SUPER-GENERALIST: GENERALIST–SPECIALIST SYNERGY FOR ACCURATE 3D COMPUTED TOMOGRAPHY DIAGNOSIS AND LESION GROUNDING ACROSS DIVERSE DISEASES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models offer a promising generalist approach to medical diagnosis, providing broad disease coverage and zero-shot diagnostic capabilities. In 3D computed tomography (CT), however, these generalist models often lack fine-grained anatomical and lesion awareness for reliable diagnosis and spatial interpretability. In contrast, supervised specialist models achieve strong performance on specific tasks but typically lack generalization across diseases and anatomies. In this work, we present SuG, a Super-Generalist framework for 3D CT diagnosis that unifies generalist vision–language learning with specialist objectives, enabling broad disease coverage, specialist-competitive diagnostic performance, and zero-shot lesion grounding. We perform specialist-enhanced vision-language alignment in SuG by incorporating spatial priors from multiple segmentation experts, including anatomy, class-specific lesion and class-agnostic lesion segmentors that capture lesions beyond anatomies annotated during training. To improve lesion grounding capability, we leverage lesion masks as spatial priors to calibrate text-conditioned visual attention, encouraging disease-related semantics to focus on clinically relevant regions. We evaluate SuG on extensive chest and abdominal CT benchmarks, including CT-RATE, Merlin, MedVL-CT69K, and several in-house tumor datasets. SuG achieves state-of-the-art performance across a wide range of disease diagnosis tasks and surpasses specialist models on several critical tumor diagnosis benchmarks. Furthermore, SuG demonstrates strong lesion grounding capability, including robust generalization to lesion types lacking class-specific supervision. Code and models will be released upon acceptance.

1 INTRODUCTION

Reliable medical image diagnosis requires not only coverage across diverse diseases but also high diagnostic accuracy and spatially grounded evidence (Dogra et al., 2025; Moor et al., 2023; Azad et al., 2024; Jin et al., 2025). This challenge is particularly pronounced in 3D computed tomography (CT), where diverse diseases must be recognized across anatomical regions and clinically relevant lesions may occupy only a small fraction of each volume. An effective 3D CT diagnostic model should therefore combine three complementary capabilities: zero-shot multi-disease diagnosis, specialist-competitive diagnostic performance, and interpretable lesion grounding. However, as illustrated in Fig. 1, existing approaches typically prioritize only a subset of these capabilities, leaving a persistent trade-off between diagnostic breadth, accuracy, and spatial interpretability (Azad et al., 2024; Cai et al., 2020; Xie et al., 2021; Isensee et al., 2018; Ardila et al., 2019; Cao et al., 2023; Zhang et al., 2023a; Guo et al., 2025; Zhang et al., 2022; Tiu et al., 2022a; Blankemeier et al., 2024).

Across medical image analysis, existing methods largely originate from two complementary regimes: *Specialist* and *Generalist* models. In this work, we use Generalist specifically to refer to CLIP-style vision–language models capable of zero-shot diagnosis across multiple diseases. By leveraging vision–language alignment, these models offer broad diagnostic coverage and zero-shot flexibility (Zhang et al., 2022; Tiu et al., 2022a; Boecking et al., 2022; Blankemeier et al., 2024). In contrast, Specialist refers to supervised models trained with task-specific voxel-level annotations for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

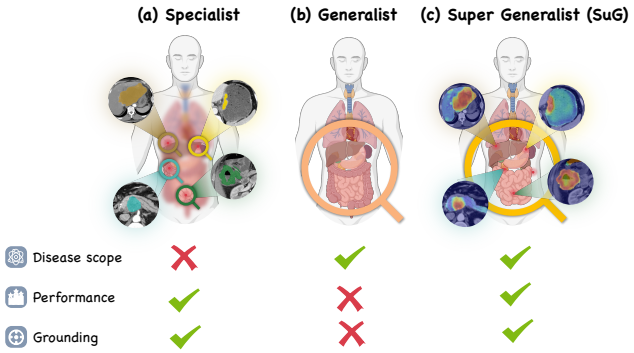

Figure 1: **Comparison of Specialist, Generalist, and Super-Generalist (SuG) approaches for 3D CT diagnosis.** We evaluate these approaches across three critical dimensions: *disease scope* (range of tasks), *performance* (diagnostic accuracy), and *grounding* (lesion localization). (a) Specialists excel in performance and grounding but are limited to a narrow disease scope. (b) Conversely, Generalists support a wide scope but suffer from inferior performance and a lack of precise grounding. (c) Our proposed SuG overcomes these trade-offs, achieving high diagnostic accuracy and precise lesion localization across diverse diseases.

fine-grained segmentation and localization. Although Specialist models achieve strong performance and precise spatial localization on predefined tasks, their reliance on task-specific annotations restricts them to a narrow set of diseases and anatomical targets, limiting their scalability to new clinical tasks (Azad et al., 2024; Cai et al., 2020; Xie et al., 2021; Isensee et al., 2018; Ardila et al., 2019; Cao et al., 2023). Consequently, within 3D CT, a gap remains between the broad diagnostic coverage of Generalists and the lesion-level precision of Specialists: existing Generalist models primarily emphasize zero-shot diagnosis, while fine-grained lesion-aware diagnosis and lesion grounding remain insufficiently explored (Chen et al., 2025; Zhou et al., 2025b; Saha et al., 2025).

To bridge these complementary limitations, we propose the Super-Generalist (SuG) framework for 3D CT diagnosis, which establishes a Generalist–Specialist synergy specifically for diagnosis and lesion grounding. SuG combines the broad zero-shot diagnostic coverage of generalist vision–language models with fine-grained anatomical and lesion knowledge from specialist supervision. Unlike prior methods that primarily incorporate organ- or lesion-level supervision through auxiliary segmentation branches (Zhang et al., 2026; Lin et al., 2024; Hu et al.), SuG goes beyond auxiliary segmentation by using lesion masks to calibrate report-conditioned visual attention, explicitly coupling diagnostic semantics with localized lesion evidence to enhance fine-grained diagnosis. This design also enables lesion-aware spatial knowledge learned from a limited set of voxel-annotated categories to transfer to held-out categories without corresponding class-specific mask supervision. Consequently, SuG unifies zero-shot multi-disease diagnosis, specialist-competitive diagnostic performance, and lesion grounding across diverse diseases.

SuG serves as a unified architecture that bridges the gap between generalist disease scope and specialist-level performance through three synergistic components: First, to achieve specialist-competitive performance, we leverage multi-task segmentation objectives with voxel-level annotations across anatomy, class-specific lesion, and class-agnostic lesion targets. This compels the model to encode high-resolution spatial details that are otherwise ignored by conventional generalist architectures. Second, to extend disease scope and further boost diagnostic performance, we propose a specialist-enhanced vision–language alignment approach. By injecting specialist-derived anatomical and lesion features into generalist embeddings, we enrich cross-modal representations with fine-grained details, significantly strengthening the alignment between visual patterns and clinical text. Finally, to ensure lesion grounding, we implement an attention calibration strategy that utilizes lesion masks as spatial priors. This steers text-derived queries toward clinically relevant visual tokens, thereby enhancing interpretability and diagnostic traceability.

Overall, SuG achieves state-of-the-art performance in broad-scope disease diagnosis and outperforms specialist models in critical tumor tasks, while demonstrating trustworthy grounding for pan lesions. Our contributions are:

- We propose *Super-Generalist* (SuG), a 3D CT diagnosis framework that establishes Generalist–Specialist synergy by unifying zero-shot multi-disease diagnosis, specialist-competitive diagnostic performance, and interpretable lesion grounding.
- We introduce specialist-enhanced vision–language alignment and lesion-guided attention calibration, coupling diagnostic semantics with localized lesion evidence and generalizing grounding to held-out categories without class-specific mask supervision.
- Experiments on 10 chest and abdominal CT benchmarks demonstrate state-of-the-art broad-coverage diagnosis, competitive or superior performance to specialist models on several tumor tasks, and strong generalization of lesion grounding.

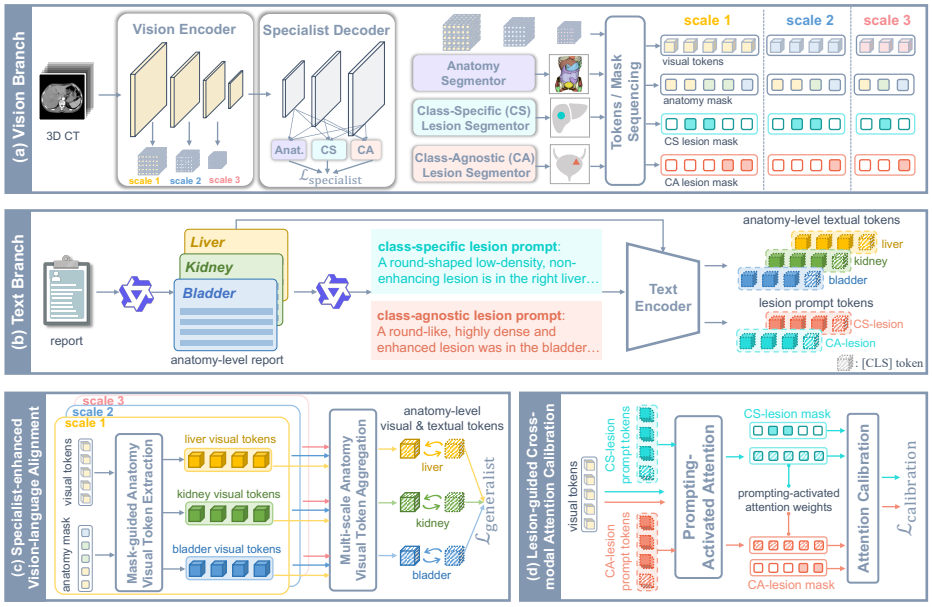

Figure 2: **An overview of the proposed SuG framework.** SuG consists of two fundamental branches and two synergistic mechanisms: (a) Vision branch extracts multi-scale features and performs anatomy/lesion segmentation via a specialist decoder. (b) Text branch encodes reports into anatomy-level tokens and lesion prompt tokens. (c) Specialist-enhanced vision-language alignment leverages predicted anatomy masks to guide the alignment between multi-scale visual and anatomy-level textual tokens, optimizing the generalist representation. (d) Lesion-guided cross-modal attention calibration dynamically refines text-conditioned visual attention through lesion masks to achieve precise, pathology-aware localization.

2 METHOD

2.1 OVERVIEW OF SUG

SuG integrates specialist supervision with generalist vision–language learning to achieve comprehensive and accurate understanding of medical images. As shown in Fig. 2, the framework consists of a vision encoder ψ, a specialist decoder ϕ, and a text encoder $g_{\theta_{\text{text}}}$. The specialist pathway is trained with voxel-level annotations to learn anatomy-aware and lesion-aware visual representations, whereas the generalist pathway is trained with paired images and radiology reports to enable broad cross-modal diagnosis without task-specific labels.

Given a specialist dataset D_S and a generalist dataset D_G, SuG is designed to (i) inject anatomical and lesion cues derived from specialist supervision into visual embeddings, (ii) align anatomy-specific image and text representations through contrastive learning, and (iii) calibrate text-conditioned attention with lesion localization priors. Together, these components enable SuG to combine broad diagnostic coverage with fine-grained lesion understanding and grounding.

2.2 SPECIALIST SEGMENTATION ACROSS ANATOMIES, LESIONS, AND BEYOND

The specialist pathway learns fine-grained anatomical and lesion structures from the voxel-labeled dataset $D_S = \{(I_j, M_j)\}_{j=1}^{N}$. Given an input CT volume I_j, the vision encoder and specialist decoder produce dense features that are processed by three segmentation heads.

Anatomy segmentation. An anatomy segmentation head ω_{anat} predicts anatomy logits

$$Z_j^{\text{anat}} = f_{\psi,\phi,\omega_{\text{anat}}}(I_j), \qquad Z_j^{\text{anat}} \in \mathbb{R}^{C_{\text{anat}} \times D \times H \times W}, \tag{1}$$

where C_{anat} denotes the number of anatomy classes, including the background class. The anatomy loss is defined as

$$\mathcal{L}_{\text{anat}} = \mathcal{L}_{\text{seg}}\big(Z_j^{\text{anat}}, M_j^{\text{anat}}\big), \tag{2}$$

where $\mathcal{L}_{\text{seg}}$ is the standard segmentation loss composed of cross-entropy and Dice terms.

Class-specific lesion segmentation. For anatomies with lesion annotations, a lesion segmentation head ω_{CS} predicts lesion logits

$$Z_j^{\text{CS}} = f_{\psi,\phi,\omega_{\text{CS}}}(I_j), \qquad Z_j^{\text{CS}} \in \mathbb{R}^{C_{\text{CS}} \times D \times H \times W}. \tag{3}$$

Since each sample provides annotation for only one lesion class, we use a binary class indicator $\mathcal{C}_j \in \{0,1\}^{C_{\text{CS}}}$ to mask irrelevant classes:

$$\mathcal{L}_{\text{CS}} = \mathcal{L}_{\text{seg}}\big((\mathcal{C}_j \odot Z_j^{\text{CS}}),\, (\mathcal{C}_j \odot M_j^{\text{CS}})\big). \tag{4}$$

Class-agnostic lesion segmentation. To extend lesion modeling beyond anatomies with lesion labels, we introduce a class-agnostic lesion segmentation head ω_{CA}:

$$Z_j^{\text{CA}} = f_{\psi,\phi,\omega_{\text{CA}}}(I_j), \qquad Z_j^{\text{CA}} \in \mathbb{R}^{2 \times D \times H \times W}. \tag{5}$$

Its target M_j^{CA} merges all annotated lesions into a single foreground class. Because lesion annotations are anatomy-specific, voxels outside annotated anatomies are often unlabeled rather than truly negative. We therefore optimize this branch within a binary validity mask $\mathcal{V}_j$, which retains regions with trustworthy supervision and excludes unlabeled anatomies from the loss (Fig. A1):

$$\mathcal{L}_{\text{CA}} = \mathcal{L}_{\text{seg}}\big(Z_j^{\text{CA}} \odot \mathcal{V}_j,\, M_j^{\text{CA}} \odot \mathcal{V}_j\big). \tag{6}$$

The overall specialist loss is $\mathcal{L}_{\text{specialist}} = \mathcal{L}_{\text{anat}} + \mathcal{L}_{\text{CS}} + \mathcal{L}_{\text{CA}}$.

2.3 SPECIALIST-ENHANCED VISION–LANGUAGE ALIGNMENT

We next transfer anatomical knowledge learned from specialist supervision to the generalist vision–language alignment process. The main idea is to extract anatomy-specific visual tokens with masks predicted by the specialist pathway, aggregate them across scales, and align them with anatomy-specific textual descriptions extracted from the paired report.

