M3D-QAdapter: 3D Medical VQA with Lesion-Level Finding-Segmentation Alignment and Query-Driven Adaptive Token Reduction

Hong Liu[2,3], Dong Wei[3], Hongze Zhu[4], Yinghao Zhang[5], 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
{zhuhongze,zhengyefeng}@westlake.edu.cn
[5] Monash Medical AI, Faculty of Information Technology, Monash University, Melbourne, VIC, Australia
yinghao.zhang1@monash.edu

Abstract. Volumetric medical visual question answering (Med-VQA) for 3D imaging, such as CT, holds great potential for clinical diagnosis support. Existing approaches based on multimodal large language models often employ predefined token compression to reduce computational cost without considering question context, and none have investigated the impact of fine-grained, lesion-level vision-language alignment on 3D Med-VQA. In this work, we propose M3D-QAdapter, a novel framework with lesion-level finding-segmentation alignment and query-driven adaptive token reduction. To overcome the issues of incomplete annotations and extreme negative-positive imbalance for finding-segmentation alignment, a newly proposed *safe negative mining* strategy heuristically identifies likely unannotated findings and excludes them from loss computation, thereby preventing the suppression of unannotated lesions while reducing negative samples. To adaptively compress visual tokens with query awareness, our framework dynamically *plans* which anatomical regions to scrutinize for a given question. It then *acts* to invoke two *tools* in order: 1) a text-prompted segmentation tool to obtain 3D masks of the planned regions, and 2) a hybrid geometry-semantics farthest point sampling tool to select a compact set of representative tokens from the masked regions. Extensive experiments on the 3D-RAD benchmark demonstrate that M3D-QAdapter achieves state-of-the-art results while maintaining low computational complexity, and the efficacy of its novel designs.

Keywords: 3D Medical VQA · Multimodal Large Language Models · Query-Driven Adaptive Workflow · Fine-Grained Vision-Language Alignment · Hybrid Farthest Point Sampling.

H. Liu and D. Wei——Contributed equally; H. Liu contributed to this work during an internship at Tencent.

1 Introduction

Driven by the rapid development of multimodal large language models (MLLMs) [4, 21, 24], recent progress in medical visual question answering (Med-VQA) demonstrates great potential for supporting clinical diagnosis. Early works focused on 2D images or 2D slices derived from 3D scans [17, 23, 35]. While advancing the field, 2D Med-VQA cannot capture volumetric details of routine clinical 3D imaging, e.g., CT. Pioneer works [2, 5, 15, 30] explored 3D MLLMs by aligning 3D vision encoders with LLMs. However, 3D Med-VQA faces two prominent challenges: high computational costs and extreme negative-positive imbalance. A CT volume may yield several thousand tokens after encoded by a vision encoder, causing high computational costs, memory requirements, and latency when fed to an LLM decoder. Meanwhile, unlike natural images where objects of interest often dominate the scene, lesions (e.g., nodules) in 3D medical imaging are typically sparse—sometimes occupying less than 1% of the volume [1]. This severe imbalance may cause critical diagnostic signals to be overwhelmed by the vast amount of healthy tissue, leading to biased predictions that prefer "no findings".

To reduce computational costs, some works adopted slice-based pipelines [8, 21, 29, 31] or 3D spatial pooling [2, 7]. However, these strategies lost 3D context, continuity, and volumetric detail, resulting in significant information loss. To enhance feature representation of CT volumes, some works employed contrastive learning [6, 25] or utilized organ segmentation for guidance [9, 20]. Additionally, [9, 20, 25] reduced the number of visual tokens at organ-level to improve efficiency. Notwithstanding, these methods performed predefined token reduction without considering question contexts in VQA, ignoring valuable clues about correct answers. Photon introduced instruction-conditioned token scheduling, leveraging query-vision interactions to regulate visual token reduction adaptively [11]. To the best of our knowledge, no prior peer-reviewed work has examined the impact of fine-grained, lesion-level vision-language alignment on 3D Med-VQA. Recently, the ReXGroundingCT dataset [1] provides correspondences between free-text findings and 3D lesion segmentations for chest CT scans. However, the lesions are subject to incomplete annotations due to a limited annotation budget.

