using the group relative policy optimization (GRPO) [35]. Specifically, for each CT volume $\mathbf{X}$, GRPO samples a group of outputs $\{o_1, o_2, ..., o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and then optimize the policy model π_θ by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}} = \mathbb{E}_{\{o_l\} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{X})} \frac{1}{G} \sum_{l=1}^{G} \Big[\min\big(R_l a_l, \text{clip}(R_l, 1-\epsilon, 1+\epsilon)\, a_l\big) - \beta\, \mathbb{D}_{\text{KL}} \Big], \tag{8}$$

where $R_l = \frac{\pi_\theta(o_l|\mathbf{X})}{\pi_{\theta_{\text{old}}}(o_l|\mathbf{X})}$, a_l is the advantage computed using the group of rewards $\{r_1, r_2, ..., r_G\}$ corresponding to the outputs in each group:

$$a_l = \big(r_l - \text{mean}(r_1, \ldots, r_G)\big) \big/ \text{std}(\{r_1, \ldots, r_G\}), \tag{9}$$

clip() is a clipping function:

$$\text{clip}(R_l, 1-\epsilon, 1+\epsilon) = \min\big(\max(R_l, 1-\epsilon), 1+\epsilon\big), \tag{10}$$

$\mathbb{D}_{\text{KL}}$ is a Kullback–Leibler term preventing catastrophic forgetting, and $\epsilon, \beta \geq 0$ are hyperparameters. We set $\epsilon = 0.2$, $\beta = 0.04$ following the original setting [35].

3.6 Training Pipeline

The training of our framework proceeds in four stages. First, we train the abnormality-proposal LLM D_{pro} with the chained next-token prediction objective (similar to Eqn. (6)) for the candidate abnormality proposal. Then, we train the CT-ViT encoder, structure-specific visual queries $\mathbf{Q}^v$, and AP-LTF (with D_{pro} frozen) with the combined loss:

$$\mathcal{L}_{\text{com}} = \mathcal{L}_{\text{so-pre}} + \mathcal{L}_{\text{sparse}} + \mathcal{L}_{\text{filt}}. \tag{11}$$

Next, we freeze the above components and train the full report generator D_{ful} with $\mathcal{L}_{\text{rg}}$ in Eqn. (6). Finally, D_{pro} goes through GRPO-based collaborative refinement by optimizing $\mathcal{J}_{\text{GRPO}}$ in Eqn. (8).

4 Experiments

Datasets and Evaluation Metrics. Following [26], we evaluate our framework on two public CTRG datasets. **1) CT-RATE [17]** contains 25,692 non-contrast chest CT volumes with reports from 21,304 unique patients. Corresponding clinical metadata is made available for a subset of 10,355 patients. All volumes are resampled to a uniform voxel spacing of $0.75\times0.75\times1.5$ mm^3 and center-cropped or zero-padded to a fixed resolution of $480\times480\times240$ voxels. We use the official split: the training set comprises 24,128 volumes (20,000 patients) and the test set comprises 1,564 volumes (1,304 patients). From the training set, we split a validation set of 300 volumes exclusively for model optimization (e.g., hyperparameters). The test set is reserved for final evaluation and comparison with prior works. **2) CTRG-Chest-548K [38]** comprises 1,804 chest CT volumes with corresponding reports but no clinical metadata. Following [26], we split the data into training/validation/test sets according to the ratio of 60/20/20, and resize each volume to $256\times256\times64$ voxels. For each case, we use the CT image and the Findings section of the English report.

Following common practice in the literature (e.g., [13, 21, 30]), we employ both natural language generation (NLG) and clinical efficacy (CE) metrics for evaluation. The former include BLEU [32], METEOR [3], and ROUGE-L [25],

Table 1: Performance comparison in CE and NLG metrics, on the test sets of CT-RATE [17] (left) and CTRG-Chest-548K [38] (right). *: cited from the original paper. Best results in **bold**, second-best underlined.

