# SOAR: Semantic Organ-Aware Pretraining for 3D CT Image Understanding

**Huidong Xie**[1,2] **Wei Ji**[3] **Sunan He**[1] **Zhihao Tang**[1] **Wen Tang**[1]
**Jun Zhao**[1*] **Hongming Shan**[1,2*]

[1] JD Health International Inc. [2] Institute of Science and Technology for Brain-inspired Intelligence, MOE Frontiers Center for Brain Science, Fudan University
[3] Department of Radiology and Biomedical Imaging, Yale University School of Medicine
huidong.xie@aya.yale.edu, zhaojun10@jd.com, hmshan@fudan.edu.cn

## Abstract

Vision-language pretraining has advanced multimodal medical AI, but its direct application to 3D computed tomography (CT) remains limited by a fundamental mismatch between dense volumetric anatomy and sparse diagnostic reports. Existing 3D CT vision-language methods typically rely on global image-report alignment or align anatomical regions with decomposed report descriptions. However, a CT scan captures rich, spatially distributed information across many organs, whereas a report summarizes only a small subset of clinically relevant findings. As a result, report-based alignment provides incomplete or weak supervision, leaving many local anatomical structures underrepresented. Motivated by this limitation, we introduce **SOAR**, a semantic organ-aware pretraining framework for 3D CT image understanding. The key innovation of SOAR lies in its integration of fine-grained, organ-level structural priors, enabling report-free organ-aware supervision for visual pretraining without relying on radiologist annotations or coarse global image-report matching. SOAR combines three complementary objectives: (**i**) *organ-level masked reconstruction* to learn localized anatomical context, (**ii**) *organ-level vision-language alignment* to associate organ-specific visual features with organ-name text embeddings, and (**iii**) *organ-level supervision* to preserve voxel-level structural detail. SOAR integrates easily with existing LLMs and is evaluated on five public/in-house benchmarks, improving multiple-choice VQA accuracy by 3.53 percentage points and disease screening/abnormality detection AUC by 4.25 points on average over all benchmarks/baseline methods. Code and model weights are available at https://github.com/HuidongXie/SOAR_CT.

## 1 Introduction

Medical image understanding is a critical yet demanding component of clinical workflow, particularly for 3D modalities such as computed tomography (CT) [1, 2]. In routine CT interpretation, radiologists must review hundreds of slices spanning multiple anatomical regions, identify subtle disease-related findings, and integrate them for diagnostic decision making [3]. Meanwhile, the increasing volume of CT examinations [4], together with the shortage of experienced radiologists [5], highlights the need for AI systems that can improve diagnostic accuracy, efficiency, and consistency.

Since the vision encoder serves as the representational foundation for downstream CT image understanding tasks, effective pretraining is essential for learning transferable 3D anatomical representations from volumetric images. Existing vision encoder pretraining methods mainly follow two paradigms:

*Co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

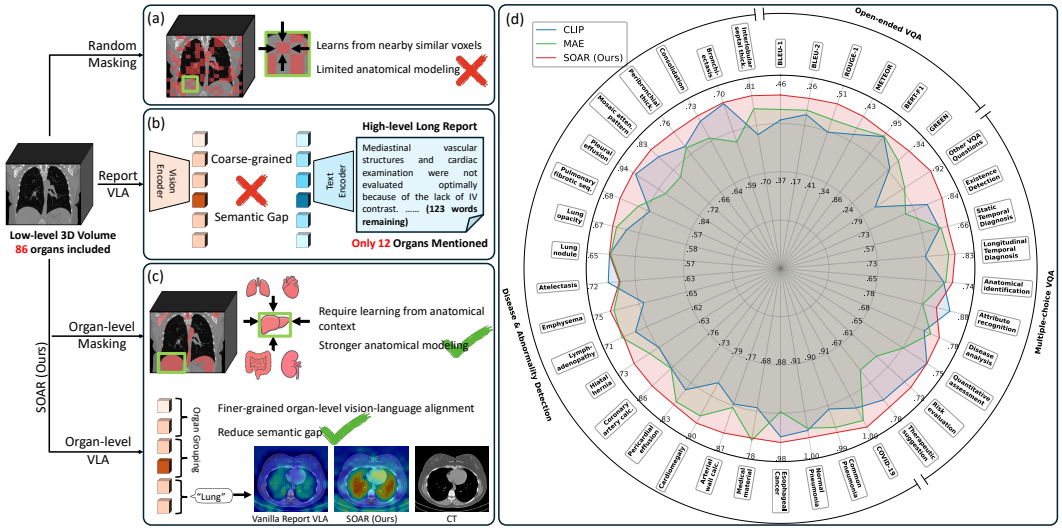



Figure 1: **Motivation and quantitative summary.** (a) Random patch masking in MAE may rely on nearby visual context, limiting anatomical modeling. (b) Vanilla VLA aligns a 3D CT volume with a sparse report, leading to coarse-grained semantics. (c) SOAR uses organ-aware masking and organ-level VLA to promote anatomical modeling and reduce the semantic gap. (d) Radar plot comparing CLIP, MAE, and SOAR across downstream evaluations. Closed-ended VQA is measured by accuracy and disease screening and abnormality detection are measured by AUC.

reconstruction-based masked image modeling and contrastive vision-language alignment, commonly represented by Masked Autoencoders (MAE) [6] and Contrastive Language-Image Pretraining (CLIP) [7], respectively. Although both paradigms have been adapted to medical imaging [2, 8], their standard formulations do not fully exploit the anatomical priors inherent to 3D CT volumes. For MAE-style pretraining shown in Fig. 1(a), random patch masking often leads to a relatively local reconstruction task: neighboring patches within the same organ usually share similar intensity, texture, and morphology, allowing the model to recover missing regions largely from nearby voxels rather than from broader anatomical context. For CLIP-style pretraining shown in Fig. 1(b), a different limitation arises from the semantic mismatch between dense volumetric images and their accompanying textual descriptions. A 3D CT volume contains many organs and anatomical structures, whereas a diagnostic report typically summarizes only a small subset of clinically relevant findings. As illustrated in Fig. 1(b), the example CT volume contains 86 segmented organs/regions, but the corresponding report mentions only 12 of them. Consequently, directly aligning an entire CT volume with an entire report provides coarse and incomplete supervision, leaving many local anatomical structures weakly grounded. These limitations motivate an organ-aware pretraining strategy that explicitly incorporates anatomical structure into both reconstruction-based and vision-language representation learning.

To address these challenges, we propose **SOAR**, a semantic organ-aware pretraining framework for 3D CT image understanding, as outlined in Fig. 1(c). Motivated by the success of medical image segmentation models [9–13], SOAR leverages segmentation-derived organ-level anatomical priors for 3D visual encoder pretraining. These priors provide fine-grained anatomical supervision without requiring radiologist annotations or paired diagnostic reports, which are often costly or unavailable at scale. The key innovation of SOAR lies in using organ-level anatomical structures as anchors, enabling fine-grained organ-aware supervision for 3D visual pretraining.

**SOAR's novelty lies in three complementary designs.** *First*, instead of randomly masking isolated patches, SOAR performs organ-level masked reconstruction by masking and reconstructing randomly selected organ regions. Since voxels within the same organ often share similar local texture and intensity patterns, this design encourages the encoder to infer missing anatomical structures from global anatomical context rather than relying on local continuity alone. *Second*, for vision-language alignment, SOAR aligns organ-level visual tokens with corresponding anatomical text embeddings, providing explicit semantic grounding for each anatomical structure and reducing the semantic gap between dense 3D visual content and coarse report-level supervision. *Third*, SOAR introduces a voxel-level anatomical structure constraint to encourage the encoder to preserve fine-

grained anatomical information. Together, these designs enable SOAR to learn more anatomically meaningful representations for downstream CT image understanding tasks.

**SOAR is conceptually simple.** Rather than relying on an architectural redesign, SOAR organizes standard pretraining objectives around anatomical units. By reformulating masked reconstruction, vision-language alignment, and structural supervision at organ level, SOAR provides a simple yet effective way to connect voxel-level visual patterns and organ-level semantics for downstream tasks.

**Contributions.** The main contributions are summarized as follows. (**i**) We present SOAR, a simple yet effective organ-level pretraining strategy that organizes masked reconstruction, vision-language alignment, and structural supervision around anatomical structures. By progressively injecting voxel-level structure and organ-level semantics, SOAR improves performance across diverse 3D medical image understanding tasks. (**ii**) We evaluate SOAR across multiple downstream tasks, including disease screening, abnormality detection, vision-language understanding, and medical visual question answering (VQA), as outlined in Fig. 1(d). Extensive results show that SOAR achieves stronger overall performance than competing 3D medical pretraining baselines, with an average improvement of 3.53 percentage points in multiple-choice VQA accuracy and 4.25 AUC points in disease screening and abnormality detection, both computed by averaging gains over all baselines and benchmarks.

## 2 Related Work

**Self-supervised pretraining for 3D medical visual encoders.** Self-supervised pretraining aims to learn transferable representations from unlabeled medical volumes without paired reports or manual annotations. Existing methods mainly rely on pretext tasks such as transformation prediction [14, 15], context restoration [16], global-local contrastive learning [17, 18], masked reconstruction [6, 8], and geometric context modeling [19]. Early methods such as Rubik's Cube and 3D self-supervised proxy tasks learn volumetric representations by predicting spatial transformations, recovering cube order, or solving 3D patch-location tasks [14, 15]. Models Genesis learns generic 3D medical representations by restoring transformed medical volumes [16]. Other methods introduce contrastive objectives to learn discriminative representations by preserving global-local consistency or contextual relationships, such as PCRL and GVSL [17, 18]. More recently, MAE-style methods learn visual representations by reconstructing masked image patches [6, 8]. MedVAE [20] introduces large-scale 2D and 3D medical image autoencoders that learn compact latent representations while preserving clinically relevant features for downstream tasks. VoCo leverages the relatively stable geometric context of organs in 3D medical images to learn semantic representations without textual supervision [19]. Concurrent work MASS [21] explores mask-guided self-supervised learning by using automatically generated class-agnostic masks for an in-context segmentation objective.

Despite their effectiveness, most existing self-supervised objectives provide limited explicit anatomical modeling. This restricts their ability to learn organ-specific representations, as many clinical findings are localized to specific anatomical regions. In contrast, SOAR addresses these limitations by introducing organ-level anatomical priors into visual encoder pretraining, combining organ-level masked reconstruction for localized anatomical context modeling and organ-level vision-language alignment for associating organ-specific visual features with anatomical text embeddings. SOAR does not rely on manual anatomical annotations but instead uses automatically derived anatomical priors from medical image segmentation foundation models [22, 13].

**Vision-language pretraining for 3D medical visual encoders.** Vision-language pretraining learns visual representations by aligning images with paired text, as commonly represented by CLIP [7]. This paradigm has been effective in natural images and 2D medical imaging, where images are aligned with short captions [23–30]. However, directly extending this paradigm to 3D CT remains challenging because whole-volume and whole-report alignment creates a semantic gap between dense volumetric anatomy and sparse textual supervision.

Recent 3D CT vision-language methods have attempted to reduce this semantic gap through finer-grained anatomy-aware alignment. For example, fVLM [31] and ViSD [32] use segmentation masks to group image tokens into anatomy-specific visual representations and align them with report-derived anatomical descriptions. However, because radiology reports are usually coarser than segmentation masks, these methods often need to merge multiple fine-grained anatomical regions into broader anatomical groups to improve image-text consistency, such as grouping left and right lung regions as lung-level representations. As stated in the fVLM paper [31], such anatomical grouping entails a trade-

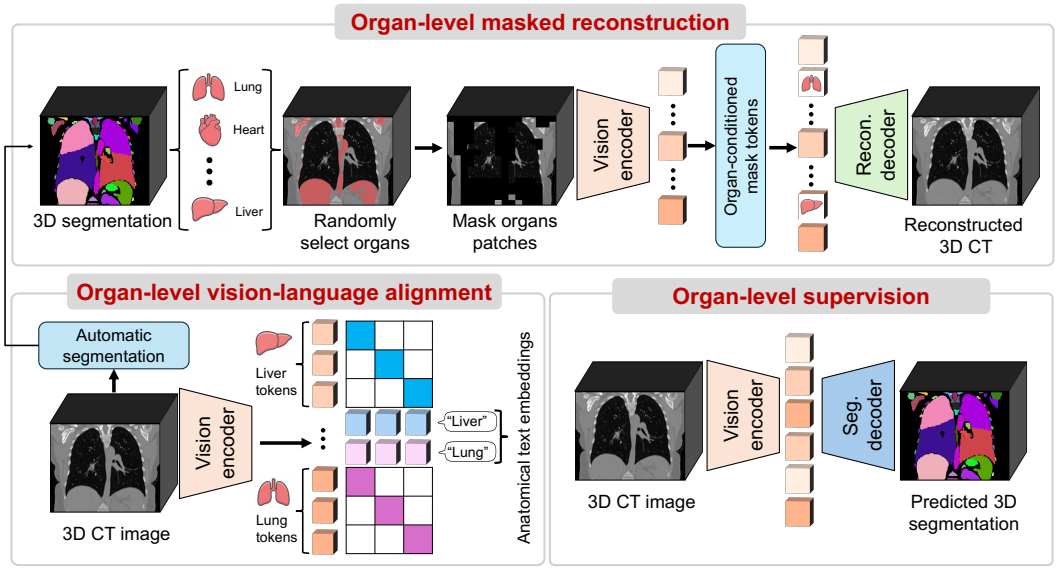



Figure 2: **Overview of SOAR.** SOAR uses automatically derived 3D anatomical priors to pretrain a 3D visual encoder with three organ-aware objectives. In organ-level masked reconstruction (organ-recon), patches from randomly selected organ regions are masked and reconstructed with organ-conditioned mask tokens. In organ-level vision-language alignment (organ-align), visual tokens are grouped by organ-level anatomical regions and aligned with corresponding anatomical text embeddings. In organ-level supervision (organ-super), the encoder is encouraged to preserve voxel-level anatomical organization, enabling fine-grained anatomical representation learning.

off among analytical granularity, image-text consistency, and data balance. Moreover, reports often mention only abnormal or clinically salient findings rather than all visible organs, so report-derived supervision may remain incomplete for learning comprehensive organ-level visual representations even after anatomy-level report decomposition.

