Anatomy-Activated Mixture-of-Experts for 3D Medical Vision-Language Pre-training

Szymon Płotka[1]* **Gizem Mert**[2,3,4] **Pedro R. A. S. Bassi**[5,6,7] **Wenxuan Li**[5,6,7]
Zongwei Zhou[5] **Anna Marcinkiewicz**[8] **Wiktoria Romańczyk**[9] **Jarosław B. Ćwikła**[10]
Łukasz Struski[1] **Ewa Szczurek**[2,4,11,12] **Jacek Tabor**[1] **Arkadiusz Sitek**[6,7]
[1] Jagiellonian University [2] Helmholtz Munich [3] Technical University of Munich
[4] Munich Center for Machine Learning [5] Johns Hopkins University
[6] Massachusetts General Hospital [7] Harvard Medical School
[8] National Medical Institute of the Ministry of the Interior and Administration
[9] Medical University of Białystok [10] GAMMED [11] University of Warsaw [12] LMU Munich

Abstract

Medical Vision-Language Pre-training (VLP) leverages the semantic richness of radiology reports to learn 3D representations, yet current architectures remain fundamentally anatomy-agnostic. While clinical diagnostics rely on organ-specific context, standard encoders apply a uniform, dense parameterization to all volumetric patches, and existing Mixture-of-Experts (MoE) routers gate on isolated local features without explicit access to the organ composition of the scan or crop. We propose Anatomy-Activated Mixture-of-Experts (A^2MoE), a framework that incorporates anatomical inductive biases within the model's computational path. A^2MoE introduces two key components: (1) a Histogram Representation Router (HRR) that conditions expert selection on global visual context and a predicted anatomical composition histogram, encouraging anatomy-dependent expert weighting; and (2) an Organ-Query Multi-Scale Projector (OQMSP) that utilizes learnable queries to extract organ-level embeddings for fine-grained alignment without requiring manual annotations at inference. By routing at the patch level rather than the token level, A^2MoE reduces routing complexity by orders of magnitude while satisfying load balancing by construction. Pre-trained on over 60,000 multi-phase CT scans, A^2MoE achieves state-of-the-art performance across segmentation, classification, and detection benchmarks. Furthermore, we demonstrate that our anatomically grounded inductive biases support robust generalization to MRI and ultrasound, establishing a scalable foundation for medical representation learning. The code is available at github.com/gmum/A2MoE.

1 Introduction

Vision-Language Pre-training (VLP) has redefined multimodal representation learning by aligning visual features with semantic language cues in a shared embedding space [1, 6, 20, 23, 26, 46, 47, 52, 58, 78, 91, 95]. In the medical domain, Medical Vision-Language Models (MedVLMs) [16, 24, 33, 45, 51, 57, 69, 75, 76, 89, 94, 96] leverage radiology reports to supervise 3D visual encoders, reducing dependence on scarce voxel-level annotations. Moreover, 3D medical images have an anatomical hierarchy that current encoders rarely exploit: local image patterns arise within organs, organs appear within broader scan-level anatomical compositions, and reports describe findings at corresponding organ-specific and global levels. Current MedVLM backbones largely ignore this structure, applying the same dense parameters to all volumetric patches regardless of their anatomical

*Correspondence to: Szymon Płotka (SZYMON.PLOTKA@UJ.EDU.PL)

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

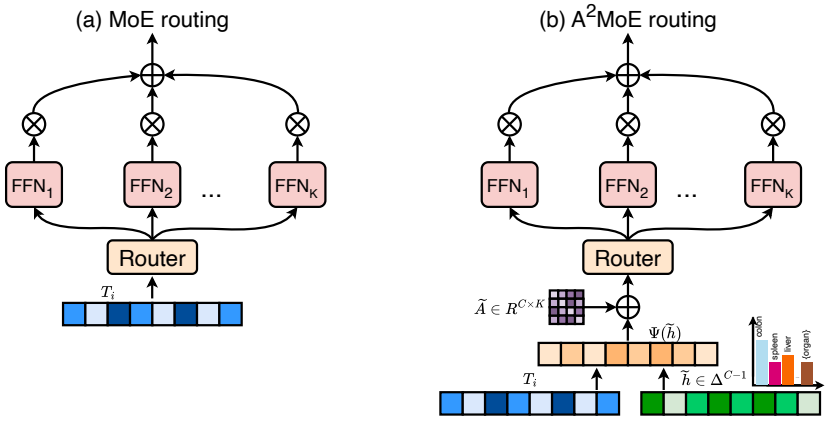

Figure 1: **Comparison of MoE routing strategies.** (a) MoE routing: The input feature map x is pooled to form a global visual context, which is passed through an MLP router to produce logits over K experts. Softmax-normalized weights are then used to compute a convex combination of expert outputs $E_k(x)$. Routing depends solely on visual features. (b) In addition to visual context, a foreground-normalized organ-composition histogram $\widetilde{\mathbf{h}}$ is embedded and fused with visual features to jointly condition the router. Each foreground entry represents the corresponding organ's share of the total foreground voxel mass in the input crop. A learnable class-to-expert affinity matrix $\mathbf{A}^{(s)}$ introduces a structural prior via a log-probability term $\log(\widetilde{\mathbf{h}}\widetilde{\mathbf{A}}^{(s)} + \epsilon)$, biasing expert selection toward anatomically consistent assignments. The resulting logits are softmax-normalized to obtain routing weights, and the final output is computed as a weighted sum of expert responses.

context. This is suboptimal because visually similar patterns can carry different clinical meanings in different organs or anatomical neighborhoods. We argue that medical VLP should therefore move beyond post-hoc image–text alignment and make anatomical context part of the model's computation path during feature extraction.

Recent advances in MedVLMs have recognized the importance of anatomical grounding, but have primarily addressed it through refined *alignment objectives*. For instance, fVLM [76] matches CT regions to specific report snippets, while ViSD-Boost [16] enhances semantic density via organ-aware contrastive loss. However, these methods focus primarily on *what* is aligned to text, but they do not fundamentally change *how* visual features are computed inside the encoder. Current 3D backbones [13, 43, 87] treat every voxel with the same dense parameterization, regardless of whether a feature arises in the liver, kidney, spleen, bowel, or background. Explicit anatomical conditioning offers a complementary direction; in clinical practice, identical visual patterns often hold divergent meanings depending on their anatomical context. A "dark spot" (hypodensity), for example, carries distinct diagnostic weight in the liver versus the spleen or kidney, motivating an explicit anatomical input alongside learned visual features.

Mixture-of-Experts (MoE) [28, 38, 74] provides a principled mechanism for conditional computation, yet its standard formulation in computer vision remains "context-blind." Current MoE routers typically dispatch tokens based on isolated local patch embeddings, ignoring the global anatomical topology of the 3D volume. This leads to a suboptimal "semantic collision": patches with high textural similarity but distinct anatomical identities, such as a liver–kidney boundary versus a liver–stomach interface, are frequently mapped to the same expert parameters. Such interference dilutes the model's capacity to learn specialized, organ-specific feature transforms. In addition to MoEs, existing cross-modal projection heads either rely on global pooling, which collapses local anatomical nuances, or depend on ground-truth masks to isolate organ-level evidence, a dependency that renders them impractical for real-world inference where such masks can be unavailable.

Motivated by these limitations in both expert routing and cross-modal projection, we propose **A**natomy-**A**ctivated **M**ixture-**o**f-**E**xperts (**A^{2}MoE**), a volumetric medical VLP framework that incorporates anatomical conditioning into the encoder and organ-indexed aggregation into the projector. Our central hypothesis is that *explicitly conditioning model capacity on global anatomical compo-*

sition allows the network to learn more discriminative, organ-specific representations than dense or context-blind backbones. By grounding the computation path in structural priors, the model can effectively decouple the complex, overlapping signals present in 3D medical scans.

To realize this, we introduce two core components. First, our Histogram Representation Router (HRR) jointly embeds global visual context with an anatomical composition histogram $\mathbf{h}$ and a learnable class-to-expert affinity prior (see Figure 1). This conditions expert weighting on crop-level anatomical context. Second, we address the challenge of mask-free inference with an Organ-Query Multi-Scale Projector (OQMSP). Instead of requiring explicit segmentation, OQMSP utilizes a set of learnable organ queries that cross-attend to MoE-refined features across multiple scales. This allows the model to "probe" for organ-specific evidence and extract per-organ embeddings based on learned attention weights, providing a scalable mechanism for fine-grained image–text alignment without masks at inference. While pre-trained on CT, the structural inductive biases of A^2MoE are modality-agnostic, enabling seamless transfer to MRI and ultrasound where anatomical consistency remains a primary diagnostic pillar. Our main contributions are:

1. **Anatomy-Activated MoE Routing.** We introduce the HRR, a novel routing mechanism that uses an anatomical composition histogram to guide conditional computation, enabling the model to allocate specialized parameters to specific anatomical structures.
2. **Mask-Free Organ-Level Alignment.** We propose OQMSP, a projector that extracts per-organ visual embeddings through learned queries, allowing for fine-grained VLP alignment without the need for segmentation masks during inference.
3. **Empirical Validation of Conditional VLP.** We show that anatomy-conditioned routing and organ-level alignment improve 3D medical representation learning and support robust transfer from CT pre-training to MRI and ultrasound.

We demonstrate that A^2MoE serves as a robust foundation for multimodal representation learning. By pre-training on over 60,000 CT-report pairs, we achieve state-of-the-art performance across segmentation, classification, and detection tasks. Notably, A^2MoE exhibits scalability across diverse structural granularities, from fine-grained anomalies to large-scale anatomical systems, and maintains high generalization performance on out-of-distribution benchmarks. Ultimately, A^2MoE provides a scalable mechanism for aligning unstructured semantic priors with dense 3D representations, facilitating effective cross-modal knowledge transfer.

2 Related Work

Mixture-of-Experts. MoE enables conditional computation by routing inputs to specialized sub-networks, scaling model capacity without proportionally increasing computational cost [28, 40, 71, 74]. In vision, V-MoE [71] and Soft MoE [68] route image tokens to experts based on local patch embeddings, while LIMoE [61] extends sparse MoE to contrastive language-image pre-training. In medical imaging, HoME [67] combines Mamba state-space blocks with soft expert routing for linear-time 3D segmentation, and Swin SMT [66] integrates Soft MoE into Swin UNETR for whole-body CT segmentation. PAMoE [83] routes tissue patches to specialized experts for gigapixel whole-slide image analysis, adapting capacity to heterogeneous histological regions, while ERMoE [21] reparameterizes experts in a learned eigenbasis and shows that the resulting routing yields anatomically interpretable specializations in 3D brain MRI. In the multimodal domain, MoHAVE [41] and MoME [17] apply hierarchical and Matryoshka-style MoE to audio-visual speech recognition, while M3-JEPA [44] uses a multi-gate MoE predictor for general multimodal alignment. While current MoE models perform routing based on low-level feature heuristics, they remain agnostic to the global anatomical landscape. Our work addresses this fundamental gap: the Histogram Representation Router is the first to condition expert selection on a high-level, structured representation of anatomical distribution, ensuring that routing is biologically informed rather than purely data-driven.

Anatomical conditioning and organ-level alignment. Anatomical priors have been incorporated into both expert routing and medical vision–language representation learning. OAMoE [56] organizes segmentation experts into anatomically guided pools with adaptive expert selection. HAR-MoE [59] combines task-conditioned expert screening with organ-conditioned refinement and learnable affinity priors for unified ultrasound segmentation and classification, sharing image-level routing decisions across spatial tokens. For cross-modal learning, ASG [48] aligns anatomical regions with corre-

sponding report sentences, while Anatomy-VLM [30] combines anatomical localization, structured clinical knowledge, and multi-scale alignment for chest radiograph interpretation. In volumetric CT, fVLM [76] performs anatomy-level contrastive pre-training with disease-aware pairing; ViSD-Boost [16] strengthens anatomical representations through normal–abnormal discrimination and anatomy normality modeling; and RADAR [92] scales anatomy-wise image–text learning for generalist abdominal CT diagnosis. Building on these directions, A^{2}MoE couples crop-composition-conditioned encoder routing with multi-scale organ-level alignment: an anatomy histogram and learned class-to-expert affinity prior condition expert mixtures, while shared organ queries aggregate features across encoder stages without masks at inference.

Medical Vision–Language Pre-training. Early methods such as GLoRIA [36] and LoVT [60] combine global image–report alignment with local cross-modal objectives to learn transferable representations from chest radiographs. Subsequent work extends language-supervised representation learning to volumetric imaging through complementary strategies. Merlin [13] combines CT volumes, radiology reports, and structured electronic health records to support diverse downstream tasks. HLIP [96] enables pre-training on uncurated clinical studies through hierarchical attention across slices, scans, and studies, while SPECTRE [24] combines self-supervised and cross-modal pre-training for volumetric CT transformers. MG-3D [62] incorporates multi-grained report knowledge through intra-patient cross-modal learning and inter-patient semantic alignment, while ExGra-Med [55] explores graph-based cross-modal alignment for semantic grounding. Beyond representation alignment, R-Super [8, 12] converts radiology reports into voxel-wise supervision for tumor segmentation, extending report-derived knowledge to dense prediction. Addressing volumetric representation efficiency, BTB3D [33] learns compact, frequency-aware 3D tokens for report generation and text-conditioned CT synthesis. Complementing these methodological advances, CARES [86] benchmarks the trustworthiness of medical vision–language models. A^{2}MoE builds on this progress by coupling anatomy-conditioned expert mixtures within the visual encoder with shared organ queries for multi-scale image–text alignment.

3 Preliminaries

Selective State-Space Models (Mamba). Our architecture builds on the Mamba layer [29], a structured state-space model designed for long-sequence modeling with linear computational complexity. In contrast to Transformer architectures, whose attention mechanism incurs a quadratic cost in sequence length, Mamba achieves linear-time processing via a discrete-time recurrence derived from continuous-time state-space dynamics. Given a feature sequence $z = \{z_t\}_{t=1}^{N} \in \mathbb{R}^{B\times N\times d}$, the hidden state h_t and output y_t evolve according to:

$$h_t = A(z_t)\, h_{t-1} + B(z_t)\, z_t, \qquad y_t = C(z_t)\, h_t, \tag{1}$$

where $A(\cdot), B(\cdot), C(\cdot)$ denote learnable, input-dependent linear maps. This parameterization enables the model to capture both short- and long-range dependencies with computational cost that scales linearly with sequence length for fixed feature and state dimensions $\mathcal{O}(Nd)$. Within our network, the Mamba layer acts as the backbone for volumetric sequence modeling, providing an efficient mechanism for capturing spatial dependencies in 3D imaging data.

Dense Mixtures-of-Experts with Soft Gating. A Mixture-of-Experts (MoE) [38] processes inputs using a bank of E independent experts. Unlike sparse MoE [61, 71], which uses top-k gating [77, 97] to activate a subset of experts, dense soft gating computes a differentiable weighted combination of all E expert outputs:

$$y_{b,n} = \sum_{e=1}^{E} p_{b,n,e}\, \mathrm{FFN}_e(x_{b,n}), \tag{2}$$

where FFN_e denotes the e-th expert and $p_{b,n,e} \geq 0$, with $\sum_e p_{b,n,e} = 1$, denotes its mixture weight. This dense output mixture differs from the slot-based token dispatch used in Soft MoE [68].

Token-level gating requires routing-logit evaluations proportional to the number of tokens. Our anatomy-activated MoE instead predicts crop-specific mixture weights that are shared across spatial positions within each encoder stage and conditioned on anatomical composition. All experts still execute; spatially shared gating reduces routing-logit evaluations rather than sparsifying expert computation. Here, anatomy-activated refers to anatomical conditioning of the mixture weights, without requiring equal expert utilization or a one-to-one correspondence between experts and organs.

4 Methodology

4.1 Anatomy-aware image–text pairs

We denote a multimodal dataset as $\{(\mathbf{X}_n^I, \mathbf{X}_n^T)\}_{n=1}^N$, where $\mathbf{X}_n^I$ is the n-th CT volume, $\mathbf{X}_n^T$ its radiology report, and N the number of pairs. Following [76], we decompose each pair into anatomy-level components:

$$(\mathbf{X}_n^I, \mathbf{X}_n^T) \longmapsto \left\{(\mathbf{X}_{n,c}^I, \mathbf{X}_{n,c}^T)\right\}_{c\in\mathcal{C}_n}, \tag{3}$$

where $\mathbf{X}_{n,c}^I$ and $\mathbf{X}_{n,c}^T$ are the volumetric region and free-text findings of organ class c, and $\mathcal{C}_n \subseteq \{1,\dots,C-1\}$ is the set of anatomical classes present and described for the n-th study. Anatomical structures are extracted via TotalSegmentator [81]. Diagnostic reports are decomposed into organ-level findings with Qwen3-30B-A3B-Instruct [90]. Subsequently, the masks and text snippets are reviewed by experienced radiologists to improve anatomical precision (see Appendix A). These masks and organ-level snippets are used to construct pre-training supervision only; at inference, A^{2}MoE requires only the raw CT volume as input.