Anatomy vision embedding. Let $\mathbf{f}_j^{(s)}$ denote the feature map at scale $s \in \{1, \ldots, S\}$. For anatomy category t, we use the predicted anatomy mask to select the corresponding tokens:

$$\mathcal{A}_{j,t}^{(s)} = \{\mathbf{f}_{j,p}^{(s)} \mid p \in \mathcal{R}_{j,t}^{s}\},$$

where $\mathcal{R}_{j,t}^{s}$ denotes the masked spatial positions. We then average the selected tokens at each scale and normalize the pooled feature:

$$\bar{\mathbf{a}}_{j,t}^{(s)} = \frac{1}{|\mathcal{R}_{j,t}^{s}|} \sum_{p \in \mathcal{R}_{j,t}^{s}} \mathbf{f}_{j,p}^{(s)}, \qquad \tilde{\mathbf{a}}_{j,t}^{(s)} = \frac{\bar{\mathbf{a}}_{j,t}^{(s)}}{\|\bar{\mathbf{a}}_{j,t}^{(s)}\|_2}.$$

The anatomy-aware visual embedding is obtained by averaging the normalized features across scales:

$$\mathbf{a}_{j,t} = \frac{1}{S} \sum_{s=1}^{S} \tilde{\mathbf{a}}_{j,t}^{(s)}.$$

This normalization is important because pooled features at different scales may have different magnitudes; otherwise, large-norm scales would dominate the multi-scale embedding.

Anatomy text embedding. For each anatomy category t, we extract the corresponding textual description $r_{j,t}$ from report R_j using Qwen LLM (Bai et al., 2023) (see Appendix Section B.1), and encode it with text encoder:

$$\mathbf{t}_{j,t} = g_{\theta_{\text{text}}}(r_{j,t}).$$

Anatomy-wise contrastive learning. For a mini-batch with T anatomy categories, let N_t denote the number of valid samples for anatomy category t, i.e., samples with both anatomy-specific visual and textual embeddings. We denote the visual and textual embeddings of the j-th valid sample for anatomy category t as $\mathbf{a}_{j,t}$ and $\mathbf{t}_{j,t}$, respectively. For each anatomy category, contrastive matching is performed only within its valid samples. The image-to-text matching probability is

$$p^{i2t}_{j,t,k} = \frac{\exp\big(\text{sim}(\mathbf{a}_{j,t}, \mathbf{t}_{k,t})/\tau\big)}{\sum_{k'=1}^{N_t} \exp\big(\text{sim}(\mathbf{a}_{j,t}, \mathbf{t}_{k',t})/\tau\big)}, \tag{7}$$

and the text-to-image probability is defined symmetrically as

$$p^{t2i}_{j,t,k} = \frac{\exp\big(\text{sim}(\mathbf{t}_{j,t}, \mathbf{a}_{k,t})/\tau\big)}{\sum_{k'=1}^{N_t} \exp\big(\text{sim}(\mathbf{t}_{j,t}, \mathbf{a}_{k',t})/\tau\big)}. \tag{8}$$

Let $\mathbf{y}^{i2t}_{j,t}, \mathbf{y}^{t2i}_{j,t} \in \{0,1\}^{N_t}$ denote the one-hot matching labels, where the positive entry corresponds to the paired sample with index j. The loss is

$$\mathcal{L}_{\text{generalist}} = \frac{1}{2} \sum_{t=1}^{T} \frac{1}{N_t} \sum_{j=1}^{N_t} \left[\text{H}\big(\mathbf{y}^{i2t}_{j,t}, \mathbf{p}^{i2t}_{j,t}\big) + \text{H}\big(\mathbf{y}^{t2i}_{j,t}, \mathbf{p}^{t2i}_{j,t}\big) \right], \tag{9}$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity, τ is a learnable temperature, and $\text{H}(\cdot,\cdot)$ is the cross-entropy loss.

2.4 Lesion-guided cross-modal attention calibration

To improve focal lesion diagnosis and localization, we further calibrate text-conditioned attention by using lesion masks from the specialist pathway. Given a lesion-related description r^{lesion}_j extracted from report R_j (see Appendix Section B.2), we obtain a lesion prompt embedding $\mathbf{h}_j$. For a visual feature map $\mathbf{f}^{(s)}_j$ at scale s, we compute text-conditioned attention at position q as

$$\alpha^{(s)}_{j,q} = \sigma\Big(\mathbf{h}_j \cdot \mathbf{f}^{(s)}_{j,q}\Big),$$

where σ is sigmoid. We supervise this attention map with the resized lesion mask $\mathfrak{R}(\hat{M}^{\text{les}}_j)$:

$$\mathcal{L}^{(s)}_{\text{attn}} = -\frac{1}{D_s H_s W_s} \sum_{q} \Big[\mathfrak{R}(\hat{M}^{\text{les}}_j)_q \log \alpha^{(s)}_{j,q} + \big(1 - \mathfrak{R}(\hat{M}^{\text{les}}_j)_q\big) \log\big(1 - \alpha^{(s)}_{j,q}\big) \Big]. \tag{10}$$

The final calibration loss $\mathcal{L}_{\text{calibration}} = \frac{1}{S}\sum_{s=1}^{S} \mathcal{L}^{(s)}_{\text{attn}}$. To reduce noise, we apply report-conditioned filtering: if a report does not mention a lesion in a given anatomy, the corresponding lesion prediction is excluded from supervision.

2.5 Progressive multi-stage optimization

We adopt a three-stage training pipeline to stably integrate specialist segmentation, generalist alignment, and lesion-guided calibration. In Stage 1, only the specialist pathway is trained: $\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{specialist}}$. In Stage 2, training is initialized from Stage 1, and segmentation and anatomy-wise vision–language alignment are jointly optimized: $\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{specialist}} + \mathcal{L}_{\text{generalist}}$. In Stage 3, training is initialized from Stage 2, and lesion-guided attention calibration is further introduced: $\mathcal{L}_{\text{stage3}} = \mathcal{L}_{\text{specialist}} + \mathcal{L}_{\text{generalist}} + \mathcal{L}_{\text{calibration}}$. This progressive optimization strategy yields stable convergence while balancing segmentation, cross-modal alignment, and lesion grounding.

Table 1: **Overview of the experimental settings.** Each setting integrates a specialist lesion segmentation dataset with a generalist image-report dataset.

Setting	Scenario	Specialist Dataset	Generalist Dataset	Main Results
Setting 1	Abdomen CT	LesionSegAbdomen	MedVL-CT69K (Shui et al., 2025)	Table 2, 5, 6
Setting 2	Abdomen CT	LesionSegAbdomen	Merlin (Blankemeier et al., 2024)	Table 3
Setting 3	Chest CT	LesionSegLung	CT-RATE (Hamamci et al., 2024a)	Table 4

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Training and evaluation settings. We train and evaluate SuG across three experimental settings, summarized in Table 1. Our assessment covers abdominal and chest CT. For abdomen, we utilize two large-scale image–report datasets, MedVL-CT69K (Shui et al., 2025) and Merlin (Blankemeier et al., 2024), alongside four anatomy-specific tumor segmentation datasets collectively termed LesionSegAbdomen. For chest CT, we employ CT-RATE (Hamamci et al., 2024a), the external benchmark RAD-ChestCT (Draelos et al., 2021), and a lung tumor segmentation dataset, LesionSegLung. Using these datasets, we evaluate SuG's super-generalist capabilities across three dimensions: (i) general disease diagnosis, (ii) fine-grained lesion-centric diagnosis, (iii) lesion grounding. Further details on datasets, configurations, and implementation are in Appendix C.1 and C.2.

Evaluation protocols. For disease diagnosis, we follow fVLM (Shui et al., 2025) to derive diagnosis scores by applying softmax over the similarities between image embeddings and the text embeddings of paired positive/negative text prompts. For lesion grounding, we generate disease-specific attention maps by calculating the similarity between positive prompt embeddings and image patch tokens. These maps serve as predicted heatmaps for both quantitative assessment and qualitative visualization.

Comparison with state-of-the-art methods. We evaluate SuG against a broad spectrum of specialist-only (S) and generalist-only (G) methods. Additionally, we also benchmark against OpenVocabCT (Li et al., 2025b), HCFNet (Zhou et al., 2025a), and VLWS (Hossain et al., 2026), which utilize generalist architectures for specialist diagnostic tasks. We adapt these methods to the specialist–generalist (S+G) framework to ensure a fair comparison. This alignment enables all baselines to leverage both lesion-specific priors and general diagnostic knowledge under our unified evaluation protocol. Detailed descriptions of these baseline adaptations are provided in Appendix C.2.

Annotated v.s. unannotated lesions. To evaluate our diagnosis and grounding performance comprehensively, we categorize lesions into two groups based on the availability of supervision: *annotated* lesions, which possess specific category labels, and *unannotated* lesions, which lack such explicit supervision. This distinction enables a dual assessment of our model: its fine-grained performance on supervised targets and its generalization capability toward unlabeled abnormalities.

Table 2: **General disease diagnosis performance (AUC) on the MedVL-CT69K test set.** Results are evaluated across 54 diseases grouped into 15 anatomical structures, with the overall average (Avg) reported. "S", "G", and "S+G" denote specialist-only, generalist-only, and specialist-generalist synergy methods, respectively. Full Sensitivity, Specificity and AUC results are provided in Table A25. Full abbreviations are in Table A5, with detailed per-disease results in Table A27.

Method	Type	Adr	Bla	Col	Eso	GB	Hrt	Kid	Liv	Lng	Pan	PV	SI	Spl	Sto	Sac	Avg
CLIP (Radford et al., 2021)	G	63.2	65.1	65.8	67.3	59.5	44.1	59.9	72.4	88.3	65.0	78.6	74.5	71.1	68.6	47.0	68.4
LOVT (Müller et al., 2022)	G	60.6	70.9	67.5	89.3	61.2	78.3	60.2	69.3	80.9	67.8	75.5	70.5	66.1	69.1	48.9	69.4
Imitate (Liu et al., 2024)	G	60.2	74.1	68.0	95.6	62.5	70.7	59.9	69.8	86.7	64.3	80.5	77.6	71.3	66.3	29.0	70.6
BIUD (Cao et al., 2024)	G	63.4	81.0	70.0	62.6	64.2	62.1	63.7	79.2	72.1	76.9	82.2	75.1	72.3	66.1	43.8	71.4
Merlin (Blankemeier et al., 2024)	G	60.3	76.9	69.1	49.2	61.2	72.8	64.2	80.1	78.7	73.5	85.9	78.4	72.0	69.9	48.2	71.9
fVLM (Shui et al., 2025)	G	65.7	84.0	80.8	98.2	64.8	85.8	74.5	82.5	82.2	85.3	96.7	82.1	82.0	74.1	87.5	81.3
ViSD-Boost (Cao et al., 2025)	G	68.5	81.2	81.9	98.1	72.6	90.5	78.5	85.9	**92.7**	88.9	97.3	88.3	82.9	81.1	77.5	84.9
OpenVocabCT (Li et al., 2025b)	S+G	68.8	84.2	84.3	96.3	85.0	81.5	84.0	90.6	68.1	90.9	95.5	**89.8**	94.2	87.4	**88.4**	85.8
HCFNet (Zhou et al., 2025a)	S+G	71.2	89.2	83.0	**99.0**	83.5	85.6	84.4	89.3	89.3	90.2	97.4	88.3	92.6	76.6	58.7	86.7
VLWS (Hossain et al., 2026)	S+G	70.1	88.9	84.1	95.3	**85.6**	89.5	80.4	84.9	87.0	**90.8**	97.6	**89.8**	94.1	85.2	80.0	86.7
SuG (Ours)	S+G	**82.9**	**92.1**	**87.0**	97.2	84.9	**91.6**	**87.4**	**92.2**	90.9	90.4	**98.4**	88.9	**95.6**	**88.4**	88.2	**90.1**

Table 3: **General disease diagnosis performance on the Merlin test set.** Balanced metrics (AUC*, F1*) follow the official Merlin protocol; full-set metrics (AUC, F1) follow common practice. Detailed per-disease results are in Table A30 and A31.

Method	AUC*	F1*	AUC	F1
CLIP (Hamamci et al., 2024b)	64.3	62.7	64.4	41.7
Merlin (Blankemeier et al., 2024)	82.4	74.5	81.6	55.8
fVLM (Shui et al., 2025)	79.1	75.2	79.0	55.0
ViSD-Boost (Cao et al., 2025)	82.8	78.4	82.5	58.4
OpenVocabCT (Li et al., 2025b)	82.9	78.6	83.2	59.3
HCFNet (Zhou et al., 2025a)	83.2	78.7	82.9	57.2
VLWS (Hossain et al., 2026)	82.6	77.9	82.2	57.3
SuG (Ours)	**85.2**	**80.8**	**85.1**	**60.7**

Table 4: **General disease diagnosis performance on the chest CT benchmarks: CT-RATE test set and RAD-ChestCT.** † denotes CT-CLIP variants fine-tuned with supervised learning. Detailed results are in Table A28 and A29.

Method	Internal validation (CT-RATE)				External validation (Rad-ChestCT)			
	Precision	ACC	F1	AUC	Precision	ACC	F1	AUC
Random (Hamamci et al., 2024b)	18.0	50.2	57.0	50.5	26.5	50.0	55.5	49.6
CT-CLIP (Hamamci et al., 2024b)	32.6	66.9	70.8	73.3	34.1	59.9	64.7	63.2
CT-VocabFine†	35.6	70.4	73.8	76.0	35.6	62.1	66.8	65.7
CT-LiPro†	34.3	69.1	72.6	76.1	35.1	60.6	65.0	64.7
BIUD (Cao et al., 2024)	33.8	68.1	71.6	71.3	33.7	60.6	65.2	62.9
Merlin (Blankemeier et al., 2024)	33.7	67.2	70.9	72.8	34.8	61.9	66.3	64.4
fVLM (Shui et al., 2025)	37.9	71.8	75.1	77.8	**37.4**	64.7	68.8	68.0
ViSD-Boost (Cao et al., 2025)	38.7	73.1	75.9	79.0	34.2	65.2	69.3	69.4
OpenVocabCT (Li et al., 2025b)	40.2	74.1	76.9	79.5	33.6	61.9	66.4	66.6
HCFNet (Zhou et al., 2025a)	40.7	73.5	76.4	80.3	35.9	67.8	71.5	71.7
VLWS (Hossain et al., 2026)	39.6	73.0	76.0	79.5	35.3	65.5	69.7	70.7
SuG (Ours)	**40.8**	**74.4**	**77.1**	**81.1**	36.9	**68.1**	**71.9**	**73.4**

Table 5: **Fine-grained lesion-centric diagnosis performance (AUC) on four annotated tumor categories across MedVL-CT69K and LesionSegAbdomen.** "–" denotes that a specific category is not supported by the corresponding specialist model. Detailed results are provided in Table A26.