Unlike prior work, SOAR uses automatically derived organ-level anatomical priors in a report-free framework that couples organ-level masked reconstruction, organ-level vision-language alignment, and organ-level supervision. Rather than collapsing fine-grained anatomical regions into broader groups to match coarse report descriptions, SOAR preserves the original organ-level granularity to retain detailed anatomical structure. This design constructs organ-level visual representations and aligns them with anatomical text embeddings without requiring paired diagnostic reports or report decomposition. It also provides supervision for visible anatomical structures regardless of whether they are mentioned in radiology reports, offering a scalable and fine-grained organ-aware strategy for 3D CT visual encoder pretraining.

# 3 Method

## 3.1 Overview of SOAR

To address the limitations discussed above, our goal is to develop a semantic organ-aware pretraining framework for 3D medical image understanding. SOAR is motivated by the success of medical image segmentation models [10, 11, 13, 22], which naturally provide fine-grained supervision by assigning anatomical labels to image voxels. Such segmentation masks establish an *explicit* correspondence between local image regions and organ identities, making them a natural source of organ-aware supervision for representation learning.

Fig. 2 presents the overall framework of SOAR, which incorporates automatically derived anatomical priors through three organ-aware mechanisms: organ-level masked reconstruction (§3.2), organ-level vision-language alignment (§3.3), and organ-level supervision (§3.4). Organ-level masked reconstruction masks organ regions and reconstructs the missing content with organ identity priors, encouraging anatomy-aware structural learning. Organ-level vision-language alignment groups visual tokens by anatomical region and aligns them with corresponding anatomical text prompts,

promoting anatomically grounded semantic representations. In parallel, an organ-level anatomical supervision objective encourages the encoder to preserve fine-grained voxel-level spatial information. Together, these objectives progressively inject anatomical supervision from voxel-level structure to organ-level semantics, enabling SOAR to learn transferable 3D visual representations with improved anatomical awareness. Importantly, unlike report-dependent vision-language alignment methods such as fVLM [31] and ViSD [32], SOAR does not require paired diagnostic reports or manual disease annotations, which are often difficult to obtain at scale.

### 3.2 Organ-level Masked Reconstruction

As detailed in Fig. 2, to obtain anatomical priors for organ-level masked reconstruction, we use TotalSegmentator [13, 33] to automatically derive organ-level anatomical masks for pretraining. Although the pretraining CT volumes do not require manual anatomical annotations, TotalSegmentator inherits supervision from manually annotated datasets. The generated pseudo-labels provide anatomical supervision for the proposed organ-aware pretraining objectives. Given an input 3D CT volume, TotalSegmentator identifies 117 anatomical classes and assigns each voxel to an anatomical category. Let $\mathbf{S} \in \{0, 1, \ldots, C\}^{H \times W \times D}$ denote the organ segmentation mask, where 0 denotes background, $C = 117$ denotes the number of anatomical classes, and $H$, $W$, and $D$ denote the height, width, and depth of the input 3D CT volume, respectively.

Following the Vision Transformer (ViT) [34], an input 3D CT volume $\mathbf{X} \in \mathbb{R}^{H \times W \times D}$ is first divided into $N$ non-overlapping 3D patches $\{\mathbf{x}_i\}_{i=1}^{N}$, where each patch corresponds to a spatial region $\mathcal{P}_i$. Each patch is then projected into a patch token by a patch embedding layer and processed by the ViT encoder to generate visual tokens $\{\mathbf{z}_i\}_{i=1}^{N}$, where $\mathbf{z}_i \in \mathbb{R}^d$ denotes the encoded feature of patch $\mathcal{P}_i$.

Instead of randomly masking patches as in vanilla MAE, organ-level masked reconstruction performs masking at the organ level. We randomly sample a subset of organs, $\mathcal{O}_{\text{mask}} \subseteq \{1, \ldots, C\}$, and mask patches that overlap with these selected organs. In this way, the masking process is guided by anatomical structure rather than by spatially uniform random sampling. After the visible tokens are processed by the encoder, the masked positions are filled with mask tokens and passed to the decoder for reconstruction. This encourages the encoder to infer missing organ regions from global anatomical context, rather than simply relying on similar local texture or nearby voxels within the same organ.

Unlike vanilla MAE that uses a single shared mask token as the decoder placeholder for all masked positions, SOAR injects organ identity into the masked tokens to provide anatomy-specific reconstruction cues. Because 3D CT follows relatively stable anatomical organization, anatomical masks provide prior knowledge about the organs within masked regions. Conditioning reconstruction on anatomical identity and visual context encourages the encoder to learn organ-specific appearance and inter-organ spatial relationships.

Concretely, let $\mathcal{O}_i$ denote the set of organ classes present in patch $\mathcal{P}_i$, and let $w_{i,o}$ be the relative voxel occupancy of organ $o$ within that patch. SOAR maintains a learnable organ embedding table $\{\mathbf{e}_o\}_{o=0}^{C}$, where $\mathbf{e}_o \in \mathbb{R}^d$ denotes the embedding of organ class $o$. The organ-conditioned mask token for the $i$-th masked patch is defined as: $\mathbf{m}_i = \sum_{o \in \mathcal{O}_i} w_{i,o}\, \mathbf{e}_o$. These organ-conditioned mask tokens are inserted at the masked positions and passed to the ViT decoder together with the encoded visible tokens for reconstruction. As a result, each masked patch is represented by a token that reflects its anatomical composition, rather than by a generic shared mask token. The organ-level masked reconstruction loss is a simple mean squared error computed over the masked patches:

$$\mathcal{L}_{\text{organ-recon}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2, \tag{1}$$

where $\mathcal{M}$ denotes the set of masked patch indices, $|\mathcal{M}|$ is the number of masked patches, $\mathbf{x}_i$ is the original voxel-intensity patch, and $\hat{\mathbf{x}}_i$ is the reconstructed patch predicted by the decoder.

### 3.3 Organ-level Vision-language Alignment

While organ-level masked reconstruction encourages organ-aware reconstruction, it does not explicitly align visual features with organ-level semantics. To inject organ-level semantic information into the visual representations, SOAR introduces a report-free organ-level vision-language alignment objective. Instead of aligning 3D CT volumes with long diagnostic reports, SOAR constructs short

and anatomically focused image-text pairs by pairing each automatically derived organ region with a concise anatomical text prompt. This design preserves the benefit of CLIP-style contrastive alignment while avoiding the ambiguity of whole-volume/report matching.

As detailed in Fig. 2, given a CT volume $\mathbf{X}$ and its segmentation mask $\mathbf{S}$, SOAR first encodes the volume with the ViT encoder to obtain vision tokens $\{\mathbf{z}_i\}_{i=1}^{N}$. The segmentation mask $\mathbf{S}$ is then mapped to the same patch resolution. For each organ class $o$, we group all patch tokens whose corresponding regions overlap with the organ mask. Note that boundary patches can be assigned to multiple organs when they contain voxels from more than one anatomical class.

Since organs vary substantially in size and shape, the number of patch tokens assigned to each organ differs across anatomical classes. To enable batch-wise training, SOAR converts these variable-length token sets into fixed-length organ-level visual representations by introducing a learnable organ-specific query token $\mathbf{q}_o$ for each anatomical class. The query token attends to all visual tokens belonging to organ $o$, producing a single fixed-dimensional embedding that captures the organ's visual semantics regardless of its spatial extent. This module is implemented as a single Transformer block consisting of multi-head self-attention and a feed-forward network. SOAR takes the output at the query-token position as the pooled organ representation. This organ visual embedding is then projected into a shared vision-language embedding space and normalized to obtain organ-level image embeddings $\mathbf{v}_o$.

For the text branch, we use a frozen PubMedBERT [35] encoder to encode concise organ-name descriptions for SOAR. The `CLS` token output from PubMedBERT is passed through a learnable projection layer and normalized to obtain organ-level text embeddings $\mathbf{t}_o$. SOAR then applies a CLIP-style InfoNCE contrastive loss [7] over organ-level image-text pairs. Let $y_i$ denote the anatomical class label of the $i$-th organ-level image-text pair in a batch, and let $\mathcal{A}(i) = \{j \mid y_j = y_i\}$ denote the set of entries that share the same anatomical class as sample $i$. SOAR defines the organ-level semantic loss as:

$$\mathcal{L}_{\text{organ-align}} = \frac{1}{2}\left[-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\sum_{j\in\mathcal{A}(i)}\exp(s_{ij})}{\sum_{j=1}^{B}\exp(s_{ij})} - \frac{1}{B}\sum_{i=1}^{B}\log\frac{\sum_{j\in\mathcal{A}(i)}\exp(s_{ji})}{\sum_{j=1}^{B}\exp(s_{ji})}\right]. \tag{2}$$

Here, $s_{ij}$ denotes the temperature-scaled cosine similarity between the $i$-th organ visual embedding and the $j$-th organ text embedding, defined as $s_{ij} = \mathbf{v}_i^{\top}\mathbf{t}_j/\tau$. $B$ denotes the number of organ-level image-text pairs in the contrastive batch, and $\tau$ is a learnable temperature parameter. This formulation encourages visual embeddings of the same organ class across different patients to align with the corresponding organ-name text embedding, while avoiding incorrect penalization of anatomically identical organ pairs from different volumes.

### 3.4 Organ-level Supervision

Both organ-level masked reconstruction and vision-language alignment operate on patch-level tokens, where voxel-level details are compressed. We hypothesize that strong fine-grained anatomical representations should preserve sufficient voxel-level information to support segmentation. As shown in Fig. 2, SOAR therefore introduces an organ-level supervision constraint by attaching an auxiliary segmentation decoder supervised by the automatically generated masks $\mathbf{S}$, helping the encoder retain voxel-level anatomical information beyond patch-level semantics.

Concretely, the encoded visual tokens $\{\mathbf{z}_i\}_{i=1}^{N}$ are first projected into the decoder space via a linear layer and combined with positional embeddings. A ViT decoder then predicts patch-wise voxel logits, which are unpatchified to recover dense anatomical logits $\hat{\mathbf{S}}$. The segmentation decoder is optimized with a combination of voxel-wise cross-entropy loss and class-wise Dice loss:

$$\mathcal{L}_{\text{organ-super}} = \mathcal{L}_{\text{CE}}(\hat{\mathbf{S}}, \mathbf{S}) + \mathcal{L}_{\text{Dice}}(\hat{\mathbf{S}}, \mathbf{S}). \tag{3}$$

This auxiliary objective is complementary to organ-level masked reconstruction and organ-level vision-language alignment. The first reconstructs masked regions at the patch level, the second aligns organ representations with semantic labels, and the organ-level supervision objective enforces voxel-level spatial discrimination across all organ boundaries. Together, they provide supervision at different levels of granularity, encouraging the encoder to capture both fine-grained spatial detail and high-level anatomical semantics.