4.2 Anatomy-Activated Mixture-of-Experts (A^{2}MoE) layer

Dense 3D encoders route every spatial position through identical feed-forward parameters, implicitly assuming homogeneous feature statistics across the volume [53, 67]. This is suboptimal for volumetric medical images, where anatomical structures vary in scale, shape, and appearance within a single crop [79, 93]. Standard MoE layers activate different experts per input but condition gating solely on hidden visual features, which does not explicitly represent the anatomical composition of the crop. We propose **A^{2}MoE**, which conditions expert selection on a compact *anatomy representation* of the crop together with global visual context. A^{2}MoE executes four steps at each encoder stage s: (i) build a batch of routing vectors $\mathbf{v}_s$ from the feature map $\mathbf{X}_s$; (ii) form a normalized histogram representation $\widetilde{\mathbf{h}}$; (iii) map $(\mathbf{v}_s, \widetilde{\mathbf{h}})$ to expert logits and softmax weights $\mathbf{W}_s \in \mathbb{R}^{B\times K_s}$; (iv) apply the same mixture $\mathbf{W}_s$ to every spatial location of $\mathbf{X}_s$ and add a residual.

Routing context. Let $\mathbf{X}_s \in \mathbb{R}^{B\times C_s\times D_s\times H_s\times W_s}$ denote the feature volume at hierarchical encoder stage $s \in \{1,2,3,4\}$. We aggregate complementary global spatial statistics using global average and max pooling (GAP, GMP) to capture both tissue composition and salient structural boundaries, then concatenate a learnable stage embedding $\mathbf{e}_s \in \mathbb{R}^{d_s}$:

$$\mathbf{v}_s \;=\; \left[\mathrm{GAP}(\mathbf{X}_s) \,\|\, \mathrm{GMP}(\mathbf{X}_s) \,\|\, \mathbf{e}_s\right] \;\in\; \mathbb{R}^{B\times(2C_s+d_s)}. \tag{4}$$

The stage embedding $\mathbf{e}_s$ is broadcast across the B samples. GAP captures overall tissue composition, while GMP preserves small or boundary structures that averaging would suppress.

Anatomical composition. Let $\Omega = \{1,\dots,DHW\}$ denote the voxel domain of the input crop, and let class 0 denote background. From the crop's pseudo-label mask L, we compute:

$$h_{b,c} = \frac{1}{|\Omega|}\sum_{j\in\Omega}\mathbf{1}[L_{b,j}=c], \qquad c\in\{0,\dots,C-1\}. \tag{5}$$

Each stage predicts raw composition using a lightweight MLP, $\hat{\mathbf{h}}_{s,b} = \mathrm{softmax}(\Psi_{\mathrm{pres}}^{(s)}(\mathbf{v}_{s,b}))$. The head predicts crop-level voxel fractions; foreground normalization is applied only when constructing the routing representation. Let $\mathcal{T}$ denote the foreground normalization defined below, and write $\widetilde{\mathbf{h}} = \mathcal{T}(\mathbf{h})$. We supervise the raw and foreground-normalized composition using:

$$\mathcal{L}_{\mathrm{comp}}^{(s)} = \frac{1}{B}\sum_b \left[\mathrm{KL}(\mathbf{h}_b\|\hat{\mathbf{h}}_{s,b}) + \mathrm{KL}(\mathcal{T}(\mathbf{h}_b)\|\mathcal{T}(\hat{\mathbf{h}}_{s,b}))\right]. \tag{6}$$

Anatomical routing representation. For a composition vector $\mathbf{h}$ over background and foreground classes, let $m_{\mathrm{fg}}(\mathbf{h}) = \sum_{j=1}^{C-1} h_j$. The routing representation is defined as follows:

$$[\mathcal{T}(\mathbf{h})]_c = \begin{cases} h_c/m_{\mathrm{fg}}(\mathbf{h}), & c\geq 1, \quad m_{\mathrm{fg}}(\mathbf{h})>0,\\ 0, & c=0, \quad m_{\mathrm{fg}}(\mathbf{h})>0,\\ h_c, & m_{\mathrm{fg}}(\mathbf{h})=0. \end{cases} \tag{7}$$

The transformation removes background and normalizes the remaining organ fractions. When the crop contains no foreground, the raw distribution is retained; the implementation uses a small positive tolerance for this fallback. The same normalization is used for label-derived and predicted composition. During pre-training, routing uses label-derived or detached predicted composition according to the schedule in Appendix C.3; inference uses predicted composition only. The resulting representation preserves the relative extent of visible organs and supplies anatomical context to the router. Expert weights also depend on visual features and learned affinities; normalization alone does not impose equal expert usage or a one-to-one organ–expert correspondence.

Representation-aware gating. Each stage s has K_s experts and a learned class-to-expert affinity matrix $\mathbf{A}^{(s)} \in \mathbb{R}^{C\times K_s}$. Its row-wise softmax, $\widetilde{\mathbf{A}}^{(s)}$, assigns a distribution over experts to each anatomical class. Weighting these distributions by the normalized anatomical histogram $\widetilde{\mathbf{h}}$ yields the structural expert prior:

$$\mathbf{p}_s = \widetilde{\mathbf{h}}\,\widetilde{\mathbf{A}}^{(s)} \in (\Delta^{K_s-1})^B. \tag{8}$$

The gating logits combine a joint visual–anatomical route, a direct anatomical route, and the structural prior:

$$\boldsymbol{\ell}_s = \underbrace{\Psi^{(s)}_{\text{gate}}\Big([\mathbf{v}_s \| \psi^{(s)}(\widetilde{\mathbf{h}})]\Big)}_{\text{visual–anatomical route}} + \underbrace{\Psi^{(s)}_{\text{anat}}(\widetilde{\mathbf{h}})}_{\text{direct anatomical route}} + \underbrace{\beta_s \log(\mathbf{p}_s+\epsilon)}_{\text{structural affinity prior}}. \tag{9}$$

Here $\psi^{(s)}$ embeds the anatomical routing representation, $\Psi^{(s)}_{\text{gate}}$ maps the concatenated context to K_s logits, and $\Psi^{(s)}_{\text{anat}}$ provides a direct anatomy-to-logit path. The learned positive coefficient β_s controls the strength of the anatomical prior, while ϵ is a small constant ensuring numerical stability. The normalized expert mixture weights are:

$$\mathbf{W}_s = \text{softmax}(\boldsymbol{\ell}_s / T_{\text{route}}), \qquad T_{\text{route}} = 0.7. \tag{10}$$

These affinities provide anatomical conditioning without imposing a one-to-one correspondence between organs and experts.

Expert aggregation. Each stage holds K_s bottleneck MLP experts $\{E^{(s)}_k\}_{k=1}^{K_s}$ implemented as 1×1×1 convolutions over channels. With $\hat{\mathbf{X}}_s = \text{InstanceNorm}(\mathbf{X}_s)$,

$$\mathbf{Y}_{s,b} = \mathbf{X}_{s,b} + \sum_{k=1}^{K_s} W_{s,b,k}\, E^{(s)}_k(\hat{\mathbf{X}}_{s,b}). \tag{11}$$

Computational cost. Let $M_s = D_s H_s W_s$ be the number of spatial positions at stage s, and let d^{gate}_s be the width entering a linear expert-logit projection. Per sample, evaluating this projection at every token costs $\mathcal{O}(M_s d^{\text{gate}}_s K_s)$, whereas one crop-level evaluation costs $\mathcal{O}(d^{\text{gate}}_s K_s)$. These bounds concern logit evaluation after context aggregation: pooling still reads the spatial feature map, and anatomy embedding and affinity-prior computations contribute additional per-crop work. All experts are applied to the feature map, so the savings concern repeated routing computations rather than a sparse reduction in expert FLOPs.

4.3 The A^{2}MoE VLM architecture

A^{2}MoE-VLM is based on a 3D CNN-Mamba vision encoder, a frozen biomedical text encoder, and OQMSP that aligns per-organ visual features with the corresponding organ-specific report findings at every encoder stage (see Figure 2).

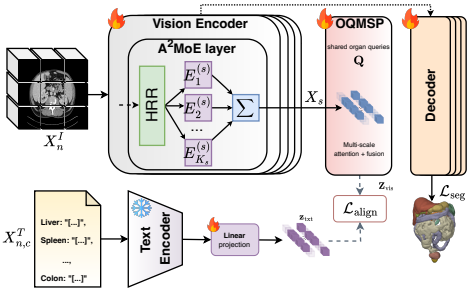

Figure 2: **Overview of the A^{2}MoE architecture.** Given a 3D CT input $\mathbf{X}^I_n$, the Histogram Representation Router (HRR) combines visual context and anatomical composition to predict stage-specific expert weights shared across spatial positions. These weights combine the outputs of all experts $\{E^{(s)}_k\}_{k=1}^{K_s}$. OQMSP uses shared organ queries $\mathbf{Q}$ to aggregate multi-scale encoder features $\mathbf{X}_s$ into organ-level visual embeddings $\mathbf{Z}_{\text{vis}}$. These are aligned with corresponding text embeddings $\mathbf{z}_{\text{txt}}$, obtained through a frozen text encoder and a trainable linear projection, using $\mathcal{L}_{\text{align}}$. A parallel decoder combines encoder features with spatial skip connections and is supervised by $\mathcal{L}_{\text{seg}}$. At inference, anatomical composition is predicted from image features.

Vision encoder. The A^{2}MoE Block (A^{2}MoEB) extends the SegMamba [88] layer with a Gated Spatial Convolution (GSC) [88] module and the A^{2}MoE layer (Sec. 4.2). Given a volume $\mathbf{X} \in \mathbb{R}^{B\times 1\times D\times H\times W}$, a stem produces the initial feature map in $\mathbb{R}^{B\times 48\times D/2\times H/2\times W/2}$, which is then processed by four A^{2}MoEBs with strided downsampling. For the ℓ-th layer of stage s,

$$\mathbf{X}'^{\ell}_s = f_{\mathrm{GSC}}(\mathbf{X}^{\ell}_s), \quad \bar{\mathbf{X}}^{\ell}_s = f_{\mathrm{Mamba}}\big(f_{\mathrm{Norm}}(\mathbf{X}'^{\ell}_s)\big) + \mathbf{X}'^{\ell}_s, \quad \mathbf{X}^{\ell+1}_s = f_{\mathrm{A^2MoE}}\big(f_{\mathrm{Norm}}(\bar{\mathbf{X}}^{\ell}_s)\big) + \bar{\mathbf{X}}^{\ell}_s, \tag{12}$$

where f_{Norm} denotes instance normalization. Features undergo spatial linearization to form a 1D sequence prior to f_{Mamba} and are subsequently restored to their volumetric dimensions before $f_{\mathrm{A^2MoE}}$. We write $\mathbf{X}_s$ for the MoE-refined output of stage s.

Text encoder. Each per-organ finding t is encoded by a frozen PubMedBERT [31] backbone. Token hidden states $\{\mathbf{u}_j\}$ are averaged with the attention mask $m_j \in \{0,1\}$ and passed through a trainable linear projection ϕ_{txt}:

$$\mathbf{z}_{\mathrm{txt}} \;=\; \mathrm{L2}\Big(\phi_{\mathrm{txt}}\Big(\tfrac{\sum_j m_j \mathbf{u}_j}{\sum_j m_j}\Big)\Big) \;\in\; \mathbb{S}^{d_e-1}, \tag{13}$$

where L2 denotes ℓ_2-normalization. Mean pooling aggregates information across content tokens, providing a pooled representation of each finding [70].

Organ-Query Multi-Scale Projector (OQMSP). Aligning volumetric features with per-organ text requires a per-organ embedding at every scale. Prior multi-scale CLIP-style projectors either pool globally, losing anatomical specificity [64], or rely on a ground-truth mask for organ-wise pooling, which can be unavailable at inference or computationally expensive. We instead use a lightweight DETR-style slot projector [18] whose forward pass does not use segmentation masks. Masks are used only to define organ-level supervision during pre-training.

We learn $C-1$ organ queries $\mathbf{Q} \in \mathbb{R}^{(C-1)\times d_e}$ (one per non-background class), shared across all four stages, together with per-stage embeddings $\mathbf{r}_s \in \mathbb{R}^{d_e}$[2]. At stage s, a stage-specific 3D convolution produces keys and values from $\mathbf{X}_s$; LN denotes layer normalization:

$$[\mathbf{K}_s, \mathbf{V}_s] \;=\; \mathrm{LN}\big(\mathrm{Flatten}(\mathrm{Conv3D}_s(\mathbf{X}_s))\big) \;\in\; \mathbb{R}^{B\times N_s\times 2d_e}, \tag{14}$$

where N_s is the number of key/value tokens after the stage-specific convolution, including any spatial downsampling. A multi-head cross-attention (MHA) block with a residual feed-forward network (FFN) then yields per-stage organ embeddings:

$$\tilde{\mathbf{Q}}_s = \mathrm{LN}\big(\mathbf{Q} + \mathbf{r}_s\big), \quad \mathbf{E}_s = \mathbf{Q} + \mathrm{MHA}\big(\tilde{\mathbf{Q}}_s, \mathbf{K}_s, \mathbf{V}_s\big), \tag{15}$$

$$\hat{\mathbf{E}}_s = \mathrm{L2}\big(\mathbf{E}_s + \mathrm{FFN}_s(\mathbf{E}_s)\big) \in \mathbb{R}^{B\times(C-1)\times d_e}. \tag{16}$$

Sharing $\mathbf{Q}$ across stages encourages consistent organ representations across resolutions, so the consistency regularizer below becomes an explicit prior on a scale-invariant organ latent. The cross-attention weights additionally provide an organ-indexed, mask-free soft localization map. Scales are combined with learnable weights normalized by softmax $\boldsymbol{\alpha} = \mathrm{softmax}(a_1, \ldots, a_4)$:

$$\mathbf{Z}_{\mathrm{vis}} \;=\; \mathrm{L2}\Big(\mathrm{LN}\Big(\textstyle\sum_{s=1}^{4} \alpha_s \hat{\mathbf{E}}_s\Big)\Big) \;\in\; \mathbb{R}^{B\times(C-1)\times d_e}. \tag{17}$$

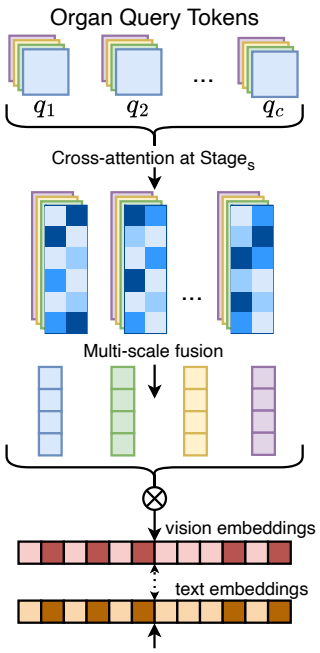

Figure 3: **An overview of OQMSP.** Learnable organ queries $\{\mathbf{q}_c\}_{c=1}^{C-1}$ cross-attend to MoE-refined feature volumes at each pyramid stage, producing stage-specific organ embeddings that are fused with learned scale weights. The query attention acts as a mask-free soft-pooling operator, allowing organ-centric visual embeddings to be aligned with text embeddings without requiring segmentation masks at inference.

The row $\mathbf{Z}_{\mathrm{vis}}[b, c]$ is the visual embedding of organ c in sample b and is produced by OQMSP without an explicit segmentation-mask or histogram input. At inference, the encoder features remain conditioned on predicted anatomical composition.

[2] At stage s the key/value $3\times3\times3$ convolution can use a stride > 1 to keep attention tractable on the finest-resolution volume.

Table 1: **Segmentation performance results for state-of-the-art methods and ablations of key components.** We evaluate across six datasets, spanning three modalities: CT, MRI, and US. We evaluate performance using DSC (↑, %) and HD95 (↓, mm), with the best results shown in **bold**. The table compares the SegMamba baseline (#1) against our proposed components, A^{2}MoE (#2) and OQMSP (#3). † indicates pre-trained models. ‡ Measured using $128 \times 128 \times 128$-voxel patches.