Method	Type	MedVL-CT69K					LesionSegAbdomen				
		Colon	Liver	Pancreas	Stomach	Average	Colon	Liver	Pancreas	Stomach	Average
nnUNet-c (Isensee et al., 2018)	S	87.2	–	–	–	–	96.0	–	–	–	–
nnUNet-l (Isensee et al., 2018)	S	–	79.5	–	–	–	–	97.4	–	–	–
nnUNet-p (Isensee et al., 2018)	S	–	–	78.6	–	–	–	–	94.0	–	–
nnUNet-s (Isensee et al., 2018)	S	–	–	–	74.7	–	–	–	–	91.5	–
nnUNet-fuse (Isensee et al., 2018)	S	87.2	79.5	78.6	74.7	80.0	96.0	97.4	94.0	91.5	94.7
CLIP (Radford et al., 2021)	G	65.9	60.0	73.0	71.6	67.6	64.6	62.5	66.4	68.9	65.6
LOVT (Müller et al., 2022)	G	68.1	58.4	74.9	69.4	67.7	83.9	55.2	70.2	54.9	66.1
Imitate (Liu et al., 2024)	G	63.5	60.3	71.4	71.9	66.8	66.7	48.3	71.7	67.9	63.7
BIUD (Cao et al., 2024)	G	66.3	59.6	77.7	73.1	69.2	73.3	65.9	66.6	57.8	65.9
Merlin (Blankemeier et al., 2024)	G	62.6	63.2	74.2	66.5	66.6	52.9	55.2	64.8	53.3	56.5
fVLM (Shui et al., 2025)	G	73.6	62.2	85.8	75.7	74.3	82.9	65.9	77.4	75.7	75.5
ViSD-Boost (Cao et al., 2025)	G	83.8	64.5	94.7	84.2	81.8	76.9	73.6	88.7	67.1	76.6
OpenVocabCT (Li et al., 2025b)	S+G	87.0	84.9	**97.4**	89.6	89.7	81.1	90.2	90.2	85.6	86.8
HCFNet (Zhou et al., 2025a)	S+G	86.0	75.1	94.8	89.3	86.3	91.3	79.8	93.8	90.9	89.0
VLWS (Hossain et al., 2026)	S+G	83.3	84.0	96.4	88.6	88.1	86.0	86.8	87.5	88.0	87.0
SuG (Ours)	S+G	**93.0**	**90.6**	96.1	**95.2**	**93.7**	**97.3**	**99.3**	**95.1**	**92.7**	**96.1**

3.2 General disease diagnosis

Abdominal CT benchmarks. Table 2 summarizes the performance on the MedVL-CT69K benchmark. While generalist models provide broad disease coverage, they consistently underperform methods that integrate specialist priors. For instance, the leading generalist baseline, ViSD-Boost (Cao et al., 2025), achieves an average AUC of 84.9%, whereas adapted S+G methods consistently reach the 85.8%–86.7% range. SuG achieves state-of-the-art performance with an average AUC of 90.1%, surpassing the strongest competitor, VLWS (Hossain et al., 2026), by 3.4%. The gain demonstrates that specialist-enhanced generalist learning can substantially boost diagnostic performance.

Table 3 shows a consistent pattern on the Merlin benchmark. SuG consistently outperforms existing methods, achieving 85.2% AUC* / 80.8% F1* under the official balanced protocol, and 85.1% AUC / 60.7% F1 on the full test set. Notably, MedVL-CT69K and Merlin are derived from distinct patient populations, roughly corresponding to Eastern and Western cohorts, respectively. The consistently superior performance of SuG across these heterogeneous datasets underscores the model's robust cross-population generalization and the stability of our specialist-enhanced generalist architecture.

Chest CT benchmarks. Table 4 further evaluates SuG on CT-RATE and the external RAD-ChestCT benchmark. SuG consistently achieves superior performance, attaining 81.1% AUC on CT-RATE and 73.4% AUC on RAD-ChestCT. Compared with the strongest competitor, HCFNet (Zhou et al., 2025a), our method yields performance gains of 0.8% and 1.7%, respectively. The more pronounced improvement on RAD-ChestCT, an external validation set, highlights SuG's enhanced robustness against cross-dataset distribution shifts. Together with the abdominal results, these findings demonstrate that SuG generalizes reliably across diverse anatomies, populations, and imaging distributions.

3.3 FINE-GRAINED LESION-CENTRIC DIAGNOSIS

Table 5 summarizes the diagnostic performance on four specific tumor categories, evaluated across both MedVL-CT69K and LesionSegAbdomen. This task poses a significant challenge for generalist models, as it necessitates precise lesion perception and localized reasoning. On MedVL-CT69K, specialist-based nnUNet baselines consistently outperform most pure generalist models, underscoring the critical role of lesion-aware supervision. While existing specialist-generalist synergy methods mitigate this gap, SuG achieves superior performance across all categories by integrating specialist-enhanced alignment with lesion-guided attention calibration. On LesionSegAbdomen, SuG further improves upon the specialist-level benchmark (nnUNet-fuse, 94.7% AUC), achieving 96.1% AUC. These results demonstrate that SuG not only preserves broad diagnostic capabilities but also achieves specialist-level, and even specialist-surpassing, performance in fine-grained lesion-centric diagnosis.

Table 6: **Quantitative lesion grounding evaluation on the LesionSegAbdomen test set.** Results are reported across four annotated tumor types using AUC, AUPR, and Cohen's d. These metrics assess the voxel-level correspondence between the predicted attention maps and ground-truth lesion masks.

Method	Colon			Liver			Stomach			Pancreas			Average		
	AUC	AUPR	Cohen's d(↑)	AUC	AUPR	Cohen's d(↑)	AUC	AUPR	Cohen's d(↑)	AUC	AUPR	Cohen's d(↑)	AUC	AUPR	Cohen's d(↑)
Generalist baseline	53.5	8.1	0.1	45.1	9.6	-0.2	39.5	7.2	-0.4	49.7	27.6	0.0	47.0	13.1	-0.1
+ Specialist	74.0	19.5	0.8	58.6	14.6	0.3	60.9	14.4	0.4	51.2	29.5	-0.1	61.2	19.5	0.4
SuG	**92.0**	**52.8**	**2.8**	**93.5**	**63.9**	**3.8**	**87.1**	**47.6**	**2.5**	**90.5**	**73.8**	**2.6**	**90.8**	**59.5**	**2.9**

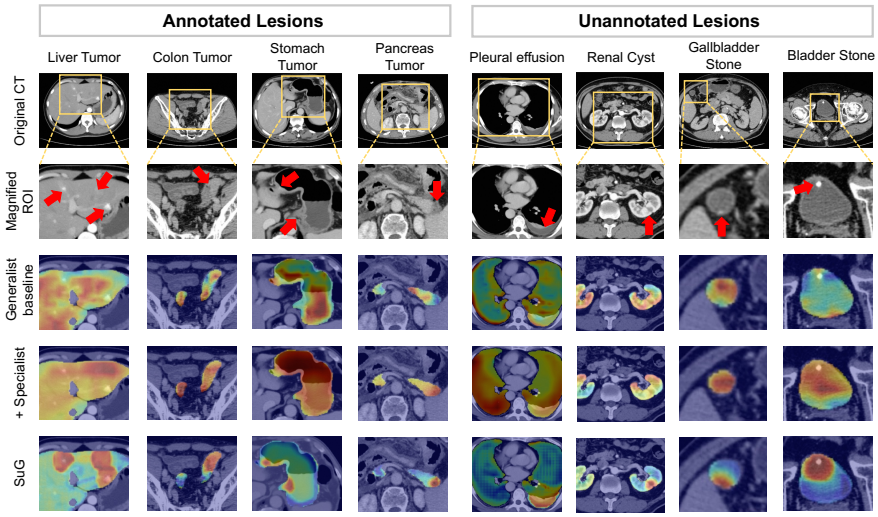

Figure 3: **Attention visualization on *annotated* and *unannotated* lesions.** The left and right panels display representative cases of four annotated (liver, colon, stomach, pancreas) and unannotated (pleural effusion, renal cyst, gallbladder stone, bladder stone) lesion categories, respectively. Yellow boxes highlight ROIs, while red arrows pinpoint lesion locations. Attention maps from our generalist baseline, the specialist-enhanced generalist model, and SuG are overlaid on CT images.

3.4 LESION GROUNDING

We compare three incremental configurations. The *Generalist baseline* configuration is a pure generalist vision-language model that shares the same encoders as SuG but lacks specialist supervision and lesion-guided attention calibration. The *+Specialist* configuration further incorporates the specialist branch and enhanced cross-modal alignment, yet still omits lesion-guided attention calibration. Finally, *SuG* denotes the full model integrated with all proposed components.

Annotated lesion grounding. Table 6 reports voxel-level grounding performance across four annotated lesion categories. AUC measures lesion–background discriminability, AUPR assesses localization under class imbalance, and Cohen's d quantifies the separation between lesion and non-lesion heatmap responses. Higher scores indicate better grounding. The generalist baseline

achieves an average AUC of 47.0%, AUPR of 13.1%, and Cohen's d of -0.1. While specialist supervision provides modest gains, SuG achieves 90.8% AUC, 59.5% AUPR, and Cohen's d of 2.9, demonstrating substantially improved lesion-focused grounding.

Qualitative analysis. Fig. 3 visualizes attention maps for annotated and unannotated lesions. Compared with the baseline and the specialist-enhanced generalist without calibration, SuG yields sparser, highly localized attention responses while effectively suppressing irrelevant regions. Taking stomach tumors as an illustrative case, the baseline tends to fixate on gastric contents rather than the wall. Although specialist supervision shifts some attention to the gastric wall, it fails to fully decouple from the contents. In contrast, SuG precisely focuses on the thickened gastric wall corresponding to the lesion, effectively filtering out irrelevant background noise.

Generalization to unannotated lesions. The right panel of Fig. 3 further demonstrates that SuG effectively localizes diverse lesions, including pleural effusion, renal cyst, gallbladder stone, and bladder stone, without direct voxel-level supervision for these categories. Quantitative evaluation on the fully external KiTS23 dataset further confirms SuG's grounding capability for kidney tumors and kidney cysts, whose voxel-level annotations are not used during training (Table A9). This suggests that SuG acquires transferable lesion-aware representations, moving beyond simple memorization of the training distribution. Additional visualizations and transfer results are in Fig. A4.

3.5 Ablation study

Table 7 evaluates the contribution of each component within SuG on the MedVL-CT69K validation set spanning three diagnostic paradigms: general disease diagnosis, lesion-centric diagnosis, and generalization to unannotated localized lesion types (details in Table A4).

Building upon the configurations defined in Section 3.4, we start from the *Generalist baseline* and incorporate the specialist branch, which yields substantial gains and confirms that specialist supervision provides strong lesion-aware priors. Introducing multi-scale feature aggregation and ℓ_2 normalization further enhances performance, demonstrating the benefits of stable multi-scale fusion for both general and lesion-centric diagnosis. Finally, lesion-guided attention calibration introduces the most significant improvements in lesion-sensitive tasks: class-specific lesion attention calibration optimizes performance on supervised categories, while class-agnostic lesion attention calibration substantially strengthens generalization to unannotated localized lesions. Collectively, these results validate our holistic SuG design and demonstrate that each component contributes in a mutually complementary manner.

Table 7: **Ablation study (diagnostic AUC) on the MedVL-CT69K validation set.** The effectiveness of each proposed component is evaluated by cumulative integration into the baseline.

Variant	general diseases	four annotated lesion-centric diseases	unannotated localized lesion diseases
Generalist baseline	80.2	75.1	74.9
+ Specialist	84.7	81.6	81.7
+ Multi-scale feat.	85.7	82.3	82.0
+ ℓ_2 norm	87.4	84.4	84.8
+ Class-specific lesion attn.	87.3	87.1	86.1
+ Class-agnostic lesion attn.	**89.5**	**87.6**	**89.2**

4 Conclusion

We presented SuG, a unified framework toward super-generalist medical image understanding. By integrating specialist-derived anatomical and lesion-aware priors into vision–language learning and calibrating cross-modal attention with lesion masks, SuG combines broad diagnostic coverage with fine-grained lesion-centric understanding and reliable grounding. Extensive experiments on abdominal and chest CT benchmarks demonstrate that SuG improves general disease diagnosis, achieves specialist-competitive performance on fine-grained lesion-centric diagnosis, and provides more faithful lesion grounding. These findings support the promise of transferring specialist knowledge into generalist medical AI.

Limitations and future work. First, while currently focused on CT benchmarks, the proposed framework can be extended to accommodate diverse imaging modalities. Furthermore, we anticipate that the class-agnostic lesion grounding can be further improved and validated as larger-scale, high-fidelity voxel-level annotations become available.

AI USE STATEMENT

We used generative AI to extract anatomy-specific descriptions and disease/core-sign text from radiology reports as part of the proposed method, as detailed in Appendix B. We also used generative AI to improve the language and readability of the manuscript. We developed the research hypotheses and methodology, interpreted the results, and drew the main scientific conclusions without generative AI assistance. We take responsibility for the final content of this work.

ETHICS STATEMENT

This study used retrospective clinical CT images and their associated radiology reports. The use of these data was approved by the relevant ethics committee(s); identifying details have been withheld from the manuscript for anonymous peer review. All clinical data were de-identified before analysis and managed in accordance with institutional data governance requirements. Public datasets were used in accordance with their respective access requirements.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the main experimental results, including the model design, training objectives, three-stage optimization pipeline, dataset settings, evaluation tasks, metrics, and implementation details. These are described in the main paper and appendix, especially in the Method section and Appendix sections on datasets, implementation details, and evaluation protocols. The code and models will be released upon acceptance.

REFERENCES

Carina Albuquerque, Roberto Henriques, and Mauro Castelli. Deep learning-based object detection algorithms in medical imaging: Systematic review. *Heliyon*, 11(1), 2025.

Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6): 954–961, 2019.

Yuki Arita, Christian Roest, Thomas C Kwee, Ramesh Paudyal, Alfonso Lema-Dopico, Stefan Fransen, Daisuke Hirahara, Eichi Takaya, Ryo Ueda, Lisa Ruby, et al. Advancements in artificial intelligence for prostate cancer: Optimizing diagnosis, treatment, and prognostic assessment. *Asian Journal of Urology*, 2025.

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, pp. rs–3, 2024.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11):713, 2020.

Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. Medianomaly: A comparative study of anomaly detection in medical images. *Medical Image Analysis*, 102:103500, 2025.

Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12):3033–3043, 2023.

Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11238–11247, 2024.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23041–23050, 2025.

Dengbo Chen, Ziwei Zhao, Kexin Zhang, Shishuang Zhao, Junjie Hou, Yaqian Wang, Nianxi Liao, Anlan Sun, Fei Gao, Jia Ding, et al. Umind-vl: A generalist ultrasound vision-language model for unified grounded perception and comprehensive interpretation. *arXiv preprint arXiv:2511.22256*, 2025.

Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1970–1980, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Siddhant Dogra, Xiaoman Zhang, Ezequiel Silva III, and Pranav Rajpurkar. The financial, operational, and clinical advantages of generalist radiology ai. *Radiology*, 316(3):e242362, 2025.

Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857, 2021.

Heng Guo, Jianfeng Zhang, Jiaxing Huang, Tony CW Mok, Dazhou Guo, Ke Yan, Le Lu, Dakai Jin, and Minfeng Xu. Towards a comprehensive, efficient and promptable anatomic structure segmentation model using 3d whole-body ct scans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3247–3256, 2025.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*, 5, 2024a.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*, 2024b.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, Xi Lin, Mingxiang Wu, Yifan Peng, George Shih, Ziyang Xu, Xian Wu, Qiong Wang, Ronald Cheong Kin Chan, Xiaohui Duan, Varut Vardhanabhuti, Winnie Chiu Wing Chu, Yefeng Zheng, Pranav Rajpurkar, Kang Zhang, and Hao Chen. Towards generalizable ai in medicine via generalist-specialist collaboration. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01653-3. URL https://www.nature.com/articles/s41551-026-01653-3.

Nazia Hossain, Xintong Jiang, Yu Tian, Philippe Seguin, O Grant Clark, and Shangpeng Sun. Vision-language semantic grounding for multi-domain crop-weed segmentation. *arXiv preprint arXiv:2602.23677*, 2026.

Yuqi Hu, Xufang Luo, and Zilong Wang. Segmentation helps understanding: Mask-infused vision-language pre-training for 3d medical images.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Amna Ikram and Azhar Imran. Resvit fusionnet model: An explainable ai-driven approach for automated grading of diabetic retinopathy in retinal images. *Computers in Biology and Medicine*, 186:109656, 2025.