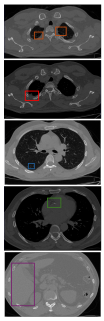
CT-RATE [17]							
	CE Metrics			**NLG Metrics**			
Method	**Pre.**	**Rec.**	**F1**	**BL-1**	**BL-4**	**MTR**	**RG-L**
R2Gen [11]	0.158	0.057	0.066	0.418	0.228	0.414	0.327
R2GenCMN [10]	0.192	0.103	0.098	0.488	0.285	0.448	0.366
M2KT [43]	0.302	0.149	0.172	0.496	0.285	0.454	0.366
R2GenGPT [39]	0.304	0.216	0.217	0.433	0.242	0.399	0.323
PromptMRG [21]	0.364	0.297	0.299	0.486	0.214	0.438	0.389
3D-CT-GPT [8]	0.242	0.153	0.166	0.459	0.253	0.422	0.361
CT2Rep [18]	0.234	0.121	0.141	0.461	0.278	0.425	**0.436**
M3D [2]	0.407	0.090	0.148	0.436	0.245	0.400	0.326
Reg2RG* [12]	0.423	0.181	0.253	0.473	0.249	0.441	0.367
Liu et al. [26]	0.337	0.366	0.315	**0.540**	0.289	0.472	0.385
Ours	**0.467**	**0.429**	**0.415**	0.535	**0.305**	**0.481**	0.396

CTRG-Chest-548K [38]							
	CE Metrics			**NLG Metrics**			
Method	**Pre.**	**Rec.**	**F1**	**BL-1**	**BL-4**	**MTR**	**RG-L**
R2Gen [11]	0.211	0.124	0.146	0.441	0.292	0.468	0.495
R2GenCMN [10]	0.152	0.108	0.115	0.471	0.299	0.482	0.491
M2KT [43]	0.227	0.114	0.146	0.493	0.314	0.497	0.490
R2GenGPT [39]	0.273	0.130	0.175	0.421	0.286	0.473	0.493
PromptMRG [21]	0.311	0.347	0.301	0.477	0.283	0.485	0.494
3D-CT-GPT [8]	0.133	0.145	0.140	0.482	0.274	0.459	0.468
CT2Rep [18]	0.223	0.135	0.125	0.452	0.297	0.489	0.491
M3D [2]	0.399	0.140	0.137	0.475	0.311	0.489	0.491
Reg2RG* [12]	-	-	-	0.496	0.320	0.497	0.477
Liu et al. [26]	0.434	0.413	0.393	**0.545**	0.326	0.509	0.497
Ours	**0.457**	**0.420**	**0.399**	0.532	**0.335**	**0.512**	**0.504**

and the latter include precision, recall, and F1 score. For CT-RATE, we use the official text classifier to extract 18 distinct types of abnormalities, enabling the calculation of CE metrics. For CTRG-Chest-548K, following [26], we utilize a pretrained report labeler, CheXbert [36], to extract labels. Additionally, following the recent trend [1, 24], we also employ two domain-specific metrics designed for radiology report evaluation: the RaTEScore based on medical named entity recognition (NER) [47] and the LLM-based GREEN [31], for a more comprehensive evaluation.

Implementation. The framework is implemented in PyTorch (2.0.0) [33]. We fine-tune Llama-3.2-3B [16] with LoRA (rank $r = 128$ and scaling factor $\alpha = 16$; 0.68% parameters fine-tuned) for the two LLMs. All three training stages use the AdamW [28] optimizer with a learning rate of 10^{-4}. Training proceeds for 100k, 30k, and 10k iterations with batch sizes of 8, 8, and 1, respectively. We apply a 10% warmup followed by a linear learning rate schedule. Following [26, 45], the anatomical structures of interest include 10 regions: lung, trachea and bronchi, mediastinum and heart, esophagus, pleura, bone, thyroid, breast, abdomen, and an "others" category (e.g., thoracic cavity, prostate), which appear only in a few reports. For the ground-truth shortlist of abnormalities, we use the region-grounded abnormalities provided in [45] for the CT-RATE dataset, and employ the open-source LLM Qwen-3-8B [41] to extract this information for the CTRG-Chest-548K dataset. We set $\eta = 0.001$ and $\lambda = 0.5$ based on preliminary experiments. Our implementation is available at: https://github.com/ccarliu/CTRG.