This paper proposes a two-stage **M**edical **3D** V**QA** framework with an adaptive plan-and-act workflow (M3D-QAdapter). Stage I conducts image-report alignment enhanced by *lesion-level finding-segmentation correspondence* to train the 3D vision encoder. To overcome the issues of extreme negative-positive imbalance and incomplete lesion annotations, we propose *safe negative mining* to identify likely unannotated findings heuristically and exclude them from loss computation, preventing the suppression of unannotated lesions while reducing negative samples. Stage II fine-tunes an LLM for Med-VQA tasks. To reduce computational costs, our framework dynamically *plans* which anatomical regions to scrutinize for a question. It then *acts* to invoke two tools in order: 1) a text-prompted segmentation tool to obtain 3D masks of planned regions, and 2) a hybrid geometry-semantics farthest point sampling (FPS) tool to select a compact set of representative visual tokens from the masked regions. This adaptive workflow is dynamic, planning different regions of interest (ROI) and

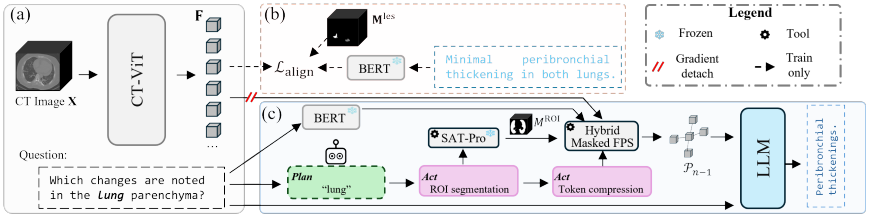

Fig. 1. Overview of M3D-QAdapter. (a) Visual token extraction. (b) Image-report alignment enhanced by lesion-level finding-segmentation correspondence. (c) Adaptive plan-and-act workflow for question-dependent, adaptive visual token compression.

selecting distinct token sets based on the question. Finally, the selected tokens are fed to the LLM for answer generation. Experiments on the 3D-RAD benchmark demonstrate state-of-the-art (SOTA) performance of M3D-QAdapter and validate its novel design.

2 Method

Fig. 1 shows an overview of our framework. Let $\mathbf{X} \in \mathbb{R}^{D\times H\times W}$ denote a 3D CT image, where D, H, and W are the number of slices, width, and height, respectively. We employ a CT-ViT [14] to extract image features $\mathbf{F} \in \mathbb{R}^{N_v\times D_v}$ (Fig. 1(a)), where N_v is the number of tokens after patchification and D_v is the feature dimension. The framework is trained in two stages. Stage I trains CT-ViT via image-report alignment *enhanced by lesion-level correspondences between radiology findings and 3D segmentations* (Fig. 1(b)). Stage II compresses the visual tokens and fine-tunes an LLM for Med-VQA tasks (Fig. 1(c)), with CT-ViT frozen. Motivated by the *ReAct* paradigm [33], we implement token compression with an adaptive plan-and-act workflow. Specifically, our framework dynamically *plans* the anatomical ROIs based on the input question. Then, it *acts* to invoke two tools in order: 1) a text-prompted segmentation tool to produce 3D ROI masks, and 2) a hybrid masked FPS tool to select a compact set of visual tokens, which is fed to the LLM for answer generation.

2.1 Stage I: Image-Report Alignment Enhanced by Lesion-level Finding-Segmentation Correspondence

As a starting point, we employ the vision-language contrastive pretraining pipeline in [25] for basic image-report alignment on the CT-RATE [15] dataset (due to space limitations, we refer readers to [25] for details, and focus on our novel additions below). In this work, we propose enhancing the alignment with lesion-level correspondences between radiology findings and 3D segmentation. This enhancement relies on the ReXGroundingCT [1] dataset, which, for the first time, links free-text radiology findings to expert-monitored 3D segmentation in CT scans.

However, ReXGroundingCT is a very limited subset (∼12%) of CT-RATE, and is subject to incomplete annotations and severe class imbalance.