### 3.5 Overall Pretraining Loss

The overall SOAR pretraining loss is a combination of the three losses described above:

$$\mathcal{L} = \mathcal{L}_{\text{organ-recon}} + \mathcal{L}_{\text{organ-align}} + \mathcal{L}_{\text{organ-super}}. \tag{4}$$

In practice, these losses are activated through staged training to stabilize optimization and progressively inject anatomical information into the visual representations (details in Appendix A). When multiple losses are active, SOAR uses two forward passes for each input volume: a masked pass for organ-level masked reconstruction and an unmasked pass shared by organ-level vision-language alignment and organ-level supervision. After pretraining, only the ViT encoder is retained and transferred to downstream tasks.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** We use five datasets in this study: CT-RATE [2], 3D-RAD [36], AMOS-MM [37], a public CT pneumonia screening dataset, and an in-house esophageal cancer screening dataset. CT-RATE is used for encoder pretraining. It contains 25,692 non-contrast chest CT scans from 21,304 patients, reconstructed into 50,188 3D volumes. For medical VQA evaluation, we use CT-RATE, 3D-RAD, and AMOS-MM. CT-RATE contains over 2.7 million report-derived VQA pairs, 3D-RAD contains 170,105 question-answer pairs across six chest CT sub-tasks, and AMOS-MM contains 19,562 closed-ended abdominal CT VQA pairs across six task categories. For disease screening and abnormality detection, we use the CT-RATE multi-abnormality labels, a public CT pneumonia dataset with CT data from 3,777 patients for pneumonia screening [38], and an in-house esophageal cancer dataset with 2,648 chest CT studies for binary cancer screening (IRB approval was obtained).

**Implementation details.** For encoder pretraining, CT images are clipped into the range of $[-1000, 1000]$ HU, normalized to $[0, 1]$, and resized to $256\times256\times128$. Each volume is tokenized into non-overlapping 3D patches of size $16\times16\times8$. The vision encoder is a 12-layer ViT with a hidden size of 768 and 12 attention heads, and the text encoder for organ-level image-text alignment is a frozen PubMedBERT [35]. We use TotalSegmentator [13, 33] to generate anatomical segmentation masks for SOAR pretraining. The resulting organ labels are converted into natural-language anatomical text before being used for organ-level vision-language alignment (see Appendix Table 5). For downstream medical VQA fine-tuning, we append the visual tokens to LLaMA-3.1-8B [39] and fine-tune the model using LoRA [40]. For disease screening and abnormality detection, we attach the pretrained vision encoder to a two-layer MLP classification head.

**Evaluation metrics.** We evaluate disease screening and abnormality detection tasks using AUC. For closed-ended VQA, we use accuracy as the primary metric. For open-ended VQA, we report BLEU [41], ROUGE [42], METEOR [43], BERT-F1 [44], and GREEN [45]. BLEU, ROUGE, METEOR, and BERT-F1 assess general language similarity, while GREEN is an LLM-based radiology-focused metric for evaluating clinical faithfulness in the generated text. We therefore prioritize GREEN when interpreting open-ended VQA performance.

**Comparison methods.** We compare SOAR with representative self-supervised and vision-language pretraining methods, including MAE [6], CLIP [7], fVLM [31], ViSD [32], and VoCo [19]. For MAE and CLIP, we train the models ourselves using the same CT-RATE pretraining data and the same 3D ViT-B backbone as SOAR. For fVLM, ViSD, and VoCo, we use their publicly released pretrained checkpoints. These checkpoints are pretrained with data resources that are no smaller than CT-RATE and include CT-RATE as part of their pretraining data. All methods are evaluated using the same downstream protocols described above. Additional experimental details are provided in Appendix A. We additionally evaluate MedVAE [20] and the pretrained TotalSegmentator encoder, with detailed comparisons provided in Appendix B.7.

### 4.2 Medical Imaging VQA

Table 1 presents the main VQA results on three benchmarks. Across the three medical VQA benchmarks, SOAR achieves overall stronger performance compared to SOTA 3D medical visual encoders. On CT-RATE, SOAR achieves the best closed-ended VQA accuracy of 89.87%, improving

over all baselines by 0.67–8.80 percentage points. For open-ended CT-RATE VQA, SOAR obtains the highest GREEN score of 46.83, with gains of 0.42–3.42 points over the baselines, and also achieves the best BLEU-1, ROUGE-1, and BERT-F1 scores. On 3D-RAD, SOAR achieves the best result on most reported metrics, with an average gain of 0.71 points over the next-best method across the 21 reported metrics and average gains of 1.40–3.19 points over the individual baselines. On AMOS-MM, SOAR achieves the highest overall accuracy of 75.38%, improving over all baselines by 0.39–5.45 percentage points. These results show that SOAR improves both closed-ended reasoning and open-ended medical answer generation across chest and abdominal 3D CT VQA benchmarks.

Fig. 3 shows a qualitative open-ended VQA example. In this case, the ground-truth answer identifies linear atelectatic changes in the left upper lobe inferior lingular segment and the right middle lobe. SOAR generates the most anatomically grounded response, correctly capturing both the abnormality type and its lung-region localization. In contrast, the comparison methods either provide coarser localization, such as referring to both lungs or only to lower lobes, or introduce unsupported findings, including nonspecific nodules, mosaic attenuation, or diffuse ground-glass densities. This example illustrates that SOAR better preserves fine-grained anatomical information in visual token representations for 3D medical VQA. Additional qualitative examples for open-ended VQA and report generation are provided in Appendix B. To further evaluate the effectiveness of SOAR, we additionally compare it with MedVAE [20] and the pretrained TotalSegmentator encoder [13], representing autoencoder-based and segmentation-based pretraining approaches, respectively. TotalSegmentator is particularly relevant as it is also used to generate the anatomical pseudo-labels for SOAR. Detailed results are provided in Appendix B.7.

Overall, these results show that SOAR learns stronger 3D CT visual representations for both multiple-choice medical reasoning and clinically grounded open-ended answer generation.

Table 1: Results across three medical VQA benchmarks. Best results are shown in **bold** and second-best results are underlined. **Δ** shows the difference between SOAR and the best baseline performer.

| Bench. | Task | Metric | MAE | CLIP | fVLM | ViSD | VoCo | SOAR | Δ |
|---|---|---|---|---|---|---|---|---|---|
| CT-RATE | Closed-ended VQA | Acc. (%) | 85.76 | 81.07 | 85.65 | 87.78 | 89.20 | **89.87** | ↑0.67 |
| | Open-ended VQA | BLEU-1 | 47.76 | 47.72 | 48.58 | 48.20 | 48.95 | **49.27** | ↑0.32 |
| | | BLEU-2 | 36.94 | 36.57 | 37.79 | 37.21 | **38.30** | 38.19 | ↓0.11 |
| | | ROUGE-1 | 53.40 | 52.86 | 54.01 | 53.68 | 54.63 | **54.65** | ↑0.02 |
| | | METEOR | 50.22 | 50.11 | 50.63 | 50.45 | **51.22** | 51.15 | ↓0.07 |
| | | BERT-F1 | 91.58 | 91.66 | 91.85 | 91.88 | 91.84 | **92.00** | ↑0.12 |
| | | GREEN | 46.18 | 43.41 | 45.98 | 44.80 | 46.41 | **46.83** | ↑0.42 |
| 3D-RAD | Image Observation | BLEU-1 | 43.66 | 43.64 | 42.76 | 43.13 | 40.48 | **44.56** | ↑0.90 |
| | | BLEU-2 | 27.05 | 27.19 | 26.32 | 26.78 | 24.71 | **28.86** | ↑1.67 |
| | | ROUGE-1 | 47.46 | 47.07 | 46.67 | 46.88 | 44.15 | **49.10** | ↑1.64 |
| | | METEOR | 44.43 | 44.66 | 43.92 | 44.26 | 42.51 | **45.62** | ↑0.96 |
| | | BERT-F1 | 92.01 | 92.06 | 91.99 | 91.98 | 91.78 | **92.38** | ↑0.32 |
| | | GREEN | 36.65 | 37.59 | 36.43 | 35.73 | 35.05 | **38.09** | ↑0.50 |
| | Anomaly Detection | BLEU-1 | 40.64 | 39.78 | 39.78 | 39.87 | 39.58 | **41.50** | ↑0.86 |
| | | BLEU-2 | 25.01 | 24.22 | 24.37 | 24.29 | 24.51 | **25.61** | ↑0.60 |
| | | ROUGE-1 | 46.34 | 44.55 | 45.84 | 45.07 | 44.41 | **47.96** | ↑1.62 |
| | | METEOR | 42.94 | 42.17 | 41.99 | 43.15 | 42.99 | **44.42** | ↑1.27 |
| | | BERT-F1 | 91.95 | 91.88 | 91.95 | 92.01 | 92.03 | **92.18** | ↑0.15 |
| | | GREEN | 33.43 | 31.64 | 33.16 | 32.65 | 32.82 | **35.52** | ↑2.09 |
| | Medical Computation | BLEU-1 | 39.90 | 36.57 | 40.06 | 36.38 | 41.77 | **42.47** | ↑0.70 |
| | | BLEU-2 | 2.25 | 1.54 | 1.60 | 1.66 | **3.25** | 3.24 | ↓0.01 |
| | | ROUGE-1 | 40.79 | 37.95 | 41.29 | 37.63 | 42.93 | **43.19** | ↑0.26 |
| | | METEOR | 21.41 | 19.27 | 21.10 | 19.20 | 22.57 | **23.09** | ↑0.52 |
| | | BERT-F1 | 95.81 | 95.54 | 95.58 | 95.40 | 95.83 | **96.00** | ↑0.17 |
| | | GREEN | 5.69 | 4.69 | 6.99 | 5.39 | **8.58** | 7.39 | ↓1.19 |
| | Exist. Detect. | Acc. (%) | 77.18 | 79.78 | 76.91 | 59.76 | 80.11 | **81.54** | ↑1.43 |
| | Static Temp. Diag. | Acc. (%) | 63.28 | 63.14 | 62.76 | 59.21 | 63.70 | **63.91** | ↑0.21 |
| | Longit. Temp. Diag. | Acc. (%) | 80.86 | 78.42 | 81.27 | 80.75 | 80.96 | **81.50** | ↑0.23 |
| AMOS-MM | Anatomical Ident. | Acc. (%) | 71.25 | 71.05 | 65.91 | 66.74 | 67.76 | **71.87** | ↑0.62 |
| | Attribute Recognition | Acc. (%) | 82.76 | **85.64** | 81.69 | 79.71 | 80.07 | 84.20 | ↓1.44 |
| | Disease Analysis | Acc. (%) | 74.83 | 74.14 | 72.54 | 71.62 | 72.31 | **75.74** | ↑0.91 |
| | Quant. Assessment | Acc. (%) | 70.26 | **73.15** | 65.11 | 66.08 | 67.68 | **73.15** | 0.00 |
| | Risk Evaluation | Acc. (%) | 64.63 | 69.02 | 62.44 | 63.66 | 63.90 | **70.00** | ↑0.98 |
| | Therapeutic Suggest. | Acc. (%) | 70.44 | 74.82 | 71.17 | 71.53 | 72.63 | **75.91** | ↑1.09 |
| | Overall Accuracy | Acc. (%) | 72.84 | 74.99 | 69.93 | 69.97 | 70.83 | **75.38** | ↑0.39 |

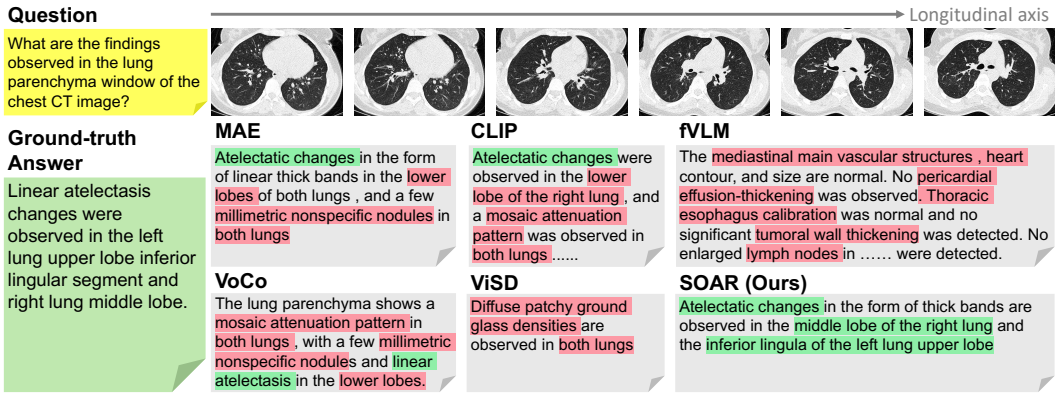



Figure 3: Qualitative VQA comparison across different visual encoders with correct and unsupported information highlighted. Some predictions and CT slices were trimmed to save space. CT display window is [-1000, 100] HU. Full version in Appendix B.2 (Case 2).

### 4.3 Disease Screening and Abnormality Detection

Table 2 presents the disease screening and abnormality detection results. SOAR achieves the best AUC across all evaluated tasks. For pneumonia screening, SOAR obtains an average AUC of 97.99, improving over all baselines by 0.71–9.68 points. For esophageal cancer screening, SOAR achieves an AUC of 95.89, improving over all baselines by 0.59–12.88 points. For CT-RATE abnormality detection, SOAR achieves the best average AUC of 75.38, improving over all baselines by 1.69–12.45 points. These results indicate that the representations learned by SOAR transfer effectively to disease screening and multi-abnormality detection tasks.