Method	Params	VRAM‡	AMOS-CT [39]		PANORAMA [2]		AMOS-MRI [39]		GEPAR3D [80]		MRISegmenter [98]		MVSeg [19]	
			DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
VoCo† [84, 85]	53.0	4.4	87.1	16.4	71.6	15.9	80.1	10.5	83.5	6.8	73.0	7.7	84.3	5.1
uC 3DU-Net [37]	21.7	3.1	83.1	16.0	76.6	8.4	74.4	16.9	83.8	12.5	76.7	5.9	83.9	4.7
SuPreM† [50]	62.2	4.4	87.1	7.7	76.6	**4.7**	75.8	19.5	80.6	11.0	70.8	8.4	84.3	4.9
Mamba-HoME [67]	170.1	3.7	86.3	18.7	77.5	4.8	77.7	9.9	86.1	6.5	75.2	6.5	84.5	4.8
Merlin† [13]	252.3	3.0	85.6	29.8	77.5	9.4	79.1	10.4	86.5	**2.7**	75.8	6.0	84.1	**3.9**
fVLM† [76]	110.0	0.6	85.2	16.4	76.8	9.8	78.1	10.9	84.1	10.1	72.0	8.1	80.2	5.0
BTB3D† [33]	101.1	1.4	86.1	18.3	77.3	9.2	77.8	11.1	84.0	9.8	76.9	6.2	84.6	4.0
#1: SegMamba [88]	66.8	2.9	85.5	17.3	76.0	8.6	76.7	9.3	85.3	10.6	74.5	6.8	83.8	5.8
#2: + A^{2}MoE	93.3	3.0	87.5	7.6	78.2	7.1	78.5	12.2	86.1	7.9	77.6	5.9	84.5	4.4
#3: **+ OQMSP†**	104.9	3.1	**88.7**	**6.2**	**79.1**	5.8	**82.3**	**8.9**	**87.7**	8.1	**79.8**	**5.5**	**85.4**	4.0

Pre-training objective. We jointly train with a weighted objective that combines spatial segmentation, anatomy-conditioned routing, organ–text alignment, cross-stage consistency, and query grounding:

$$\mathcal{L} = \lambda_{\text{seg}}\,\mathcal{L}_{\text{seg}} + \lambda_{\text{MoE}}\,\mathcal{L}_{\text{MoE}} + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}} + \lambda_{\text{stage}}\,\mathcal{L}_{\text{stage}} + \lambda_{\text{ground}}\,\mathcal{L}_{\text{ground}}. \tag{18}$$

Here, $\mathcal{L}_{\text{seg}}$ is Dice–cross-entropy supervision for the parallel decoder, and $\mathcal{L}_{\text{MoE}}$ combines crop-composition supervision with regularization of the anatomy-conditioned router. $\mathcal{L}_{\text{align}}$ is symmetric organ-level InfoNCE between OQMSP visual embeddings and their corresponding report findings. $\mathcal{L}_{\text{stage}}$ penalizes cosine disagreement between same-organ embeddings from different encoder stages, while $\mathcal{L}_{\text{ground}}$ uses a KL loss to align normalized query attention with downsampled organ pseudo-masks. Pseudo-masks supervise the decoder and query attention during pre-training; the router and OQMSP do not require masks at inference. Full objective definition details appear in Appendix C.3.

5 Experiments

5.1 Experimental setup

Datasets. We pre-train A^{2}MoE on a curated dataset of 67,331 multi-center, multi-phase CT scans from 37,989 patients. These data are sourced from the public Merlin [9, 13] and CancerVerse [49] datasets and one proprietary dataset. All scans are paired with corresponding radiology reports. We provide comprehensive dataset specifications in Appendix B.

Evaluation metrics. We compare the performance of state-of-the-art pre-training methods for various downstream tasks, including multi-disease classification, detection, and segmentation. For classification and detection, following [34, 76], we evaluate using specificity, sensitivity, and ROC-AUC. For segmentation, we use the Dice similarity coefficient (DSC) and the 95th-percentile Hausdorff distance (HD95).

Implementation details. Following [13], we resample each CT scan to $1.5{\times}1.5{\times}3.0$ mm^3 anisotropic resolution. For intensity normalization, all CT scans are clipped to the Hounsfield unit (HU) window of $[-1000, 1000]$ and scaled linearly to $[0, 1]$. During pre-training, $192 \times 192 \times 96$ foreground-biased crops are used as model inputs. Anatomy histograms are computed from the visible label voxels inside each crop; organ-text contrastive pairs are formed only for organs with sufficient crop support, while the presence head is trained on the crop-level histogram. Detailed pre-training and fine-tuning settings are provided in Appendix C.

5.2 Comparison with state-of-the-art methods

We compare our method with several state-of-the-art segmentation models, including VoCo [84, 85], SegMamba [88], Mamba-HoME [67], SuPreM [50], and uC 3DU-Net [37]. We also evaluate against MedVLM-based methods, specifically Merlin [13], fVLM [76], BTB3D [33], RadGPT [11], and CT-CLIP [34]. For classification and detection tasks, we compare against CT-CHAT [34], CT2Rep

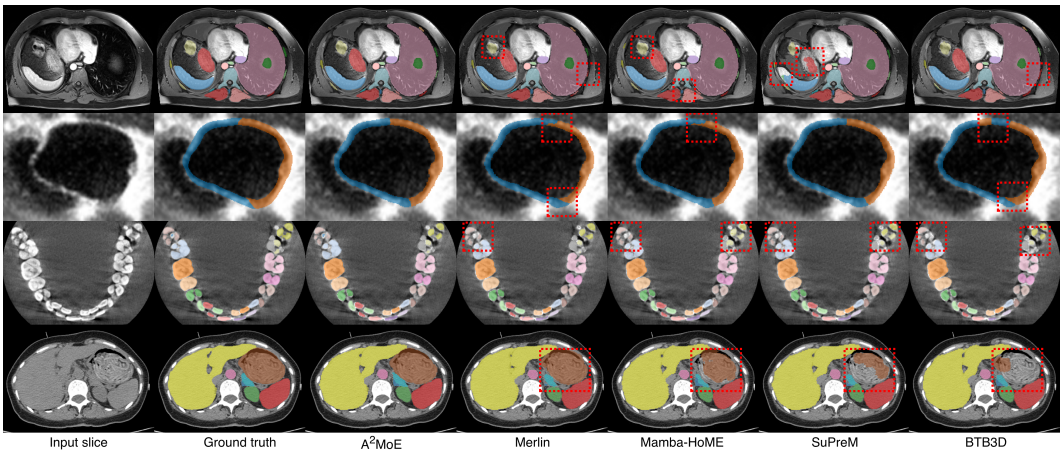

Figure 4: **Qualitative segmentation results.** Comparison of A^{2}MoE against the top four baseline methods. Key differences are highlighted in red bounding boxes (see Appendix F for more samples).

[32], M3D [5], and RadFM [82]. To ensure a fair comparison, we trained each model using consistent hardware settings and data augmentations [27]. All hyperparameters were set according to their original papers.

5.3 Downstream tasks

To demonstrate the robustness and capability of our pre-trained encoder, we conduct additional experiments on several downstream tasks, including segmentation, classification, and detection.

Segmentation. We integrate the pre-trained encoder with a UNETR-based decoder [35]. These experiments utilize both in-distribution and out-of-distribution datasets across multiple modalities, including AMOS-CT/MRI [39], PANORAMA [2], GEPAR3D [80], MRISegmenter [98], and MVSeg [19]. Table 1 shows quantitative segmentation performance results. A^{2}MoE demonstrates superior performance across diverse imaging modalities, consistently outperforming established state-of-the-art models. Quantitative evaluation indicates that the architecture is uniquely capable of segmenting both fine-grained anatomical details and large-scale structures with high precision.

To assess the robustness of the observed performance gains, we use critical-difference diagrams (see Figure 15). We consider differences statistically significant at $p < 0.05$. Qualitatively, A^{2}MoE produces anatomically consistent results that are largely free from the artifacts typically found in competing methods (see Figure 4). Furthermore, the architecture maintains high reliability across both in-distribution and out-of-distribution datasets, demonstrating a robust ability to generalize across varying object scales and data acquisition protocols.

Classification. Figure 8 and Table 25 evaluate abdominal trauma classification on RATIC [73]. A^{2}MoE achieves the best mean ROC-AUC among the evaluated encoders and improves most strongly on bowel and liver injury, where organ-localized context is clinically important. These results support the use of anatomy-conditioned representations beyond segmentation, while kidney and spleen results show that the gains are not uniform across every organ and metric.

Detection. Figure 9 and Table 26 evaluate tumor detection on AbdomenAtlas 3.0 [11]. A^{2}MoE is most competitive for small lesions ($\leq$2 cm), improving sensitivity for kidney and liver tumors and matching the best pancreatic sensitivity among the evaluated models. The specificity–sensitivity trade-off differs by organ, especially for kidney tumors, so we interpret these results as evidence that anatomy-conditioned features improve small-lesion sensitivity rather than as a uniformly dominant detector.

5.4 Ablation studies

We investigate the impact of several factors on the performance of A^{2}MoE, focusing on: (1) key architectural components; (2) histogram-aware routing; and (3) the number of experts per encoder

stage. For a fair comparison, we maintain the same training dataset and hyperparameters across all experimental configurations to isolate the impact of each component. Detailed ablation studies are provided in Appendix D.

Effect of the key components. We evaluate the contribution of the two main components of our segmentation model: A^{2}MoE routing and contrastive alignment. We compare three configurations: (i) the SegMamba baseline [88] (#1); (ii) SegMamba with A^{2}MoE (#2); and (iii) the full model combining A^{2}MoE and contrastive alignment (OQMSP). As shown in Table 1, A^{2}MoE improves over the baseline across datasets, while the full model gives the strongest overall performance. These results indicate that anatomy-activated routing and contrastive alignment provide complementary benefits for segmentation tasks.

Effect of histogram-aware routing. We ablate the two anatomy-conditioned components of the router: the histogram embedding $\psi^{(s)}(\widetilde{\mathbf{h}})$ and the structural affinity prior $\beta_s \log(\widetilde{\mathbf{h}}\widetilde{\mathbf{A}}^{(s)} + \epsilon)$. Table 2 shows that both components improve segmentation consistently across datasets. The gains in DSC are moderate but stable, while the reductions in HD95 are larger, indicating that histogram-aware routing primarily improves boundary robustness and reduces distant false positives rather than only increasing volumetric overlap. Adding the affinity prior gives the strongest improvement, suggesting that explicit class-to-expert structure provides useful anatomical regularization for MoE routing.

Table 2: **Ablation of histogram-aware MoE routing.** We report DSC (%) with HD95 (mm) in parentheses, with the best **bolded**.

Config	AMOS-CT	AMOS-MRI	GEPAR3D	MVSeg
MoE (Baseline)	86.0 (14.4)	77.4 (14.5)	85.6 (9.8)	84.0 (5.0)
+ $\psi^{(s)}(\widetilde{\mathbf{h}})$	87.0 (12.0)	77.9 (14.0)	85.8 (8.5)	84.3 (4.8)
+ $\widetilde{\mathbf{h}}\widetilde{\mathbf{A}}^{(s)}$(**A^{2}MoE**)	**87.5 (7.6)**	**78.5 (12.2)**	**86.1 (7.9)**	**84.5 (4.4)**

Effect of the number of experts. We evaluate how expert count and stage-wise allocation impact performance. As shown in Table 3, the uniform $[16, 16, 16, 16]$ configuration offers the best balance, achieving state-of-the-art results on three datasets while remaining competitive across others. Smaller uniform setups lack sufficient capacity, and pyramid allocations exhibit instability across modalities. Consequently, we adopt the $[16, 16, 16, 16]$ configuration as our default setup.

Table 3: **Ablation study results with various numbers of experts per stage.** We report mDSC (%). The best values are **bolded**.

#	Experts	Params	mDSC (%)					
			A-CT	PANO	A-MRI	GEP	MRISeg	MVSeg
1	[4, 4, 4, 4]	83.8	**87.5**	77.9	78.0	85.9	**77.7**	84.8
2	[8, 8, 8, 8]	87.0	87.3	75.1	77.9	85.0	77.5	**85.0**
3	[4, 8, 16, 32]	102.3	**87.5**	77.4	77.4	86.1	77.6	84.5
4	[16, 16, 16, 16]	93.3	86.9	**78.2**	**78.5**	**86.2**	77.4	84.9

6 Conclusions

We introduced A^{2}MoE, a 3D MedVLM that incorporates explicit anatomical priors into both conditional computation and cross-modal alignment. The proposed HRR grounds expert selection in crop-level class histograms combined with a learnable class-to-expert affinity matrix, producing a gating distribution per crop and stage. Furthermore, the OQMSP aligns MoE-refined visual features with organ-specific textual representations through scale-shared queries. At inference, anatomical composition is predicted from image features, allowing the visual branch to operate without segmentation masks or report text. Pre-trained on a large-scale cohort of over 60,000 CT scans, A^{2}MoE achieves state-of-the-art performance on external benchmarks for segmentation, classification, and detection. Moreover, it demonstrates strong transfer to MRI and ultrasound, suggesting that anatomical grounding serves as a modality-agnostic prior.

Limitations and future directions. Despite its performance, A^{2}MoE has certain limitations. First, the current framework relies on a fixed set of anatomical classes for the composition histogram; while this covers major organ systems, it may require adaptation for highly granular or rare pathological structures not represented in the pre-training prior. Second, while the model exhibits strong cross-modality transfer, its performance is inherently tied to the quality and diversity of the initial CT-report pairs. Finally, A^{2}MoE focuses on static 3D volumes; extending this anatomical conditioning to longitudinal data remains an open challenge for future research.

Acknowledgments

This research was funded in part by National Science Center, Poland 2025/57/B/ST6/03100. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018792.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[2] Natalia Alves, Megan Schuurmans, Dawid Rutkowski, Anindo Saha, Pierpaolo Vendittelli, Nancy Obuchowski, Marjolein H Liedenbaum, Ingfrid S Haldorsen, Anders Molven, Derya Yakar, et al. Artificial intelligence and radiologists in pancreatic cancer detection using standard of care ct scans (panorama): an international, paired, non-inferiority, confirmatory, observational study. *The Lancet Oncology*, 27(1):116–124, 2026.

[3] Nils Alves, Merlijn Schuurmans, Dominik Rutkowski, Derya Yakar, Ingfrid Haldorsen, Marjolein Liedenbaum, Anders Molven, Phillip Vendittelli, Geert Litjens, Jurgen Hermans, et al. The panorama study protocol: Pancreatic cancer diagnosis-radiologists meet ai. *Zenodo*, 2024.

[4] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

[5] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[6] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[7] Pedro RAS Bassi, Sergio SJ Dertkigil, and Andrea Cavalli. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. *Nature Communications*, 15(1):291, 2024.

[8] Pedro RAS Bassi, Wenxuan Li, Jieneng Chen, Zheren Zhu, Tianyu Lin, Sergio Decherchi, Andrea Cavalli, Kang Wang, Yang Yang, Alan L Yuille, et al. Learning segmentation from radiology reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–315. Springer, 2025.

[9] Pedro RAS Bassi, Wenxuan Li, Szymon Płotka, Ruby Honjol, Jakub Prządo, Xinze Zhou, Kang Wang, Yang Yang, Malte Jensen, Akshay S Chaudhari, et al. Merlin plus: A large-scale, multi-cancer, image-mask-report dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 194–204. Springer, 2026.

[10] Pedro RAS Bassi, Qilong Wu, Wenxuan Li, Sergio Decherchi, Andrea Cavalli, Alan Yuille, and Zongwei Zhou. Label critic: Design data before models. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025.

[11] Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, et al. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23720–23730, 2025.

[12] Pedro RAS Bassi, Xinze Zhou, Wenxuan Li, Szymon Płotka, Jieneng Chen, Qi Chen, Zheren Zhu, Jakub Prządo, Ibrahim E Hamacı, Sezgin Er, et al. Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks. *arXiv preprint arXiv:2510.14803*, 2025.

[13] Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652(8112):1318–1328, 2026.

[14] Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Kevin Marchesini, Niels Van Nistelrooij, Pieter Van Lierop, Tong Xi, Yusheng Liu, et al. Segmenting the inferior alveolar canal in cbcts volumes: the toothfairy challenge. *IEEE Transactions on Medical Imaging*, 44(4):1890–1906, 2024.

[15] Federico Bolelli, Kevin Marchesini, Niels van Nistelrooij, Luca Lumetti, Vittorio Pipoli, Elisa Ficarra, Shankeeth Vinayahalingam, and Costantino Grana. Segmenting maxillofacial structures in cbct volumes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5238–5248, 2025.

[16] 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*, pages 23041–23050, 2025.