Egwonor Loveth Irede, Omowunmi Rebecca Aworinde, Ogunnaike Korede Lekan, Osemudiamhen D Amienghemhen, Tochukwu Perpetua Okonkwo, Asishana Paul Onivefu, and Ikhazuagbe H Ifijen. Medical imaging: a critical review on x-ray imaging for the detection of infection. *Biomedical Materials & Devices*, 4(1):1–45, 2026.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Haibo Jin, Haoxuan Che, Sunan He, and Hao Chen. A chain of diagnosis framework for accurate and explainable radiology report generation. *IEEE Transactions on Medical Imaging*, 2025.

KiTS Challenge Organizers. The 2023 kidney and kidney tumor segmentation challenge (kits23), 2023. URL `https://kits-challenge.org/kits23/`.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *Medical Image Analysis*, pp. 103749, 2025.

Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 4652–4660, 2025a.

Yuheng Li, Yuxiang Lai, Maria Thor, Deborah Marshall, Zachary Buchwald, David S. Yu, and Xiaofeng Yang. Towards universal text-driven ct image segmentation. *arXiv preprint arXiv:2503.06030*, 2025b.

Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Kai Cao, Le Lu, Jiebo Luo, and Ling Zhang. Ct-glip: 3d grounded language-image pretraining with ct scans and radiology reports for full-body scenarios. *arXiv preprint arXiv:2404.15272*, 2024.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. Imitate: Clinical prior guided hierarchical vision-language pre-training. *IEEE Transactions on Medical Imaging*, 2024.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European conference on computer vision*, pp. 685–701. Springer, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Aranya Saha, Tanvir Ahmed Khan, Ismam Nur Swapnil, and Mohammad Ariful Haque. Clarify: A specialist-generalist framework for accurate and lightweight dermatological visual question answering. *arXiv preprint arXiv:2508.18430*, 2025.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*, 2025.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022a.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022b.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems*, 35:33536–33549, 2022.

Zhikang Wang, Yumeng Zhang, Yingxue Xu, Seiya Imoto, Hao Chen, and Jiangning Song. Histo-genomic knowledge association for cancer prognosis from histopathology whole slide images. *IEEE Transactions on Medical Imaging*, 44(5):2170–2181, 2025.

Liang Wen, Hui Sun, Guobiao Liang, and Yue Yu. A deep ensemble learning framework for glioma segmentation and grading prediction. *Scientific Reports*, 15(1):4448, 2025.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 171–180. Springer, 2021.

Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Zhaohu Xing, and Lei Zhu. Diffmic-v2: Medical image classification via improved diffusion network. *IEEE Transactions on Medical Imaging*, 44 (5):2244–2255, 2025.

Jianpeng Zhang, Xianghua Ye, Jianfeng Zhang, Yuxing Tang, Minfeng Xu, Jianfei Guo, Xin Chen, Zaiyi Liu, Jingren Zhou, Le Lu, et al. Parse and recall: Towards accurate lung nodule malignancy prediction like radiologists. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 199–209. Springer, 2023a.

Peng Zhang, Jingjing Yang, Xiaolong Zhong, Heloisa Sobreiro Selistre-de Araujo, Stergios Boussios, Yongneng Ma, and Hua Fang. A novel pd-1/pd-l1 pathway-related seven-gene signature for the development and validation of the prognosis prediction model for breast cancer. *Translational cancer research*, 13(3):1554–1566, 2024.

Qi Zhang, Jianpeng Zhang, Weiwei Cao, Zilin Lu, Wanxing Chang, Haonan Ding, Cao Chen, Zhi Li, Xing Xue, Sinuo Wang, et al. An expert-level generalist ai for abdominal ct diagnosis. *Science*, 393(6817):eaec6129, 2026.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023b.

Yichi Zhang, Bohao Lv, Le Xue, Wenbo Zhang, Yuchen Liu, Yu Fu, Yuan Cheng, and Yuan Qi. Semisam+: rethinking semi-supervised medical image segmentation in the era of foundation models. *Medical Image Analysis*, pp. 103733, 2025.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*, 2023.

Xichuan Zhou, Qianqian Song, Jing Nie, Yujie Feng, Haijun Liu, Fu Liang, Lihui Chen, and Jin Xie. Hybrid cross-modality fusion network for medical image segmentation with contrastive learning. *Engineering Applications of Artificial Intelligence*, 144:110073, 2025a.

Yukun Zhou, Paul Nderitu, Jocelyn Hui Lin Goh, Justin Engelmann, Siegfried K Wagner, Anran Ran, Hongyang Jiang, Lie Ju, Ke Zou, Sahana Srinivasan, et al. Generalist versus specialist vision foundation models for ocular disease and oculomics. *arXiv preprint arXiv:2509.03421*, 2025b.

A Related work

A.1 Medical specialist AI

Specialist models have long been the dominant paradigm in medical image analysis, typically designed for disease-specific supervised learning tasks such as segmentation (Koleilat et al., 2025; Li et al., 2025a; Zhang et al., 2025), detection (Cai et al., 2025; Irede et al., 2026; Albuquerque et al., 2025), grading (Ikram & Imran, 2025; Yang et al., 2025; Wen et al., 2025), and prognosis prediction (Zhang et al., 2024; Arita et al., 2025; Wang et al., 2025). Among these tasks, segmentation and localization of anatomical structures and lesions are fundamental challenges. U-Net (Ronneberger et al., 2015) and its variants (Zhou et al., 2021; Isensee et al., 2018; Xie et al., 2021; Hatamizadeh et al., 2021) form the foundation of most specialist models. In particular, nnUNet (Isensee et al., 2018) has become one of the most successful examples, achieving highly competitive performance on a wide range of medical image segmentation benchmarks. However, developing specialist models inherently requires high-quality annotated datasets. Taking segmentation as an example, training such models typically demands large-scale voxel-level annotations. While specialist models excel in their dedicated tasks and can achieve very high accuracy, they face two key limitations. First, extending to new tasks incurs high costs, as it often requires collecting new annotations and retraining from scratch. Second, their task scope is narrow, with limited cross-anatomy and cross-disease transferability in open-world settings.

A.2 Medical generalist AI

Vision–language models (VLMs) have emerged as a promising pathway toward generalist medical AI. A mainstream approach follows the CLIP-style contrastive learning paradigm, where paired medical images and radiology reports are used to align visual and textual representations (Zhang et al., 2022; Tiu et al., 2022a; Boecking et al., 2022; Huang et al., 2023; Zhou et al., 2023; Zhang et al., 2023b; Lin et al., 2023). Once trained, such models exhibit broad task generalization capabilities, enabling zero-shot or few-shot diagnosis across diverse conditions. Due to the limited availability of large-scale paired medical data, early explorations focused primarily on 2D chest X-ray scenarios. Pioneer studies (Zhang et al., 2022; Tiu et al., 2022b) demonstrated that this paradigm can endow models with medical image understanding and diagnostic capabilities without additional annotated labels, leveraging only image–report pairs. To improve cross-modal alignment quality, subsequent methods have incorporated local alignment into contrastive learning. For example, LoVT (Müller et al., 2022) and MGCA (Wang et al., 2022) explicitly model correspondences between local image regions and corresponding report phrases or sentences. This alleviates the semantic loss caused by purely global alignment, capturing fine-grained clinical details. More recently, researchers have extended medical VLMs to 3D volumetric data such as CT, exploring generalist AI in 3D medical image understanding (Blankemeier et al., 2024; Hamamci et al., 2024b; Shui et al., 2025; Cao et al., 2025). Although GSCo (He et al., 2026) proposes a collaborative framework for generalist and specialist models, it limits the specialist's role to an external advisor that provides inference-time guidance, rather than achieving a deep architectural integration. In contrast, our approach internalizes specialist knowledge directly within the model architecture. This structural synergy allows us to simultaneously satisfy three core objectives, i.e. broad diagnostic coverage, specialist-competitive accuracy, and interpretable lesion grounding, which remain fundamentally different from GSCo's retrieve-and-predict paradigm. As a result, these generalist approaches still remain behind specialist models on tasks that require fine-grained lesion perception and localization, particularly when interpretable lesion-grounded diagnosis is needed.

A.3 Language-driven medical image segmentation

Another related line of research explores the use of pretrained vision–language alignment knowledge for medical image segmentation. These methods are closely related to our setting at the architectural level, as they also integrate visual encoding, textual encoding, and dense prediction within a unified framework. However, their objective is fundamentally different from ours. Existing language-driven segmentation methods are primarily designed to exploit such cross-modal alignment to improve segmentation performance, i.e., to build a stronger specialist model. Their main emphasis is typically on how to better transfer pretrained VLM knowledge to segmentation, for example by designing decoder architectures that more effectively fuse image and text features to enhance segmentation

Table A1: **List of anatomical structures used in prompts for disease and core imaging sign extraction.**

adrenal gland	aorta	erector spinae muscle
brain	clavicle	large bowel
duodenum	esophagus	face
femur	gallbladder	gluteus muscle
heart	hip joint	humerus
iliac artery	iliac vena	iliopsoas muscle
inferior vena cava	kidney	liver
lung	pancreas	portal vein
pulmonary artery	rib	sacrum
scapula	small bowel	spleen
stomach	trachea	bladder
cervical vertebrae	lumbar vertebrae	thoracic vertebrae

quality (Li et al., 2025b; Zhou et al., 2025a; Hossain et al., 2026). In contrast, our goal is not to use language supervision mainly to improve segmentation itself. Instead, we use segmentation-derived specialist knowledge to retain the broad diagnostic capability of a generalist model while improving its ability to recognize and localize focal lesions.

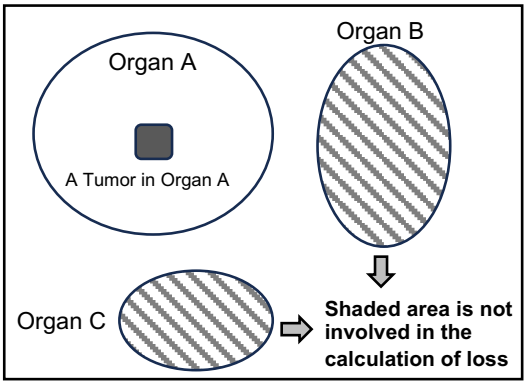

Figure A1: **Illustration of the masked loss strategy for class-agnostic lesion segmentation.**

B Prompt Templates for Report Parsing

B.1 Prompt template for anatomy-level report description extraction

For reproducibility, we provide the prompt template used to extract anatomy-level textual descriptions from CT reports using Qwen. These extracted descriptions are used to construct the anatomy-specific text input.

In the template, `{TARGET_ANATOMY}` denotes the queried anatomical region, and `{REPORT_TEXT}` denotes the input report.

B.2 Prompt template for disease and core imaging sign extraction

We also provide the prompt used to extract diseases and their supporting core imaging signs from CT reports. This prompt is used to derive lesion-related descriptions for attention calibration and to support report-conditioned filtering.

In the template, `{TARGET_ANATOMY_LIST}` refers to the predefined target anatomy list (see Table A1), and `{REPORT_TEXT}` denotes the input report.

```
You are a professional radiologist. Please extract the descriptive
information about the specific anatomy from this English CT diagnostic
report.

Target anatomy: {TARGET_ANATOMY}

Requirements:
    • Extract only the content directly related to the target
      anatomy.
    • If the report mentions specific subregions, parts, involved
      sites, or laterality of the target anatomy, preserve these
      anatomical details in the extracted description.
    • Keep the extracted text concise and faithful to the original
      report.
    • Do not add explanations, background knowledge, or inferred
      findings that are not explicitly stated in the report.
    • Even if the target anatomy has multiple independent parts,
      treat it as a whole and return only one integrated description.
Output format:
    • Output exactly one line in the format: {TARGET_ANATOMY}:
      description.
    • Always use the queried target anatomy as the unified prefix.
    • If no relevant information is found, output: {TARGET_ANATOMY}:
      no relevant finding.
    • Do not return multiple entries, JSON, or additional comments.
CT report input: {REPORT_TEXT}
```

Figure A2: **Prompt template for anatomy-specific report extraction.** We query the LLM independently for each target anatomy to ensure high consistency and quality in the extracted descriptions, which also simplifies the parsing process compared to multi-anatomy extraction.

C Experiment setting Details

C.1 Dataset

C.1.1 Abdomen scenario

For the abdominal setting, we use two large-scale image–report datasets, MedVL-CT69K (Shui et al., 2025) and Merlin (Blankemeier et al., 2024), together with four anatomy-specific tumor segmentation datasets, which we collectively denote as LesionSegAbdomen. The image–report datasets contain abdominal CT scans paired with detailed radiology reports and are mainly used to support generalist vision–language learning and broad-coverage diagnosis. In contrast, LesionSegAbdomen provides voxel-level lesion annotations and is used for specialist supervision as well as lesion-centric evaluation. Across all abdominal and chest settings, there is no overlap between the segmentation datasets and the image–report datasets used within the same experimental setting.

MedVL-CT69K. MedVL-CT69K (Shui et al., 2025) is a large-scale abdominal CT dataset containing 272,124 CT scans from 69,086 patients, each paired with a corresponding radiology report. Its validation and test sets are annotated by expert radiologists with 36 and 54 disease labels, respectively. Following the official fVLM split (Shui et al., 2025), we divide the dataset into training, validation, and test sets with 64,476, 1,151, and 3,459 patients, respectively.

Merlin. Merlin (Blankemeier et al., 2024) is a public abdominal CT dataset collected from the Stanford Hospital Emergency Department between 2012 and 2018. It contains 25,494 CT scans from 18,317 patients, each paired with a radiology report. Following the original benchmark setting (Blankemeier et al., 2024), we use 11,010 patients for training, 3644 patients for validation, and 3,667 patients for testing. The dataset provides ground-truth labels for 30 disease categories, which we use to evaluate diagnostic performance.

```
 You are a medical imaging report analysis assistant.  Extract
diseases and their supporting core imaging signs from an English CT
report.

Target anatomies:  {TARGET_ANATOMY_LIST}

Requirements:
- Each disease must be mapped to exactly one target anatomy.
- Discard diseases that cannot be clearly mapped to the target anatomy
list.
- Output keys must follow:  anatomy-english_disease_name-lesion_type.
- If IMPRESSION / CONCLUSION / ASSESSMENT exists, use it as the primary
diagnosis list.
- Otherwise, convert abnormal FINDINGS statements into disease/problem
labels.

Lesion types:
- Localized Lesion:  focal abnormality, e.g., lesion, mass, nodule,
tumor, cyst, calcification, thrombus, aneurysm, stricture, focal wall
thickening.
- Diffuse Lesion:  generalized or anatomy-wide abnormality, e.g.,
diffuse inflammation, steatosis, hepatomegaly, edema, generalized
dilatation.
- If uncertain, default to Diffuse Lesion.

Procedure:
1.  Extract explicit abnormal imaging findings from
FINDINGS/DESCRIPTION.
2.  Keep only generic core signs; remove size, exact location,
laterality, measurements, and comparison terms.
3.  Identify diseases from IMPRESSION/CONCLUSION if present; otherwise
derive disease/problem labels from abnormal FINDINGS.
4.  For each disease, collect the supporting core signs from FINDINGS.
5.  Map each disease to exactly one target anatomy using predefined
rules and standard medical knowledge.
6.  Build a comma-separated English core sign string.

Representative anatomy mapping rules:
- large bowel:  colon, rectum, cecum, appendix, anal canal
- small bowel:  jejunum, ileum
- duodenum:  duodenum only
- gallbladder:  gallbladder, cystic duct, and biliary duct terms if no
separate duct anatomy exists
- portal vein:  portal vein system, including splenic vein
- Discard anatomies not in the target list, e.g., uterus, ovaries,
prostate, lymph nodes

Output format:
- Output only a single JSON dictionary, with no extra text.
- Key format:  anatomy-english_disease_name-lesion_type
- lesion_type must be exactly "Localized Lesion" or "Diffuse Lesion"
- Value:  the English core sign string (comma-separated).

CT report input:  {REPORT_TEXT}
```

Figure A3: **Prompt template designed for disease and core imaging sign extraction from CT reports.**

KiTS23. KiTS23 (KiTS Challenge Organizers, 2023) is a public external kidney CT benchmark with voxel-level masks for kidneys, kidney tumors, and kidney cysts. We use its tumor and cyst masks only for the unannotated lesion grounding evaluation in Appendix E.2.