Table 2: Disease and abnormality screening results across pneumonia classification, esophageal cancer screening, and CT-RATE abnormality detection benchmarks. All reported values are AUC (%). Best results are shown in **bold**. Δ shows SOAR's improvement over the next-best method.

| Task | MAE | CLIP | fVLM | ViSD | VoCo | SOAR | Δ |
|---|---|---|---|---|---|---|---|
| COVID-19 | 98.08 | 96.74 | 97.50 | 92.21 | 87.14 | **98.79** | ↑0.71 |
| Common Pneumonia | 96.94 | 94.98 | 95.72 | 91.18 | 85.77 | **97.47** | ↑0.53 |
| Normal Class | 96.83 | 97.04 | 96.62 | 94.11 | 92.03 | **97.72** | ↑0.68 |
| Avg. | 97.28 | 96.25 | 96.61 | 92.50 | 88.31 | **97.99** | ↑0.71 |
| Esophageal Cancer | 92.67 | 95.30 | 94.73 | 92.50 | 83.01 | **95.89** | ↑0.59 |
| CT-RATE Abnormality Avg. | 73.69 | 73.36 | 69.07 | 69.88 | 62.93 | **75.38** | ↑1.69 |

### 4.4 Ablation Studies

Table 3 reports the ablation results of SOAR on CT-RATE VQA. Replacing random patch masking with random organ masking improves multiple-choice accuracy from 85.76% to 88.64%, confirming the benefit of organ-level masked reconstruction. Masking 30 randomly selected organs performs best and is used as the default setting. Adding organ-level supervision and organ-conditioned mask tokens further improves accuracy to 89.37% and 89.63%, respectively. The full model achieves the best performance, showing that the proposed organ-aware components provide complementary benefits.

Table 3: Ablation study of SOAR components on the CT-RATE VQA benchmark, including the vanilla MAE baseline. Numbers in parentheses in the organ-level recon column indicate the number of masked organs used during pretraining.

| Components | | | | Multiple-choice Acc. (%) | Open-ended VQA | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Organ-level Recon (# masked organs) | Organ-level Super | Organ-cond. Masked Tokens | Organ-level Align | | BLEU-1 | BLEU-2 | ROUGE-1 | METEOR | BERT-F1 | GREEN |
| Vanilla MAE | | | | 85.76 | 47.76 | 36.94 | 53.40 | 50.22 | 91.58 | 46.18 |
| ✓ (10) | | | | 84.44 | 48.37 | 37.25 | 53.57 | 50.46 | 91.86 | 45.44 |
| ✓ (30) | | | | 88.64 | 48.88 | 38.02 | 54.36 | 50.76 | 91.94 | 46.53 |
| ✓ (60) | | | | 85.79 | 48.16 | 37.63 | 53.82 | 50.46 | 91.86 | 46.37 |
| ✓ (30) | ✓ | | | 89.37 | 49.05 | 38.17 | 54.56 | 50.88 | 91.92 | 46.22 |
| ✓ (30) | ✓ | ✓ | | 89.63 | 48.78 | 37.94 | 54.24 | 50.71 | 91.90 | 46.14 |
| ✓ (30) | ✓ | ✓ | ✓ | **89.87** | **49.27** | **38.19** | **54.65** | **51.15** | **92.00** | **46.83** |

## 5 Conclusion

We present **SOAR**, a conceptually simple yet effective vision pretraining framework for 3D CT image understanding. SOAR progressively injects anatomical information into visual representations by first learning organ-level structural context through masked reconstruction, then preserving fine-grained voxel-level spatial details through organ-level supervision, and finally introducing semantic grounding through report-free organ-level vision-language alignment. Across various downstream tasks, SOAR learns more transferable 3D CT representations than representative self-supervised and vision-language pretraining baselines. These results suggest that anatomical structure provides an effective and scalable supervision source for bridging voxel-level visual patterns, organ-level semantics, and downstream report-style reasoning in 3D medical image pretraining.

We further discuss the limitations and future directions of SOAR in Appendix C, and its potential societal impacts in Appendix D.

## Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62601185, T2596013, and 62471148), the National Science and Technology Major Project for Noncommunicable Chronic Diseases (Grant Nos. 2025ZD0544800, 2025ZD0544801, and 2025ZD0544802), the STI2030-Major Projects (No. 2021ZD0200204), and the Shanghai Center for Brain Science and Brain-inspired Technology.

## References

[1] L. Blankemeier, A. Kumar, J. P. Cohen, J. Liu, L. Liu, D. Van Veen, S. J. S. Gardezi, H. Yu, M. Paschali, Z. Chen, J.-B. Delbrouck, E. Reis, R. Holland, C. Truyts, C. Bluethgen, Y. Wu, L. Lian, M. E. K. Jensen, S. Ostmeier, M. Varma, J. M. J. Valanarasu, Z. Fang, Z. Huo, Z. Nabulsi, D. Ardila, W.-H. Weng, E. Amaro Junior, N. Ahuja, J. Fries, N. H. Shah, G. Zaharchuk, M. Willis, A. Yala, A. Johnston, R. D. Boutin, A. Wentland, C. P. Langlotz, J. Hom, S. Gatidis, and A. S. Chaudhari, "Merlin: a computed tomography vision–language foundation model and dataset," *Nature*, vol. 652, no. 8112, pp. 1318–1328, 2026.

[2] I. E. Hamamci, S. Er, C. Wang, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, O. F. Durugol, B. Hou, S. Shit, W. Dai, M. Xu, H. Reynaud, M. F. Dasdelen, B. Wittmann, T. Amiranashvili, E. Simsar, M. Simsar, E. B. Erdemir, A. Alanbay, A. Sekuboyina, B. Lafci, A. Kaplan, Z. Lu, M. Polacin, B. Kainz, C. Bluethgen, K. Batmanghelich, M. K. Ozdemir, and B. Menze, "Generalist foundation models from a multimodal dataset for 3D computed tomography," *Nature Biomedical Engineering*, vol. 10, no. 8, pp. 1610–1628, 2026.

[3] R. Alexander, S. Waite, M. A. Bruno, E. A. Krupinski, L. Berlin, S. Macknik, and S. Martinez-Conde, "Mandating Limits on Workload, Duty, and Speed in Radiology," *Radiology*, vol. 304, no. 2, pp. 274–282, Aug. 2022.

[4] R. Smith-Bindman, P. W. Chu, H. Azman Firdaus *et al.*, "Projected lifetime cancer risks from current computed tomography imaging," *JAMA Internal Medicine*, vol. 185, no. 6, pp. 710–719, 2025.

[5] A. Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *BMJ*, vol. 359, p. j4683, Oct. 2017.

[6] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 000–16 009.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8748–8763.

[8] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image classification and segmentation," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–6.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, vol. 9351. Springer, 2015, pp. 234–241.

[10] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[11] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 574–584.

[12] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of Swin transformers for 3D medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 730–20 740.

[13] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, "TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. e230024, 2023.

[14] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3D medical images by playing a Rubik's Cube," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer, 2019, pp. 420–428.

[15] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3D self-supervised methods for medical imaging," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 18 158–18 172.

[16] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models Genesis," *Medical Image Analysis*, vol. 67, p. 101840, 2021.

[17] H.-Y. Zhou, C. Lu, S. Yang, X. Han, and Y. Yu, "Preservational learning improves self-supervised medical image models by reconstructing diverse contexts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3499–3509.

[18] Y. He, G. Yang, R. Ge, Y. Chen, J.-L. Coatrieux, B. Wang, and S. Li, "Geometric visual similarity learning in 3D medical image self-supervised pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9538–9547.

[19] L. Wu, J. Zhuang, and H. Chen, "Large-scale 3D medical image pre-training with geometric context priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 3, pp. 3801–3818, 2026.

[20] M. Varma, A. Kumar, R. van der Sluijs, S. Ostmeier, L. Blankemeier, P. J. M. Chambon, C. Bluethgen, J. Prince, C. Langlotz, and A. S. Chaudhari, "MedVAE: Efficient automated interpretation of medical images with large-scale generalizable autoencoders," in *Proceedings of The 8th International Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, vol. 301. PMLR, 2026, pp. 1597–1626. [Online]. Available: https://proceedings.mlr.press/v301/varma26a.html

[21] Y. Gao, Y. Zhang, C. Wang, J. Liu, M. Varma, J.-B. Delbrouck, A. Chaudhari, and C. Langlotz, "Learning generalizable 3D medical image representations from mask-guided self-supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026, pp. 13 744–13 754.

[22] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang *et al.*, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[23] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, "Long-CLIP: Unlocking the long-text capability of CLIP," in *European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 310–325.

[24] M.-A. Lavoie, A. Mahmoud, A. Zaimi, A. F. Tchango, and S. L. Waslander, "CLIP Is Shortsighted: Paying Attention Beyond the First Sentence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026, pp. 9524–9534.

[25] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3876–3887.

[26] D. Ma, J. Pang, M. B. Gotway, and J. Liang, "A fully open AI foundation model applied to chest radiography," *Nature*, vol. 643, pp. 488–498, 2025.

[27] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Proceedings of the 7th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, Z. Lipton, R. Ranganath, M. Sendak, M. Sjoding, and S. Yeung, Eds., vol. 182. PMLR, 05–06 Aug 2022, pp. 2–25.

[28] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3942–3951.

[29] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 33 536–33 549.

[30] C. Liu, C. Ouyang, S. Cheng, A. Shah, W. Bai, and R. Arcucci, "G2D: From Global to Dense Radiography Representation Learning via Vision-Language Pre-training," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[31] Z. Shui, J. Zhang, W. Cao, S. Wang, R. Guo, L. Lu, L. Yang, X. Ye, T. Liang, Q. Zhang, and L. Zhang, "Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding," in *The Thirteenth International Conference on Learning Representations*, 2025.

[32] W. Cao, J. Zhang, Z. Shui, S. Wang, Z. Chen, X. Li, L. Lu, X. Ye, Q. Zhang, T. Liang, and L. Zhang, "Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 23 041–23 050.

[33] J. Wasserthal, "Dataset with Segmentations of 117 Important Anatomical Structures in 1228 CT Images," Zenodo, 2023, version 2.0.1. DOI: 10.5281/zenodo.10047292.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[35] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2021.

[36] X. Gai, J. Liu, Y. Li, Z. Meng, J. Wu, and Z. Liu, "3D-RAD: A Comprehensive 3D Radiology Med-VQA Dataset with Multi-Temporal Analysis and Diverse Diagnostic Tasks," in *Advances in Neural Information Processing Systems*, 2025, Datasets and Benchmarks Track.

[37] Y. Ji, C. Ge, R. Li, and P. Luo, "AMOS-MM: Abdominal Multimodal Analysis Challenge: Structured description of the challenge design," 2024, MICCAI 2024 Challenge.

[38] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang *et al.*, "Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433.e11, 2020.

[39] A. Grattafiori *et al.*, "The Llama 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024.

[40] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318.

[42] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[43] S. Banerjee and A. Lavie, "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*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72.

[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *International Conference on Learning Representations*, 2020.

[45] S. Ostmeier, J. Xu, Z. Chen, M. Varma, L. Blankemeier, C. Bluethgen, A. E. Michalson, M. Moseley, C. Langlotz, A. S. Chaudhari, and J.-B. Delbrouck, "GREEN: Generative radiology report evaluation and error notation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 374–390.

[46] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

# Appendix


A. Experimental Details

B. Additional Method Comparisons and Experimental Results

B.1 Systematic Comparison with Related 3D CT Pretraining Methods

B.2 Qualitative Medical VQA Analysis

B.3 Qualitative Radiology Report Generation Analysis

B.4 Anatomical Label Set for Organ-Level Vision-Language Alignment

B.5 Reconstruction and Segmentation Visualization

B.6 VQA Results with Statistical Testing

B.7 Additional Baseline Comparisons

B.8 Ablation Study of Organ-Level Semantic Embeddings

B.9 Sensitivity to Distorted Anatomical Pseudo-Labels

C. Limitations and Future Work

D. Societal Impacts


## A Experimental details

**SOAR Pretraining.** Pretraining is performed only on the official CT-RATE training split, which contains 20,000 patients and 47,149 reconstructed CT volumes. The vision encoder is a 12-layer Vision Transformer (ViT) with a hidden size of 768, 12 attention heads, an MLP dimension of 3072, and perceptron-based positional embeddings. Each input CT volume is resized to $256 \times 256 \times 128$ and tokenized into non-overlapping 3D patches of size $16 \times 16 \times 8$. The text encoder is a frozen PubMedBERT [35]. The MAE decoder and segmentation decoder use the same ViT-based decoder architecture, with a hidden dimension of 1080. For organ-level masked reconstruction in SOAR, we randomly select and mask 30 organs in each training sample. For the MAE baseline, we use a 75% random patch masking ratio. The model is optimized using AdamW [46] with a weight decay of 0.1. The learning rate is linearly warmed up to $1 \times 10^{-4}$ during the first 3% of training steps and then decayed to zero using a cosine schedule. Pretraining is performed with a batch size of 48 on 8 H200 GPUs. During training, we apply random $90^{\circ}$ rotations, random flips along each spatial axis with a probability of 10% per axis, random intensity scaling, and random intensity shifting. For the organ-level masked reconstruction part, because we mask organs rather than a fixed number of patches, organ size variation leads to sample-specific numbers of visible and masked patches; we therefore process each volume independently through the encoder instead of padding token sequences to a fixed batch length.