[17] Umberto Cappellazzo, Minsu Kim, Pingchuan Ma, Honglie Chen, Xubo Liu, Stavros Petridis, and Maja Pantic. Mome: Mixture of matryoshka experts for audio-visual speech recognition. *Advances in Neural Information Processing Systems*, 38:40013–40040, 2026.

[18] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[19] Patrick Carnahan, John Moore, Daniel Bainbridge, Mehdi Eskandari, Elvis CS Chen, and Terry M Peters. Deepmitral: Fully automatic 3d echocardiography segmentation for patient specific mitral valve modelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 459–468. Springer, 2021.

[20] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

[21] Anzhe Cheng, Shukai Duan, Shixuan Li, Chenzhong Yin, Mingxi Cheng, Heng Ping, Tamoghna Chattopadhyay, Sophia I Thomopoulos, Shahin Nazarian, Paul Thompson, et al. Ermoe: Eigen-reparameterized mixture-of-experts for stable routing and interpretable specialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12997–13006, 2026.

[22] Maciej Chrabaszcz, Hubert Baniecki, Piotr Komorowski, Szymon Plotka, and Przemyslaw Biecek. Aggregated attributions for explanatory analysis of 3d segmentation models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 160–171, 2025.

[23] Yuetan Chu, Xinghua Ma, Xinran Jin, Gongning Luo, and Xin Gao. Medtri: Structured medical report normalization for enhanced vision–language pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 366–376. Springer, 2026.

[24] Cris Claessens, Christiaan Viviers, Giacomo D'Amicantonio, Egor Bondarev, and Fons van der Sommen. Scaling self-supervised and cross-modal pretraining for volumetric ct transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13636–13647, 2026.

[25] Zhiming Cui, Yu Fang, Lanzhuju Mei, Bojun Zhang, Bo Yu, Jiameng Liu, Caiwen Jiang, Yuhang Sun, Lei Ma, Jiawei Huang, et al. A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images. *Nature Communications*, 13(1):2096, 2022.

[26] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in Neural Information Processing Systems*, 35:32942–32956, 2022.

[27] Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu D Tizabi, Fabian Isensee, Tim J Adler, Sharib Ali, Vincent Andrearczyk, Marc Aubreville, Ujjwal Baid, et al. Why is the winner the best? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19955–19966, 2023.

[28] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[29] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[30] Difei Gu, Yunhe Gao, Mu Zhou, and Dimitris Metaxas. Anatomy-vlm: a fine-grained vision-language model for medical interpretation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2838–2847, 2026.

[31] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[32] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024.

[33] Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. *Advances in Neural Information Processing Systems*, 38:135074–135102, 2026.

[34] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19, 2026.

[35] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[36] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[37] Xingru Huang, Jian Huang, Tianyun Zhang, HE HONG, Shaowei Jiang, Yaoqi Sun, et al. Upping the game: How 2d u-net skip connections flip 3d segmentation. *Advances in Neural Information Processing Systems*, 37:87282–87309, 2024.

[38] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[39] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35:36722–36732, 2022.

[40] Masahiro Kada, Ryota Yoshihashi, Satoshi Ikehata, Rei Kawakami, and Ikuro Sato. Teacher-guided routing for sparse vision mixture-of-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6198–6208, 2026.

[41] Sungnyun Kim, Kangwook Jang, Sangmin Bae, Sungwoo Cho, and Se-Young Yun. MoHAVE: Mixture of hierarchical audio-visual experts for robust speech recognition. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 30113–30131. PMLR, 13–19 Jul 2025.

[42] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[43] Changsun Lee, Sangjoon Park, Cheong Il Shin, Woo Hee Choi, Hyun Jeong Park, Jeong Eun Lee, and Jong Chul Ye. Read like a radiologist: Efficient vision-language model for 3d medical imaging interpretation. *Medical Image Analysis*, page 104077, 2026.

[44] Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Huazhen Huang, Qingqing Gu, Yetao Wu, and Luo Ji. M3-JEPA: Multimodal alignment via multi-gate MoE based on the joint-embedding predictive architecture. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 33902–33917. PMLR, 13–19 Jul 2025.

[45] Jiayu Lei, Ziqing Fan, Yanyong Zhang, Weidi Xie, Ya Zhang, and Yanfeng Wang. Versatile vision-language model for 3d computed tomography. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 5945–5954, 2026.

[46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023.

[47] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul 2022.

[48] Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 80–90. Springer, 2024.

[49] Wenxuan Li, Pedro RAS Bassi, Xinze Zhou, Jakob Wasserthal, Ibrahim E Hamamci, Sezgin Er, Bjoern Menze, Gulhan E Akan, Szymon Płotka, Jakub Prządo, et al. Cancerverse: A fully open longitudinal and multimodal dataset for multicancer screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 77–87. Springer, 2026.

[50] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised 3d models transfer to medical imaging tasks? In *International Conference on Learning Representations*, volume 2024, pages 16389–16404, 2024.

[51] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(1):96–107, 2023.

[52] Xinjie Liang, Xiangyu Li, Fanding Li, Jie Jiang, Qing Dong, Wei Wang, Kuanquan Wang, Suyu Dong, Gongning Luo, and Shuo Li. Medfilip: Medical fine-grained language-image pre-training. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3587–3597, 2025.

[53] Yujie Lu, Jingwen Li, Sibo Ju, Yanzhou Su, He Yao, Yisong Liu, Min Zhu, and Junlong Cheng. Segmote: Token-level mixture of experts for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 36332–36342, 2026.

[54] Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Elisa Ficarra, and Costantino Grana. Enhancing patch-based learning for the segmentation of the mandibular canal. *IEEE Access*, 12:79014–79024, 2024.

[55] Duy MH Nguyen, Nghiem Diep, Trung Nguyen, Hoang-Bao Le, Tai Nguyen, Anh-Tien Nguyen, TrungTin Nguyen, Nhat Ho, Pengtao Xie, Roger Wattenhofer, et al. Exgra-med: Extended context graph alignment for medical vision-language models. *Advances in Neural Information Processing Systems*, 38:109568–109611, 2026.

[56] Hancang Mi, Hong Seng Gan, Dong Ma, and Mustafa Alper Selver. Organ-aware mixture-of-experts framework for generalized pan-tumor segmentation. *Biomedical Signal Processing and Control*, 126:110937, 2026.

[57] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.

[58] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR, 2023.

[59] Yajun Mu, Yong Chen, Shuhan Yang, Jun Gao, Qicheng Lao, Kang Li, and Qingbo Kang. Harmoe: Hierarchical affinity-aware routing mixture-of-experts for unified multi-organ ultrasound analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 240–250. Springer, 2026.

[60] 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*, pages 685–701. Springer, 2022.

[61] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multi-modal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

[62] Xuefeng Ni, Linshan Wu, Jiaxin Zhuang, Qiong Wang, Mingxiang Wu, Varut Vardhanabhuti, Lihai Zhang, Hanyu Gao, and Hao Chen. Mg-3d: Multi-grained knowledge-enhanced vision-language pre-training for 3d medical image analysis. *Medical Image Analysis*, page 104027, 2026.

[63] Krzysztof Ociepa, Remigiusz Kinas, Krzysztof Wróbel, Adrian Gwoździej, et al. Bielik 11b v3: Multilingual large language model for european languages. *arXiv preprint arXiv:2601.11579*, 2025.

[64] Jun Yeong Park, JunYoung Seo, Minji Kang, and Yu Rang Park. Moeclip: patch-specialized experts for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 35534–35544, 2026.

[65] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Szymon Płotka, Maciej Chrabaszcz, and Przemyslaw Biecek. Swin smt: Global sequential modeling for enhancing 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 689–698. Springer, 2024.

[67] Szymon Płotka, Gizem Mert, Maciej Chrabaszcz, Ewa Szczurek, and Arkadiusz Sitek. Mamba goes home: Hierarchical soft mixture-of-experts for 3d medical image segmentation. *Advances in Neural Information Processing Systems*, 38:97871–97909, 2026.

[68] Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, volume 2024, pages 28435–28445, 2024.

[69] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[70] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[71] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

[72] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 556–564. Springer, 2015.

[73] Jeffrey D Rudie, Hui-Ming Lin, Robyn L Ball, Sabeena Jalal, Luciano M Prevedello, Savvas Nicolaou, Brett S Marinelli, Adam E Flanders, Kirti Magudia, George Shih, et al. The rsna abdominal traumatic injury ct (ratic) dataset. *Radiology: Artificial Intelligence*, 6(6):e240101, 2024.

[74] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

[75] Bowen Shi, Weiwei Cao, Ruifeng Yuan, Wanxing Chang, Wenrui Dai, Hongkai Xiong, Ling Zhang, and Jianpeng Zhang. Disease-centric vision-language pretraining with hybrid visual encoding for 3D computed tomography. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 111726–111742. PMLR, 2026.

[76] 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. In *International Conference on Learning Representations*, volume 2025, pages 24094–24107, 2025.

[77] Łukasz Struski, Michal B. Bednarczyk, Igor T. Podolak, and Jacek Tabor. LapSum - one method to differentiate them all: Ranking, sorting and top-k selection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 56990–57007. PMLR, 13–19 Jul 2025.

[78] Ke Su, Xingxing Zhang, Siyang Zhang, Jun Zhu, and Bo Zhang. To boost zero-shot generalization for embodied reasoning with vision-language pre-training. *IEEE Transactions on Image Processing*, 33:5370–5381, 2024.

[79] Tomasz Szczepański, Michal K Grzeszczyk, Szymon Płotka, Arleta Adamowicz, Piotr Fudalej, Przemysław Korzeniowski, Tomasz Trzciński, and Arkadiusz Sitek. Let me decode you: Decoder conditioning with tabular data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 228–238. Springer, 2024.

[80] Tomasz Szczepański, Szymon Płotka, Michal K Grzeszczyk, Arleta Adamowicz, Piotr Fudalej, Przemysław Korzeniowski, Tomasz Trzciński, and Arkadiusz Sitek. Gepar3d: Geometry prior-assisted learning for 3d tooth segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 218–228. Springer, 2025.

[81] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[82] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

[83] Junxian Wu, Minheng Chen, Xinyi Ke, Tianwang Xun, Xiaoming Jiang, Hongyu Zhou, Lizhi Shao, and Youyong Kong. Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5144–5153, 2025.

[84] Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22873–22882, 2024.

[85] Linshan Wu, Jiaxin Zhuang, and Hao Chen. Large-scale 3d medical image pre-training with geometric context priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[86] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365, 2024.

[87] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[88] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–588. Springer, 2024.

[89] Huimin Yan, Xian Yang, Liang Bai, and Jiye Liang. Local alignment for medical vision-language pre-training. *IEEE Transactions on Image Processing*, 34:7321–7334, 2025.

[90] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[91] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[92] 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.

[93] Xinru Zhang, Ni Ou, Berke Doga Basaran, Marco Visentin, Mengyun Qiao, Renyang Gu, Paul M Matthews, Yaou Liu, Chuyang Ye, and Wenjia Bai. A foundation model for lesion segmentation on brain mri with mixture of modality experts. *IEEE Transactions on Medical Imaging*, 44(6):2594–2604, 2025.

[94] 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*, pages 2–25. PMLR, 2022.

[95] Ziyang Zhang, Yang Yu, Yucheng Chen, Xulei Yang, and Si Yong Yeo. Medunifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29744–29755, 2025.

[96] Chenhui Zhao, Yiwei Lyu, Asadur Zaman Chowdury, Edward S Harake, Akhil Kondepudi, Akshay T Rao, Xinhai Hou, Honglak Lee, and Todd C Hollon. Towards scalable language-image pre-training for 3d medical imaging. *Transactions on Machine Learning Research*, 2026.

[97] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

[98] Yan Zhuang, Tejas Sudharshan Mathai, Pritam Mukherjee, Brandon Khoury, Boah Kim, Benjamin Hou, Nusrat Rabbee, Abhinav Suri, and Ronald M Summers. Mrisegmenter: A fully accurate and robust automated multiorgan and structure segmentation tool for t1-weighted abdominal mri. *Radiology*, 315(1):e241979, 2025.

Appendix

Table of Contents

A Data pre-processing **20**
- A.1 Anatomy parsing . . . 20
- A.2 Prompts . . . 25

B Datasets **28**
- B.1 Training . . . 28
- B.2 Fine-tuning . . . 29

C Experimental setup **32**
- C.1 Hardware and software setup . . . 32
- C.2 Pre-training . . . 32
- C.3 Pre-training objective and routing details . . . 32
- C.4 Downstream tasks . . . 33
- C.5 Evaluation metrics . . . 34

D Ablation studies **35**
- D.1 Effect of the key components . . . 35
- D.2 Effect of the number of Experts per stage . . . 37

E Quantitative results **39**
- E.1 Segmentation . . . 39
- E.2 Classification . . . 42
- E.3 Detection . . . 43

F Qualitative results **44**

G Statistical significance **47**

H Generalizability **48**

I Limitations and future work **49**

J Impact statement **50**

K Notation **51**

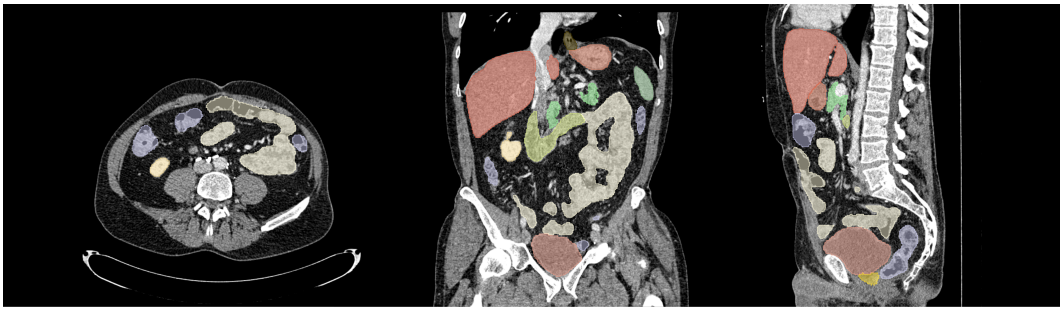

Figure 5: Anatomical segmentation results obtained with TotalSegmentator [81] across axial, coronal, and sagittal planes for 14 diverse abdominal and pelvic structures. The visualization demonstrates the model's spatial consistency across orthogonal views for organs of varying scales, ranging from large structures (e.g., liver, colon) to small, high-variance targets (e.g., adrenal glands, duodenum). These high-fidelity masks serve as a structured anatomical prior, providing localized supervision for image-text alignment and ensuring that the joint embedding space preserves the underlying biological topology.

A Data pre-processing

A.1 Anatomy parsing

We leverage the pre-trained TotalSegmentator [81] to generate voxel-wise pseudo-labels (see Figure 5). Note that the pseudo-labels and radiology organ snippets were verified by experienced radiologists; see Sec. A.1.5 for details. Although the original model encompasses 117 anatomical structures, we directly extract 14 key abdominal organs, including *spleen, left kidney, right kidney, gallbladder, liver, stomach, pancreas, left adrenal gland, right adrenal gland, duodenum, small bowel, colon, urinary bladder, and prostate* (see Table 4). This pruning process serves two primary purposes: first, it filters out structures irrelevant to our domain-specific radiology reports (e.g., thoracic organs); and second, it removes low-prevalence classes (e.g., skeletal and vascular structures) that introduce unnecessary noise. This semantic alignment bridges the gap between high-resolution volumetric segmentations and the coarser, descriptive terminology utilized in clinical practice. By consolidating these structures, we reduce computational complexity and ensure the segmentation output is clinically actionable, facilitating better integration with standardized reporting frameworks (see Figure 6).

Table 4: **Anatomical pseudo-labels.** Overview of the 14 anatomical classes extracted through TotalSegmentator [81] to serve as auxiliary supervision. These classes encompass a diverse range of abdominal and pelvic structures with varying morphologies. These masks spatially ground the model, aligning radiological tokens with anatomical regions to preserve biological topology in the latent space.

Class	Anatomy
0	Background
1	Spleen
2	Left kidney
3	Right kidney
4	Gallbladder
5	Liver
6	Stomach
7	Pancreas
8	Left adrenal gland
9	Right adrenal gland
10	Duodenum
11	Small bowel
12	Colon
13	Urinary bladder
14	Prostate

The adoption of these masks also addresses a fundamental challenge in medical vision-language modeling: the inherent scarcity and spatial ambiguity of lesion localization in clinical text. Pathological findings in specific anatomical regions often follow a long-tailed distribution, leading to an acute class imbalance during training. Without such anatomical grounding, many training iterations would consist of "normal" samples, potentially leading to representation collapse or a bias toward the majority class. By incorporating these 14 structures as an auxiliary spatial prior, we provide the model with a robust framework to learn meaningful diagnostic features even in the presence of sparse or imbalanced clinical labels.