Visual Token to Lesion-Level Finding Correspondence Estimation. Assume a CT image $\mathbf{X}$ in ReXGroundingCT has K radiology findings (in texts) and corresponding segmentation masks (in voxels). We obtain an embedding $\boldsymbol{t}_k \in \mathbb{R}^{1\times D_t}$, where $k \in [1, K]$ and D_t is the feature dimension, for each finding with the CT-CLIP BERT encoder [16]. Then, we compute cosine similarity between the embedding and visual tokens: $s_{k,i} = \tau \cdot \langle \|\boldsymbol{t}_k\|, \|\ln(\boldsymbol{f}_i)\| \rangle$, where τ is a learnable temperature parameter, $\|\cdot\|$ is L2 normalization, $\ln(\cdot)$ is a learnable linear projection, and $\boldsymbol{f}_i \in \mathbb{R}^{1\times D_v}$ is the i^{th} token in $\mathbf{F}$ with $i \in [1, N_v]$. Next, we apply sigmoid activation to obtain $p_{k,i} = \text{sigmoid}(s_{k,i})$. Intuitively, $p_{k,i}$ indicates the probability of the i^{th} visual token corresponding to the k^{th} radiology finding.

Safe Negative Mining. For effective use of the ReXGroundingCT dataset, we face two major issues. The first is incomplete annotations: to reduce the workload, ReXGroundingCT limited the annotation to at most three instances per finding (e.g., only three nodules are annotated in a volume when more exist). Treating all unannotated regions as definite negatives introduces potential label noise, penalizing the model for correctly detecting valid but unannotated lesions. The second is the extreme negative-positive imbalance: lesions typically occupy only a small portion of the volume, compounding the first issue. In this case, training can be biased by the overwhelmingly easy negatives (healthy tissue), leading the model to frequently predict the trivial solution of "no findings".

To handle potentially incomplete annotations while also reducing the number of negative samples, we propose *safe negative mining*. To do so, we first convert the voxel-wise lesion segmentation masks to token-wise counterparts $\mathbf{M}^{\text{les}} \in \{0,1\}^{K\times N_v}$. Specifically, $\mathbf{M}^{\text{les}}_{k,i} = 1$ indicates that the i^{th} visual token belongs to the k^{th} finding, and 0 indicates the opposite. In practice, we set $\mathbf{M}^{\text{les}}_{k,i} = 1$ if $\geq$ 10% of the i^{th} token's patch volume overlaps with the k^{th} finding's ground truth volume, and 0 otherwise. Then, if the model predicts a high possibility $p_{k,i} > \delta$ for a visual token with $\mathbf{M}^{\text{les}}_{k,i} = 0$, where δ is a threshold and empirically set to 0.8 based on preliminary experiments, we consider it a "potentially unannotated" lesion, and exclude it from contributing to the loss computation as a negative sample. Therefore, the loss on negative samples is:

$$\mathcal{L}_{\text{neg}} = -\frac{1}{N_{\text{neg}}} \sum_{k=1}^{K} \sum_{i=1}^{N_v} (1 - \mathbf{M}^{\text{les}}_{k,i})) \cdot \mathbb{1}(p_{k,i} \leq \delta) \cdot \log(1 - p_{k,i}), \tag{1}$$

where N_{neg} is the total number of negative samples excluding "potentially unannotated" lesions, and $\mathbb{1}(x)$ is an indicator function which equals 1 if x is true and 0 otherwise. The threshold δ acts as a guard, preventing the suppression of likely unannotated lesions. Meanwhile, the loss on positive samples is $\mathcal{L}_{\text{pos}} = -\frac{1}{N_{\text{pos}}} \sum_{k=1}^{K} \sum_{i=1}^{N_v} \mathbf{M}^{\text{les}}_{k,i} \log(p_{k,i})$, where N_{pos} is the number of positive samples. The loss normalization by N_{neg} and N_{pos} helps balance negative and positive sample learning. The complete alignment loss for Stage I is $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}} + \mathcal{L}_{\text{so-pre}}$, where $\mathcal{L}_{\text{so-pre}}$ is the vision-language contrastive pretraining objective in [25]. Note that $\mathcal{L}_{\text{pos}}$ and $\mathcal{L}_{\text{neg}}$ are computed only for samples

in ReXGroundingCT, whereas $\mathcal{L}_{\mathrm{so-pre}}$ is computed for all training samples of CT-RATE (including the ReXGroundingCT subset).