SOAR training proceeds in three stages to stabilize optimization. In the first stage, the model is pretrained with the organ-level masked reconstruction objective alone to establish anatomy-aware reconstruction capability. In the second stage, the segmentation loss is introduced alongside organ-level masked reconstruction, encouraging the encoder to further retain fine-grained spatial discrimination. In the third stage, the organ-level CLIP objective is added, aligning the learned visual representations with organ-level semantic embeddings.

**Fine-tuning Details for Medical VQA Benchmarks.** For downstream medical VQA fine-tuning, we use the official train/test split of each benchmark. We append the visual tokens to LLaMA-3.1-8B [39] and fine-tune the model using LoRA [40]. LoRA is applied with a rank of 16, a scaling factor of 32, and a dropout rate of 0.05. Visual features are directly projected into the LLM embedding space without token compression, allowing the evaluation to focus on the quality of the learned visual token representations. We optimize the trainable parameters with an initial learning rate of $5 \times 10^{-5}$, while the vision encoder is updated with a 10-times smaller learning rate of $5 \times 10^{-6}$.

**Disease Screening and Abnormality Detection.** For disease screening and CT-RATE abnormality detection, we freeze the pretrained vision encoder and train a two-layer MLP classification head. The tasks are formulated according to their label spaces: 3-class classification for pneumonia detection, binary classification for esophageal cancer screening, and multi-label classification for CT-RATE abnormality detection. For pneumonia detection and esophageal cancer screening, we use 20% of the available labeled data to train the classification head. For CT-RATE abnormality detection, we use the

same official train/test split as encoder pretraining. Across all tasks, the classification head is trained for 10 epochs using AdamW [46] with a learning rate of $1 \times 10^{-4}$, a weight decay of 0.05, a dropout rate of 0.3, and a batch size of 5. The learning rate is decayed by a factor of 0.5 every 2 epochs.

# B Additional Method Comparisons and Experimental Results

## B.1 Systematic Comparison with Related 3D CT Pretraining Methods

Table 4 summarizes the key methodological differences between SOAR and representative 3D CT pretraining approaches. Unlike MAE and VoCo, which do not explicitly align anatomical regions with text, SOAR introduces organ-level vision-language alignment using anatomical text embeddings. Unlike CLIP, which operates at the global image-report level, and fVLM and ViSD, which use report-derived regional descriptions, SOAR does not require paired diagnostic reports.

Importantly, SOAR retains the original 117-class anatomical granularity provided by TotalSegmentator, rather than merging fine-grained structures into broader report-aligned regions. By organizing masked reconstruction, semantic alignment, and structural supervision around these anatomical units, SOAR provides complementary anatomical supervision at both organ and voxel levels. The key distinction is therefore not the introduction of entirely new pretraining objectives, but their integration around fine-grained anatomical structures as the fundamental units of supervision, enabling report-free semantic grounding while preserving detailed anatomical information.

Table 4: Systematic comparison of representative 3D CT visual encoder pretraining methods in terms of anatomical granularity, language supervision, and dependence on paired diagnostic reports.

| Method | Anatomical granularity | Language supervision | Reports required |
|---|---|---|---|
| MAE | None | None | No |
| CLIP | Global | Whole radiology report | Yes |
| fVLM | Coarse anatomical groups | Report-derived regional text | Yes |
| ViSD | Coarse anatomical groups | Report-derived regional text | Yes |
| VoCo | Spatial context | None | No |
| **SOAR** | **117 fine-grained anatomical structures** | **Organ-name embeddings** | **No** |

## B.2 Qualitative Medical VQA Analysis

We present eight qualitative examples from the medical VQA task to compare different vision encoder pretraining strategies.

For visualization, **green** shading highlights findings in the ground-truth answers and correct model predictions, while **red** and **gray** shading indicate incorrect or unsupported predictions and partially correct but anatomically imprecise predictions, respectively.

**Case 1: Atelectasis localization.**

> **Question:** Are there any areas of atelectasis in the lungs?
> **Ground Truth:** Yes, there are linear **atelectasis** areas in the **left lung upper lobe lingular segment** and **right lung middle lobe medial segment**.
> **MAE:** Yes, there are areas of **atelectasis** in **both lungs**.
> **CLIP:** Yes, there are areas of **atelectasis** in the **lower lobes** of **both lungs**.
> **fVLM:** The **nodules** are millimetric and nonspecific.
> **ViSD:** Yes, there are areas of **atelectasis** in **both lungs**.
> **VoCo:** Yes, there are areas of linear **atelectasis** in **both lungs**.
> **SOAR (ours):** Yes, there are areas of linear **atelectasis** in the **right lung middle lobe medial segment** and **left lung upper lobe lingular segment**.

In this example, SOAR closely matches the ground truth by identifying both the right lung middle lobe medial segment and the left lung upper lobe lingular segment. In contrast, other pretraining

strategies either provide generic bilateral lung findings, incorrectly localize the finding to the lower lobes, or generate an unrelated response.

**Case 2: Lung parenchyma findings.**

> **Question:** What are the findings observed in the lung parenchyma window of the Chest CT image?
>
> **Ground Truth:** **Linear atelectasis changes** were observed in the **left lung upper lobe inferior lingular segment** and **right lung middle lobe**.
>
> **MAE:** **Atelectatic changes** in the form of linear thick bands in the **lower lobes** of **both lungs**, and a few **millimetric nonspecific nodules** in **both lungs**.
>
> **CLIP:** **Atelectatic changes** were observed in the **lower lobe of the right lung**, and a **mosaic attenuation pattern** was observed in **both lungs** (small airway disease? small vessel disease?).
>
> **fVLM:** The **mediastinal main vascular structures**, **heart contour**, and size are normal. No **pericardial effusion-thickening** was observed. **Thoracic esophagus** calibration was normal and no significant **tumoral wall thickening** was detected. No enlarged **lymph nodes** in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.
>
> **ViSD:** **Diffuse patchy ground glass densities** are observed in **both lungs**.
>
> **VoCo:** The lung parenchyma shows a **mosaic attenuation pattern** in **both lungs**, with a few **millimetric nonspecific nodules** and **linear atelectasis** in the **lower lobes**.
>
> **SOAR (ours):** **Atelectatic changes** in the form of thick bands are observed in the **middle lobe of the right lung** and the **inferior lingula of the left lung upper lobe**.

SOAR correctly captures the main finding of atelectatic change and localizes it to the right middle lobe and left upper lobe inferior lingula. The baseline pretraining strategies either assign the finding to the lower lobes, introduce additional unsupported findings, or describe unrelated normal mediastinal structures.

**Case 3: Atelectasis detection and localization.**

> **Question:** Are there any atelectasis in the lungs?
>
> **Ground Truth:** **Atelectasis** are observed in the **middle lobe of the right lung** and the **lingular segment of the left lung upper lobe**.
>
> **MAE:** Yes, **atelectasis** is observed in the **lower lobe of the left lung**.
>
> **CLIP:** Yes, **atelectasis** is observed in the **lower lobe of the left lung**.
>
> **fVLM:** Yes, there are **nonspecific nodules** in **both lungs**, with the largest being **5 mm in diameter in the right lung**.
>
> **ViSD:** Yes, **atelectasis** is observed in the **lower lobes** of **both lungs**.
>
> **VoCo:** Yes, there are linear **atelectasis** in **both lungs**.
>
> **SOAR (ours):** Yes, linear **atelectasis** is observed in the **right lung middle lobe medial segment** and **left lung upper lobe lingular segment**.

This case further demonstrates SOAR's ability to recover specific anatomical locations even when the question does not provide location cues. SOAR localizes atelectasis to the right middle lobe medial segment and left upper lobe lingular segment, while the baseline pretraining strategies provide incorrect lower-lobe localization, generic bilateral findings, or unrelated nodule descriptions.

**Case 4: Consolidation or atelectasis.**

> **Question:** Are there any signs of consolidation or atelectasis in the lungs?
>
> **Ground Truth:** Yes, appearances that can be evaluated in favor of **consolidations or atelectasis** are observed in the **right lung middle lobe** and **left lung upper lobe lingular segment**.
>
> **MAE:** **No**, no **consolidation or atelectasis** was detected in the lung parenchyma.
>
> **CLIP:** **No**, no **consolidation or atelectasis** was detected in the lung parenchyma.
>
> **fVLM:** Yes, **minimal emphysematous changes** are observed in **both lungs**.
>
> **ViSD:** Yes, there are areas of **consolidation and atelectasis** in the **lower lobes** of **both lungs**.

**VoCo:** **No**, there is no finding in favor of **consolidation or atelectasis** in the lung parenchyma.
**SOAR (ours):** Yes, **consolidation and atelectasis** are observed in the **right lung middle lobe medial segment** and **left lung upper lobe lingular segment**.

SOAR identifies the presence of the relevant abnormality and localizes it to the right middle lobe medial segment and left upper lobe lingular segment. MAE, CLIP, and VoCo incorrectly report no consolidation or atelectasis, while ViSD detects the abnormality but localizes it to the lower lobes. Although the ground truth uses the broader phrase "consolidations or atelectasis," SOAR provides the closest anatomical match among the competing pretraining strategies.

**Case 5: Pleuroparenchymal sequelae density increases.**

**Question:** Were any pleuroparenchymal sequelae density increases observed in the lungs?
**Ground Truth:** Yes, **increases** were observed in the **middle lobe of the right lung** and the **inferior lingular segment of the left lung**.
**MAE:** Yes, they were observed in **both lungs**.
**CLIP:** Yes, they were observed in **both lungs**.
**fVLM:** **No**, **pericardial thickening-effusion** was not detected.
**ViSD:** Yes, they were observed in **both lungs**.
**VoCo:** Yes, **pleuroparenchymal sequelae density increases** were observed in **both lungs apical**.
**SOAR (ours):** Yes, **pleuroparenchymal sequelae density increases** were observed in the **right lung middle lobe medial segment** and **left lung inferior lingular segment**.

SOAR accurately identifies the presence of pleuroparenchymal sequelae density increases and provides the correct lobe-level localization. The baseline pretraining strategies mostly return generic bilateral lung findings, an irrelevant negative cardiac finding, or an incorrect apical localization.

**Case 6: Upper-abdominal findings and unsupported abnormalities.**

**Question:** What is observed in the upper abdomen sections?
**Ground Truth:** **No feature was observed** in the upper abdomen sections.
**MAE:** A **3, mm diameter calculus** is observed in the **upper pole of the left kidney**.
**CLIP:** A **16, mm diameter nodular lesion** is observed in the **left adrenal gland**, which is evaluated in favor of **adenoma**.
**fVLM:** A **2, mm diameter calculus** is observed in the **middle part of the left kidney**.
**ViSD:** A **2, mm diameter calculus** was observed in the **upper pole of the left kidney**.
**VoCo:** A **3, mm diameter calculus** is observed in the **middle zone of the left kidney**.
**SOAR (ours):** **No features were detected** in the upper abdomen sections.

The reference answer reports no upper-abdominal findings. In contrast, MAE, fVLM, ViSD, and VoCo describe renal calculi, while CLIP describes a left adrenal nodule interpreted as a possible adenoma. None of these abnormalities is supported by the reference answer. SOAR instead correctly produces a negative response, illustrating reduced unsupported positive findings in this example.

**Case 7: Bilateral adrenal gland assessment.**

**Question:** What type of changes were observed in the bilateral adrenal glands?
**Ground Truth:** Bilateral adrenal gland calibration was **normal**.
**MAE:** **Diffuse thickening** was observed in the bilateral adrenal glands.
**CLIP:** **Diffuse thickening** was observed in both adrenal glands.
**fVLM:** **Diffuse thickening** was observed in the bilateral adrenal glands.
**ViSD:** **Diffuse thickening** was observed in both adrenal glands.
**VoCo:** **Nodular thickening** was observed in the bilateral adrenal glands.
**SOAR (ours):** Bilateral adrenal glands were **normal**, and **no space-occupying lesion was detected**.

In this case, all five baselines incorrectly describe adrenal gland thickening despite the normal appearance stated in the reference answer: MAE, CLIP, fVLM, and ViSD report diffuse thickening, whereas VoCo reports nodular thickening. By contrast, SOAR correctly identifies normal bilateral adrenal glands and reports no space-occupying lesion. This example further illustrates a reduction in unsupported abnormalities rather than merely a gain in positive-lesion recognition.