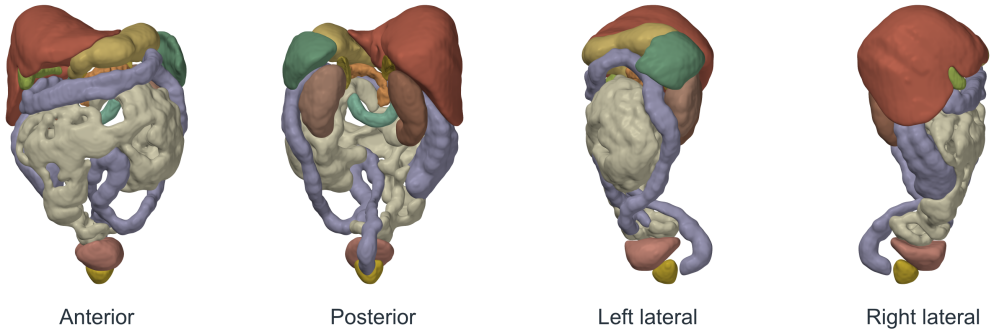

Figure 6: Visualization of the 3D projections of the pseudo-labels generated with TotalSegmentator [81], shown from the anterior, posterior, left lateral, and right lateral views. These projections highlight the global structural consistency and anatomical coherence of the pseudo-labels (see Table 4). The multi-view renderings provide a qualitative assessment of the overall segmentation quality across the full 3D volume.

A.1.1 Report decomposition

Raw radiology reports are cleaned and decomposed at the organ level based on the selected anatomical structures. We clean these reports by removing unnecessary information, such as dates, details regarding modalities other than CT, and findings related to previous examinations. Following [76], the decomposition is performed as follows:

1. We develop domain-knowledge prompts with guidance from experienced radiologists to ensure comprehensive terminology coverage, preventing missed organ structures arising from synonym variation in clinical reports (see Section A.2);
2. We employ Qwen3-30B-A3B-Instruct [90] for image-text alignment at the organ level using the prompts presented in Section A.2. To accelerate inference, we employ vLLM [42];
3. We then align the descriptions of each anatomical structure from the *findings* and *impression* sections,
4. Lastly, for an organ described as *normal*, we use a template describing the absence of pathology (see Section A.1.3).

A.1.2 Diagnosis extraction

To automate the labeling process, we utilized an LLM (Qwen3-30B-A3B-Instruct), prompted with a domain-specific heuristic to distinguish between clinical pathology and benign anatomical descriptions for each organ (see Table 4). The core challenge in radiological labeling is preventing "false positives" triggered by descriptive text that does not indicate active disease.

We established a strict hierarchy for the *Pathological (1)* label: it is reserved for intrinsic tissue abnormalities, including mass effects, inflammatory processes (e.g., dilatation, stones), or persistent lesions (including stable cysts). Conversely, the *Normal/Unremarkable (0)* label accounts for three specific clinical scenarios:

1. Anatomical and Vascular Variants: Non-pathological deviations such as "replaced" arterial supplies,
2. Stable Post-Surgical States: The absence of an organ or expected post-operative changes without secondary complications,
3. External Proximity: Displacement or compression of an organ by an adjacent mass (e.g., "modeling") is excluded from the pathological count unless the report explicitly notes tissue infiltration or parenchymal damage.

A.1.3 Normal templates

To mitigate the representation collapse often observed in medical domain embeddings, we move beyond the use of monolithic labels like "Normal." Such simplistic categorizations fail to capture the multimodal distribution of healthy anatomical states, leading to an underspecified latent space that lacks the granularity required for robust contrastive learning.

Instead, we synthesize structured normal templates to serve as high-fidelity prototypical anchors. By embedding rich, organ-specific descriptive text rather than a binary state, we force the model to align its representations with detailed physiological features (e.g., organ axis, attenuation, and morphology). This approach ensures that during the contrastive optimization, specifically when minimizing the distance between positive pairs, the model learns a structured embedding space where "normality" is defined by a dense cluster of anatomical constraints. Consequently, pathological deviations are not just "not normal," but are explicitly represented as semantic residuals relative to a highly defined physiological baseline, enhancing the model's ability to generalize across heterogeneous clinical manifestations.

Normal Templates

```
{
    "Spleen": "The spleen is normal in size and homogeneous, with no
    splenomegaly or focal lesions identified.",
    "Kidney left": "The kidney left is normal in size and axis.
    No hydronephrosis, renal calculi, or solid masses.",
    "Kidney right": "The kidney right is normal in size and axis.
    No hydronephrosis, renal calculi, or solid masses.",
    "Gallbladder": "The gallbladder has a normal caliber. No wall
    thickening, pericholecystic fluid, or gallstones.",
    "Liver": "The liver is normal in size with smooth contours. No focal
    hepatic lesions, biliary dilatation, or steatosis.",
    "Stomach": "The stomach shows normal wall thickness with no
    distension or masses.",
    "Pancreas": "The pancreas demonstrates normal attenuation.
    No pancreatic duct dilatation or focal masses.",
    "Adrenal gland left": "The adrenal gland left have normal morphology
    with no nodules or masses.",
    "Adrenal gland right": "The adrenal gland right have normal morphology
    with no nodules or masses.",
    "Small bowel": "The small bowel loops are normal in caliber
    without wall thickening or obstruction.",
    "Colon": "The colon is normal in caliber with no wall thickening,
    obstruction, or free air.",
    "Urinary bladder": "The urinary bladder is well-distended with
    no wall thickening or masses.",
    "Prostate": "The prostate is normal in size with no enlargement
    or focal nodules."
}
```

Original radiology report

FINDINGS: Lower thorax: Bibasilar dependent opacities, greater on the right likely reflect dependent atelectasis. Liver and biliary tree: Numerous subcentimeter hypodensities are too small to definitively characterize but are likely to represent benign entity such as hamartomas or small cysts. Mild to moderate intraand extrahepatic biliary ductal dilation is likely related to postcholecystectomy state. Gallbladder: Surgically absent. Spleen: Normal. Pancreas: Normal. Adrenal glands: Normal. Kidneys and ureters: Bilateral subcentimeter hypodensities are indeterminant, likely representing small cysts. Symmetric enhancement without hydronephrosis. Gastrointestinal tract: Multiple dilated upstream small bowel loops are present to the level of the left upper quadrant where 2 adjacent transition points are noted. Bowel loops between these areas of transition are thickened and have hazy margins. There is marked infiltration of the mesentery suggesting venous compromise and ischemia. Distal small bowel loops are collapsed. No pneumatosis. Colonic diverticulosis without evidence of diverticulitis. Peritoneal cavity: Small moderate volume of ascites in the right lower quadrant and pelvis. No free air. Bladder: Normal. Prostate and seminal vesicles: Normal. Vasculature: Calcified atherosclerotic disease without evidence of aneurysm. Lymph nodes: Normal. Abdominal wall: Normal. Musculoskeletal: Mild degenerative spine change.



Structured radiology report


```
{
  "study_id": "AC423ce5b",
  "cleaned_report": "[... The full cleaned report text is here ...]",
  "organs": {
    "Spleen": "The spleen is normal in size and homogeneous, with no
    splenomegaly or focal lesions identified.",
    "Kidney left": "The kidney left is normal in size and axis.
    No hydronephrosis, renal calculi, or solid masses.",
    "Kidney right": "The kidney right is normal in size and axis.
    No hydronephrosis, renal calculi, or solid masses.",
    "Gallbladder": "The gallbladder is surgically absent.",
    "Liver": "The liver is normal in size with smooth contours. No focal
    hepatic lesions, biliary dilatation, or steatosis.",
    "Stomach": "Unremarkable",
    "Pancreas": "The pancreas demonstrates normal attenuation.
    No pancreatic duct dilatation or focal masses.",
    "Adrenal gland left": "The adrenal gland left have normal morphology
    with no nodules or masses.",
    "Adrenal gland right": "The adrenal gland right have normal morphology
    with no nodules or masses.",
    "Small bowel": "Multiple dilated upstream small bowel loops present
    to the level of the left upper quadrant where two adjacent transition points
    are noted. Bowel loops between these areas of transition are thickened
    and have hazy margins. Marked infiltration of the mesentery suggesting
    venous compromise and ischemia. Distal small bowel loops are collapsed.",
    "Colon": "The colon is normal in caliber with no wall thickening,
    obstruction, or free air.",
    "Urinary bladder": "The urinary bladder is well-distended with no wall
    thickening or masses.",
    "Prostate": "The prostate is normal in size with no enlargement
    or focal nodules."
  }
}
```

A.1.4 Radiology report translation

We translate our proprietary radiology reports from Polish to English using Bielik-11B-v3.0-Instruct [63].

A.1.5 Verification of anatomy-aware snippets

Pseudo-label correction. To maximize segmentation reliability, we implement a multi-layered verification framework that integrates clinical ground truth with automated quality-control heuristics. The first layer involves cross-referencing pseudo-masks with the original radiology report to ensure anatomical plausibility. For instance, when a report explicitly documents a "status post-nephrectomy" or "surgically absent" organ, any voxels erroneously detected by TotalSegmentator [81] for that structure are programmatically reclassified as background. This report-guided correction directly addresses the most common systematic failure mode of deep-learning segmenters: the tendency to misidentify scar tissue, surgical clips, or adjacent structures as an organ that is no longer present.

Building upon this clinical validation, we adopt the methodology of "Label Critic" [10] as a formal quality-control mechanism to manage label uncertainty. This framework identifies likely incorrect labels in single-annotation settings by flagging outliers that deviate from learned anatomical priors. Furthermore, in scenarios where multiple segmenters or iterations produce competing outputs, the mechanism acts as an arbiter to select the superior annotation based on spatial consistency and boundary fidelity. By combining NLP-driven anatomical pruning with these formal QC metrics, the pipeline ensures that downstream quantitative analysis, such as radiomics or volumetry, is performed on masks that are verified against the patient's specific surgical history and radiological reality.

Radiology report snippets. To ensure the clinical accuracy of the dataset, each anatomy-aware snippet underwent manual review by a senior radiologist with over 30 years of clinical experience. The verification process focused on confirming that the extracted segments maintained anatomical integrity and clinical relevance relative to the original imaging reports.

A.2 Prompts

A.2.1 Raw report cleaning

This cleaning pseudo-prompt serves as a specialized preprocessing heuristic designed to isolate "point-in-time" clinical observations from the longitudinal noise characteristic of radiology narratives. It enforces an aggressive filtration logic that programmatically de-identifies temporal data and excises comparative metadata, such as cross-modality references (e.g., prior MRI or PET/CT) and stability descriptors, which frequently induce contextual bias in downstream extraction models.

By distilling the narrative into a pure anatomical state-of-health summary, the prompt ensures that the model's attention is localized strictly to current physiological findings and objective measurements. This standardized text-normalization protocol, including its complete rule-set and edge-case handling, is available in our code repository.

Protocol 1: Structured Report Normalization Pipeline

```
ALGORITHM: Radiology Text Normalization
INPUT: Raw CT findings (text)
OUTPUT: Cleaned CT findings (text)

INSTRUCTION SET:
1.  STRIP_TEMPORAL:
    - Remove <DATE>, YEAR, [0-9]{4}, 'now', 'interval', 'current'.
2.  STRIP_LONGITUDINAL_REFERENCES:
    - Remove all clauses containing: 'prior', 'stable', 'unchanged',
      'comparison', 'compared to', 'no change'.
3.  STRIP_CROSS_MODAL_NOISE:
    - Remove references to: 'PET/CT', 'MRI', 'ultrasound', 'X-ray',
    'scintigraphy'.
4.  STRIP_NON_SEMANTIC_SYNTAX:
    - Remove all content within ( ), [ ], < >, { }.
    - Remove billing codes and meta-annotations.
5.  CORE_FINDING_RETAIN:
    - Extract current anatomical findings.
    - Preserve clinical measurements (e.g., '4 mm').
    - Maintain medical terminology and morphology descriptions.

EXECUTION:
- Apply aggressive filtering to prune phrases containing comparative language.
- Ensure natural linguistic flow of remaining medical tokens.
- Output high-density clinical text only.
```

A.2.2 Organ-aware report decomposition

Protocol 2: Hierarchical Organ-Specific Feature Extraction

```
ALGORITHM: Segment-then-Map Anatomical Extraction
INPUT: Cleaned CT findings (text), Target Organ List
OUTPUT: Structured JSON Map (Organ: Finding)

INSTRUCTION SET:
1.  LOCALIZE_CONTEXT:
    - Identify sentences specific to each Target Organ.
    - Validate finding ownership (e.g., exclude "peri-pancreatic" from Pancreas).
2.  LATERALIZATION_RESOLVER:
    - Map plural descriptors (e.g., "Kidneys") to both lateral child nodes.
    - Isolate unilateral findings strictly to the specific side (Right/Left).
3.  EVIDENCE_HIERARCHY_DECISION:
    - IF PATHOLOGY_DETECTED:
        - Identify: Calcification, cyst, nodule, stone, surgical change.
        - ACTION: Extract verbatim clinical snippet.
    - ELSE IF EXPLICIT_NORMAL:
        - Identify: "normal", "unremarkable", "no findings".
        - ACTION: Inject EXACT string from Standardized Template Pool.
    - ELSE:
        - ACTION: Assign "Unremarkable" (Indicates omission/absence).
4.  PATHOLOGY_GUARDRAIL:
    - Explicitly inhibit Template Injection if any secondary findings
      (e.g., fat infiltration, resections, clips) are identified.

EXECUTION:
- Iterate through Target Organ List.
- Perform high-fidelity mapping of sub-segments (e.g., Rectum -> Colon).
- Output strictly validated JSON schema.
```

A.2.3 Diagnosis extraction

Protocol 3: Binary Pathological Classification and Scoring

```
ALGORITHM: Binary Pathology Labeling (0/1)
INPUT: Structured JSON Findings Map
OUTPUT: Binary JSON Map (Organ: Int)

SCORING CRITERIA (STRICT):
1.  CLASS_0 (STRICTLY_NORMAL):
    - Match must be 100% identical to the 'Normal Template'.
    - Apply if finding is explicitly 'Unremarkable'.
2.  CLASS_1 (ANY_FINDING):
    - Apply for any deviation from physiological baseline.
    - INCIDENTIALOMAS: Tiny cysts, granulomas, nodules.
    - CALCIFICATIONS: Vascular or parenchymal calcifications.
    - SURGICAL_STATES: Resections, cholecystectomy, clips, stents.
    - MORPHOLOGY: Atrophy, organ absence, wall thickening.

EXECUTION:
- Evaluate each organ entry in the findings map.
- Trigger CLASS_1 for any text not matching the predefined Normal Pool.
- Return a condensed JSON map mapping Organ names to binary integers.
```

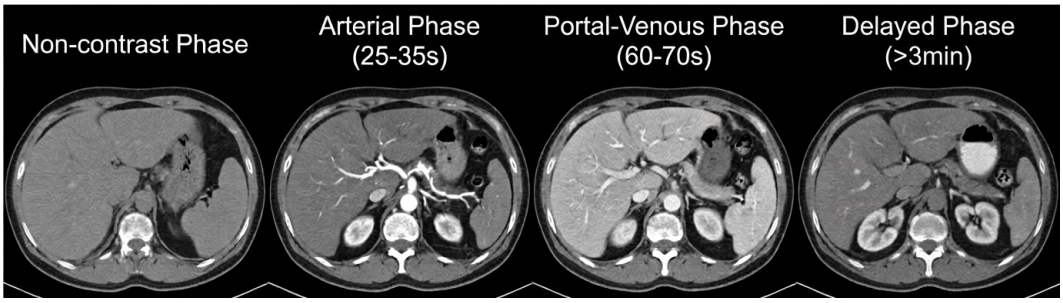

Figure 7: **Multi-phase contrast-enhanced CT scans.** In our pre-training and evaluation, we utilize multi-phase CT volumes encompassing non-contrast, arterial (25–35 s), portal-venous (60–70 s), and delayed (>3 min) phases. We demonstrate the diversity of hemodynamic enhancement patterns across these temporal windows; as illustrated, each phase provides unique voxel intensities for the same anatomical structures. This multi-temporal alignment enables the model to learn robust anatomy-normality representations and capture temporal-spectral dependencies that are typically absent in single-phase medical vision-language datasets.

B Datasets

In this section, we provide a detailed description of the datasets used in the pre-training, testing, and fine-tuning stages.

B.1 Training

To train our model, we curate a large-scale, multimodal, multi-center dataset comprising both public and proprietary sources. This collection includes 67,331 multi-phase CT scans across 53,643 exams from 37,989 patients, each paired with its corresponding radiology report. Detailed descriptions of the individual cohorts are provided below, and Table 5 summarizes the dataset distributions.