2.2 Stage II: Adaptive Workflow of Query-Driven Dynamic Token Compression for LLM Fine-tuning

Previous works often compress the visual tokens of whole CT volumes either by spatial pooling [2, 7] or static top-K selection [25], without considering varying query context. In contrast, we propose query-driven dynamic token compression to select tokens most relevant to a question. The process is dynamic, as we do not know what the question asks about beforehand; thus, we implement an adaptive plan-and-act workflow for flexibility. The framework first *plans* by identifying which anatomic regions the question is querying about, then *acts* to invoke two *tools* in order: 1) a text-prompted segmentation tool to obtain 3D masks of the planned regions, and 2) a mask-constrained hybrid FPS tool to select a compact set of representative visual tokens. This design resonates with human observers' cognitive processes in Med-VQA: first planning which anatomies to investigate based on the question, and then acting to extract key information therein.

Anatomical Planning. Given a question, our framework attempts to identify the queried anatomic regions (e.g., lung, mediastinum) by keyword matching against a predefined dictionary. If no match is found (∼2% of the time), it then prompts a local LLM (Qwen-2.5-7B [32]) to infer the implied ROIs from the question. The dictionary and prompt will be released along with our code. When a question spans multiple anatomical regions, the planner outputs all relevant targets, and their masks are merged into a unified ROI for token selection.

Tool 1: Text-Prompted ROI Segmentation. After planning the ROI, the framework invokes a text-prompted (e.g., "lung", "mediastinum") radiology scan segmentation model, SAT-Nano [36], to obtain a corresponding 3D mask. Then, a binary indicator map $M^{\mathrm{ROI}} \in \{0,1\}^{1\times N_v}$ is constructed for all visual tokens: for the i^{th} token, $M_i^{\mathrm{ROI}} = 1$ if $\geq 10\%$ of its patch volume overlaps with the ROI mask, and 0 otherwise. Intuitively, M^{ROI} indicates which visual tokens are relevant to the specific question and which are not, enabling smart allocation of computational budget to the planned ROI.

Tool 2: Query-Guided Hybrid Masked FPS. This action aims to sample the most representative visual tokens from all ROI tokens in M^{ROI}. A straightforward FPS based solely on geometric distance may miss semantically critical tokens. To address this issue, we design a query-guided hybrid FPS tool accounting for both geometry and semantics, as detailed below.

1) Query-Aware Feature Fusion. We extract the embedding of the input question with the CT-CLIP BERT [16], denoted by $\boldsymbol{q} \in \mathbb{R}^{1\times D_t}$. $\boldsymbol{q}$ is fused with each visual token $\boldsymbol{f}$ by a multilayer perceptron (MLP), yielding query-aware visual features $\boldsymbol{f'} = \mathrm{fuse}(\boldsymbol{f}, \boldsymbol{q}) \in \mathbb{R}^{1\times D_v}$. This fusion enables the tokens to reflect the context contained in the question, along with the visual features of the CT image.

2) Semantics-Aware Initialization. Naive FPS typically initializes at a random point or at the geometric centroid, which may introduce undesirable randomness or place it far from semantically preferable locations. Instead, we learn the

semantic importance w of each fused token with a scoring MLP: $w = \text{score}(\boldsymbol{f}')$, and initiate the FPS with the most important token in the ROI mask: $pt_0 = \text{argmax}_{i \in \{i | M_i^{\text{ROI}}=1\}} w_i$. Given that $\boldsymbol{f}'$ combines query and visual information, pt_0 is a token with high query-visual relevancy.

3) Hybrid Geometry-Semantics FPS. Starting from pt_0, we iteratively select the remaining $N-1$ visual tokens. Here, the token budget N is a hyperparameter, which will be empirically studied later. In the n^{th} iteration ($n \in [1, N-1]$), we select a new token that is farthest from the set of presiously selected tokens $\mathcal{P}_{n-1} = \{pt_0, \ldots, pt_{n-1}\}$, according to the hybrid geometry-semantics distance:

$$\mathcal{D}(i, \mathcal{P}_{n-1}) = \min_{j \in \mathcal{P}_{n-1}} \big[\underbrace{0.5 \cdot \|\boldsymbol{c}_i - \boldsymbol{c}_j\|_2}_{\text{Geometry}} + \underbrace{0.5 \cdot (1 - \text{sim}(\boldsymbol{f}'_i, \boldsymbol{f}'_j))}_{\text{Semantics}} \big], \tag{2}$$

where $\boldsymbol{c}$ denotes normalized 3D coordinates of the token patch centers, $\|\cdot\|_2$ is L2 distance, and $\text{sim}(\boldsymbol{f}'_i, \boldsymbol{f}'_j)$ computes cosine similarity after applying L2 normalization to $\boldsymbol{f}'_i$ and $\boldsymbol{f}'_j$. Thus, the n^{th} token is selected to maximize the hybrid distance within the planned ROI: $pt_n = \text{argmax}_i \, \mathcal{D}(i, \mathcal{P}_{n-1})$, where $i \in \{i | M_i^{\text{ROI}} = 1\}$ and $i \notin \mathcal{P}_{n-1}$. Then, pt_n is added to the selected point set $\mathcal{P}_n = \mathcal{P}_{n-1} \cup \{pt_n\}$. This process continues until the token budget is reached. The hybrid FPS makes the sampled tokens geometrically diverse (broadly covering the ROI) while maintaining high semantic relevance in the query-guided feature space.

Answer Generation. After FPS, the set of sampled visual tokens, along with a global visual token $\boldsymbol{f}^{\text{global}}$ (i.e., the class token in a ViT [10]; omitted in Fig. 1 for simplicity) and the tokens of the input question, is fed to the LLM for answer generation. The training loss is $\mathcal{L}_{\text{QA}} = \sum_{l=1}^{L} \log p_\theta(y_l | y_{<l}, \{\boldsymbol{f}'_i | i \in \mathcal{P}_{N-1}\}, \boldsymbol{f}^{\text{global}}, Q)$, where θ indicate trainable parameters, L is the length of the target answer, y_l is the l^{th} answer token, $y_{<l}$ includes all answer tokens prior to l, and Q is the question. Note that, before feeding to the LLM, the visual tokens are projected into the LLM's input space. Following the simple design principle of [24], we use an MLP with two linear projection layers and a ReLU activation in between. Thus, θ includes the parameters of three MLPs (two in the FPS tool and one for the projection) plus LoRA [18] parameters for the LLM.

3 Experiments

Datasets and Evaluation Metrics. *Stage I* (alignment) uses two datasets for training. **CT-RATE** [15] comprises 25,692 non-contrast chest CT volumes of 21,304 unique patients, with corresponding radiology reports. **ReXGroundingCT** [1] provides lesion-level 3D segmentations paired with lung and pleura findings (in free texts) for 3,142 volumes in CT-RATE. *Stage II* fine-tunes the LLM on **3D-RAD** [12], a large-scale 3D CT VQA dataset—also derived from CT-RATE. It consists of 136,195 question-answer (QA) pairs for training, and 33,910 QA pairs for benchmarking. 3D-RAD encompasses six clinically relevant tasks: anomaly detection, image observation, medical measurement, existence detection, static temporal diagnosis, and longitudinal temporal diagnosis, supporting both open- and closed-ended questions for rigorous and comprehensive

Table 1. Comparison of zero-shot/fine-tuned performance on the 3D-RAD [12] benchmark. Stat./Longt. Temp. Diag.: static/longitudinal temporal diagnosis. Compared performance is from [11] under the same data split.