**Case 8: Fine-grained renal lesion localization.**

> **Question:** Where is the hypodense, well-circumscribed lesion located?
> **Ground Truth:** **Upper pole of the right kidney**.
> **MAE:** **Lower pole** of the **right kidney**.
> **CLIP:** **Left kidney**.
> **fVLM:** **Right kidney**.
> **ViSD:** **Right kidney**.
> **VoCo:** **Left adrenal gland**.
> **SOAR (ours):** **Right kidney upper pole**.

This example highlights suborgan-level localization. MAE identifies the correct kidney but incorrectly places the lesion in the lower pole. CLIP selects the contralateral kidney, while VoCo selects the left adrenal gland. fVLM and ViSD correctly identify the right kidney but omit its upper-pole location. In comparison, SOAR matches both the organ and the precise anatomical subregion in the reference answer.

Overall, these examples suggest that SOAR more faithfully represents anatomical information in 3D CT images.

### B.3 Qualitative Radiology Report Generation Analysis

We provide a full report-generation example to qualitatively compare SOAR with baseline encoders. In the representative case shown below, the ground-truth report identifies minimal emphysematous changes in both lungs as the primary impression, with additional findings including patent central airways, right middle-lobe linear atelectasis, absence of lung mass or infiltrative lesion, and no pleural or pericardial effusion. SOAR correctly preserves the primary abnormality of minimal bilateral emphysematous changes and maintains an anatomically organized description across the airway, lung parenchyma, mediastinum, heart, pleural space, esophagus, upper abdomen, and osseous structures. In contrast, MAE, fVLM, and VoCo over-normalize the case as a normal thoracic CT examination, while CLIP and ViSD introduce unsupported pneumonia-related findings.

Notably, SOAR also generates several additional organ-specific negative findings that are not explicitly stated in the ground-truth report, including the absence of an upper abdominal mass, preserved intervertebral disc spaces, and no lytic or destructive osseous lesions. This observation suggests that SOAR tends to produce broader organ-context descriptions than the baseline encoders, which may reflect the effect of organ-aware pretraining in encouraging the LLM to incorporate more anatomical context during report generation. However, although this qualitative example suggests improved preservation of abnormal findings and reduced severe hallucinations compared with baseline encoders, extensive clinical validation is still needed to rigorously assess the clinical faithfulness and reliability of the generated reports.

**Ground Truth Report.**

> **Findings:**
> Trachea and both main bronchi are **open**. There is **no obstructive pathology** in the trachea and both main bronchi.
> There are **minimal emphysematous changes in both lungs**. **Linear atelectasis** was observed in the **middle lobe of the right lung**.
> **No mass or infiltrative lesion** was detected in both lungs. Mediastinal structures cannot be evaluated optimally because contrast material is not given.
> As far as can be observed: Heart contour and size are **normal**. **No pleural or pericardial effusion** was detected.
> The widths of the mediastinal main vascular structures are **normal**. **No pathologically enlarged lymph nodes** were detected in the mediastinum and hilar regions.

**No pathological wall thickness increase** was observed in the esophagus within the sections. **No upper abdominal free fluid-collection** or **pathologically enlarged lymph nodes** were observed in the sections.
Thoracic vertebral corpus heights, alignments and densities are **normal**. There are **millimetric osteophytes** in the vertebra corpus corners. The neural foramina are **open**.
**Impression:**
**Minimal emphysematous changes in both lungs**.

### MAE Prediction.

**Findings:**
Trachea, both main bronchi are **open**. Mediastinal main vascular structures, heart contour, size are **normal**. Thoracic aorta diameter is **normal**.
**Pericardial effusion**-thickening was not observed. Thoracic esophagus calibration was normal and **no significant tumoral wall thickening** was detected.
**No enlarged lymph nodes** in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.
When examined in the lung parenchyma window; **Aeration of both lung parenchyma is normal** and no nodular or **infiltrative lesion** is detected in the lung parenchyma.
**Pleural effusion**-thickening was not detected. Upper abdominal organs included in the sections are normal.
No space-occupying lesion was detected in the liver. Bilateral adrenal glands were normal and no space-occupying lesion was detected.
Bone structures in the study area are natural. Vertebral corpus heights are **preserved**.
**Impression:**
**Thoracic CT examination within normal limits**.

### CLIP Prediction.

**Findings:**
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. Esophagus is **within normal limits**. **Pleural effusion**-thickening was not detected in both hemithorax.
In the evaluation of both lung parenchyma; **Patchy**, **peripheral-subpleural**, **ground glass density**, **crazy paving appearances** and **consolidations** were observed in both lungs.
**Viral pneumonia?** There are **cylindrical bronchiectasis** and **vascular enlargement** in the affected areas. **CT involvement score** was evaluated as **mild**.
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.
There are **degenerative changes** in bone structures.
**Impression:**
**Viral pneumonia?** Outlooks include **classic or probable findings for COVID**.
**Atherosclerosis** Note: Other infectious agents such as influenza, parainfluenza, mycoplasma, other organized pneumonias such as drug toxicity, connective tissue diseases should be consided in the differential diagnosis as they may cause similar appearances.

### fVLM Prediction.

**Findings:**
Trachea, both main bronchi are **open**. Mediastinal main vascular structures, heart contour, size are **normal**. Thoracic aorta diameter is **normal**.
**Pericardial effusion**-thickening was not observed. Thoracic esophagus calibration was normal and **no significant tumoral wall thickening** was detected.
**No enlarged lymph nodes** in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.
When examined in the lung parenchyma window; **Aeration of both lung parenchyma is normal** and no nodular or **infiltrative lesion** is detected in the lung parenchyma.
**Pleural effusion**-thickening was not detected. Upper abdominal organs included in the sections are normal.
No space-occupying lesion was detected in the liver. Bilateral adrenal glands were normal and no space-occupying lesion was detected.

Bone structures in the study area are natural. Vertebral corpus heights are **preserved**.

**Impression:**
**Thoracic CT examination within normal limits**.

### ViSD Prediction.

**Findings:**
Trachea, both main bronchi are **open**. Mediastinal main vascular structures, heart contour, size are **normal**. Thoracic aorta diameter is **normal**.
**Pericardial effusion**-thickening was not observed. Thoracic esophagus calibration was normal and **no significant tumoral wall thickening** was detected.
**No enlarged lymph nodes** in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.
When examined in the lung parenchyma window; In both lung parenchyma, there are **ground glass densities** with a tendency to merge, being more prominent in the **lower lobes and posterior parts**.
Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver.
Bilateral adrenal glands were normal and no space-occupying lesion was detected. Bone structures in the study area are natural. Vertebral corpus heights are **preserved**.

**Impression:**
**Findings consistent with Covid pneumonia in both lungs**.

### VoCo Prediction.

**Findings:**
Trachea, both main bronchi are **open**. Mediastinal main vascular structures, heart contour, size are **normal**. Thoracic aorta diameter is **normal**.
**Pericardial effusion**-thickening was not observed. Thoracic esophagus calibration was normal and **no significant tumoral wall thickening** was detected.
**No enlarged lymph nodes** in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.
When examined in the lung parenchyma window; **Aeration of both lung parenchyma is normal** and no nodular or **infiltrative lesion** is detected in the lung parenchyma.
**Pleural effusion**-thickening was not detected. Upper abdominal organs included in the sections are normal.
No space-occupying lesion was detected in the liver. Bilateral adrenal glands were normal and no space-occupying lesion was detected.
Bone structures in the study area are natural. Vertebral corpus heights are **preserved**.

**Impression:**
**Thoracic CT examination within normal limits**.

### SOAR (Ours) Prediction.

**Findings:**
Trachea and both main bronchi are **open**. **No occlusive pathology** was detected in the trachea and both main bronchi.
There are **minimal emphysematous changes in both lungs**. **No mass or infiltrative lesion** was detected in both lungs.
Mediastinal structures cannot be evaluated optimally because contrast material is not given. As far as can be observed: Heart contour and size are **normal**.
**No pleural or pericardial effusion** was detected. The widths of the mediastinal main vascular structures are **normal**.
**No pathologically enlarged lymph nodes** were detected in the mediastinum and hilar regions. **No pathological wall thickness increase** was observed in the esophagus within the sections.
**No upper abdominal free fluid-collection** was detected in the sections. **No enlarged lymph nodes in pathological dimensions** were detected.
In the upper abdominal organs within the sections, there is no mass with distinguishable borders as far as it can be observed within the borders of non-enhanced CT.
Thoracic vertebral corpus heights, alignments and densities are **normal**. Intervertebral disc distances are preserved. The neural foramina are **open**.
No lytic-destructive lesions were detected in the bone structures within the sections.

**Impression:**

**Minimal emphysematous changes in both lungs**.

### B.4 Anatomical Label Set for Organ-Level Vision-Language Alignment

Table 5 summarizes the anatomical labels used for organ-level vision-language alignment. The mapping is derived from TotalSegmentator and includes background, major organs, lung lobes, vessels, bones, muscles, and other anatomical structures. These organ identities are used to define the anatomical classes for organ-level vision-language alignment training, enabling SOAR to associate segmentation-derived organ regions with their corresponding organ names.

Table 5: Organ ID and organ name mapping derived from TotalSegmentator and used for organ-level CLIP training.

| ID | Organ name | ID | Organ name | ID | Organ name | ID | Organ name |
|---|---|---|---|---|---|---|---|
| 0 | Background | 30 | L2 vertebra | 60 | Right brachiocephalic vein | 90 | Brain |
| 1 | Spleen | 31 | L1 vertebra | 61 | Left atrial appendage | 91 | Skull |
| 2 | Right kidney | 32 | T12 vertebra | 62 | Superior vena cava | 92 | Left rib 1 |
| 3 | Left kidney | 33 | T11 vertebra | 63 | Inferior vena cava | 93 | Left rib 2 |
| 4 | Gallbladder | 34 | T10 vertebra | 64 | Portal vein and splenic vein | 94 | Left rib 3 |
| 5 | Liver | 35 | T9 vertebra | 65 | Left iliac artery | 95 | Left rib 4 |
| 6 | Stomach | 36 | T8 vertebra | 66 | Right iliac artery | 96 | Left rib 5 |
| 7 | Pancreas | 37 | T7 vertebra | 67 | Left iliac vein | 97 | Left rib 6 |
| 8 | Right adrenal gland | 38 | T6 vertebra | 68 | Right iliac vein | 98 | Left rib 7 |
| 9 | Left adrenal gland | 39 | T5 vertebra | 69 | Left humerus | 99 | Left rib 8 |
| 10 | Left upper lobe of lung | 40 | T4 vertebra | 70 | Right humerus | 100 | Left rib 9 |
| 11 | Left lower lobe of lung | 41 | T3 vertebra | 71 | Left scapula | 101 | Left rib 10 |
| 12 | Right upper lobe of lung | 42 | T2 vertebra | 72 | Right scapula | 102 | Left rib 11 |
| 13 | Right middle lobe of lung | 43 | T1 vertebra | 73 | Left clavicle | 103 | Left rib 12 |
| 14 | Right lower lobe of lung | 44 | C7 vertebra | 74 | Right clavicle | 104 | Right rib 1 |
| 15 | Esophagus | 45 | C6 vertebra | 75 | Left femur | 105 | Right rib 2 |
| 16 | Trachea | 46 | C5 vertebra | 76 | Right femur | 106 | Right rib 3 |
| 17 | Thyroid gland | 47 | C4 vertebra | 77 | Left hip | 107 | Right rib 4 |
| 18 | Small bowel | 48 | C3 vertebra | 78 | Right hip | 108 | Right rib 5 |
| 19 | Duodenum | 49 | C2 vertebra | 79 | Spinal cord | 109 | Right rib 6 |
| 20 | Colon | 50 | C1 vertebra | 80 | Left gluteus maximus muscle | 110 | Right rib 7 |
| 21 | Urinary bladder | 51 | Heart | 81 | Right gluteus maximus muscle | 111 | Right rib 8 |
| 22 | Prostate gland | 52 | Aorta | 82 | Left gluteus medius muscle | 112 | Right rib 9 |
| 23 | Left renal cyst | 53 | Pulmonary vein | 83 | Right gluteus medius muscle | 113 | Right rib 10 |
| 24 | Right renal cyst | 54 | Brachiocephalic trunk | 84 | Left gluteus minimus muscle | 114 | Right rib 11 |
| 25 | Sacrum | 55 | Right subclavian artery | 85 | Right gluteus minimus muscle | 115 | Right rib 12 |
| 26 | S1 vertebra | 56 | Left subclavian artery | 86 | Left paraspinal muscle | 116 | Sternum |
| 27 | L5 vertebra | 57 | Right common carotid artery | 87 | Right paraspinal muscle | 117 | Costal cartilages |
| 28 | L4 vertebra | 58 | Left common carotid artery | 88 | Left iliopsoas muscle | | |
| 29 | L3 vertebra | 59 | Left brachiocephalic vein | 89 | Right iliopsoas muscle | | |

To examine anatomical coverage and potential class imbalance in CT-RATE, we analyze the organ-presence statistics derived from TotalSegmentator across 50,185 valid CT-RATE volumes. Table 6 reports the frequency of each of the 117 anatomical structures. Overall, 69 structures appear in at least 98% of volumes, and 82 appear in at least 90%. While most thoracic and abdominal organs exhibit high coverage, structures outside the typical CT acquisition field of view are less frequently observed, reflecting the anatomical distribution of the dataset.