Table 5: **Distribution of training CT datasets.** We summarize the number of scans, exams, and unique patients across the three cohorts.

Dataset	Scans	Exams	Patients
Merlin [13]	25,494	25,494	18,317
CancerVerse [49]	24,422	22,344	14,599
Proprietary set	17,415	5,805	4,004
Total	67,331	53,643	37,989

B.1.1 Merlin

Merlin [13] consists of 25,494 publicly available[3] contrast-enhanced computed tomography (CECT) scans from 18,317 patients. Each CT exam is paired with the corresponding radiology report. The dataset includes abdominal and pelvic CT exams conducted between 2012 and 2018 at the Stanford Hospital Emergency Departments.

B.1.2 CancerVerse

CancerVerse [49] is a publicly available longitudinal and multimodal CT dataset for multicancer screening[4]. The current release contains 24,422 CT scans from 14,599 patients, acquired at hospitals in Switzerland and Turkey between 2010 and 2025. It covers the chest, abdomen, and pelvis and pairs CT imaging with radiology reports, complemented by pathology results and clinical variables. The dataset provides expert voxel-wise tumor annotations for 13 cancer types and includes a verified

[3]Merlin is publicly available at https://stanfordaimi.azurewebsites.net/datasets/60b9c7ff-877b-48ce-96c3-0194c8205c40

[4]CancerVerse is publicly available at https://huggingface.co/datasets/BodyMaps/CancerVerse

healthy cohort for evaluating false-positive detections. Patient histories span up to 14.4 years, with up to 26 examination dates per patient, supporting longitudinal analysis of imaging findings and their associated reports.

B.1.3 Proprietary set

We utilized a nationwide cohort from Poland comprising 17,415 multi-phase CT scans across 5,805 examinations from 4,004 patients. The dataset was collected from over 300 hospitals across the country, with each scan paired with its corresponding radiology report. Ethical approval for this study was granted by the Bioethics Committee at the Faculty of Medicine, Collegium Medicum, University of Warmia and Mazury in Olsztyn (Approval No. 9/2023, March 30, 2023).

B.2 Fine-tuning

B.2.1 AMOS-CT

The AMOS-CT dataset[5] [39] consists of a total of 500 CT scans collected from diverse clinical centers, ensuring significant variability in scanner types and patient demographics. However, the ground-truth labels are publicly accessible only for the training and validation cohorts, comprising 200 and 100 scans, respectively. Each volume provides high-quality, voxel-level manual annotations for 15 distinct abdominal organs: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus. For our experimental setup, we utilize the 200 training scans for model optimization and the 100 validation scans as our test set. The dataset is particularly challenging due to its large intra-class variation; for instance, the number of axial slices per scan ranges from 68 to 353, and the voxel spacing varies significantly from $[0.45, 1.07] \times [0.45, 1.07] \times [1.25, 5.0]$ mm^3. The mean spacing is $0.70 \times 0.70 \times 4.20$ mm^3. Beyond its scale, AMOS-CT is recognized for including both healthy and pathological cases, requiring models to be robust against anatomical distortions and varied fields of view typical in real-world clinical environments.

B.2.2 AMOS-MRI

AMOS-MRI is the MRI subset of the AMOS dataset [39], comprising 100 3D abdominal MRI scans with voxel-wise annotations of 15 anatomical structures[6]. The annotations cover major abdominal organs and vessels, including the liver, spleen, kidneys, pancreas, stomach, aorta, and inferior vena cava. Although the training annotations include 15 foreground anatomical classes, the AMOS-MRI test-set annotations do not include the urinary bladder or the prostate/uterus class. We therefore evaluate segmentation performance on the remaining 13 foreground classes. The benchmark includes 40 scans for training, 20 for validation, and 40 for testing, with the test set further divided into 11 in-distribution and 29 out-of-distribution scans. Its heterogeneous acquisition settings support the evaluation of segmentation accuracy and generalization across scanner domains.

B.2.3 PANORAMA

The PANORAMA dataset [3] comprises 2,238 multi-center contrast-enhanced computed tomography (CECT) scans acquired in a single portal-venous phase. It includes 1,964 newly acquired scans from five European centers, as well as publicly available data from two medical datasets in the United States, namely NIH [72] and MSD Pancreas [4]. The dataset consists of 676 PDAC (pancreatic ductal adenocarcinoma) cases and 1,562 non-PDAC cases. While the dataset contains six voxel-wise labels (PDAC lesion, veins, arteries, pancreatic parenchyma, pancreatic duct, and common bile duct), we use only the PDAC lesion and merge the pancreatic parenchyma and pancreatic duct classes into a single *pancreas* label. Each CT image consists of 37–1,572 slices, with a resolution ranging from 512×512 to 1024×1024 pixels. The voxel spacing ranges over $[0.31, 1.03] \times [0.31, 1.03] \times [0.45, 5.0]$ mm^3 with a mean of $0.75 \times 0.75 \times 2.0$ mm^3.

[5] AMOS-CT dataset is publicly available at https://zenodo.org/records/7262581
[6] AMOS-MRI is publicly available at https://zenodo.org/records/7262581

B.2.4 GEPAR3D

We utilize the GEPAR3D dataset[7] [80], which is derived from 98 publicly available cone-beam computed tomography (CBCT) scans originally released by [25]. While the initial iteration of these data provided only binary masks (distinguishing tooth structure from background), it presented significant limitations for tasks requiring individual tooth identification or pathological analysis. To address this, the GEPAR3D authors performed an extensive manual reannotation process [80]. Experienced clinicians labeled the dataset into 32 distinct anatomical classes following the Universal Numbering System, effectively transforming it into a benchmark for fine-grained multi-class dental segmentation. The dataset is characterized by its high anatomical complexity, featuring diverse dental conditions including restorations and implants that introduce metallic artifacts, varying stages of tooth decay and bone loss, and impacted teeth and irregular dental alignments. The GEPAR3D test set consists of 46 publicly available CBCT scans from four external centers as part of the GEPAR3D dataset [80]. This cohort includes two subsets sourced from public repositories: 15 scans from [25], which also constitute a subset of the ToothFairy2 dataset [14, 15, 54], and two newly released centers provided to the research community. The remaining scans originate from two independent clinical centers, ensuring a diverse evaluation of model generalization across different scanning protocols and anatomical variations.

B.2.5 MRISegmenter

We utilize the MRISegmenter abdominal MRI cohort[8] [98], introduced as part of a retrospective study of patients acquired at the U.S. National Institutes of Health (NIH) Clinical Center. The resource comprises 195 patients with 780 3D T1-weighted abdominal MRI series in total, corresponding to four acquisitions per patient: axial pre-contrast T1 and contrast-enhanced T1 in arterial, portal venous, and delayed phases. Each series is accompanied by dense, voxel-level reference segmentations of abdominal organs and anatomic structures, providing a fine-grained multi-class benchmark for volumetric abdominal MRI segmentation beyond coarse organ-only labels. The cohort is clinically heterogeneous across adult patients (training split: 135 subjects; internal test split: 60 subjects) and spans the full multi-phase dynamic-liver protocol, which introduces realistic variability in parenchymal enhancement, vascular conspicuity, and partial-volume effects across phases.

B.2.6 MVSeg

We utilize the MVSeg2023 benchmark[9] [19], released in conjunction with the MICCAI 2023 workshop on *Mitral Valve Segmentation from 3D Transesophageal Echocardiography*. The goal is automatic voxel-wise delineation of the mitral valve leaflets from *single-frame* three-dimensional transesophageal echocardiography (3D TEE) volumes acquired at end-diastole in a mid-esophageal, en-face mitral view. Volumes were collected on a Philips Epiq cardiac ultrasound system according to clinical standard of care at King's College Hospital (London, UK), converted from vendor DICOM to Cartesian coordinates with Philips QLab, and distributed as NIfTI files with spatial metadata limited to image size and spacing. Each case consists of one patient-specific 3D TEE volume with dense annotations of the anterior and posterior mitral leaflets (three labels: background, posterior leaflet, anterior leaflet), yielding a fine-grained yet low-cardinality segmentation task that is representative of intra-operative imaging noise, acoustic shadowing, and thin-structure ambiguity. The public release is partitioned into 105 training, 30 validation, and 40 test volumes for standardized comparison.

B.2.7 AbdomenAtlas 3.0

We utilize AbdomenAtlas 3.0[10] [11], a large-scale abdominal CT resource as part of the RadGPT pipeline. The cohort aggregates 9,262 isotropic three-dimensional CT volumes in NIfTI format (2,789,975 axial slices) drawn from 17 previously released public abdominal CT datasets that originally shipped without paired radiology reports. Each study is distributed as a *triplet* of (i) the

[7] GEPAR3D dataset is publicly available at https://zenodo.org/records/15739014

[8] The MRISegmenter segmentation tool, model weights, and accompanying dataset are publicly available at https://github.com/rsummers11/MRISegmenter under the data-use terms described in the original publication.

[9] The MVSeg2023 challenge data are catalogued on Hugging Face at https://huggingface.co/datasets/pcarnahan/MVSeg2023; the official challenge portal is https://www.synapse.org/MVSeg2023. Access may require accepting the host platform's data-use conditions.

[10] AbdomenAtlas 3.0 and the RadGPT release are publicly available at https://github.com/MrGiovanni/RadGPT.

CT volume, (ii) dense voxel-wise annotations of liver, kidney, and pancreatic tumors together with ancillary structures used for staging and localization—including Couinaud liver sub-segments (1–8), pancreatic sub-regions (head, body, tail), and peripancreatic vasculature relevant to pancreatic ductal adenocarcinom (PDAC) T1–T4 staging, and (iii) expert-reviewed radiology text. In total 3,955 scans contain at least one annotated tumor; the curators report a $4.2\times$ expansion of publicly available tumor masks relative to the union of the source releases, with 3,011 additional tumors identified after revision of automatic segmentations by board-certified radiologists. The textual modality comprises structured, narrative, and (where applicable) enhanced reports generated and verified under radiologist supervision with RadGPT, documenting per-lesion size, attenuation (HU), volume, anatomic location within sub-segments, and resectability-related attributes, for a total of 1,843,262 report tokens. AbdomenAtlas 3.0 is the first public distribution in this line that couples every scan with both high-resolution tumor (and staging-relevant) segmentations and paired, tumor-focused radiology text, making it a natural benchmark for 3D abdominal vision–language pre-training and segmentation-assisted report generation.

B.2.8 RATIC

We utilize the RSNA Abdominal Traumatic Injury CT (RATIC) cohort[11] [73], curated by the Radiological Society of North America (RSNA) in collaboration with the American Society of Emergency Radiology and the Society of Abdominal Radiology. The resource comprises 4,274 consecutive adult ($\geq$18 years) abdominal–pelvic CT studies, totaling 6,481 image series contributed from 23 institutions across 14 countries on six continents—at the time of its release, the largest publicly available multicenter collection explicitly labeled for blunt abdominal traumatic injury on CT. Each study ships in DICOM format together with comma-separated tabular metadata: patient demographics, series-level acquisition descriptors (imaging phase and anatomic coverage), and study-level injury labels for the liver, kidneys, spleen, bowel/mesentery, and active contrast extravasation. Solid-organ injuries are graded on the American Association for the Surgery of Trauma (AAST) organ-injury scale, with low-grade (grades I–III) versus high-grade (grades IV–V) injuries dichotomized to reduce inter-rater noise; bowel/mesenteric injury and extravasation receive complementary image-level annotations because of their heterogeneous appearance across phases. In addition, voxel-wise segmentations of the liver, spleen, left and right kidneys, and a composite gastrointestinal label (esophagus through colon) are publicly released in NIfTI format for a 206-series pilot subset of the training portion, enabling joint segmentation and injury-classification benchmarks. The dataset was used for the RSNA 2023 Abdominal Trauma Detection challenge and is intended to stress-test algorithms under realistic protocol heterogeneity (e.g. mono- versus multi-phase acquisition, variable slice thickness), reflecting the clinical variability encountered in emergency CT workflows.

[11] RATIC is distributed for non-commercial research through the RSNA AI Challenge platform at https://imaging.rsna.org/dataset/5 and the companion Kaggle release at https://www.kaggle.com/c/rsna-2023-abdominal-trauma-detection.

C Experimental setup

We provide a detailed description of our experimental setup, covering data preprocessing, model implementation specifics, and evaluation metrics.

C.1 Hardware and software setup

All experiments are conducted on a high-performance computing cluster utilizing 4 $\times$ NVIDIA GH200 GPUs, each with 96 GB of memory. All networks are implemented using Python 3.11, PyTorch 2.7 [65], and the medical imaging framework MONAI 1.3.0. Training stability and efficiency are ensured by employing the PyTorch Distributed Data-Parallel (DDP) framework. For DDP training, we use PyTorch Lightning Fabric.

C.2 Pre-training

We utilize TotalSegmentator [81] to segment anatomical structures from the 3D CT scans. For downstream analysis, we focus on the primary anatomical structures within the abdomen and pelvis, resulting in $C - 1 = 14$ foreground classes and $C = 15$ total classes (see Table 4 for the full list).

We resample all CT scans to $1.5 \times 1.5 \times 3.0$ mm^3 anisotropic resolution. All CT scans are normalized to a range of $[0, 1]$. We first apply a clipping operation to the raw Hounsfield Units (HU) to focus on clinically relevant intensities and remove artifacts. The normalized intensity I_{norm} is calculated as:

$$I_{\text{norm}} = \frac{\text{clamp}(I, HU_{\min}, HU_{\max}) - HU_{\min}}{HU_{\max} - HU_{\min}} \tag{19}$$

where the clipping bounds are set to $[-1000, 1000]$.

C.3 Pre-training objective and routing details

The implementation described here uses the anatomical routing representation defined in the main text. Each stage predicts raw crop composition. During training, label-derived or detached predicted composition is selected independently for each sample; for zero-based epoch e, the probability of predicted routing is $p_e = 0.5 + 0.5 \min(e/3, 1)$. At inference, routing uses predicted composition only. No segmentation mask or report is required.

Routing supervision. The routing objective sums the composition loss and router penalties over stages:

$$\mathcal{L}_{\text{MoE}} = \sum_{s=1}^{4} \left[\mathcal{L}_{\text{comp}}^{(s)} + 0.5\, \mathbb{E}_b\, \text{KL}\Big(\mathbf{W}_{s,b} \,\|\, \text{sg}\Big[\mathcal{T}(\mathbf{h}_b) \widetilde{\mathbf{A}}^{(s)} \Big] \Big) + \mathcal{R}_s \right]. \tag{20}$$

Here sg stops gradients through the alignment target. The usage regularizer $\mathcal{R}_s$ combines rolling anatomical-prior matching, a normalized per-sample entropy-band penalty with bounds $[0.25, 0.8]$, and foreground affinity coverage. Its outer weight is 0.1 and the relative coverage weight is 0.1. These terms encourage controlled mixture usage, without guaranteeing equal weights or organ-exclusive experts. The positive prior strength is parameterized as $\beta_s = \text{softplus}(b_s)$, and the log-prior uses $\epsilon = 10^{-6}$.

Organ–text supervision. The symmetric within-organ InfoNCE objective uses structured report findings. Exact duplicates after case folding and whitespace normalization form multiple positives. Negatives are different-case, different-text findings with opposite known normal/abnormal report labels; other pairs are masked. Each direction uses current-batch anchors and current plus queued candidates. Detached per-organ FIFO queues of length 64 are cleared at the start of each training epoch and are not used for validation. An anchor contributes only when it has both a positive and an eligible negative. Missing findings are excluded. Organ–text pairs require nonempty findings and visible organ support of at least 0.001 of the crop; no additional whole-organ coverage filter is applied in this recipe. Report labels are weak supervision and do not establish clinical incompatibility between every negative pair.

Cross-stage consistency. Organ-indexed embeddings are encouraged to agree across the four scales:

$$\mathcal{L}_{\text{stage}} = \frac{1}{6} \sum_{1 \le s < s' \le 4} \mathbb{E}\Big[1 - \cos(\hat{\mathbf{E}}_s, \hat{\mathbf{E}}_{s'})\Big]. \tag{21}$$

Query grounding. For each valid organ–stage target, its binary pseudo-mask is average-pooled to the attention grid and normalized to a distribution $\mathbf{m}_{s,b,c}$. With normalized query attention $\mathbf{a}_{s,b,c}$, we use

$$\mathcal{L}_{\text{ground}} = \frac{1}{|\mathcal{G}|} \sum_{(s,b,c) \in \mathcal{G}} \text{KL}(\mathbf{m}_{s,b,c} \| \mathbf{a}_{s,b,c}), \tag{22}$$

where $\mathcal{G}$ contains eligible targets with nonzero mask mass. The term is zero when no target is valid. Masks supervise attention during training but are not inputs to OQMSP's forward computation. This encourages organ-level attention rather than establishing lesion localization.