		Zero-shot						Fine-tuned			
Task	**Metric**	Qwen2.5-VL [3] 3B	RadFM [30] 13B	M3D-L2 [2] 7B	M3D-P3 [2] 4B	OmniV [19] 1.5B	Lingshu [31] 7B	M3D-L2 [2] 7B	M3D-P3 [2] 4B	Photon [11] 3B	Ours 3.6B
Existence Detection	Accuracy	19.52	29.20	18.00	40.25	28.66	59.60	81.09	82.43	83.07	**84.91**
Stat. Temp. Diag.	Accuracy	0.00	44.11	25.47	25.40	22.96	6.02	51.20	49.30	52.86	**57.96**
Longt. Temp. Diag.	Accuracy	0.29	42.99	24.17	24.31	24.23	12.13	74.78	74.77	77.01	**79.57**
Medical Measurement	BLEU	1.78	3.34	15.95	2.55	2.52	2.50	30.54	20.85	37.74	**37.80**
	ROUGE	4.30	6.62	23.24	5.63	7.88	5.45	36.06	30.38	39.36	**39.54**
	BERTScore	84.20	86.85	91.50	85.74	85.66	83.81	94.65	93.25	96.14	**96.28**
Image Observation	BLEU	3.51	13.48	10.69	16.31	16.42	5.34	31.28	39.66	51.59	**55.69**
	ROUGE	8.84	19.14	20.82	23.19	26.69	12.25	39.12	50.52	56.66	**60.13**
	BERTScore	84.53	87.16	86.61	86.92	88.29	85.67	90.00	92.19	93.62	**94.06**
Anomaly Detection	BLEU	2.93	11.00	9.10	15.06	13.47	3.71	25.25	33.28	42.33	**43.54**
	ROUGE	9.17	17.62	18.64	23.19	25.72	9.46	33.76	42.45	47.50	**47.68**
	BERTScore	84.47	86.76	86.07	87.11	88.21	84.81	89.16	90.72	91.96	**92.19**

benchmarking. As ReXGroundingCT and 3D-RAD strictly follow the original training/validation split of CT-RATE, we train the model only on the training splits of all datasets to safely avoid test data leakage. For preprocessing, we resample all volumes to the voxel spacing of $0.75\times0.75\times1.5$ mm^3, and center-crop or zero-pad them to a fixed resolution of $480\times480\times240$ voxels.

Following [11], we evaluate performance using metrics tailored to each task's nature. For free-text tasks, we employ BLEU [27], ROUGE [22], and BERTScore [34] to assess fluency, lexical overlap, and semantic similarity between generated and reference answers. For category prediction tasks, we report accuracy.

Implementation. Our framework is implemented using PyTorch (2.0.0) [28] and trained on four NVIDIA Tesla V100 GPUs with 32GB of memory each. The optimizer is AdamW [26], with learning rates of 10^{-4} and 10^{-4} for Stage I and Stage II. We set the warm-up ratio to 10% and use the linear learning rate scheduler after warm-up. The model is trained for 100K and 100K steps for Stage I and Stage II, with a batch size of 8. For answer generation, we fine-tune Llama-3.2-3B [13] with LoRA (rank $r = 128$ and scaling factor $\alpha = 16$; 0.68% parameters fine-tuned) [18]. Our implementation is available at: https://github.com/ccarliu/M3D-QAdapter.

Comparison to SOTA Methods. Table 1 shows that our proposed M3D-QAdapter achieves the best performance across *all* tasks and metrics. Notably, it improves upon the previous SOTA, Photon, by 9.6% in static temporal diagnosis, and by 7.9% (BLEU) and 6.1% (ROUGE) in image observation. Also, longitudinal temporal diagnosis improves by 3.3%. The performance gains in these diagnosis-oriented tasks highlight the benefits of lesion-level alignment between free-text radiology findings and 3D segmentations, as well as the query-aware adaptive workflow in M3D-QAdapter. Meanwhile, M3D-QAdapter's consistently superior performance to existing approaches demonstrates its strong generalization across tasks of various natures, including descriptive (e.g., observation), quantitative (e.g., measurement), and deductive (e.g., temporal diagnosis).

Ablation Studies. Table 2(a) shows that setting the visual token budget N to 25 saturates the performance. Table 2(b) shows that while simply introduc-

ing the lesion-level alignment data (ReXGroundingCT) slightly improves performance, our safe negative mining efficiently uses the data and boosts performance by mitigating the class imbalance and incomplete annotation issues. As ReXGroundingCT is a small subset (∼12%) of CT-RATE, we expect that increasing the amount of fine-grained alignment data would further benefit our M3D-QAdapter.