Table 6: Anatomical structure coverage across CT-RATE volumes, based on automatically generated TotalSegmentator masks. Coverage (Cov.) indicates the percentage of volumes containing each structure, and # Vol. denotes the corresponding number of volumes. The analysis includes the full dataset rather than only the pretraining split.

| ID | Organ | Cov. (%) | # Vol. | ID | Organ | Cov. (%) | # Vol. | ID | Organ | Cov. (%) | # Vol. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | spleen | 98.18 | 49272 | 40 | vertebrae_T4 | 98.36 | 49363 | 79 | spinal_cord | 99.96 | 50164 |
| 2 | kidney_right | 93.91 | 47131 | 41 | vertebrae_T3 | 98.39 | 49376 | 80 | gluteus_maximus_left | 0.54 | 270 |
| 3 | kidney_left | 96.10 | 48230 | 42 | vertebrae_T2 | 98.39 | 49377 | 81 | gluteus_maximus_right | 0.28 | 143 |
| 4 | gallbladder | 88.02 | 44172 | 43 | vertebrae_T1 | 98.39 | 49379 | 82 | gluteus_medius_left | 0.42 | 210 |
| 5 | liver | 98.36 | 49361 | 44 | vertebrae_C7 | 97.68 | 49020 | 83 | gluteus_medius_right | 0.41 | 208 |
| 6 | stomach | 98.31 | 49336 | 45 | vertebrae_C6 | 73.44 | 36856 | 84 | gluteus_minimus_left | 0.16 | 82 |
| 7 | pancreas | 97.84 | 49101 | 46 | vertebrae_C5 | 28.59 | 14349 | 85 | gluteus_minimus_right | 0.17 | 83 |
| 8 | adrenal_gland_right | 97.74 | 49050 | 47 | vertebrae_C4 | 5.09 | 2554 | 86 | autochthon_left | 98.57 | 49469 |
| 9 | adrenal_gland_left | 97.29 | 48827 | 48 | vertebrae_C3 | 2.40 | 1204 | 87 | autochthon_right | 98.53 | 49449 |
| 10 | lung_upper_lobe_left | 98.45 | 49407 | 49 | vertebrae_C2 | 2.12 | 1062 | 88 | iliopsoas_left | 88.44 | 44382 |
| 11 | lung_lower_lobe_left | 98.41 | 49388 | 50 | vertebrae_C1 | 2.09 | 1050 | 89 | iliopsoas_right | 90.44 | 45387 |
| 12 | lung_upper_lobe_right | 98.43 | 49398 | 51 | heart | 98.37 | 49368 | 90 | brain | 3.25 | 1631 |
| 13 | lung_middle_lobe_right | 98.37 | 49369 | 52 | aorta | 98.40 | 49384 | 91 | skull | 43.07 | 21617 |
| 14 | lung_lower_lobe_right | 98.41 | 49385 | 53 | pulmonary_vein | 98.32 | 49341 | 92 | rib_left_1 | 98.22 | 49293 |
| 15 | esophagus | 98.44 | 49403 | 54 | brachiocephalic_trunk | 98.20 | 49282 | 93 | rib_left_2 | 98.24 | 49304 |
| 16 | trachea | 98.49 | 49428 | 55 | subclavian_artery_right | 98.28 | 49320 | 94 | rib_left_3 | 98.39 | 49375 |
| 17 | thyroid_gland | 98.36 | 49361 | 56 | subclavian_artery_left | 98.29 | 49329 | 95 | rib_left_4 | 98.37 | 49367 |
| 18 | small_bowel | 89.55 | 44941 | 57 | common_carotid_artery_right | 98.24 | 49303 | 96 | rib_left_5 | 98.36 | 49364 |
| 19 | duodenum | 92.56 | 46450 | 58 | common_carotid_artery_left | 98.36 | 49360 | 97 | rib_left_6 | 98.37 | 49366 |
| 20 | colon | 95.72 | 48036 | 59 | brachiocephalic_vein_left | 98.26 | 49310 | 98 | rib_left_7 | 98.36 | 49364 |
| 21 | urinary_bladder | 1.29 | 649 | 60 | brachiocephalic_vein_right | 98.27 | 49319 | 99 | rib_left_8 | 98.35 | 49358 |
| 22 | prostate | 0.14 | 69 | 61 | atrial_appendage_left | 98.13 | 49246 | 100 | rib_left_9 | 98.34 | 49352 |
| 23 | kidney_cyst_left | 3.16 | 1586 | 62 | superior_vena_cava | 98.32 | 49340 | 101 | rib_left_10 | 98.32 | 49340 |
| 24 | kidney_cyst_right | 4.20 | 2110 | 63 | inferior_vena_cava | 98.34 | 49353 | 102 | rib_left_11 | 98.23 | 49298 |
| 25 | sacrum | 1.76 | 882 | 64 | portal_vein_and_splenic_vein | 96.67 | 48513 | 103 | rib_left_12 | 95.01 | 47681 |
| 26 | vertebrae_S1 | 0.37 | 184 | 65 | iliac_artery_left | 0.76 | 380 | 104 | rib_right_1 | 98.22 | 49293 |
| 27 | vertebrae_L5 | 1.00 | 500 | 66 | iliac_artery_right | 0.85 | 425 | 105 | rib_right_2 | 98.25 | 49305 |
| 28 | vertebrae_L4 | 5.72 | 2871 | 67 | iliac_vena_left | 0.46 | 231 | 106 | rib_right_3 | 98.38 | 49371 |
| 29 | vertebrae_L3 | 32.96 | 16539 | 68 | iliac_vena_right | 0.39 | 198 | 107 | rib_right_4 | 98.37 | 49369 |
| 30 | vertebrae_L2 | 75.66 | 37968 | 69 | humerus_left | 98.14 | 49252 | 108 | rib_right_5 | 98.37 | 49365 |
| 31 | vertebrae_L1 | 94.30 | 47323 | 70 | humerus_right | 98.16 | 49262 | 109 | rib_right_6 | 98.37 | 49366 |
| 32 | vertebrae_T12 | 98.00 | 49182 | 71 | scapula_left | 98.47 | 49416 | 110 | rib_right_7 | 98.36 | 49363 |
| 33 | vertebrae_T11 | 98.28 | 49324 | 72 | scapula_right | 98.46 | 49413 | 111 | rib_right_8 | 98.36 | 49360 |
| 34 | vertebrae_T10 | 98.32 | 49344 | 73 | clavicula_left | 98.44 | 49400 | 112 | rib_right_9 | 98.34 | 49351 |
| 35 | vertebrae_T9 | 98.35 | 49356 | 74 | clavicula_right | 98.44 | 49403 | 113 | rib_right_10 | 98.32 | 49344 |
| 36 | vertebrae_T8 | 98.34 | 49351 | 75 | femur_left | 1.59 | 796 | 114 | rib_right_11 | 98.24 | 49300 |
| 37 | vertebrae_T7 | 98.35 | 49356 | 76 | femur_right | 0.66 | 329 | 115 | rib_right_12 | 95.10 | 47728 |
| 38 | vertebrae_T6 | 98.33 | 49348 | 77 | hip_left | 6.06 | 3040 | 116 | sternum | 98.38 | 49372 |
| 39 | vertebrae_T5 | 98.35 | 49356 | 78 | hip_right | 3.24 | 1625 | 117 | costal_cartilages | 98.40 | 49380 |

### B.5 Reconstruction and Segmentation Visualization

SOAR includes both reconstruction and segmentation decoders during pretraining. Here we visualize its organ-level reconstruction outputs and segmentation predictions.

Fig. 4 qualitatively visualizes the reconstruction and segmentation outputs produced by SOAR. In this case, SOAR reconstructs the masked organ regions with plausible anatomical structure, and the predicted segmentation masks closely match the automatic TotalSegmentator masks used as pseudo supervision, demonstrating that the encoded visual tokens preserve fine-grained spatial information needed to recover dense organ segmentation masks.

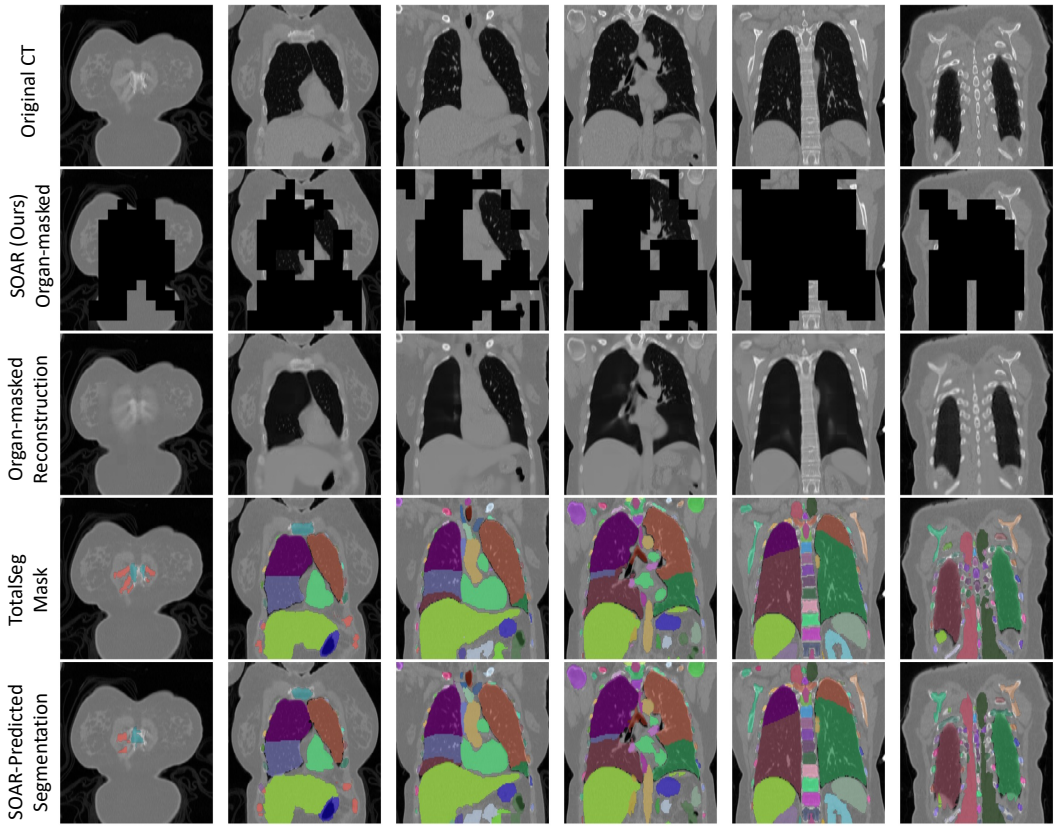



Figure 4: Qualitative visualization of SOAR outputs using reconstruction and segmentation decoders. Columns show representative slices from the same CT volume. The organ token refers to the organ-conditioned masked token used to provide anatomical identity information during pretraining.

Fig. 5 shows another representative case where the automatic TotalSegmentator masks contain visible imperfections. For example, the liver mask does not fully cover the entire liver, and the right lung should contain three lobes, but TotalSegmentator does not segment them out in this case. In contrast, the segmentation masks predicted from SOAR visual tokens appear smoother and more anatomically coherent. This qualitative example suggests that SOAR does not merely copy the noisy pseudo labels, but may learn regularized organ-level representations from the joint pretraining objectives. We hypothesize that this is partly because SOAR includes an organ-level vision-language alignment step, which aligns organ-level visual tokens with organ-name text embeddings, whereas TotalSegmentator does not use text input. However, without manually annotated reference masks, this qualitative comparison cannot establish improved segmentation accuracy.