Table A2: **Statistics of lesion segmentation datasets.** Listed are the counts of CT scans and patients across train, test, and total sets for each dataset.

Scenario	Dataset	Train		Test		Total	
		CT num.	Patient num.	CT num.	Patient num.	CT num.	Patient num.
Abdomen	LesionSegC	3385	847	435	100	3820	947
	LesionSegS	3462	1865	178	100	3640	1965
	LesionSegL	6631	1843	734	200	7365	2043
	LesionSegP	6888	2296	2277	759	9165	3055
Chest	LesionSegLung	7755	4393	395	100	8150	4493

LesionSegAbdomen. We refer to the union of four anatomy-specific abdominal tumor segmentation datasets as LesionSegAbdomen. It includes LesionSegC (colorectal), LesionSegS (stomach), LesionSegL (liver), and LesionSegP (pancreas), each providing voxel-level lesion annotations for its corresponding anatomy. In total, these four datasets contain 23,990 CT scans from 8,010 patients. We perform train and test splitting independently for each sub-dataset, resulting in 6,851 patients for training and 1,159 patients for testing after aggregation (see Table A2). Each CT scan is associated with a tumor mask annotated by senior radiology experts.

C.1.2 CHEST SCENARIO

For the chest setting, we use three CT datasets to evaluate the generalization of SuG to chest imaging: the large-scale image–report dataset CT-RATE (Hamamci et al., 2024a), the external multi-disease benchmark RAD-ChestCT (Draelos et al., 2021), and a lung tumor segmentation dataset, denoted as LesionSegLung, for specialist supervision.

CT-RATE and RAD-ChestCT. CT-RATE (Hamamci et al., 2024a) contains 25,692 non-contrast chest CT volumes from 21,304 unique patients, each paired with a radiology report. Following the benchmark protocol, we use 20,000 patients for training and 1,304 patients for testing. To assess external generalization, we further evaluate on the RAD-ChestCT dataset (Draelos et al., 2021), which contains 3,630 chest CT scans. Following prior work (Shui et al., 2025), we report diagnostic performance on 16 diseases for CT-RATE and 13 diseases for RAD-ChestCT.

LesionSegLung. For specialist supervision in the chest setting, we collect a lung tumor segmentation dataset, LesionSegLung. It contains 4,493 patients, randomly split into 4,393 training cases and 100 test cases (see Table A2). Each CT scan is annotated with a lesion mask by senior radiology experts.

C.1.3 EXPERIMENTAL SETTINGS

We consider three experimental settings, each corresponding to a SuG model trained using the same progressive three-stage pipeline. The three settings differ only in the pairing between the specialist and generalist training datasets. Specifically, **Setting 1** corresponds to the abdominal CT scenario using LesionSegAbdomen together with MedVL-CT69K; **Setting 2** also targets the abdominal CT scenario, but pairs LesionSegAbdomen with Merlin; and **Setting 3** corresponds to the chest CT scenario using LesionSegLung together with CT-RATE. In each setting, the lesion segmentation dataset provides voxel-level specialist supervision for learning anatomy- and lesion-aware representations, while the image–report dataset supports generalist vision–language learning. This design allows us to evaluate SuG under both abdominal and chest scenarios, as well as across different image–report sources.

C.2 BASELINES AND IMPLEMENTATION DETAILS

C.2.1 BASELINES

We compare SuG against three categories of methods: specialist segmentation models, generalist vision–language models, and specialist–generalist synergy (S+G) methods that integrate segmentation-based specialist modeling with vision–language learning. For the specialist baseline, we adopt nnU-Net (Isensee et al., 2018), a widely used supervised segmentation framework, and train it for anatomy-specific lesion-centric diagnosis. The procedure used to derive diagnostic probabilities

from its segmentation outputs is described in Appendix C.3. The generalist baselines include CLIP (Radford et al., 2021), LOVT (Müller et al., 2022), Imitate (Liu et al., 2024), BIUD (Cao et al., 2024), Merlin (Blankemeier et al., 2024), fVLM (Shui et al., 2025), and ViSD-Boost (Cao et al., 2025). As representative specialist–generalist synergy baselines, we include three methods: OpenVocabCT (Li et al., 2025b), HCFNet (Zhou et al., 2025a), and VLWS (Hossain et al., 2026). It is worth noting that existing specialist–generalist synergy methods are mainly developed to enhance specialist, rather than to strengthen the broad diagnostic capability of generalist models. We therefore regard these methods as the closest available baselines for specialist–generalist collaboration, although they are not specifically designed for the super-generalist setting considered in this work. This gap also highlights the value of super-generalist that explicitly unifies broad diagnostic coverage with specialist-level lesion understanding. All baseline methods are re-trained on the same training data used by SuG for fair comparison.

C.2.2 Implementation details

Stage 1: Specialist pretraining. In the first stage, we train the specialist pathway within the nnU-Net framework. The architecture is automatically configured by nnU-Net. To maintain consistency with the subsequent vision–language alignment stage, we use a target spacing of $[5.0, 1.0, 1.0]$ and a patch size of $96 \times 256 \times 384$ (Shui et al., 2025; Cao et al., 2025). Training is performed for 1000 epochs using SGD with a batch size of 2. The learning rate is initialized at 1×10^{-2} and decayed to 0 over the course of training. This stage is trained on a single NVIDIA H20 GPU. **Stage 2: Specialist-enhanced vision–language alignment**. In the second stage, we initialize the visual encoder from the pretrained specialist model obtained in Stage 1 and perform specialist-enhanced vision–language learning. A pretrained BERT (Devlin et al., 2019) is adopted as the text encoder. We jointly optimize the segmentation objective and the vision–language alignment objective for 30 epochs using AdamW with a batch size of 48. The learning rate is initialized at 1×10^{-4} and decayed to 1×10^{-6} with cosine annealing. This stage is trained on 24 NVIDIA H20 GPUs. **Stage 3: Lesion-guided attention calibration**. In the final stage, we initialize the model from Stage 2, introduce the lesion-guided attention calibration loss, and continue training for 5 epochs. Unless otherwise specified, all remaining training settings are identical to those used in Stage 2. We implement all models in Python 3.11 using PyTorch 2.6.0 and CUDA 12.4.

C.3 Diagnostic score derivation for specialist baselines

Our specialist baselines are built on nnU-Net segmentation models, which produce voxel-wise lesion predictions rather than scan-level disease scores. To evaluate these models under classification metrics, we convert their segmentation outputs into diagnostic scores. For an input CT scan, the segmentation model outputs both a segmentation map and the corresponding class probability map. We first perform connected-component analysis on the segmentation map to identify all spatially independent lesion regions for the target class. For each connected lesion region, the lesion score is defined as the average probability of the target class over all voxels within that region. If a CT scan contains multiple disconnected lesions of the same class, we use the maximum lesion score among all such regions as the final diagnostic score for that disease in the scan. This strategy allows voxel-wise specialist models to be directly compared with other methods under classification metrics.

C.4 Evaluation tasks and metrics

C.4.1 Evaluation tasks

We design five complementary evaluation tasks to assess whether SuG satisfies the three key requirements of a super-generalist model: broad diagnostic coverage, specialist-competitive lesion-centric diagnosis, and reliable lesion grounding.

General disease diagnosis. We first evaluate the broad-coverage diagnostic capability of SuG on multi-disease test sets to examine whether it can serve as a strong generalist model. This evaluation is conducted on MedVL-CT69K, Merlin, CT-RATE, and RAD-ChestCT, covering both abdominal and chest CT scenarios.

Fine-grained lesion-centric diagnosis. We further evaluate SuG on lesion-centric disease categories that require fine-grained lesion perception and are traditionally challenging for generalist models. We

Table A3: **Label mapping from MedVL-CT69K to the four tumor categories used for evaluation.**

Four-tumor category	Original labels
Large bowel tumor	rectal cancer, colon cancer
Liver tumor	liver cyst, liver cancer, liver metastatic tumor
Stomach tumor	gastric cancer
Pancreas tumor	pancreatic cancer

include four lesion-centric categories from LesionSegAbdomen, as well as four lesion-related disease categories from the 54 test diseases in MedVL-CT69K. Detailed label mappings are provided in the Table A3.

Annotated lesion grounding. To evaluate whether SuG can localize lesions that receive supervision during training, we assess lesion grounding primarily in the abdominal setting. Specifically, we evaluate grounding performance on four annotated lesion categories using the LesionSegAbdomen test set.

Generalization to lesions without category-specific voxel supervision. Beyond the supervised lesion categories, we evaluate whether SuG can generalize to lesions that lack category-specific specialist annotations. MedVL-CT69K provides qualitative examples and diagnosis-based evaluation for such categories. We additionally perform quantitative grounding on the external KiTS23 dataset for kidney tumors and kidney cysts, neither of which appears in the voxel-level annotations used for specialist training.

Radiology report generation. Finally, we evaluate whether the specialist-enhanced visual representations learned by SuG can transfer to cross-modal medical image interpretation beyond diagnosis and grounding. To do so, we pair the pretrained vision encoder with a text decoder and assess its performance on the downstream radiology report generation task.

C.4.2 Evaluation metrics

Disease diagnosis tasks: Following previous works (Shui et al., 2025; Blankemeier et al., 2024; Hamamci et al., 2024a; Cao et al., 2025), we report AUC, F1-score, sensitivity (SE), and specificity (SP). **Annotated and unannotated lesion grounding:** We perform qualitative and quantitative evaluations. Qualitatively, we visualize image–text similarity maps (attention maps) to examine whether the model attends to lesion-relevant regions. Quantitatively, we compare the predicted heatmaps with lesion masks using three voxel-level metrics. Specifically, AUC measures overall discriminability between lesion and background voxels across all thresholds; AUPR is particularly informative under class imbalance and reflects the precision–recall trade-off for lesion localization; and Cohen's d quantifies the standardized separation between heatmap responses on lesion versus non-lesion voxels. Higher values indicate more accurate lesion localization and clearer separation between lesion and background regions. **Radiology report generation:** We report both clinical efficacy metrics and conventional NLG metrics. The former include precision, recall, F1-score, and GREEN, while the latter include BLEU4, ROUGE-L, METEOR, and CIDEr.

D Merlin Training and Inference Protocol

Merlin contains both anatomy-grounded disease labels that naturally map to one of our anatomical targets and scan-level labels that are not naturally associated with a single anatomy. To handle both cases in a unified framework, we use a hybrid training and inference protocol.

D.1 Alternating training objectives on Merlin

In addition to the anatomy-wise contrastive learning described in the main paper, Merlin training alternates between two alignment modes:

Table A4: **Disease subsets defined for validation-set evaluation.** The "General diseases" set comprises all 36 disease labels. The "Unannotated localized lesions" subset consists of diseases not explicitly covered by the four-tumor specialist supervision, while the "Annotated lesions" subset corresponds to the four tumor categories used for lesion-centric evaluation.

Subset	Disease labels
General diseases	liver_cyst, liver_fatty liver, liver_cirrhosis, liver_intrahepatic bile duct dilatation, liver_metastatic tumor, liver_Glisson's sheath fluid, liver_liver cancer, portal vein_embolism, portal vein_portal hypertension, kidney_cyst, kidney_stone, kidney_hydronephrosis, kidney_atrophy, adrenal gland_thickening, adrenal gland_nodule, spleen_splenomegaly, spleen_infarction, spleen_hemangioma, gallbladder_adenomyomatosis, gallbladder_stone, gallbladder_inflammation, stomach_gastric cancer, stomach_wall thickening, pancreas_atrophy, pancreas_pancreatic duct dilatation, pancreas_inflammation, pancreas_pancreatic cancer, large bowel_appendicitis, large bowel_rectal cancer, large bowel_obstruction, large bowel_colon cancer, large bowel_fluid accumulation, large bowel_gas accumulation, small bowel_obstruction, small bowel_gas accumulation, small bowel_fluid accumulation
Unannotated localized lesions	liver_Glisson's sheath fluid, portal vein_embolism, kidney_cyst, kidney_stone, adrenal gland_nodule, spleen_infarction, spleen_hemangioma, gallbladder_adenomyomatosis, gallbladder_stone, large bowel_appendicitis
Annotated lesions	large bowel_tumor, liver_tumor, stomach_tumor, pancreas_tumor

Table A5: **Full names of anatomical abbreviations used in Table 2.**

Abbreviation	Full name
Adr	Adrenal gland
Bla	Bladder
Col	Colon
Eso	Esophagus
GB	Gallbladder
Hrt	Heart
Kid	Kidney
Liv	Liver
Lng	Lung
Pan	Pancreas
PV	Portal vein
SI	Small intestine
Spl	Spleen
Sto	Stomach
Sac	Sacrum
Avg	Average

(1) Anatomy-wise alignment. For anatomy-grounded diseases, we use the same anatomy-wise contrastive strategy as in the main paper: anatomy-specific visual embeddings extracted from segmented anatomy regions are aligned with anatomy-level textual descriptions.

(2) Whole-body image–report alignment. For whole-body diseases, we additionally perform whole-body image–report contrastive learning. In this branch, the full CT volume is encoded into a global visual representation and aligned with the complete report text. The same bidirectional image–text contrastive objective is used, but applied at scan level rather than anatomy level. In practice, these two alignment modes are alternated during Merlin training. This enables the model to learn both anatomy-specific and scan-level semantics.

D.2 Inference on Merlin

At test time, we select the inference mode according to the disease subset (see Table A6):
Anatomy-level inference. For anatomy-wise diseases, we extract the visual embedding of the relevant anatomy and compute similarity to the corresponding disease prompt embedding.
Whole-body inference. For whole-body diseases, we compute a global visual embedding from the full CT and compare it with the disease prompt embedding.

Table A6: **Partition of Merlin diseases for the hybrid training and inference protocol.** Diseases are categorized into "Anatomy-wise" and "Whole-body" subsets, which are evaluated using anatomy-level and whole-body inference, respectively.