Comparison with SOTA Methods. We compare our framework with various SOTA methods, including those proposed for chest X-rays: R2Gen [11], R2GenCMN [10], M2KT [43], R2GenGPT [39], PromptMRG [21]; and specialized CTRG methods: 3D-CT-GPT [8], CT2Rep [18], M3D [2], Reg2RG [12], and [26]. Unless otherwise specified, we use the codes released by the authors. For methods originally designed for 2D images, we replace their image encoders with the same encoder as ours to extract CT volume representations. We optimize the performance of all compared methods on the validation data. Ta-



Ground-Truth: 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**. Pleuroparenchymal sequelae changes were observed **in both lung apex**. Millimetric nonspecific nodules were observed **in both lungs**. There is no mass or infiltrative lesion in both lungs. Heart contour and size are normal. The widths of the mediastinal main vascular structures are normal. There are atheromatous plaques **in the aorta and coronary arteries**. There are no pathologically enlarged lymph nodes in the mediastinum and hilar regions. No pathological wall thickness increase was observed in the esophagus within the sections. There is a decrease in the parenchymal density of the liver in line with the adiposity. No upper abdominal free fluid-collection was detected in the sections. Vertebral corpus heights, alignments and densities within the sections are normal.

Liu et al. **:** The trachea was in the midline of both main bronchi and no obstructive pathology was detected in the lumen. Calcified atheroma plaques were observed **in the thoracic aorta and coronary arteries**. Heart contour size is normal. Pericardial effusion-thickening was not observed. Sliding type hiatal hernia was observed at the lower end of the esophagus. When examined in the lung parenchyma window; Passive atelectatic changes were observed in the right lung middle lobe medial and left lung inferior lingular segment. Millimetric nonspecific parenchymal nodules was observed **in the middle lobe of the right lung**. Apart from this no mass lesion-active infiltration with distinguishable borders was detected in both lungs. Upper abdominal organs included in the sections are normal. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Vertebral corpus heights were preserved.

Ours: Trachea and main bronchi are open. No pathological lymph node was detected in the mediastinum. The heart is in natural appearance. Calcific atheroma plaques were observed **in the main vascular structures and coronary arteries**. Esophagus is within normal limits. Pleural effusion-thickening was not detected in both hemithorax. In the evaluation of both lung parenchyma; Sequelae changes were observed **in the apex of both lungs.** There are millimetric nonspecific nodules **in both lungs**. Emphysematous changes were observed **in both lungs**. In the sections passing through the upper part of the abdomen, the bilateral adrenal glands appear natural. No significant pathology was detected in the abdominal sections. No obvious pathology was detected in bone structures.

Fig. 3: Ground truth and example reports generated by the previous best-performing method [26] and our method for a case in CT-RATE [17]. Colored texts and color- and order-matching boxes in CT slices indicate clinic-relevant findings in the ground truth, underline indicates hallucinated sentences. CT slices are displayed with varying window levels to highlight lesions. Compared with [26], our method generates more comprehensive and clinically accurate reports, with fewer hallucinations (false positives) and omissions (false negatives).

ble 1 compares the performance in CE and NLG metrics. Across both datasets, R2Gen, R2GenCMN, and M2KT achieve reasonable NLG scores but generally poor CE metrics. R2GenGPT improves CE metrics by leveraging an LLM, yet shows notably lower NLG scores, which we conjecture is due to its global alignment strategies overlooking critical CT subregions and thus missing fine-grained features. PromptMRG yields substantial gains in CE metrics by incorporating diagnostic classification outputs. Among models explicitly tailored for CTRG, 3D-CT-GPT, CT2Rep, and M3D perform poorly in CE metrics—likely a consequence of their straightforward image representation strategies. In contrast, Reg2RG and [26] obtain more balanced CE and NLG metrics, and generally better CE metrics, likely due to their extraction of organ- and structure-specific image representations, respectively. Finally, our proposed method achieves the best CE metrics among all methods on both datasets, as well as competent NLG metrics. Notably, our method substantially boosts the previous SOTA F1 score by 32 % (0.315 to 0.415), recall by 17% (0.366 to 0.429), and precision by 10% (0.423 to 0.467) on the CT-RATE dataset. It is also among the top two models in all NLG metrics on CT-RATE. These results demonstrate the strong capability and generalizability of our method for CTRG.