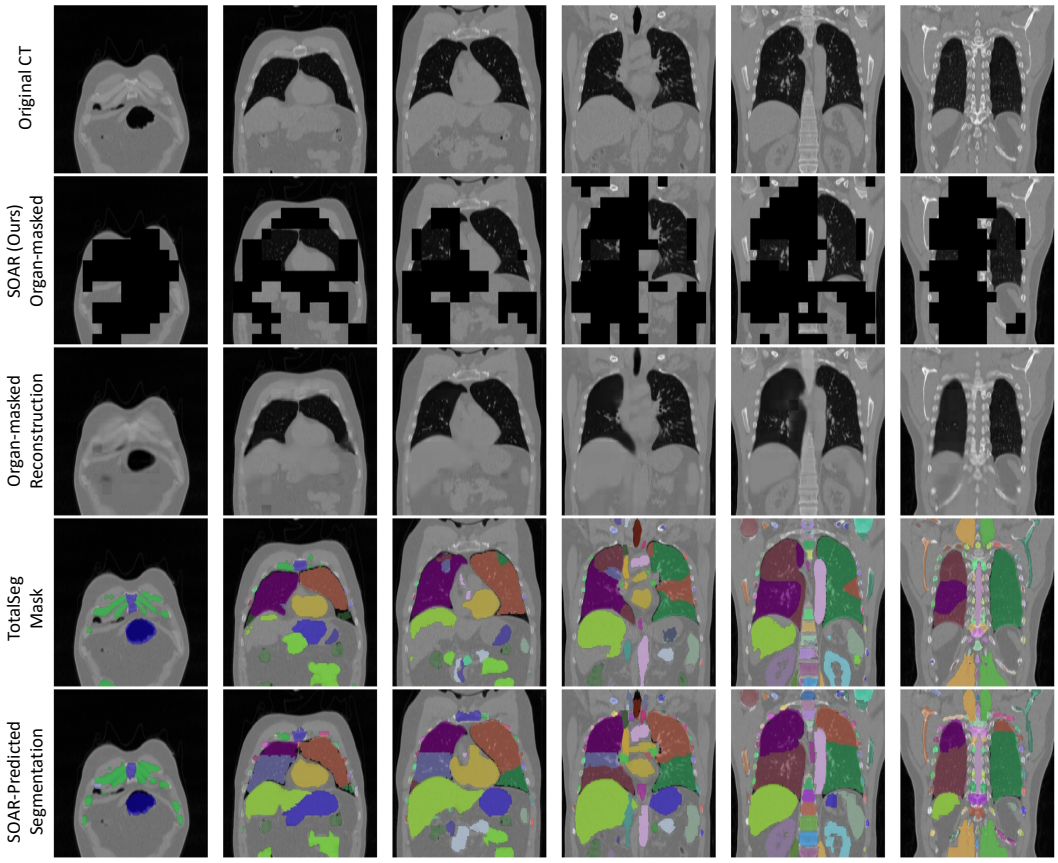



Figure 5: Qualitative visualization of SOAR outputs in a case with visibly imperfect segmentation pseudo-labels. Columns show representative slices from the same CT volume. The automatic TotalSegmentator masks used as pseudo-label supervision contain visible imperfections, while the segmentation masks predicted from SOAR visual tokens appear smoother and more anatomically coherent.

### B.6 VQA Results with Statistical Testing

Table 7 reports the combined open-ended VQA performance across the two benchmarks, CT-RATE and 3D-RAD. SOAR achieves the best overall results across nearly all metrics, including BLEU-1, ROUGE-1, METEOR, BERT-F1, and GREEN. The paired t-test results further show that the improvements of SOAR over most comparison methods are statistically significant, demonstrating the robustness of its advantage in open-ended medical VQA.

Table 7: Open-ended VQA results across two medical VQA benchmarks (3D-RAD + CT-RATE combined). Best results are shown in **bold** and second-best results are underlined. Significance levels from paired $t$-test vs. SOAR: $^{*}p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.00005$.

| Metric | MAE | CLIP | fVLM | ViSD | VoCo | SOAR |
|---|---|---|---|---|---|---|
| BLEU-1 | 47.33*** | 47.21*** | 48.05*** | 47.65*** | 48.38*** | **48.77** |
| BLEU-2 | 35.87*** | 35.49*** | 36.63* | 36.09*** | **37.10** | 37.09 |
| ROUGE-1 | 52.87*** | 52.27*** | 53.42*** | 53.04*** | 53.93* | **54.11** |
| METEOR | 49.47*** | 49.31*** | 49.80** | 49.66*** | 50.39 | **50.40** |
| BERT-F1 | 91.66*** | 91.73*** | 91.91*** | 91.93*** | 91.90*** | **92.06** |
| GREEN | 45.01*** | 42.35*** | 44.83*** | 43.67*** | 45.22*** | **45.71** |

### B.7 Additional Baseline Comparisons

To further evaluate the effectiveness of SOAR, we compare it against two additional pretrained visual encoders: MedVAE [20] and TotalSegmentator [13]. MedVAE represents an autoencoder-based pretraining approach, while TotalSegmentator provides a segmentation-pretrained encoder and is particularly relevant because it is used to generate the anatomical pseudo-labels for SOAR. All encoders are evaluated on CT-RATE VQA using the same downstream fine-tuning and evaluation protocols.

As shown in Table 8, SOAR consistently outperforms both baselines across all reported CT-RATE VQA metrics. For closed-ended VQA, SOAR achieves an accuracy of 89.87%, exceeding MedVAE and TotalSegmentator by 14.17 and 2.75 percentage points, respectively. For open-ended VQA, SOAR improves upon MedVAE by 0.55–5.31 points and TotalSegmentator by 1.60–7.30 points across the five evaluation metrics. These results demonstrate that the proposed organ-aware pretraining framework learns more effective representations for downstream CT understanding than the evaluated autoencoder- and segmentation-pretrained baselines.

### B.8 Ablation Study of Organ-Level Semantic Embeddings

To assess whether the semantic information in organ-name text embeddings contributes beyond anatomical class identity alone, we replace the PubMedBERT-derived organ-name embeddings used for organ-level alignment with fixed class tokens and retrain the encoder under otherwise identical settings.

Table 9 reports results on pneumonia classification and CT-RATE abnormality detection. Replacing organ-name embeddings with fixed class tokens decreases mean pneumonia AUC from 97.99% to 92.75% (5.24 points), and mean CT-RATE abnormality AUC from 75.38% to 70.90% (4.48 points). Improvements are also observed for COVID-19, common pneumonia, and normal-class discrimination. These results demonstrate that PubMedBERT-derived organ-name embeddings provide more effective supervision than fixed anatomical class tokens, highlighting the benefit of incorporating organ-level semantic information into 3D CT visual representation learning.

### B.9 Sensitivity to Distorted Anatomical Pseudo-Labels

To evaluate the effect of imperfect anatomical masks on organ-aware pretraining, we simulate under-segmentation errors by removing 30% of each organ's volume from its TotalSegmentator mask. Specifically, we randomly select a voxel on the organ boundary and expand a connected region within the organ until 30% of the organ is covered. The remaining organ labels are unchanged. These distorted masks are used only to determine which image patches are masked during OrganMAE pretraining. The original CT images and reconstruction targets remain unchanged.

Table 8: Additional baseline comparisons on CT-RATE VQA. Δ denotes the absolute improvement of SOAR over the corresponding baseline in metric points.

| Task | Metric | MedVAE | TotalSeg. | SOAR | $\Delta_{\text{MedVAE}}$ | $\Delta_{\text{TotalSeg}}$ |
|---|---|---|---|---|---|---|
| Closed-ended | Acc. (%) | 75.70 | 87.12 | **89.87** | +14.17 | +2.75 |
| Open-ended | BLEU-1 | 44.29 | 41.97 | **49.27** | +4.98 | +7.30 |
| | BLEU-2 | 34.21 | 33.20 | **38.19** | +3.98 | +4.99 |
| | ROUGE-1 | 50.49 | 49.76 | **54.65** | +4.16 | +4.89 |
| | METEOR | 45.84 | 46.67 | **51.15** | +5.31 | +4.48 |
| | BERT-F1 | 91.45 | 90.40 | **92.00** | +0.55 | +1.60 |

Table 9: Effect of replacing PubMedBERT-derived organ-name embeddings with fixed class tokens. All results are AUC (%). Δ denotes the absolute gain of SOAR over the fixed-class-token variant.

| Benchmark | Task/Class | Fixed class tokens | SOAR | Δ |
|---|---|---|---|---|
| Pneumonia | COVID-19 | 93.21 | **98.79** | +5.58 |
| | Common pneumonia | 90.81 | **97.47** | +6.66 |
| | Normal | 94.22 | **97.72** | +3.50 |
| | Average | 92.75 | **97.99** | +5.24 |
| CT-RATE abnormality | Average | 70.90 | **75.38** | +4.48 |

As shown in Table 10, using corrupted anatomical masks substantially degrades downstream performance. The average pneumonia classification AUC decreases from 97.99% to 87.15%, while the average CT-RATE abnormality detection AUC decreases from 75.38% to 68.13%. The degradation under corrupted masks confirms that SOAR benefits from meaningful anatomical priors rather than arbitrary structured masking.

Table 10: Sensitivity analysis under distorted anatomical pseudo-labels. All results are AUC (%). The distorted-mask variant is pretrained using corrupted pseudo-segmentation masks. Δ denotes the absolute improvement of SOAR over the distorted-mask variant in AUC points.

| Benchmark | Task/Class | Distorted masks | SOAR | Δ |
|---|---|---|---|---|
| Pneumonia | COVID-19 | 87.58 | **98.79** | +11.21 |
| | Common pneumonia | 83.55 | **97.47** | +13.92 |
| | Normal | 90.32 | **97.72** | +7.40 |
| | Average | 87.15 | **97.99** | +10.84 |
| CT-RATE abnormality | Average | 68.13 | **75.38** | +7.25 |

## C Limitations and Future Work

SOAR has several limitations. First, the current study focuses on 3D CT representation learning, and the generalizability of SOAR to contrast-enhanced CT, MRI, PET, multi-phase imaging, and whole-body multi-modal settings remains to be further evaluated. Second, SOAR uses concise organ-name prompts for organ-level vision-language alignment. While this design provides scalable anatomical grounding, it does not explicitly encode disease-specific or fine-grained pathological semantics. Future work may incorporate richer anatomy-specific and disease-aware text supervision. Third, although fVLM, ViSD, and VoCo all include CT-RATE as at least part of their pretraining data, the comparison encoders were not pretrained under fully identical data conditions. Differences in pretraining datasets and data scales may still affect downstream performance comparisons. Finally, although SOAR improves downstream VQA and report-generation performance, clinical faithfulness remains an important open question. In qualitative examples, SOAR sometimes generates broader negative organ-context descriptions beyond the ground-truth report, which may be a positive indication that SOAR captures fine-grained anatomical details, as intended by its organ-aware design. Nevertheless, future work should include radiologist evaluation, larger multi-center validation, and prospective studies to rigorously assess robustness, safety, and clinical utility.

## D Societal Impacts

SOAR may have positive societal impacts by improving 3D CT image understanding and supporting downstream clinical AI tasks, including disease screening, abnormality detection, medical VQA, and report generation. By learning from unlabeled CT volumes with automatically derived anatomical priors, SOAR may reduce reliance on costly manual annotations and support the development of more scalable medical AI systems. If rigorously validated, such systems could assist radiologists, improve workflow efficiency, and promote more consistent interpretation of volumetric medical images.

However, SOAR also carries potential risks. Errors in automatic segmentation masks, dataset bias, domain shift, or limited validation across institutions, scanners, patient populations, and imaging protocols may lead to unreliable predictions in underrepresented clinical settings. In report generation and VQA, incorrect, incomplete, or unsupported outputs could increase the risk of over-reliance, automation bias, or clinically misleading interpretation. Therefore, SOAR should not be used for clinical decision-making without rigorous multi-center validation, radiologist evaluation, uncertainty assessment, privacy protection, fairness analysis, and appropriate safeguards for clinical accountability.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction accurately represent the contributions and scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This work mainly includes empirical contributions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental configurations in Section 4 and Appendix A. Code and model weights are available at https://github.com/HuidongXie/SOAR_CT.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Most experiments are conducted using publicly available datasets. Detailed descriptions of the datasets, model architectures, training configurations, and evaluation protocols are provided in Section 4 and Appendix A to facilitate reproducibility. Source code and pretrained model weights are available at https://github.com/HuidongXie/SOAR_CT. The in-house esophageal cancer screening dataset is subject to institutional data-sharing restrictions but may be made available to qualified researchers through a standard data transfer agreement. For data access requests, please contact Jun Zhao at zhaojun10@jd.com.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details are illustrated in Section 4 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Statistical significance tests comparing the proposed method with each comparison method were conducted, and the results are reported in Appendix B.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work is conducted in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Refer to Appendix D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: This work does not focus on content generation and uses clinically verified datasets for all experiments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original publications for all existing datasets, models, and tools used in the study, and all assets are used in accordance with their publicly released licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: There is no new asset released in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This study uses retrospective medical imaging datasets, including publicly available datasets and an in-house esophageal cancer dataset for which IRB approval was obtained. No participants were directly recruited, and no crowdsourcing, prospective human-subject experiments, or participant interactions were conducted. Therefore, participant instructions and compensation are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: IRB approval for the esophageal cancer screening patients was obtained.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.