Subset	Disease list
Anatomy-wise	`aortic_valve_calcification`, `pleural_effusion`, `splenomegaly`, `abdominal_aortic_aneurysm`, `hepatic_steatosis`, `bowel_obstruction`, `hepatomegaly`, `cardiomegaly`, `pancreatic_atrophy`, `hydronephrosis`, `surgically_absent_gallbladder`, `atherosclerosis`, `submucosal_edema`, `hiatal_hernia`, `atelectasis`, `biliary_ductal_dilation`, `renal_cyst`, `renal_hypodensities`, `gallstones`, `fracture`, `appendicitis`
Whole-body	`free_air`, `metastatic_disease`, `lymphadenopathy`, `anasarca`, `ascites`, `coronary_calcification`, `thrombosis`, `prostatomegaly`, `osteopenia`

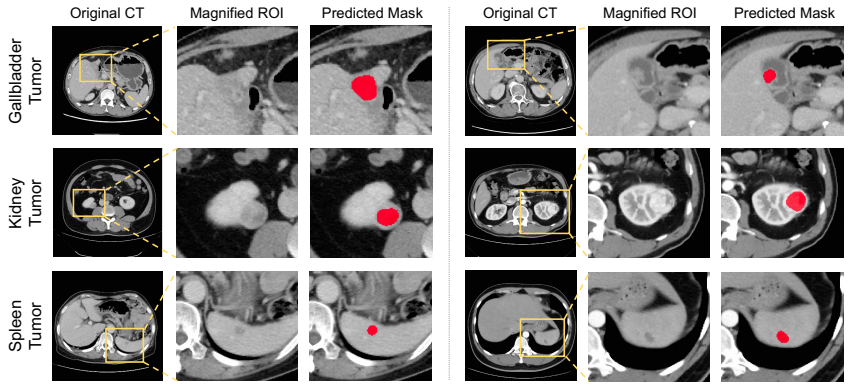

Figure A4: **Visualization of unannotated lesion segmentation.** For each representative case, the original CT slice, corresponding magnified ROI, and predicted lesion mask are presented.

E Additional Experiments and Robustness Analyses

This section presents additional evaluations of unified chest and abdominal training, grounding generalization to unannotated diseases, and segmentation annotation efficiency. It also examines robustness to report-parsing variations and report noise. Owing to computational resource constraints, the model-training experiments in this section use a batch size of 32.

E.1 Joint training on chest and abdominal CT scans

For the primary experiments, we use separate models for the abdominal and chest settings following the established protocol for each benchmark. To assess cross-region optimization, we jointly train a unified SuG model on abdominal CT scans from Merlin and chest CT scans from CT-RATE. The unified model improves all reported metrics on Merlin, CT-RATE, and the external RAD-ChestCT benchmark (Tables A7 and A8). These results demonstrate the effectiveness of SuG for joint training across chest and abdominal CT data.

Table A7: **Region-specific and unified SuG models on the Merlin test set.** AUC* and F1* follow the official balanced Merlin protocol, whereas AUC and F1 are computed on the full test set.

Model	AUC*	F1*	AUC	F1
Region-specific SuG	85.2	80.8	85.1	60.7
Unified SuG	**87.1**	**81.7**	**87.0**	**62.7**

Table A8: **Region-specific and unified SuG models on the chest CT benchmarks.**

Dataset	Model	AUC	ACC	F1	Precision
CT-RATE	Region-specific SuG	81.1	74.4	77.1	40.8
	Unified SuG	**82.2**	**75.4**	**78.1**	**42.1**
RAD-ChestCT	Region-specific SuG	73.4	68.1	71.9	36.9
	Unified SuG	**74.7**	**69.4**	**72.9**	**38.5**

E.2 Unannotated lesion grounding

We evaluate the grounding capability of SuG for unannotated diseases on the KiTS23 dataset (KiTS Challenge Organizers, 2023). The evaluated SuG model is trained using MedVL-CT69K and LesionSegAbdomen. Notably, no voxel-level annotations of kidney tumors or kidney cysts are used during training and KiTS23 is also a fully external dataset. As shown in Table A9, SuG achieves the best grounding performance for both categories. Compared with the + *Specialist* configuration, SuG improves the average AUC from 57.5 to 77.7, the average AUPR from 18.5 to 26.8, and the average Cohen's d from 0.2 to 1.4. These results confirm the generalizability of SuG's grounding capability to unannotated diseases on a fully external dataset.

Table A9: **Quantitative lesion grounding evaluation on the external KiTS23 dataset.**

Method	Kidney tumor			Kidney cyst			Average		
	AUC	AUPR	Cohen's d	AUC	AUPR	Cohen's d	AUC	AUPR	Cohen's d
Generalist baseline	43.6	16.6	-0.3	40.2	4.3	-0.3	41.9	10.4	-0.3
+ Specialist	67.2	32.2	0.7	47.9	4.8	-0.2	57.5	18.5	0.2
SuG	**80.3**	**40.6**	**1.4**	**75.0**	**13.0**	**1.3**	**77.7**	**26.8**	**1.4**

E.3 Segmentation annotation efficiency

Given the high cost of voxel-level segmentation annotations, we evaluate whether SuG maintains its performance with reduced segmentation supervision. Specifically, we train SuG with 10%, 25%, 50%, or 100% of the available segmentation training data. The abdominal model is evaluated on Merlin, while the chest model is evaluated on CT-RATE and the external RAD-ChestCT benchmark. All other training and evaluation settings remain unchanged within each dataset. Across Merlin, CT-RATE, and RAD-ChestCT, none of the reduced-data settings shows a statistically significant difference from the 100% reference on any reported metric (Tables A10, A11, and A12). These results demonstrate SuG's segmentation annotation efficiency across the evaluated abdominal and chest CT settings.

Table A10: **Segmentation annotation efficiency on the Merlin test set.** AUC* and F1* follow the official balanced Merlin protocol, whereas AUC and F1 are computed on the full test set. Significance is assessed against the 100% reference using two-sided paired permutation tests at the patient level. All reduced-data comparisons are not significant (n.s.).

Segmentation data	AUC*	F1*	AUC	F1
10%	84.30 (n.s.)	78.70 (n.s.)	84.30 (n.s.)	59.90 (n.s.)
25%	84.50 (n.s.)	79.30 (n.s.)	84.70 (n.s.)	60.20 (n.s.)
50%	85.07 (n.s.)	**80.49** (n.s.)	85.41 (n.s.)	**61.47** (n.s.)
100%	**85.10**	80.37	**85.49**	61.42

Table A11: **Segmentation annotation efficiency on the CT-RATE test set.** Significance is assessed against the 100% reference using two-sided paired permutation tests at the patient level. All reduced-data comparisons are not significant (n.s.).

Segmentation data	AUC	ACC	F1	Precision
10%	80.8 (n.s.)	**74.8** (n.s.)	**77.3** (n.s.)	**41.9** (n.s.)
25%	81.1 (n.s.)	74.5 (n.s.)	77.2 (n.s.)	41.3 (n.s.)
50%	81.2 (n.s.)	74.7 (n.s.)	**77.3** (n.s.)	41.3 (n.s.)
100%	**81.3**	74.6	**77.3**	41.2

Table A12: **External evaluation of the segmentation annotation efficiency on RAD-ChestCT.** Significance is assessed against the 100% reference using two-sided paired permutation tests at the patient level. All reduced-data comparisons are not significant (n.s.).

Segmentation data	AUC	ACC	F1	Precision
10%	72.4 (n.s.)	67.5 (n.s.)	71.2 (n.s.)	36.1 (n.s.)
25%	**73.6** (n.s.)	68.9 (n.s.)	72.4 (n.s.)	37.3 (n.s.)
50%	**73.6** (n.s.)	68.5 (n.s.)	72.1 (n.s.)	37.2 (n.s.)
100%	**73.6**	**69.0**	**72.5**	**37.4**

E.4 Robustness to report-parsing variations

Following the report-processing strategy of fVLM (Shui et al., 2025), we adopt a standardized pipeline to extract anatomy-specific information from CT reports. We fix the parsing LLM, prompt template, anatomy taxonomy, decoding configuration, and post-processing rules for all reports. This ensures that every report is processed using the same configuration and minimizes variability during data preparation.

Despite this standardized design, we further examine whether reasonable variations in report parsing affect the extracted reports and downstream SuG performance. Qwen3.7-Max with prompt P0 and the default decoding seed of 1234 serves as the reference configuration. We independently evaluate two alternative LLMs (`deepseek-v3.2` and `glm-5.2`), two prompt formulations (P1 and P2), and two alternative decoding seeds (12 and 34). All other parsing settings remain fixed within each comparison.

To assess the effects on parsed reports, we use the same fixed subset of 150 CT reports for all parsing configurations. **Extraction accuracy** compares each extraction with the original report using atomic fact-level precision, recall, and F1, together with a rubric-based report score assigned by GPT-5.6-terra at low reasoning effort. **Reference similarity** evaluates agreement with the reference extraction using GPT-5.6-terra semantic similarity and NLG metrics, including chrF++, ROUGE-L F1, and BLEU-4. As shown in Tables A13 and A14, the extractions produced under all parsing variations faithfully preserve the clinical content of the original reports and remain highly similar to the reference extraction.

To assess downstream stability, we apply each LLM and prompt variant to the complete training-report corpus. We retrain SuG using all resulting reports and evaluate both diagnosis and lesion grounding. The seed variants show the highest agreement with P0, so downstream retraining focuses on the LLM and prompt factors. Table A15 shows no statistically significant decrease in any diagnostic or grounding metric relative to P0. These results demonstrate that SuG maintains stable performance across variations in the report-parsing strategy.

Table A13: **Extraction accuracy under report-parsing variations.** Statistical significance is assessed against the reference using two-sided paired permutation tests at the patient level. $^{*}p < 0.05$; n.s., not significant.

LLM	Prompt	Seed	Evaluable reports	Report score (0–10)	F1	Precision	Recall
Qwen3.7-Max	P0	1234	150	9.580	0.915	0.891	0.941
DeepSeek-v3.2	P0	1234	150	9.529 (n.s.)	0.881 (*)	0.860 (n.s.)	0.903 (*)
GLM-5.2	P0	1234	150	9.579 (n.s.)	0.898 (n.s.)	0.877 (n.s.)	0.921 (n.s.)
Qwen3.7-Max	**P1**	1234	150	9.589 (n.s.)	0.904 (n.s.)	0.883 (n.s.)	0.926 (n.s.)
Qwen3.7-Max	**P2**	1234	149	9.532 (n.s.)	0.908 (n.s.)	0.883 (n.s.)	0.935 (n.s.)
Qwen3.7-Max	P0	**12**	150	9.578 (n.s.)	0.909 (n.s.)	0.885 (n.s.)	0.935 (n.s.)
Qwen3.7-Max	P0	**34**	150	9.521 (n.s.)	0.901 (n.s.)	0.873 (n.s.)	0.930 (n.s.)

Table A14: **Agreement between each parsing variation and the fixed reference extraction.** Similarity is scored from 0 to 10. High Agreement is the proportion of evaluable reports with a report-level score of at least 8/10.

LLM	Prompt	Seed	Evaluable reports	Similarity	High Agreement	chrF++	ROUGE-L F1	BLEU-4
DeepSeek-v3.2	P0	1234	150	9.617	98.0%	93.47	94.90	90.71
GLM-5.2	P0	1234	150	9.565	96.0%	92.22	94.52	89.70
Qwen3.7-Max	**P1**	1234	150	9.692	100.0%	93.84	95.73	91.99
Qwen3.7-Max	**P2**	1234	149	9.727	99.3%	94.94	96.40	93.29
Qwen3.7-Max	P0	**12**	150	9.913	100.0%	98.82	99.10	98.36
Qwen3.7-Max	P0	**34**	150	**9.918**	100.0%	**99.01**	**99.24**	**98.60**

Table A15: **Effects of report-parsing variations on downstream diagnosis and lesion grounding.** AUC* and F1* follow the official balanced Merlin protocol, whereas AUC and F1 are computed on the full test set. Significance is assessed against the reference using two-sided paired permutation tests at the patient level. $^{*}p < 0.05$; $^{***}p < 0.001$; n.s., not significant.

LLM	Prompt	Diagnosis				Grounding		
		AUC*	F1*	AUC	F1	AUC	AUPR	Cohen's d
Qwen3.7-Max	P0	85.10	80.37	85.49	61.42	88.07	58.63	2.65
DeepSeek-v3.2	P0	85.32 (n.s.)	80.59 (n.s.)	85.52 (n.s.)	62.48 (n.s.)	87.91 (n.s.)	60.85 (*)	2.69 (n.s.)
GLM-5.2	P0	85.46 (n.s.)	80.61 (n.s.)	85.67 (n.s.)	62.03 (n.s.)	88.62 (n.s.)	60.85 (n.s.)	2.81 (*)
Qwen3.7-Max	**P1**	85.71 (n.s.)	81.10 (n.s.)	85.81 (n.s.)	61.19 (n.s.)	88.78 (n.s.)	62.02 (***)	2.68 (n.s.)
Qwen3.7-Max	**P2**	85.21 (n.s.)	80.86 (n.s.)	85.21 (n.s.)	60.77 (n.s.)	87.63 (n.s.)	60.80 (n.s.)	2.57 (n.s.)

E.5 Robustness to report noise

The training of SuG relies on aligning CT scans with their corresponding radiology reports. The effect of report quality on model performance is therefore worth investigating. During data collection, three common forms of report noise may occur: missing pairs, incomplete reports, and mismatched pairs. Missing pairs can be readily detected and excluded during data quality control. We therefore investigate incomplete reports and mismatched pairs, which are more difficult to identify.

For **incomplete reports**, we vary both the proportion of affected reports and the extent of content loss. We randomly select 30%, 50%, or 80% of the Merlin training reports and remove 10% to 30% of the content from each selected report. For **mismatched pairs**, we evaluate mismatch rates of 10%, 30%, and 50%. We randomly select the corresponding proportion of training pairs and reassign their reports to different CT scans. These settings simulate progressively higher levels of report incompleteness and image-report misalignment. SuG is retrained for each noise setting, while all other experimental settings remain unchanged.

Table A16 shows that partial content loss within individual reports has a limited effect on performance. Even when 80% of the reports are affected, AUC* and F1* decrease by only 0.83 and 0.71 points, respectively. Neither decrease is statistically significant. In contrast, CT-report mismatch causes progressively greater degradation at mismatch rates of 30% and 50%. Mismatched reports associate CT scans with unrelated clinical findings and directly disrupt the image-report alignment central to SuG.

Table A16: **Robustness to simulated report noise on Merlin.** Values in parentheses are absolute changes from clean training. AUC* and F1* follow the official balanced Merlin protocol, whereas AUC and F1 are computed on the full test set.

Corruption	Affected reports	AUC*	F1*	AUC	F1
Clean	–	85.10	80.37	85.49	61.42
Incomplete	30%	84.51 (-0.59)	80.14 (-0.23)	84.71 (-0.78)	60.64 (-0.78)
Incomplete	50%	84.40 (-0.70)	79.96 (-0.41)	84.89 (-0.60)	59.53 (-1.89)
Incomplete	80%	84.27 (-0.83)	79.66 (-0.71)	84.25 (-1.24)	59.81 (-1.61)
Mismatch	10%	84.36 (-0.74)	79.81 (-0.56)	84.21 (-1.28)	60.25 (-1.17)
Mismatch	30%	82.42 (-2.68)	78.11 (-2.26)	82.44 (-3.05)	57.76 (-3.66)
Mismatch	50%	77.27 (-7.83)	70.01 (-10.36)	77.11 (-8.38)	50.68 (-10.74)

F Radiology report generation transfer

To further assess whether the diagnosis-aware and lesion-sensitive visual representations learned by SuG can transfer beyond classification and grounding, we evaluate SuG on a downstream radiology report generation task. Specifically, we pair the pretrained vision encoder with a text decoder and assess the resulting model under two transfer settings: *Frozen*, where the visual encoder is fixed, and *Finetuning*, where the visual encoder is jointly optimized with the text decoder. For comparison, we include visual pretraining baselines (Supervised and MAE), generalist vision–language baselines (CLIP, BIUD, Merlin, fVLM, and ViSD-Boost), and specialist–generalist synergy baselines (OpenVocabCT, HCFNet, and VLWS).