Table 2(c) shows the ablation of Stage II components. (i)–(ii) Replacing all Stage II operations with random sampling or naive FPS yields the worst results, validating the overall effectiveness of our design. (iii) Ablating anatomic planning while keeping other components also leads to notably inferior performance, indicating the necessity of the modular tool invocation of ROI segmentation, which boosts performance by constraining token sampling in relevant regions. (iv) Random token sampling from the ROI (without FPS) deteriorates all metrics. Therefore, both modular tool invocations (ROI segmentation and FPS) jointly contribute to M3D-QAdapter's excellent performance. (v) Ablating query-aware fusion leads to a 1-point drop in accuracy, suggesting the benefit of incorporating the contextual information of the questions. (vi) Replacing our semantics-aware initialization with a random one causes slight performance drops in accuracy and BLEU. (vii) Lastly, implementing a pure geometry-based FPS yields substantially worse results in all metrics than our hybrid geometry-semantics FPS. This is because the hybrid FPS enhances the pure geometry-based with a semantic dissimilarity term, ensuring the selected tokens are both spatially dispersed and semantically diverse. Furthermore, Table 2(c-viii) shows that the results of using external resources alone (ReXGroundingCT data without SNM + SAT segmentation without hybrid FPS) substantially inferior to our full model. These results validate our novel framework design.

Complexity Analysis. Compared to the baseline in Table 2(c-i), M3D-QAdapter incurs reasonable increases in inference complexity (mainly due to the segmentation tool; Table 2(d)) but notably boosts performance. Thus, we consider the adaptive workflow a worthwhile trade-off between complexity and performance. In addition, M3D-QAdapter retains drastically fewer visual tokens for answer generation than the previous SOTA, Photon [11] (25 versus 400), while achieving comprehensive performance improvements in Table 1. Notably, the SAT-Nano segmentation runs once per CT volume and can be shared across all questions for that volume. In the 3D-RAD benchmark where >15 questions are asked per volume on average, the amortized segmentation cost is only ∼0.13 s/question, reducing the effective per-question latency to ∼2.6s.

Analysis of Query Awareness. Thanks to the flexible adaptive workflow, our visual token compression is truly adaptive to the questions. For example, for lung-related questions, *all* selected tokens belong to lung regions for M3D-QAdapter. In contrast, the ratios are only ∼10% for the variants using random token sampling (Table 2(c-i)) and non-masked FPS (Table 2(c-ii)).

Clinical Error Analysis. To assess clinical correctness beyond automatic metrics, we conduct an LLM-assisted error analysis on 500 randomly sampled open-ended cases. The results show: correct 34.1%, partially correct 23.1%, localization

Table 2. (a)–(c) Ablation studies. We report the mean accuracy, BLEU, and BERTScore of all metrics in Table 1. **(d) Complexity comparison.**

(a) Token budget N

N	Acc.	BLEU	BERTScore
20	73.5	45.5	94.1
25	**74.2**	**45.7**	**94.2**
30	**74.2**	45.6	94.1

(b) Stage I ablation

Method	Acc.	BLEU	BERTScore
Baseline	72.5	44.0	94.0
+ ReX (w/o SNM*)	73.1	44.3	94.0
+ ReX (w SNM)	**74.2**	**45.7**	**94.2**

**SNM*: safe negative mining.

(c) Stage II ablation

Variant	ROI Seg.	Acc.	BLEU	BERTScore
(i) Random sampling	×	72.1	43.2	93.7
(ii) Naive FPS	×	72.2	43.8	93.9
(iii) Hybrid FPS w/o ROI seg.	×	72.5	44.7	94.0
(iv) w/o FPS	✓	73.1	45.0	93.9
(v) w/o Query-aware fusion	✓	73.2	**45.7**	**94.2**
(vi) w/o Semantic initialization	✓	73.8	45.6	**94.2**
(vii) w/o Hybrid distance	✓	72.8	44.3	93.8
(viii) Ext. resources only	✓	72.5	44.5	94.0
M3D-QAdapter	✓	**74.2**	**45.7**	**94.2**

(d) Inference complexity analysis

		Backbone	Random sampling	Adaptive workflow			
Method	**Metric**	Visual encoder & LLM	-	Plan	Seg. (SAT-Nano)	FPS (Including BERT)	**Total**
Baseline	Latency (s)	2.3	0.04	-	-	-	2.34
	Memory (GB)	11.9	-	-	-	-	11.9
M3D-QAdapter	Latency (s)	2.3	-	0.1	2.0	0.06	4.46
	Memory (GB)	11.9	-	-	24	0.5	36.4

error 32.1%, missed finding 8.8%, and hallucinated finding 2.0%. Notably, the hallucination rate—the most clinically dangerous error type—is only 2.0%.