Fig. 3 shows examples of generated CT reports for a case in the CT-RATE dataset. Compared with [26], our method generates more comprehensive and clinically accurate reports, with fewer hallucinations (false positives) and omissions (false negatives). More examples are provided in supplementary material.

Additional Metrics. Table 3 compares the performance of our approach with the top-performing CTRG method listed in Table 1 (Liu et al. [26]), using two radiology-specific evaluation metrics: RaTEScore [47] and GREEN [31]. We also include a recent CTRG method, μ^2tokenizer [24], that employs a DPO optimiza-

Table 2: Ablation study on the CT-RATE dataset [17]. Best results in **bold**, second-best underlined.

Ablation config.	CE Metrics ↑			NLG Metrics ↑			
	Pre.	Rec.	F1	BL-1	BL-4	MTR	RG-L
(a) Baseline	0.439	0.343	0.356	0.379	0.199	0.317	0.327
(b) + $\mathcal{L}_{\text{sparse}}$	0.464	0.367	0.377	0.375	0.192	0.317	0.322
(c) + AP-SFT	**0.468**	0.397	0.394	0.524	0.282	0.477	0.389
(d) + AP-LTF	0.461	0.418	0.391	0.514	0.289	0.447	0.377
(e) + GRPO (**Full**)	0.467	**0.429**	**0.415**	**0.535**	**0.305**	**0.481**	**0.396**
(f) Single-LLM variant	0.451	0.390	0.381	0.424	0.302	0.393	0.366

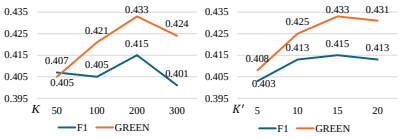


Fig. 4: Performance curves of F1 and GREEN scores with respect to varying values of K (left; with K' fixed to 15) and K' (right; with K fixed to 200).

tion strategy specifically tailored to the GREEN score. Our method achieves the highest RaTEScore and GREEN score, confirming its superiority over existing SOTA approaches across diverse metrics.

Ablation Studies. We conduct an ablation study on CT-RATE [17] (Table 2) to validate the efficacy of our framework's novel components, by incrementally adding each component to build the complete model. The baseline (a) directly feeds $K = 10$ high-attention local patch tokens along with the global visual tokens of all structures to the full report-generation LLM D_{ful} (Section 3.1), and yields the worst CE metrics alongside poor NLG metrics. Row (b) adds the mask-guided sparse negative entropy loss $\mathcal{L}_{\text{sparse}}$ (Section 3.2) and improves all three CE metrics, indicating the positive effect of concentrating the patch tokens' attention despite the faults in text-prompted segmentation masks. Row (c) further incorporates the abnormality-prefixed supervised fine-tuning (AP-SFT; Section 3.4), resulting in consistent improvements in the CE and NLG metrics. These results validate that the prefixed ground-truth-parsed abnormalities help boost the report-generation model's awareness of clinically relevant findings. Row (d) introduces the abnormality-prompted local token filter (AP-LTF), which first increases K to 200 and then filters K' most informative local patch tokens prompted by the candidate abnormalities (Section 3.3). However, this configuration deteriorates five of seven total metrics. We conjecture this is because during training, D_{ful} receives abnormalities parsed from ground-truth reports, whereas during inference it receives candidate abnormalities predicted by D_{pro}, thereby resulting in a training-inference distributional gap. To close this gap, we propose a GRPO-based refinement to align D_{pro}'s predictions with D_{ful}'s expectation (Section 3.5). Thus, our full model (e) with the GRPO refinement achieves the best performance for six of the seven metrics, and the second-best for the remaining one.

Table 2(f) shows that using a single LLM in our framework notably worsens all metrics. This could be due to limited model capacity (3B) or insufficient training samples for effective multitask training, warranting future exploration.