Table A17 shows that SuG achieves strong performance across both clinical efficacy metrics and natural language generation metrics. Under the Frozen setting, SuG obtains the best F1-score of 36.1, demonstrating that its pretrained visual encoder can effectively captures highly transferable clinical information. Under the Finetuning setting, SuG further improves the clinical quality of the generated reports, raising the F1-score from 36.1 to 42.2, while remaining competitive on the other evaluation metrics. Overall, these results indicate that SuG could provides clinically informative visual representations, enabling faithful downstream report generation.

Table A17: **Radiology report generation performance on the MedVL-CT69K test set.** We compare two transfer settings: *Frozen* (fixed visual encoder) and *Finetuning* (jointly optimized visual encoder and report decoder). "Supervised" refers to a classification model trained on 54 diseases, while "MAE" and "Supervised" represent 3D models pre-trained on the MedVL-CT69K training set. **Bold** and underlined values denote the best and second-best scores, respectively.

Encoder	Init	Persion	Recal	F1	GREEN	BLEU4	ROUGE-L	METEOR	CIDEr
Frozen	Supervised	19.1	18.6	13.2	25.9	12.8	40.6	30.8	6.6
	MAE (Chen et al., 2023)	8.9	5.9	4.3	21.6	13.1	41.6	30.5	6.1
	CLIP (Radford et al., 2021)	21.6	20.4	14.6	33.4	15.5	42.2	31.0	9.6
	BIUD (Cao et al., 2024)	17.0	21.4	15.9	33.7	18.9	44.2	29.1	13.9
	Merlin (Blankemeier et al., 2024)	22.6	20.9	20.7	34.2	19.0	43.8	30.0	14.3
	fVLM (Shui et al., 2025)	24.0	31.6	26.5	37.2	19.6	45.1	31.3	14.9
	ViSD-Boost (Cao et al., 2025)	34.3	**39.3**	35.2	44.4	24.7	48.7	**32.7**	27.3
	OpenVocabCT (Li et al., 2025b)	36.2	37.7	34.9	44.9	23.9	48.1	32.2	23.0
	HCFNet (Zhou et al., 2025a)	36.3	36.8	34.3	44.6	24.3	48.3	32.2	23.9
	VLWS (Hossain et al., 2026)	**36.8**	37.2	35.1	**45.3**	24.5	48.6	32.3	23.1
	SuG (Ours)	36.7	38.2	**36.1**	44.5	**25.1**	**48.8**	32.6	**28.0**
Finetuning	Supervised	18.0	28.3	20.4	35.5	17.9	43.4	30.6	11.7
	MAE (Chen et al., 2023)	13.4	14.1	10.9	29.4	15.1	42.5	30.3	8.8
	CLIP (Radford et al., 2021)	21.0	29.5	23.2	37.6	19.5	44.8	30.7	14.3
	BIUD (Cao et al., 2024)	26.1	31.6	24.2	38.8	19.0	44.7	30.9	13.9
	Merlin (Blankemeier et al., 2024)	27.5	29.9	25.8	39.2	20.9	46.0	31.1	17.2
	fVLM (Shui et al., 2025)	38.6	36.9	32.7	40.2	21.9	46.4	31.6	17.1
	ViSD-Boost (Cao et al., 2025)	39.8	44.1	40.9	46.7	**28.4**	**51.0**	34.1	**50.7**
	OpenVocabCT (Li et al., 2025b)	38.2	44.2	40.4	47.5	26.2	49.7	33.6	34.5
	HCFNet (Zhou et al., 2025a)	**39.9**	43.7	40.8	**49.1**	27.5	50.5	34.0	36.4
	VLWS (Hossain et al., 2026)	39.8	44.4	41.1	47.4	27.2	50.3	34.0	37.7
	SuG (Ours)	39.6	**46.6**	**42.2**	48.7	27.5	50.6	**34.2**	47.8

G Statistical Significance Analysis

To further validate the robustness of the performance gains in the main results, we conduct statistical significance analysis using paired bootstrap resampling. For each benchmark, we compare SuG against the *second-best* competing method under the corresponding evaluation setting. Specifically, the compared methods are HCFNet on CT-RATE, RAD-ChestCT, and Merlin, and VLWS on MedVL-CT69K and LesionSegAbdomen.

Our significance analysis is performed at the case level or patient level, following the protocol used in each dataset. For each bootstrap replicate, we resample evaluation cases with replacement in a paired manner for the two compared methods, recompute the per-dataset macro-level metrics, and record the performance difference between SuG and the second-best method. We then report the two-sided bootstrap p-value.

Tables A18–A24 summarize the statistical significance results on all evaluated benchmarks. Overall, the improvements of SuG over the second-best method are statistically significant on the vast majority of metrics. In particular, strong significance is consistently observed for AUC-related performance across datasets, and most of the corresponding sensitivity, specificity, F1, and precision gains are also significant. Only a small number of metrics, such as Precision on CT-RATE and ACC on RAD-ChestCT, do not reach statistical significance.

Table A18: **Bootstrap significance analysis on MedVL-CT69K.** We use patient-level paired bootstrap resampling to evaluate the statistical significance of macro-level performance differences between SuG and VLWS. Significance levels are denoted as follows: $*p < 0.05, **p < 0.01, ***p < 0.001$.

Metric	VLWS	SuG	p-value
AUC	86.7	90.1 (+3.36)	***
Sensitivity	81.8	85.0 (+3.20)	***
Specificity	81.9	84.9 (+3.00)	***

Table A19: **Bootstrap significance analysis on Merlin.** We use patient-level paired bootstrap resampling to evaluate the statistical significance of macro-level performance differences between SuG and HCFNet for the balanced metrics . Significance levels are denoted as follows: $*p < 0.05, **p < 0.01, ***p < 0.001$.

Metric	HCFNet	SuG	p-value
Balanced AUC	83.2	85.2 (+2.02)	**
Balanced F1	78.7	80.8 (+2.05)	**

Table A20: **Bootstrap significance analysis on Merlin for unbalanced metrics.** We use patient-level paired bootstrap resampling to evaluate the statistical significance of macro-level performance differences between SuG and HCFNet for the unbalanced metrics. Significance levels are denoted as follows: $*p < 0.05, **p < 0.01, ***p < 0.001$

Metric	HCFNet	SuG	p-value
AUC	82.9	85.1 (+2.19)	***
F1	57.2	60.7 (+3.44)	***

Table A21: **Bootstrap significance analysis on CT-RATE.** We use patient-level paired bootstrap resampling to evaluate the statistical significance of macro-level performance differences between SuG and HCFNet. Significance levels are denoted as follows: $*p < 0.05, **p < 0.01, ***p < 0.001$.

Metric	HCFNet	SuG	p-value
AUC	80.3	81.1 (+0.80)	***
ACC	73.5	74.4 (+0.84)	***
F1	76.4	77.1 (+0.69)	***
Precision	40.7	40.8 (+0.15)	

Table A22: **Bootstrap significance analysis on RAD-ChestCT.** We use case-level paired bootstrap resampling to evaluate the statistical significance of macro-level performance differences between SuG and HCFNet on RAD-ChestCT. Significance levels are denoted as follows: $*p < 0.05, **p < 0.01, ***p < 0.001$.

Metric	HCFNet	SuG	p-value
AUC	71.7	73.4 (+1.69)	***
ACC	67.8	68.1 (+0.31)	
F1	71.5	71.9 (+0.36)	*
Precision	35.9	36.9 (+0.99)	***

Table A23: **Bootstrap significance analysis on MedVL-CT69K for the four tumor categories.** We use patient-level paired bootstrap resampling across the four tumor categories to evaluate the statistical significance of macro-level performance differences between SuG and VLWS. Significance levels are denoted as follows: $*p < 0.05, **p < 0.01, ***p < 0.001$.

Metric	VLWS	SuG	p-value
AUC	88.1	93.7 (+5.63)	***
Sensitivity	81.4	86.8 (+5.34)	*
Specificity	82.3	88.0 (+5.72)	***

Table A24: **Bootstrap significance analysis on LesionSegAbdomen.** We use patient-level paired bootstrap resampling across the four tumor categories to evaluate the statistical significance of macro-level performance differences between SuG and VLWS. Significance levels are denoted as follows: $*p < 0.05, **p < 0.01, ***p < 0.001$.

Metric	VLWS	SuG	p-value
AUC	87.0	96.1 (+9.10)	***
Sensitivity	78.9	89.4 (+10.50)	***
Specificity	81.9	91.7 (+9.80)	***

Table A25: **General disease diagnosis performance on the MedVL-CT69K test set.** Results are reported for 15 individual anatomical structures, with the final column ("Average") representing the aggregated performance across all 54 diseases. SE, SP, and AUC denote sensitivity, specificity, and area under the receiver operating characteristic curve, respectively. "G", "S", and "S+G" represent generalist-only, specialist-only, and specialist-generalist synergy methods, respectively.

Method	Type	Adrenal gland			Bladder			Colon			Esophagus			Gallbladder			Heart			Kidney			Liver		
		SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC
CLIP (Radford et al., 2021)	G	66.6	55.4	63.2	57.7	67.6	65.1	64.4	63.2	65.8	65.1	78.3	67.3	55.5	59.9	59.5	36.2	77.1	44.1	55.5	61.6	59.9	70.8	65.4	72.4
LOVT (Müller et al., 2022)	G	61.8	54.6	60.6	71.4	62.3	70.9	69.6	58.5	67.5	84.2	85.1	89.3	65.2	51.8	61.2	84.6	64.9	78.3	62.2	54.7	60.2	68.3	60.2	69.3
Imitate (Liu et al., 2024)	G	64.3	55.9	60.2	72.6	67.9	74.1	69.2	60.7	68.0	98.1	89.4	95.6	60.0	59.7	62.5	71.5	75.1	70.7	60.7	55.8	59.9	66.5	65.4	69.8
BIUD (Cao et al., 2024)	G	64.9	56.0	63.4	79.8	73.5	81.0	70.4	64.8	70.0	54.1	91.9	62.6	60.6	61.1	64.2	68.2	56.1	62.1	60.8	61.8	63.7	72.6	74.0	79.2
Merlin (Blankemeier et al., 2024)	G	58.9	57.9	60.3	70.8	73.0	76.9	71.1	62.0	69.1	41.5	88.5	49.2	64.6	53.5	61.2	69.6	75.1	72.8	58.6	64.5	64.2	73.6	75.9	80.1
fVLM (Shui et al., 2025)	G	63.0	63.9	65.7	76.2	77.3	84.0	76.1	75.1	80.8	94.4	96.1	98.2	64.9	58.8	64.8	87.2	75.8	85.8	67.9	72.5	74.5	77.2	76.0	82.5
ViSD-Boost (Cao et al., 2025)	G	63.5	64.9	68.5	75.0	74.4	81.2	78.2	75.7	81.9	94.4	95.4	98.1	65.6	69.7	72.6	84.6	82.6	90.5	72.4	74.5	78.5	78.4	80.3	85.9
OpenVocabCT (Li et al., 2025b)	S+G	60.2	67.6	68.8	81.0	74.4	84.2	77.0	81.1	84.3	88.7	98.3	96.3	77.6	78.6	85.0	75.0	81.1	81.5	77.3	79.2	84.0	84.5	83.3	90.6
HCFNet (Zhou et al., 2025a)	S+G	70.1	63.7	71.2	81.0	87.2	89.2	75.6	77.9	83.0	93.7	97.9	99.0	80.5	74.8	83.5	82.9	76.2	85.6	75.5	81.8	84.4	81.4	85.2	89.3
VLWS (Hossain et al., 2026)	S+G	68.2	64.5	70.1	81.0	82.8	88.9	79.6	79.3	84.1	93.7	94.7	95.3	80.9	78.5	85.6	86.6	80.6	89.5	71.8	79.1	80.4	79.4	79.9	84.9
SuG (Ours)	S+G	71.7	82.6	82.9	89.3	84.5	92.1	80.8	81.6	87.0	92.4	93.3	97.2	81.9	75.9	84.9	87.8	84.4	91.6	80.2	83.0	87.4	87.4	86.9	92.2

Method	Type	Lung			Pancreas			Portal vein			Small Intestine			Spleen			Stomach			Sacrum			Average		
		SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC
CLIP (Radford et al., 2021)	G	80.4	96.1	88.3	65.4	62.4	65.0	72.4	72.4	78.6	69.9	66.6	74.5	72.8	65.9	71.1	62.5	68.0	68.6	47.1	56.0	47.0	65.5	68.0	68.4
LOVT (Müller et al., 2022)	G	78.7	65.0	80.9	68.3	62.5	67.8	82.6	60.2	75.5	72.4	61.5	70.5	70.1	49.0	66.1	62.9	67.9	69.1	70.6	38.8	48.9	70.8	60.1	69.4
Imitate (Liu et al., 2024)	G	81.1	89.7	86.7	65.0	61.3	64.3	76.1	69.3	80.5	76.1	68.0	77.6	64.0	68.9	71.3	64.0	63.7	66.3	35.3	43.4	29.0	69.2	66.6	70.6
BIUD (Cao et al., 2024)	G	69.3	84.5	72.1	72.4	70.3	76.9	82.5	71.4	82.2	74.8	67.5	75.1	65.6	72.3	72.3	63.1	63.8	66.1	70.6	29.2	43.8	69.3	69.0	71.4
Merlin (Blankemeier et al., 2024)	G	76.9	80.1	78.7	74.2	63.8	73.5	86.2	78.0	85.9	73.4	72.1	78.4	67.3	72.2	72.0	63.3	67.7	69.9	47.1	65.8	48.2	69.2	69.7	71.9
fVLM (Shui et al., 2025)	G	74.3	78.9	82.2	75.8	80.8	85.3	90.8	93.2	96.7	74.0	78.6	82.1	76.5	78.0	82.0	69.9	67.8	74.1	88.2	83.3	87.5	75.8	76.5	81.3
ViSD-Boost (Cao et al., 2025)	G	90.2	88.9	92.7	80.7	85.1	88.9	92.7	92.7	97.3	84.1	79.1	88.3	78.2	77.1	82.9	73.1	77.4	81.1	70.6	75.7	77.5	79.4	79.7	84.9
OpenVocabCT (Li et al., 2025b)	S+G	64.3	86.4	68.1	86.9	83.8	90.9	91.7	90.6	95.5	81.4	86.7	89.8	87.1	87.9	94.2	78.2	84.3	87.4	82.4	85.8	88.4	79.4	83.2	85.8
HCFNet (Zhou et al., 2025a)	S+G	84.9	87.6	89.3	83.0	87.9	90.2	93.6	93.7	97.4	85.8	80.1	88.3	86.6	85.8	92.6	74.2	69.9	76.6	76.5	49.2	58.7	81.4	81.8	86.7
VLWS (Hossain et al., 2026)	S+G	82.3	88.4	87.0	86.0	85.0	90.8	94.5	91.7	97.6	84.8	82.4	89.8	87.2	87.4	94.1	77.1	80.7	85.2	88.2	65.7	80.0	81.8	81.9	86.7
SuG (Ours)	S+G	88.1	86.5	90.9	84.3	85.7	90.4	94.5	93.9	98.4	83.9	84.4	88.9	91.6	90.5	95.6	80.5	83.4	88.4	88.2	80.3	88.2	**85.0**	**84.9**	**90.1**

Table A26: **Fine-grained lesion-centric diagnosis performance on four tumor categories (colon, liver, pancreas, and stomach) on the MedVL-CT69K and LesionSegAbdomen test sets.** "–" indicates that a specific category is not supported by the specialist model.