C.4 Downstream tasks

In this section, we provide a detailed description of our experimental setup for downstream tasks, including segmentation, classification, and detection.

C.4.1 Segmentation

Each segmentation model architecture is trained with an input patch size of $P \times P \times P$, where $P = 128$, to accommodate GPU memory constraints while preserving sufficient spatial context. We employ the sum of Dice and cross-entropy losses as our final loss function.

We utilize the AdamW optimizer with an initial learning rate η_0 set to 1e-4. The learning rate schedule follows cosine annealing with a minimum learning rate of η_{min}, which is set to 1e-6 and a warm-up period of W epochs. We train the model for a total of T epochs with a batch size of $B = 2$ per GPU, totaling an effective batch size of $B_{\text{eff}} = 8$. Weight decay is applied with a factor of $\lambda = 1e - 5$.

AMOS-CT. All CT scans are resampled to $1.5 \times 1.5 \times 1.5$ mm^3 isotropic resolution. Then, we set the clipping window to $[-1000, 1000]$ and linearly scale the intensities to $[0, 1]$. We train each model for 800 epochs.

PANORAMA. All CT scans are resampled to $0.8 \times 0.8 \times 3.0$ mm^3 anisotropic resolution. Then, we set the clipping window to $[-175, 250]$ and linearly scale the intensities to $[0, 1]$. We train each model for 800 epochs.

GEPAR3D. All CT scans are resampled to $0.4 \times 0.4 \times 0.4$ mm^3 isotropic resolution. Then, we set the clipping window to $[0, 5000]$ and linearly scale the intensities to $[0, 1]$. We train each model for 1,000 epochs.

AMOS-MRI. All MRI scans are resampled to $1.5 \times 1.5 \times 1.5$ mm^3 isotropic resolution. Then, we set the clipping window to $[0, 1000]$ and linearly scale the intensities to $[0, 1]$. We train each model for 1,000 epochs.

MRISegmenter. All MRI scans are resampled to $1.5 \times 1.5 \times 3.0$ mm^3 anisotropic resolution. Then, we set the clipping window to $[0, 1000]$ and linearly scale the intensities to $[0, 1]$. We train each model for 1,000 epochs.

MVSeg. All US scans are resampled to $0.5 \times 0.5 \times 0.5$ mm^3 isotropic resolution. Then, we set the clipping window to $[0, 255]$ and linearly scale the intensities to $[0, 1]$. We train each model for 1,000 epochs.

C.4.2 Classification

We evaluate abdominal trauma *classification* on RATIC [73]. Each CT scan is resampled to $1.5 \times 1.5 \times 3.0$ mm^3 anisotropic spacing. For every compared model we keep only the pre-trained 3D *encoder* (the segmentation decoder is discarded), attach a case-level head, and fine-tune the encoder and head on the RATIC training split. Volumes are processed with a sliding window of $P \times P \times P$ patches (P=128): we encode all valid patches, apply global average pooling over patch embeddings to obtain a fixed-length volume descriptor, and pass it through a lightweight MLP that outputs one

logit per evaluated binary or graded task (bowel, liver, kidney, and spleen; see Figure 8 and Table 24 and Table 25). Training minimizes the sum of sigmoid cross-entropy losses over the active heads on each sample.

C.4.3 Detection

We report pancreas, kidney, and liver tumor *detection* on AbdomenAtlas 3.0 [11] (dataset details in Appendix B) using *only* the corresponding segmentation-trained models: we do not train a separate detector. Study-level predictions are obtained from the *segmentation* decoder by aggregating voxel-wise tumor-class *logits* (e.g., thresholding the maximum tumor-class probability or the volume of connected high-confidence voxels). Sensitivity is stratified by lesion long-axis diameter ($\leq$2 cm vs. $>$2 cm); specificity is computed at the study level. Results are shown in Figure 9 and Table 26; formal definitions appear in Appendix C.5.2.

C.5 Evaluation metrics

C.5.1 Segmentation

For segmentation tasks, we use the Dice similarity coefficient (DSC) and the 95th-percentile Hausdorff distance (HD95) to quantitatively assess model performance. The DSC measures the spatial overlap between the predicted segmentation P and the ground truth G, and is defined as:

$$\mathrm{DSC}(P, G) = \frac{2|P \cap G|}{|P| + |G|}, \tag{23}$$

where $|P|$ and $|G|$ denote the number of voxels in the predicted and ground-truth masks, respectively.

To evaluate boundary accuracy, we report HD95, a robust variant of the Hausdorff distance that reduces sensitivity to outliers. Let ∂P and ∂G denote the contours of the predicted and ground-truth masks. The following definition takes the maximum of the two directed 95th-percentile contour distances:

$$\mathrm{HD95}(P, G) = \max\left\{\mathrm{Percentile}_{95}\left(d(p, G)\right),\ \mathrm{Percentile}_{95}(\{d(g, \partial P) : g \in \partial G\})\right\}, \tag{24}$$

where $d(p, \partial G) = \min_{g \in \partial G} \|p - g\|_2$ and $d(g, \partial P) = \min_{p \in \partial P} \|g - p\|_2$, using physical coordinates in millimeters. Higher DSC and lower HD95 indicate better segmentation performance.

C.5.2 Classification and detection

For classification and detection tasks, we evaluate model performance using sensitivity, specificity, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). Let TP, TN, FP, and FN denote the number of true-positive, true-negative, false-positive, and false-negative predictions, respectively. Sensitivity (also called recall or the true-positive rate) and specificity (the true-negative rate) are defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{25}$$

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{26}$$

To account for class imbalance, we report the F1-score, which is the harmonic mean of precision and sensitivity. Given precision $= \frac{TP}{TP+FP}$, the F1-score is defined as:

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN}. \tag{27}$$

The ROC-AUC value ranges from 0 to 1, where an AUC of 0.5 indicates performance no better than random chance, and an AUC of 1.0 represents a perfect classifier. Higher sensitivity, specificity, F1-score, and ROC-AUC correspond to superior classification results.

D Ablation studies

D.1 Effect of the key components

Table 6: Ablation study on the effect of key components on the AMOS-CT [39] validation set. B, MoE, and T denote the baseline, A^{2}MoE, and OQMSP modules, respectively. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). Organ abbreviations: Spl (spleen), RK/LK (right/left kidney), GB (gallbladder), Eso (esophagus), Liv (liver), Sto (stomach), Aor (aorta), IVC (inferior vena cava), Pan (pancreas), RAG/LAG (right/left adrenal gland), Duo (duodenum), Bla (bladder), and P/U (prostate/uterus). The best results are **bolded**.

	Config			Organ Classes (DSC %)															Overall	
#	B	MoE	T	Spl	RK	LK	GB	Eso	Liv	Sto	Aor	IVC	Pan	RAG	LAG	Duo	Bla	P/U	mDSC	mHD95
1	✓	✗	✗	95.0	95.5	95.4	75.5	80.7	96.2	88.2	93.5	88.9	82.7	75.1	73.6	75.6	86.0	80.1	85.5	17.3
2	✓	✓	✗	95.8	95.9	94.9	82.7	83.6	96.8	91.5	94.1	90.0	85.6	75.1	75.5	79.7	89.3	81.6	87.5	7.6
3	✓	✓	✓	**96.7**	**96.1**	**95.6**	**84.8**	**84.9**	**97.7**	**92.2**	**94.3**	**90.6**	**87.3**	**77.0**	**77.6**	**82.9**	**89.5**	**83.6**	**88.7**	**6.2**

Table 7: Ablation study on the effect of key components on the PANORAMA [2] test sets (NIH, MSD, and proprietary). B, MoE, and T denote the baseline, A^{2}MoE, and OQMSP modules, respectively. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). The best results are **bolded**.

	Config			NIH	MSD Pancreas		Proprietary		Overall	
#	B	MoE	T	Pancreas	PDAC	Pancreas	PDAC	Pancreas	mDSC	mHD95
1	✓	✗	✗	92.7	47.9	88.1	54.0	81.9	76.0	8.6
2	✓	✓	✗	92.0	53.6	87.7	65.8	81.4	78.2	7.1
3	✓	✓	✓	**93.1**	**55.2**	**89.1**	**67.2**	**82.2**	**79.1**	**5.8**

Table 8: Ablation study on the effect of key components on the AMOS-MRI [39] validation set. B, MoE, and T denote the baseline, A^{2}MoE, and OQMSP modules, respectively. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). Organ abbreviations: Spl (spleen), RK/LK (right/left kidney), GB (gallbladder), Eso (esophagus), Liv (liver), Sto (stomach), Aor (aorta), IVC (inferior vena cava), Pan (pancreas), RAG/LAG (right/left adrenal gland), and Duo (duodenum). Evaluation covers 13 foreground classes; urinary bladder and prostate/uterus are excluded because their annotations are unavailable in the test set. The best results are **bolded**.

	Config			Organ Classes (DSC %)													Overall	
#	B	MoE	T	Spl	RK	LK	GB	Eso	Liv	Sto	Aor	IVC	Pan	RAG	LAG	Duo	mDSC	mHD95
1	✓	✗	✗	86.4	89.8	89.6	62.5	70.5	95.7	82.4	87.0	81.0	83.5	59.8	64.2	64.4	76.7	9.3
2	✓	✓	✗	91.0	90.6	90.7	67.3	72.3	96.5	81.4	87.5	83.4	80.6	57.4	59.5	62.8	78.5	12.2
3	✓	✓	✓	**91.5**	**90.9**	**91.2**	**78.9**	**75.0**	**97.4**	**86.8**	**87.6**	**84.4**	**86.5**	**64.5**	**67.0**	**68.1**	**82.3**	**8.9**

Table 9: Ablation study on the effect of key components on the GEPAR3D [80] test set. B, MoE, and T denote the baseline, A^{2}MoE, and OQMSP modules, respectively. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). FDI notation is used for individual teeth across quadrants (Q1–Q4). The best results are **bolded**.

	Config			Upper Right (Q1)								Upper Left (Q2)								Lower Left (Q3)								Lower Right (Q4)								Overall	
#	B	MoE	T	18	17	16	15	14	13	12	11	21	22	23	24	25	26	27	28	38	37	36	35	34	33	32	31	41	42	43	44	45	46	47	48	mDSC	mHD95
1	✓	✗	✗	68.0	87.9	90.2	85.3	85.2	91.7	**91.0**	**93.7**	93.1	90.9	89.2	86.1	90.0	93.0	85.3	70.6	66.9	78.3	84.5	89.4	89.4	85.4	87.2	**86.0**	80.3	**89.1**	**92.6**	89.1	93.3	84.8	81.4	56.6	85.3	10.6
2	✓	✓	✗	67.9	87.6	**90.8**	88.0	85.6	91.4	89.3	93.0	**93.3**	**91.7**	**89.4**	84.0	**92.0**	92.4	**91.4**	74.0	**68.1**	**88.2**	**88.6**	**91.4**	88.6	84.7	85.4	80.7	83.5	88.9	91.4	88.6	92.2	88.7	83.2	59.8	86.2	8.3
3	✓	✓	✓	**70.4**	**92.2**	89.1	**88.3**	**91.0**	**93.5**	90.4	93.6	91.5	**91.7**	88.7	**87.4**	91.3	**94.2**	90.7	**76.7**	64.3	87.1	87.5	89.8	**91.1**	**86.5**	**89.5**	**86.0**	**84.5**	88.5	87.6	**91.2**	**93.8**	**90.0**	**88.3**	**64.3**	**87.7**	**8.1**

Table 10: Ablation study on the effect of key components on the MRISegmenter [98] test set. B, MoE, and T denote the baseline, A^{2}MoE, and OQMSP modules, respectively. We evaluate segmentation performance across all 62 anatomical classes. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). The best results are **bolded**.

	Config			Anatomical Classes (DSC %) ↑																																	
#	B	MoE	T	1 Spleen	2 Kid R	3 Kid L	4 Gallb	5 Liver	6 Esoph	7 Stom	8 Aorta	9 IVC	10 P&S Vn	11 Pancr	12 Adr R	13 Adr L	14 Lung R	15 Lung L	16 Sm Bow	17 Duod	18 Colon	19 Ilia Art L	20 Ilia Art R	21 Ilia Vn L	22 Ilia Vn R	23 GluMax L	24 GluMax R	25 GluMed L	26 GluMed R	27 Auto L	28 Auto R	29 Ilio L	30 Ilio R	31 Hip L
1	✓	✗	✗	90.7	96.4	86.6	73.6	98.4	84.7	92.6	95.2	91.9	61.7	87.1	60.2	58.0	97.1	96.1	91.0	82.6	89.8	75.9	80.2	67.1	59.7	63.5	60.4	84.6	78.9	97.3	97.1	95.3	95.8	80.5
2	✓	✓	✗	91.5	96.3	91.5	78.6	98.5	85.4	93.6	**95.5**	93.1	**64.0**	88.1	61.0	59.2	97.5	96.6	92.2	84.2	91.6	76.8	77.1	66.5	58.2	70.0	59.1	86.8	79.2	**97.5**	**97.4**	95.5	96.4	83.4
3	✓	✓	✓	**92.8**	**96.7**	**93.0**	**83.1**	**98.6**	**85.6**	**94.4**	95.4	**93.2**	63.7	**89.3**	**61.3**	**60.1**	**97.6**	**96.7**	**93.9**	**84.7**	**93.7**	**81.6**	**83.9**	**78.4**	**78.1**	**79.4**	**78.5**	**90.7**	**87.8**	**97.5**	**97.4**	**95.8**	**96.5**	**86.6**

	Config			Anatomical Classes (DSC %)																																		Overall	
#	B	MoE	T	32 Hip R	33 Sacr	34 Rib L4	35 Rib L5	36 Rib L6	37 Rib L7	38 Rib L8	39 Rib L9	40 Rib L10	41 Rib L11	42 Rib L12	43 Rib R4	44 Rib R5	45 Rib R6	46 Rib R7	47 Rib R8	48 Rib R9	49 Rib R10	50 Rib R11	51 Rib R12	52 L5	53 L4	54 L3	55 L2	56 L1	57 T12	58 T11	59 T10	60 T9	61 T8	62 T7	mDSC	mHD95
1	✓	✗	✗	69.6	49.7	15.9	38.8	48.3	53.8	57.1	60.5	69.0	70.0	66.7	18.0	36.0	50.0	60.6	64.6	67.0	69.4	68.1	64.7	66.4	85.6	87.0	87.1	87.3	88.6	88.2	84.9	78.7	69.4	48.9	74.5	6.8
2	✓	✓	✗	67.5	52.1	18.0	44.1	53.7	60.2	**64.2**	**69.5**	**73.1**	**74.0**	**69.7**	19.4	47.1	56.7	65.2	67.0	68.1	**72.0**	70.6	**67.5**	76.1	87.7	88.4	88.1	87.7	88.8	**89.2**	87.7	80.5	70.2	46.1	77.6	5.9
3	✓	✓	✓	**80.6**	**70.3**	**19.1**	**49.3**	**58.5**	**61.7**	63.8	67.0	71.0	**74.0**	67.9	**21.5**	**50.5**	**60.2**	**68.4**	**70.3**	**71.5**	71.3	**71.8**	67.4	**79.7**	**89.8**	**89.9**	**89.6**	**88.8**	**88.9**	88.3	**87.9**	**80.9**	**77.9**	**58.9**	**79.8**	**5.5**

Table 11: Ablation study on the effect of key components on the MVSeg [19] validation set. B, MoE, and T denote the baseline, A^{2}MoE, and OQMSP modules, respectively. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). Organ abbreviations: PL (posterior leaflet) and AL (anterior leaflet). The best results are **bolded**.

	Config			Organ Classes (DSC %)		Overall	
#	B	MoE	T	PL	AL	mDSC	mHD95
1	✓	✗	✗	83.0	84.7	83.8	5.8
2	✓	✓	✗	84.2	85.6	84.9	4.3
3	✓	✓	✓	**84.5**	**86.0**	**85.4**	**4.0**

D.2 Effect of the number of Experts per stage

Table 12: Ablation study on the effect of the number of experts per encoder stage on the AMOS-CT [39] validation set. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). Organ abbreviations: Spl (spleen), RK/LK (right/left kidney), GB (gallbladder), Eso (esophagus), Liv (liver), Sto (stomach), Aor (aorta), IVC (inferior vena cava), Pan (pancreas), RAG/LAG (right/left adrenal gland), Duo (duodenum), Bla (bladder), and P/U (prostate in males or uterus in females). The best results are **bolded**.