In the supplementary material, Section 9 conducts further ablation studies and validates 1) the design of the combined scoring function (Eqn. (3)), and 2) the robustness to variations in the candidate abnormality proposal and anatomical structure masks. Additionally, Fig.6 provides qualitative ablation analysis.

Table 3: Performance comparison on CT-RATE [17] using medical RaTEScore and GREEN score. *: cited from [24].

Method	Year	RaTEScore [47]	GREEN [31]
Liu et al. [26]	2025	0.569	0.376
μ^2tokenizer* [24]	2025	-	0.429
Ours	2025	**0.679**	**0.433**

Table 4: Performance comparison on the CT-RATE dataset [17] of different rewards for the GRPO-based refinement.

Metric	Reward			
	BLEU-4	F1	GREEN	BLEU-4 + F1 (Ours)
BLEU-4	0.303	0.301	0.298	**0.305**
F1	0.401	0.408	0.407	**0.415**
GREEN	0.415	0.423	**0.433**	**0.433**

Table 5: Inference complexity analysis, with one NVIDIA V100 GPU of 32G RAM.

Method	LLM	Param. (B)	Mem. (G)	FLOPs (G)	Latency (s)
(a) Single-LLM variant	Llama-3.2-3B	3.6	20.4	7548	87
(b) Ours w/o AP-LTF	Llama-3.2-3B×2	6.8	26.6	7546	87
(c) Liu et al. [26]	Llama-2-7B	7.1	25.8	7616	89
(d) Ours	Llama-3.2-3B×2	6.8	26.6	7548	87

Cascade Error Analysis. As AP-LTF is conditioned on D_{pro}'s shortlist, a false negative (FN) of D_{pro} may propagate to D_{ful}. On CT-RATE, D_{pro}'s FN rate drops from 60.0% to 57.9% after the GRPO refinement, indicating effective alignment. More importantly, the cascade is not strictly unidirectional: among findings absent from D_{pro}'s shortlist, only 72.8% remain missing in the final report, i.e., D_{ful} recovers 27.2% of D_{pro}'s omissions. This recovery stems from the fact that even when D_{pro} produces a false negative, AP-LTF still retrieves informative visual tokens from the relevant anatomical region to feed D_{ful}, driven by the attention weight $\mathbf{A}^v$ of the combined score in Eqn. (3).

Optimal Number of Local Patch Tokens. Figure 4 plots the performance curves of F1 and GREEN scores with respect to the varying values of two hyperparameters K and K'. Specifically, K is the number of local patch tokens initially extracted for a specific structure. These tokens form the pool for filtering by our proposed AP-LTF. On the one hand, if K is too small, there are not sufficient non-sink tokens for effective filtering by AP-LTF. On the other hand, if K is too large, excessive noise may be introduced, which can adversely impact performance. K' is the number of local patch tokens eventually selected by AP-LTF for the final report generation. We vary the value of one of them while keeping the other fixed. The trends are similar for K and K': as the hyperparameter values increase, both scores initially rise, then peak (at $K = 200$ and $K' = 15$), and subsequently decrease. Therefore, we select $K = 200$ and $K' = 15$ for our framework. Notably, with $K' = 15$, a total of 160 visual tokens (1 global plus 15 local tokens per structure; multiplied by 10 structures of interest) are fed to the report generation LLM D_{ful}, realizing a 3.9% compression rate compared with the 4,096 tokens initially yielded by the CT-ViT encoder.

Reward Design. Our composite reward considers both clinical efficacy (F1 score) and language quality (BLEU-4) for the GRPO-based refinement (Eqn. (7)). Table 4 compares the performance of alternative rewards: BLEU-4, F1 score, and the GREEN score [31]. As we can see, our composite reward achieves

the best performance for the corresponding metrics, indicating a well-balanced combination of clinical efficacy and language quality.