MedVL-CT69K

Method	Type	Colon			Liver			Pancreas			Stomach			Average		
		SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC
nnUNet-c (Isensee et al., 2018)	S	80.1	80.8	87.2	–	–	–	–	–	–	–	–	–	–	–	–
nnUNet-l (Isensee et al., 2018)	S	–	–	–	70.1	77.5	79.5	–	–	–	–	–	–	–	–	–
nnUNet-p (Isensee et al., 2018)	S	–	–	–	–	–	–	75.9	72.5	78.6	–	–	–	–	–	–
nnUNet-s (Isensee et al., 2018)	S	–	–	–	–	–	–	–	–	–	71.8	64.3	74.7	–	–	–
nnUNet-fuse (Isensee et al., 2018)	S	80.1	80.8	87.2	70.1	77.5	79.5	75.9	72.5	78.6	71.8	64.3	74.7	74.5	73.8	80.0
CLIP (Radford et al., 2021)	G	59.6	65.4	65.9	55.0	61.5	60.0	62.1	76.5	73.0	68.4	63.4	71.6	61.3	66.7	67.6
LOVT (Müller et al., 2022)	G	92.8	21.4	68.1	64.5	49.0	58.4	72.4	67.6	74.9	89.7	32.2	69.4	79.9	42.5	67.7
Imitate (Liu et al., 2024)	G	83.1	34.1	63.5	59.8	55.6	60.3	69.0	63.6	71.4	59.8	72.5	71.9	67.9	56.4	66.8
BIUD (Cao et al., 2024)	G	83.1	34.5	66.3	57.4	56.5	59.6	75.9	72.4	77.7	65.0	73.1	73.1	70.3	59.1	69.2
Merlin (Blankemeier et al., 2024)	G	68.1	51.2	62.6	65.8	54.6	63.2	75.9	65.0	74.2	69.2	59.8	66.5	69.7	57.6	66.6
fVLM (Shui et al., 2025)	G	66.3	69.7	73.6	62.3	56.1	62.2	82.8	76.4	85.8	70.9	72.5	75.7	70.6	68.7	74.3
ViSD-Boost (Cao et al., 2025)	G	74.1	78.0	83.8	62.4	58.1	64.5	93.1	82.6	94.7	72.6	82.9	84.2	75.6	75.4	81.8
OpenVocabCT (Li et al., 2025b)	S+G	81.3	75.3	87.0	77.9	78.1	84.9	93.1	92.6	97.4	81.2	90.9	89.6	83.4	84.2	89.7
HCFNet (Zhou et al., 2025a)	S+G	77.1	78.0	86.0	67.3	73.1	75.1	89.7	94.2	94.8	84.6	80.4	89.3	79.7	81.4	86.3
VLWS (Hossain et al., 2026)	S+G	72.3	80.3	83.3	74.7	80.2	84.0	93.1	88.8	96.4	85.5	79.7	88.6	81.4	82.3	88.1
SuG (Ours)	S+G	81.9	91.0	93.0	82.3	85.2	90.6	93.1	88.1	96.1	89.7	87.7	95.2	**86.8**	**88.0**	**93.7**

LesionSegAbdomen

Method	Type	Colon			Liver			Pancreas			Stomach			Average		
		SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC
nnUNet-c (Isensee et al., 2018)	S	92.0	93.0	96.0	–	–	–	–	–	–	–	–	–	–	–	–
nnUNet-l (Isensee et al., 2018)	S	–	–	–	93.4	91.0	97.4	–	–	–	–	–	–	–	–	–
nnUNet-p (Isensee et al., 2018)	S	–	–	–	–	–	–	85.4	88.0	94.0	–	–	–	–	–	–
nnUNet-s (Isensee et al., 2018)	S	–	–	–	–	–	–	–	–	–	86.0	84.0	91.5	–	–	–
nnUNet-fuse (Isensee et al., 2018)	S	92.0	93.0	96.0	93.4	91.0	97.4	85.4	88.0	94.0	86.0	84.0	91.5	89.2	89.0	94.7
CLIP (Radford et al., 2021)	G	59.0	71.2	64.6	65.0	58.1	62.5	71.0	53.4	66.4	69.0	61.9	68.9	66.0	61.1	65.6
LOVT (Müller et al., 2022)	G	80.0	67.3	83.9	84.8	16.8	55.2	72.3	60.0	70.2	54.0	55.8	54.9	72.8	50.0	66.1
Imitate (Liu et al., 2024)	G	56.0	66.9	66.7	44.2	59.2	48.3	75.6	58.7	71.7	59.0	74.2	67.9	58.7	64.8	63.7
BIUD (Cao et al., 2024)	G	61.0	72.3	73.3	56.9	66.8	65.9	64.6	60.3	66.6	56.0	56.4	57.8	59.6	64.0	65.9
Merlin (Blankemeier et al., 2024)	G	60.0	47.3	52.9	56.3	52.1	55.2	67.6	53.0	64.8	44.0	67.0	53.3	57.0	54.9	56.5
fVLM (Shui et al., 2025)	G	80.0	75.5	82.9	61.9	64.8	65.9	70.3	68.6	77.4	71.0	76.6	75.7	70.8	71.4	75.5
ViSD-Boost (Cao et al., 2025)	G	67.0	71.0	76.9	65.0	71.4	73.6	82.0	80.1	88.7	61.0	64.7	67.1	68.7	71.8	76.6
OpenVocabCT (Li et al., 2025b)	S+G	71.0	79.6	81.1	83.2	84.6	90.2	82.6	82.6	90.2	79.0	86.4	85.6	79.0	83.3	86.8
HCFNet (Zhou et al., 2025a)	S+G	83.0	87.2	91.3	68.5	82.5	79.8	85.6	86.5	93.8	84.0	83.2	90.9	80.3	84.9	89.0
VLWS (Hossain et al., 2026)	S+G	77.0	81.2	86.0	77.2	85.6	86.8	81.2	77.9	87.5	80.0	83.1	88.0	78.9	81.9	87.0
SuG (Ours)	S+G	91.0	95.1	97.3	95.9	95.0	99.3	87.7	88.7	95.1	83.0	88.0	92.7	**89.4**	**91.7**	**96.1**

Table A27: **Per-disease diagnostic performance on the MedVL-CT69K test set.** Reported metrics include sensitivity (SE), specificity (SP), accuracy (ACC), and AUC, all presented as percentages.

Anatomy	Abnormality	SE	SP	ACC	AUC
Adrenal gland	Adrenal hypertrophy	68.8	81.3	75.0	80.9
	Nodule	74.7	83.9	79.3	85.0
Bladder	Diverticulum	85.7	78.4	82.1	88.1
	Stone	92.9	90.5	91.7	96.0
Colon	Colonic gas	81.4	81.7	81.5	89.3
	Effusion	82.0	83.2	82.6	87.8
	Obstruction	94.1	98.3	96.2	99.1
	Diverticulum	72.1	75.6	73.8	80.5
	Colon cancer	86.5	90.8	88.6	93.9
	Rectal cancer	86.3	88.5	87.4	93.4
	Appendicitis	84.2	72.6	78.4	87.9
	Appendicolith	59.5	62.1	60.8	63.8
Duodenum	Diverticulum	86.7	93.3	90.0	94.9
Esophagus	Hiatal hernia	90.0	89.4	89.7	96.5
	Varicose veins	94.9	97.2	96.0	98.0
Gallbladder	Inflammation	83.7	75.4	79.5	87.0
	Gallstone	87.0	89.8	88.4	94.3
	Adenomyomatosis	75.0	62.5	68.8	73.3
Heart	Cardiomegaly	95.0	90.1	92.5	96.8
	Pericardial effusion	80.5	78.8	79.7	86.5
Kidney	Atrophy	86.5	79.3	82.9	87.4
	Cyst	76.2	81.9	79.0	86.1
	Hydronephrosis	86.2	86.5	86.3	93.9
	Renal calculus	71.8	84.5	78.2	82.2
Liver	Fatty liver	92.4	90.6	91.5	96.5
	Glisson's capsule effusion	95.6	86.6	91.1	93.9
	Metastatic tumor	86.1	90.9	88.5	93.7
	Intrahepatic bile duct dilatation	76.9	78.4	77.6	86.3
	Cancer	93.4	95.3	94.3	98.5
	Cyst	81.8	79.0	80.4	87.4
	Abscess	83.3	80.0	81.7	83.8
	Cirrhosis	89.9	94.2	92.1	97.3
Lung	Atelectasis	94.3	98.4	96.4	98.1
	Bronchiectasis	88.9	98.4	93.7	96.2
	Emphysema	80.0	63.9	72.0	75.4
	Pneumonia	82.8	74.2	78.5	87.3
	Pleural effusion	94.3	97.5	95.9	97.6
Pancreas	Pancreatic cancer	86.2	85.2	85.7	89.0
	Atrophy	81.1	84.0	82.5	88.6
	Pancreatitis	87.0	95.4	91.2	95.8
	Pancreatic duct dilatation	85.1	80.8	82.9	90.6
	Steatosis	82.2	83.2	82.7	87.8
Portal vein	Hypertension	94.4	95.2	94.8	98.7
	Thrombosis	94.5	92.6	93.6	98.1
Sacrum	Osteitis	88.2	80.3	84.3	88.3
Small intestine	Gas accumulation	84.0	76.3	80.2	86.5
	Fluid accumulation	88.7	85.3	87.0	92.5
	Obstruction	93.4	91.6	92.5	95.7
	Intussusception	66.7	75.6	71.1	74.9
Spleen	Hemangioma	87.2	92.1	89.7	94.2
	Infarction	100.0	93.4	96.7	98.7
	Splenomegaly	87.5	85.8	86.7	93.8
Stomach	Gastric wall thickening	74.8	80.8	77.8	84.4
	Gastric cancer	86.3	86.0	86.2	92.5

Table A28: **Per-disease diagnostic performance on the CT-RATE test set.** Reported metrics include precision (Prec), accuracy (ACC), F1, and ROC AUC (AUC), all presented as percentages.

Abnormality	Prec	ACC	F1	AUC
Emphysema	34.8	68.9	71.8	75.1
Atelectasis	35.1	63.8	66.4	69.4
Lung nodule	56.9	61.7	61.8	65.6
Lung opacity	60.8	70.7	71.0	77.8
Pulmonary fibrotic sequela	33.9	54.8	57.1	62.1
Pleural effusion	66.8	93.4	93.9	96.7
Mosaic attenuation pattern	16.2	66.1	73.7	75.8
Peribronchial thickening	28.2	75.7	79.6	80.3
Consolidation	47.7	79.2	81.1	88.3
Bronchiectasis	22.1	72.1	77.0	72.9
Interlobular septal thickening	22.7	75.7	81.0	86.4
Cardiomegaly	39.1	84.5	86.8	93.8
Pericardial effusion	25.2	81.6	85.5	86.8
Coronary artery wall calcification	65.7	85.3	85.9	92.6
Hiatal hernia	28.5	70.4	74.8	79.9
Arterial wall calcification	69.6	86.1	86.5	93.7

Table A29: **Per-disease diagnostic performance on the RAD-ChestCT test set.** Reported metrics include precision (Prec), accuracy (ACC), F1, and ROC AUC (AUC), all presented as percentages.

Abnormality	Prec	ACC	F1	AUC
Emphysema	45.9	69.3	70.8	74.6
Atelectasis	36.8	57.1	59.0	60.1
Lung nodule	81.8	61.7	64.3	63.7
Lung opacity	58.8	56.5	56.5	58.0
Pulmonary fibrotic sequela	29.2	73.4	77.4	80.5
Pleural effusion	60.5	85.7	86.5	92.2
Peribronchial thickening	12.6	56.7	66.0	64.2
Consolidation	28.6	71.4	75.6	78.8
Bronchiectasis	26.1	63.9	68.7	71.0
Interlobular septal thickening	15.1	69.2	76.8	78.1
Cardiomegaly	28.6	76.1	80.3	85.3
Pericardial effusion	26.3	67.6	71.8	68.9
Hiatal hernia	29.1	77.0	80.5	78.2

Table A30: **Per-disease diagnostic performance on the Merlin test set.** Reported metrics include F1, ROC AUC (AUC), and accuracy (ACC), all presented as percentages.

Abnormality	F1	AUC	ACC
aortic_valve_calcification	95.7	98.7	95.7
pleural_effusion	92.7	97.5	92.7
splenomegaly	91.0	97.2	91.1
bowel_obstruction	93.6	98.1	93.8
abdominal_aortic_aneurysm	94.1	94.8	94.1
hepatomegaly	86.0	92.3	86.8
pancreatic_atrophy	85.2	91.1	84.8
hepatic_steatosis	88.5	93.0	88.9
surgically_absent_gallbladder	90.0	90.5	90.5
cardiomegaly	86.1	91.9	85.7
hydronephrosis	85.6	90.6	85.7
atherosclerosis	83.9	89.7	83.6
submucosal_edema	77.5	85.0	77.4
hiatal_hernia	80.3	87.6	80.1
biliary_ductal_dilation	82.5	87.3	80.8
renal_cyst	75.9	82.9	75.5
atelectasis	79.0	86.9	79.1
gallstones	68.7	75.7	69.3
renal_hypodensities	68.2	73.5	66.0
fracture	71.6	76.2	71.8
appendicitis	51.4	52.1	58.1
osteopenia	83.5	88.9	83.3
ascites	84.5	92.2	84.2
free_air	86.3	92.4	85.6
coronary_calcification	76.9	81.0	76.2
anasarca	80.2	87.3	80.1
lymphadenopathy	71.0	73.4	68.4
metastatic_disease	76.8	80.4	75.0
prostatomegaly	74.6	75.3	73.5
thrombosis	62.0	53.4	57.8

Table A31: **Per-disease *unbalanced* diagnostic performance on the Merlin test set.** Reported metrics include F1, ROC AUC (AUC), and accuracy (ACC), all presented as percentages.

Abnormality	F1	AUC	ACC
aortic_valve_calcification	95.1	98.1	95.0
pleural_effusion	93.4	97.3	92.2
splenomegaly	40.1	96.7	89.1
abdominal_aortic_aneurysm	80.0	95.9	92.6
bowel_obstruction	42.9	95.0	89.0
pancreatic_atrophy	31.2	91.0	84.4
hepatomegaly	38.8	88.8	83.8
biliary_ductal_dilation	63.5	89.5	83.6
cardiomegaly	77.8	92.3	86.2
hepatic_steatosis	92.8	92.5	86.5
surgically_absent_gallbladder	86.1	90.2	90.2
submucosal_edema	76.7	85.2	77.8
hydronephrosis	31.9	90.1	83.1
atherosclerosis	91.1	89.9	83.2
renal_cyst	63.0	83.2	75.7
hiatal_hernia	55.0	88.3	80.9
atelectasis	78.2	86.7	79.0
renal_hypodensities	26.2	75.5	69.2
gallstones	62.3	75.3	69.1
fracture	71.4	77.2	72.8
appendicitis	47.5	49.8	54.8
osteopenia	54.9	90.5	83.7
ascites	89.0	92.2	84.5
free_air	26.5	87.3	78.2
coronary_calcification	43.5	80.8	74.0
anasarca	54.7	87.7	81.7
metastatic_disease	62.9	81.2	75.3
lymphadenopathy	47.2	75.1	70.7
prostatomegaly	23.9	74.6	72.2
thrombosis	73.0	53.7	58.5