4 Conclusion

This paper proposed M3D-QAdapter, a novel framework for 3D medical visual question answering. M3D-QAdapter featured lesion-level alignment between radiology findings and 3D segmentation, and a query-driven adaptive workflow for visual token compression. Extensive experiments on the 3D-RAD benchmark demonstrated its state-of-the-art performance and validated its novel designs.

Limitations & Future Work. Due to limited data availability, we evaluated M3D-QAdapter only for CT in this work. We plan to assess its generalization to MRI in future work. While achieving encouraging results, this work did not rigorously explore the hyperparameter space, which warrants further investigation.

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).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baharoon, M., et al.: ReXGroundingCT: A 3D chest CT dataset for segmentation of findings from free-text reports. arXiv preprint arXiv:2507.22030 (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. Bai, S., et al.: Qwen2.5-VL technical report (2025), https://arxiv.org/abs/2502.13923
4. Bai, S., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)

5. Blankemeier, L., et al.: Merlin: A vision language foundation model for 3D computed tomography. Research Square pp. rs–3 (2024)
6. Cao, W., et al.: Bootstrapping chest CT image understanding by distilling knowledge from X-ray expert models. In: CVPR. pp. 11238–11247 (2024)
7. Chen, H., et al.: 3D-CT-GPT: Generating 3D radiology reports through integration of large vision-language models. arXiv preprint arXiv:2409.19330 (2024)
8. Chen, J., et al.: Towards injecting medical visual knowledge into multimodal LLMs at scale. In: EMNLP. pp. 7346–7370 (2024)
9. Chen, Z., Bie, Y., Jin, H., Chen, H.: Large language model with region-guided referring and grounding for CT report generation. IEEE TMI **44**(8), 3139–3150 (2025)
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Fang, C., et al.: Photon: Speedup volume understanding with efficient multimodal large language models. In: ICLR (2026)
12. Gai, X., et al.: 3D-RAD: A comprehensive 3D radiology Med-VQA dataset with multi-temporal analysis and diverse diagnostic tasks. In: NeurIPS Datasets and Benchmarks Track (2025)
13. Grattafiori, A., et al.: The Llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
14. Hamamci, I.E., et al.: GenerateCT: Text-guided 3D chest CT generation. CoRR (2023)
15. Hamamci, I.E., et al.: Developing generalist foundation models from a multimodal dataset for 3D computed tomography. arXiv preprint arXiv:2403.17834 (2024)
16. Hamamci, I.E., et al.: A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. CoRR (2024)
17. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
18. Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
19. Jiang, S., et al.: OmniV-Med: Scaling medical vision-language model for universal visual understanding. arXiv preprint arXiv:2504.14692 (2025)
20. Kyung, S., et al.: MedRegion-CT: Region-focused multimodal LLM for comprehensive 3D CT report generation. arXiv preprint arXiv:2506.23102 (2025)
21. Li, C., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS **36**, 28541–28564 (2023)
22. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
23. Liu, B., et al.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: IEEE ISBI. pp. 1650–1654. IEEE (2021)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)
25. Liu, H., et al.: Structure observation driven image-text contrastive learning for computed tomography report generation. In: IPMI. pp. 218–233. Springer (2025)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
28. Paszke, A., et al.: PyTorch: An imperative style, high-performance deep learning library. NeurIPS pp. 8024–8035 (2019)
29. Sellergren, A., et al.: MedGemma technical report. arXiv preprint arXiv:2507.05201 (2025)

30. Wu, C., et al.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Nature Communications **16**(1), 7866 (2025)
31. Xu, W., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
32. Yang, A., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
33. Yao, S., et al.: ReAct: Synergizing reasoning and acting in language models. In: ICLR (2022)
34. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675 (2019)
35. Zhang, X., et al.: PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
36. Zhao, Z., et al.: One model to rule them all: Towards universal segmentation for medical images with text prompts. arXiv e-prints pp. arXiv–2312 (2023)