Complexity Analysis. Above all, the bulk token compression enables fine-tuning the LLMs (Llama-3.2-3B) with LoRA on our hardware (two NVIDIA V100 GPUs; maximum ∼800 visual tokens allowed), serving as a practical prerequisite for the inference complexity analysis (Table 5). Using a single multitask LLM reduces parameters by 47% compared with two LLMs, while maintaining the same FLOPs and latency (Table 5(a)). Yet, our proposed dual, specialized LLMs perform notably better across all metrics (cf. Table 2(e) and (f)). AP-LTF has little impact on complexity (Table 5(b)). The image-text alignment, mask-guided sparse negative-entropy loss, and GRPO refinement only affect training, thus do not add inference complexity. Previous SOTA [26] used an LLM about twice the size of our single LLM, with comparable complexities to ours (Table 5(c)). To conclude, our dual-LLM design (Table 5(d)) retains comparable latency with the single-LLM variant, adds marginal latency with its new modules (e.g., AP-LTF), and has comparable complexity to previous SOTA while achieving new SOTA performance (Table 1). It requires only two V100s (32G) for training and one for inference, and is highly scalable thanks to effective token compression and low extra modular complexity.

5 Conclusion

This paper presented a novel CTRG framework that effectively addressed the challenges of information redundancy and critical feature extraction in volumetric CT scans by fully imitating the radiologists' coarse-to-fine visual search pipeline. Extensive experiments on two publicly available datasets demonstrated the superior performance and clinical efficacy of the proposed framework.

Limitations & Future Work. This work focused on structure-level representations for efficient and effective 3D CTRG. In practice, however, many clinically relevant findings are at smaller scales, e.g., nodules in the lung. In the future, we plan to further improve our framework by considering lesion-level representations [4, 5] for CTRG with finer granularity. Our framework creatively tackles the high information redundancy & computational demand of CT volumes. It has great potential for application to similar 3D clinical imaging data, e.g., MRI, which we plan to evaluate in the future.

6 Acknowledgments

This work was supported by the Ministry of Science and Technology of the People's Republic of China (STI2030-Major Projects 2021ZD0201900), the National Natural Science Foundation of China (Grant No. 62371409).

References

1. Baharoon, M., Ma, J., Fang, C., Toma, A., Wang, B.: Exploring the design space of 3D mllms for CT report generation. In: MICCAI. pp. 237–246. Springer (2025)

2. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3D: Advancing 3D medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
3. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
4. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: MAIRA-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
5. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: ECCV. pp. 1–21. Springer (2022)
6. Brady, A., Laoide, R.Ó., McCarthy, P., McDermott, R.: Discrepancy and error in radiology: concepts, causes and consequences. The Ulster Medical Journal **81**(1), 3 (2012)
7. Cao, W., Zhang, J., Xia, Y., Mok, T.C., Li, Z., Ye, X., Lu, L., Zheng, J., Tang, Y., Zhang, L.: Bootstrapping chest CT image understanding by distilling knowledge from X-ray expert models. In: CVPR. pp. 11238–11247 (2024)
8. Chen, H., Zhao, W., Li, Y., Zhong, T., Wang, Y., Shang, Y., Guo, L., Han, J., Liu, T., Liu, J., et al.: 3D-CT-GPT: Generating 3D radiology reports through integration of large vision-language models. arXiv preprint arXiv:2409.19330 (2024)
9. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: ECCV. pp. 19–35. Springer (2024)
10. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) ACL. pp. 5904–5914. Association for Computational Linguistics, Online (Aug 2021)
11. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: EMNLP. pp. 1439–1449. Association for Computational Linguistics (2020)
12. Chen, Z., Bie, Y., Jin, H., Chen, H.: Large language model with region-guided referring and grounding for CT report generation. IEEE Trans. Med. Imaging (2025)
13. Chen, Z., Luo, L., Bie, Y., Chen, H.: Dia-LLaMA: Towards large language model-driven CT report generation. In: MICCAI. pp. 141–151. Springer (2025)
14. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations
15. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
17. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Wittmann, B., Simsar, E., Simsar, M., et al.: A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. CoRR (2024)
18. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated radiology report generation for 3D medical imaging. In: MICCAI. pp. 476–486. Springer (2024)

19. Hamamci, I.E., Er, S., Simsar, E., Tezcan, A., Simsek, A.G., Almas, F., Esirgun, S.N., Reynaud, H., Pati, S., Bluethgen, C., et al.: GenerateCT: Text-guided 3D chest CT generation. CoRR (2023)
20. Hinton, G., Srivastava, N., Swersky, K.: Lecture 6A overview of mini–batch gradient descent. Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online (2012)
21. Jin, H., Che, H., Lin, Y., Chen, H.: PromptMRG: Diagnosis-driven prompts for medical report generation. In: AAAI. vol. 38, pp. 2607–2615 (2024)
22. Kyung, S., Seo, J., Lim, H., Kim, D., Park, H., Sung, J., Kim, J., Jo, W., Nam, Y., Kim, N.: MedRegion-CT: Region-focused multimodal LLM for comprehensive 3D CT report generation. arXiv preprint arXiv:2506.23102 (2025)
23. Li, K.Y., Goyal, S., Semedo, J.D., Zico Kolter, J.: Inference optimal VLMs need only one visual token but larger models. arXiv e-prints pp. arXiv–2411 (2024)
24. Li, S., Qin, P., Wu, H., Nie, D., Thirunavukarasu, A.J., Yu, J., Zhang, L.: μ^2Tokenizer: Differentiable multi-scale multi-modal tokenizer for radiology report generation. In: MICCAI. pp. 3–12. Springer (2025)
25. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
26. Liu, H., Wei, D., Peng, Q., Huang, Y., Wu, X., Zheng, Y., Wang, L.: Structure observation driven image-text contrastive learning for computed tomography report generation. In: International Conference on Information Processing in Medical Imaging. pp. 218–233. Springer (2025)
27. Liu, T., Shi, L., Hong, R., Hu, Y., Yin, Q., Zhang, L.: Multi-stage vision token dropping: Towards efficient multimodal large language model. arXiv preprint arXiv:2411.10803 (2024)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
29. Luo, J., Fan, W.C., Wang, L., He, X., Rahman, T., Abolmaesumi, P., Sigal, L.: To sink or not to sink: Visual information pathways in large vision-language models. arXiv preprint arXiv:2510.08510 (2025)
30. Nicolson, A., Dowling, J., Koopman, B.: Improving chest X-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023)
31. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: GREEN: Generative radiology report evaluation and error notation. arXiv preprint arXiv:2405.03595 (2024)
32. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
34. Peelen, M.V., Kastner, S.: A neural basis for real-world visual search in human occipitotemporal cortex. Proceedings of the National Academy of Sciences **108**(29), 12125–12130 (2011)
35. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
36. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., Lungren, M.: CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) EMNLP. pp. 1500–1519. Association for Computational Linguistics, Online (2020)

37. Tan, X., Ye, P., Tu, C., Cao, J., Yang, Y., Zhang, L., Zhou, D., Chen, T.: TokenCarve: Information-preserving visual token compression in multimodal large language models. arXiv preprint arXiv:2503.10501 (2025)
38. Tang, Y., Yang, H., Zhang, L., Yuan, Y.: Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. Expert Systems with Applications **237**, 121442 (2024)
39. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2GenGPT: Radiology report generation with frozen LLMs. Meta-Radiology **1**(3), 100033 (2023)
40. Wolfe, J.M.: Visual search: How do we find what we are looking for? Annual review of vision science **6**(1), 539–562 (2020)
41. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
42. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: VisionZip: Longer is better but not necessary in vision language models. In: CVPR. pp. 19792–19802 (2025)
43. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis **86**, 102798 (2023)
44. Zhang, S., Fang, Q., Yang, Z., Feng, Y.: LLaVA-mini: Efficient image and video large multimodal models with one vision token. arXiv preprint arXiv:2501.03895 (2025)
45. Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., Xie, W.: RadGenome-Chest CT: A grounded vision-language dataset for chest CT analysis. arXiv preprint arXiv:2404.16754 (2024)
46. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., et al.: SparseVLM: Visual token sparsification for efficient vision-language model inference. arXiv preprint arXiv:2410.04417 (2024)
47. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: RaTEScore: A metric for radiology report generation. arXiv preprint arXiv:2406.16845 (2024)
48. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: One model to rule them all: Towards universal segmentation for medical images with text prompts. arXiv preprint arXiv:2312.17183 (2023)