		Organ Classes (DSC %)															Overall	
#	Experts	Spl	RK	LK	GB	Eso	Liv	Sto	Aor	IVC	Pan	RAG	LAG	Duo	Bla	P/U	mDSC	mHD95
1	[4, 4, 4, 4]	95.6	94.4	**95.3**	**83.9**	83.5	**96.9**	91.0	94.1	89.9	**85.9**	74.2	75.7	79.6	89.2	**82.9**	**87.5**	8.4
2	[8, 8, 8, 8]	**95.8**	95.6	93.2	80.7	**83.6**	96.6	**91.8**	**94.2**	**90.0**	85.8	**75.1**	**76.2**	**79.9**	89.1	82.2	87.3	8.7
3	[4, 8, 16, 32]	**95.8**	**95.9**	94.9	82.7	**83.6**	96.8	91.5	94.1	**90.0**	85.6	**75.1**	75.5	79.7	**89.3**	81.6	**87.5**	**7.6**
4	[16, 16, 16, 16]	95.7	95.4	92.9	80.5	83.1	95.3	91.3	94.1	89.8	84.7	74.9	76.0	79.7	88.3	81.9	86.9	9.2

Table 13: Ablation study on the effect of the number of experts per encoder stage on the PANORAMA [2] test sets (NIH, MSD, and proprietary). Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). The best results are **bolded**.

		NIH	MSD Pancreas		Proprietary		Overall	
#	Experts	Pancreas	PDAC	Pancreas	PDAC	Pancreas	mDSC	mHD95
1	[4, 4, 4, 4]	**92.4**	50.7	**88.9**	63.7	**81.8**	77.9	7.6
2	[8, 8, 8, 8]	92.0	50.4	85.4	63.3	79.5	76.1	8.5
3	[4, 8, 16, 32]	92.3	51.4	88.1	60.5	81.8	77.4	8.2
4	[16, 16, 16, 16]	92.0	**53.6**	87.7	**65.8**	81.4	**78.2**	**7.1**

Table 14: Ablation study on the effect of the number of experts per encoder stage on the AMOS-MRI validation set. Evaluation covers 13 foreground classes; urinary bladder and prostate/uterus are excluded because their annotations are unavailable in the test set. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). The best results are **bolded**.

		Organ Classes (DSC %)													Overall	
#	Experts	Spl	RK	LK	GB	Eso	Liv	Sto	Aor	IVC	Pan	RAG	LAG	Duo	mDSC	mHD95
1	[4, 4, 4, 4]	90.5	90.3	90.5	**68.4**	**72.3**	**96.7**	80.9	87.0	83.1	80.3	**58.5**	55.8	60.2	78.0	13.4
2	[8, 8, 8, 8]	90.6	**90.7**	90.5	67.3	70.8	96.3	**81.5**	86.5	82.7	79.2	**58.5**	56.5	61.9	77.9	16.9
3	[4, 8, 16, 32]	90.4	**90.7**	90.5	63.3	70.9	96.4	79.6	87.2	**83.5**	**81.0**	57.7	53.6	61.6	77.4	**12.2**
4	[16, 16, 16, 16]	**91.0**	90.6	**90.7**	67.3	**72.3**	96.5	81.4	**87.5**	83.4	80.6	57.4	**59.5**	**62.8**	**78.5**	**12.2**

Table 15: Ablation study on the effect of the number of experts per encoder stage on the GEPAR3D [80] test set. FDI notation is used for individual teeth across quadrants (Q1–Q4). Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). The best results are **bolded**.

		Upper Right (Q1)								Upper Left (Q2)								Lower Left (Q3)								Lower Right (Q4)								Overall	
#	Experts	18	17	16	15	14	13	12	11	21	22	23	24	25	26	27	28	38	37	36	35	34	33	32	31	41	42	43	44	45	46	47	48	mDSC	mHD95
1	[4, 4, 4, 4]	**68.1**	89.6	88.8	86.6	84.7	91.3	90.5	**93.8**	91.4	88.5	88.7	85.5	**92.0**	**94.4**	89.4	73.1	67.6	86.0	85.5	91.0	**91.3**	86.1	**87.1**	81.7	80.8	**89.4**	91.5	88.3	91.3	85.4	**83.9**	58.0	85.9	**7.7**
2	[8, 8, 8, 8]	60.6	88.1	90.7	87.2	**88.1**	88.8	89.9	90.3	89.0	91.4	88.5	83.7	90.4	92.7	88.2	65.0	63.4	82.1	84.4	90.5	91.0	**86.1**	86.7	**86.1**	**84.3**	86.3	90.8	88.8	92.3	87.7	82.1	59.1	85.0	9.4
3	[4, 8, 16, 32]	67.4	**90.8**	86.9	86.6	85.5	**91.7**	**91.0**	93.7	93.1	90.9	89.2	**86.1**	90.0	93.0	**92.2**	73.0	**71.0**	86.5	85.9	89.8	89.0	85.1	86.4	83.1	81.0	88.8	**92.6**	**89.4**	91.4	**88.9**	83.0	**60.9**	86.1	7.9
4	[16, 16, 16, 16]	67.9	87.6	**90.8**	**88.0**	85.6	91.4	89.3	93.0	**93.3**	**91.7**	**89.4**	84.0	**92.0**	92.4	91.4	**74.0**	68.1	**88.2**	**88.6**	**91.4**	88.6	84.7	85.4	80.7	83.5	88.9	91.4	88.6	**92.2**	88.7	83.2	59.8	**86.2**	8.3

Table 16: Ablation study on the effect of the number of experts per encoder stage on the MRISegmenter [98] test set. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm) across all 62 anatomical classes. Labels are sorted strictly by ID. The best results are **bolded**.

Experts	1 Spleen	2 Kid R	3 Kid L	4 Gallb	5 Liver	6 Esoph	7 Stom	8 Aorta	9 IVC	10 P&S Vn	11 Pancr	12 Adr R	13 Adr L	14 Lung R	15 Lung L	16 Sm Bow	17 Duod	18 Colon	19 Ilia Art L	20 Ilia Art R	21 Ilia Vn L	22 Ilia Vn R	23 GluMax L	24 GluMax R	25 GluMed L	26 GluMed R	27 Auto L	28 Auto R	29 Ilio L	30 Ilio R	31 Hip L
[4, 4, 4, 4]	91.1	**96.5**	89.1	79.0	**98.5**	85.3	93.5	**95.5**	**93.1**	**65.5**	87.5	60.6	59.6	**97.5**	96.5	**92.5**	84.0	91.7	**77.4**	**79.1**	65.5	58.0	63.8	**71.6**	85.2	79.9	**97.6**	**97.4**	**95.6**	**96.4**	80.1
[8, 8, 8, 8]	**93.4**	96.4	89.4	78.0	**98.5**	85.3	93.4	95.4	92.9	64.0	**88.2**	60.8	**59.7**	97.4	96.4	92.4	**84.4**	**91.8**	76.6	77.1	65.5	56.6	72.3	56.9	84.2	**80.7**	97.4	**97.4**	**95.6**	96.3	80.9
[4, 8, 16, 32]	91.5	96.3	**91.5**	78.6	**98.5**	**85.4**	93.6	**95.5**	**93.1**	64.0	88.1	**61.0**	59.2	**97.5**	**96.6**	92.2	84.2	91.6	76.8	77.1	**66.5**	**58.2**	70.0	59.1	**86.8**	79.2	97.5	**97.4**	95.5	**96.4**	**83.4**
[16, 16, 16, 16]	91.4	**96.5**	88.7	**80.0**	**98.5**	85.3	**93.7**	**95.5**	93.0	64.5	87.9	60.9	58.0	**97.5**	96.5	92.2	83.7	91.7	76.7	77.2	60.8	56.1	**72.7**	64.5	84.9	80.3	97.5	**97.4**	**95.6**	**96.4**	83.0

Experts	32 Hip R	33 Sacr	34 Rib L4	35 Rib L5	36 Rib L6	37 Rib L7	38 Rib L8	39 Rib L9	40 Rib L10	41 Rib L11	42 Rib L12	43 Rib R4	44 Rib R5	45 Rib R6	46 Rib R7	47 Rib R8	48 Rib R9	49 Rib R10	50 Rib R11	51 Rib R12	52 L5	53 L4	54 L3	55 L2	56 L1	57 T12	58 T11	59 T10	60 T9	61 T8	62 T7	mDSC	mHD95
[4, 4, 4, 4]	70.7	54.4	15.4	**45.7**	**56.2**	**62.3**	**65.6**	67.8	72.3	**75.5**	69.5	**22.0**	42.4	57.6	65.3	**68.7**	70.6	70.0	**71.1**	67.4	73.7	87.6	88.1	88.0	87.7	88.8	88.7	87.9	80.0	69.1	43.7	**77.7**	5.8
[8, 8, 8, 8]	67.2	**57.5**	13.7	43.0	55.7	61.9	64.2	68.7	72.0	72.8	69.2	19.1	44.9	56.6	65.4	68.2	**72.0**	**73.2**	70.5	**67.7**	74.9	86.7	87.1	87.3	87.8	88.5	**89.2**	**88.4**	79.6	73.0	45.2	77.5	**5.6**
[4, 8, 16, 32]	67.5	52.1	**18.0**	44.1	53.7	60.2	64.2	**69.5**	**73.1**	74.0	**69.7**	19.4	**47.1**	56.7	65.2	67.0	68.1	72.0	70.6	67.5	**76.1**	**87.7**	88.4	88.1	87.7	88.8	**89.2**	87.7	**80.5**	70.2	**46.1**	77.6	5.9
[16, 16, 16, 16]	**70.8**	55.2	13.7	45.1	53.2	59.9	65.0	69.0	72.3	73.3	69.0	21.0	42.9	**57.7**	**65.5**	68.3	68.8	71.3	**71.1**	66.2	72.7	87.4	**88.6**	**89.1**	**89.2**	**89.8**	88.1	87.2	79.8	**73.2**	45.0	77.4	5.7

Table 17: Ablation study on the effect of the number of experts per encoder stage on the MVSeg [19] test set. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). Organ abbreviations: PL (posterior leaflet) and AL (anterior leaflet). The best results are **bolded**.

		Organ Classes (DSC %) ↑		Overall	
#	Experts	PL	AL	mDSC	mHD95
1	[4, 4, 4, 4]	84.1	85.5	84.8	4.8
2	[8, 8, 8, 8]	**84.2**	**85.7**	**85.0**	**4.2**
3	[4, 8, 16, 32]	83.4	85.5	84.5	4.4
4	[16, 16, 16, 16]	**84.2**	85.6	84.9	4.3

E Quantitative results

E.1 Segmentation

Table 18: Quantitative segmentation performance on the AMOS-CT [39] validation set. Evaluation metrics include mDSC (%) and mHD95 (mm). Organ abbreviations: Spl (spleen), RK/LK (right/left kidney), GB (gallbladder), Eso (esophagus), Liv (liver), Sto (stomach), Aor (aorta), IVC (inferior vena cava), Pan (pancreas), RAG/LAG (right/left adrenal gland), Duo (duodenum), Bla (bladder), and P/U. The best results are **bolded**. † indicates pre-trained models.

	Organ Classes (DSC %) ↑															Overall	
Method	Spl	RK	LK	GB	Eso	Liv	Sto	Aor	IVC	Pan	RAG	LAG	Duo	Bla	P/U	**mDSC**	**mHD95**
VoCo† [84, 85]	96.2	94.3	95.4	81.6	83.2	96.0	91.1	93.9	89.7	85.5	76.0	75.7	79.8	88.0	79.4	87.1	16.4
uC 3DU-Net [37]	94.8	93.1	94.7	70.1	79.6	89.2	86.7	93.0	86.9	77.8	73.1	73.3	72.6	84.5	76.6	83.1	16.0
SuPreM† [50]	96.3	95.7	95.8	81.2	83.5	97.3	90.9	93.6	89.4	85.5	74.4	75.2	79.6	88.4	80.0	87.1	7.7
SegMamba [88]	95.0	95.5	95.4	75.5	80.7	96.2	88.2	93.5	88.9	82.7	75.1	73.6	75.6	86.0	80.1	85.5	17.3
Mamba-HoME [67]	96.0	95.0	94.4	81.7	82.0	94.7	90.2	93.8	88.9	84.0	74.4	74.1	78.4	83.8	83.2	86.3	18.7
Merlin† [13]	94.9	95.3	95.1	79.5	80.0	96.7	90.4	93.4	88.4	84.0	74.1	71.6	77.2	84.9	79.2	85.6	29.8
fVLM† [76]	95.4	94.8	95.4	75.2	81.4	94.7	88.0	93.1	88.0	83.1	74.6	73.9	76.5	84.2	80.4	85.2	16.4
BTB3D† [33]	96.0	95.2	93.8	80.2	81.8	97.1	90.0	93.2	87.2	85.0	73.6	73.4	78.8	87.3	79.2	86.1	18.3
A^{2}MoE†	**96.7**	**96.1**	**95.6**	**84.8**	**84.9**	**97.7**	**92.2**	**94.3**	**90.6**	**87.3**	**77.0**	**77.6**	**82.9**	**89.5**	**83.6**	**88.7**	**6.2**

Table 19: Quantitative segmentation and detection performance on the PANORAMA [2] test sets (NIH, MSD, and proprietary). Evaluation metrics include mDSC (%) and mHD95 (mm). The best results are **bolded**. † indicates pre-trained models.

	NIH	MSD Pancreas		Proprietary		Overall	
Method	**Pancreas**	**PDAC**	**Pancreas**	**PDAC**	**Pancreas**	**mDSC**	**mHD95**
VoCo† [84, 85]	90.1	38.7	86.5	43.3	80.3	71.6	15.9
SegMamba [88]	92.7	47.9	88.1	54.0	81.9	76.0	8.6
uC 3DU-Net [37]	92.6	49.0	88.9	56.8	80.2	76.6	**4.7**
SuPreM† [50]	92.3	46.0	88.6	61.4	82.0	76.6	**4.7**
Mamba-HoME [67]	92.9	51.2	88.4	60.8	81.7	77.5	4.8
Merlin† [13]	92.5	51.0	88.1	62.7	81.2	77.5	9.4
fVLM† [76]	91.5	44.1	88.0	58.4	81.0	76.8	9.8
BTB3D† [33]	92.0	52.1	88.4	61.5	81.4	77.3	9.2
A^{2}MoE†	**93.1**	**55.2**	**89.1**	**67.2**	**82.2**	**79.1**	5.8

Table 20: Quantitative segmentation performance on the AMOS-MRI [39] validation set. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). Organ abbreviations: Spl (spleen), RK/LK (right/left kidney), GB (gallbladder), Eso (esophagus), Liv (liver), Sto (stomach), Aor (aorta), IVC (inferior vena cava), Pan (pancreas), RAG/LAG (right/left adrenal gland), and Duo (duodenum). Evaluation covers 13 foreground classes; urinary bladder and prostate/uterus are excluded because their annotations are unavailable in the test set. The best results are **bolded**. † indicates pre-trained models.

	Organ Classes (DSC %) ↑													Overall	
Method	Spl	RK	LK	GB	Eso	Liv	Sto	Aor	IVC	Pan	RAG	LAG	Duo	**mDSC**	**mHD95**
VoCo† [84, 85]	91.2	90.5	91.1	71.2	74.9	97.2	84.7	87.1	83.9	82.9	62.0	59.5	65.0	80.1	10.5
uC 3DU-Net [37]	88.7	89.7	88.7	52.0	70.2	95.7	76.6	85.4	81.2	76.4	56.3	52.2	53.8	74.4	16.9
SuPreM† [50]	89.9	89.8	89.2	67.7	67.3	95.6	79.5	85.1	79.4	79.8	55.2	54.1	58.0	75.8	19.5
SegMamba [88]	86.4	89.8	89.6	62.5	70.5	95.7	82.4	87.0	81.0	83.5	59.8	64.2	64.4	76.7	9.3
Mamba-HoME [67]	87.4	88.9	88.3	69.8	71.7	95.9	82.7	85.3	82.3	83.4	59.0	64.2	63.5	77.7	9.9
Merlin† [13]	90.2	89.5	90.1	70.2	73.9	96.2	83.7	86.1	82.9	81.9	61.0	58.5	64.0	79.1	10.4
fVLM† [76]	89.2	88.5	89.1	69.2	72.9	95.2	82.7	85.1	81.9	80.9	60.0	57.5	63.0	78.1	10.9
BTB3D† [33]	89.0	88.3	88.9	69.0	72.7	95.0	82.5	84.9	81.7	80.7	59.8	57.3	62.8	77.8	11.1
A^{2}MoE	**91.5**	**90.9**	**91.2**	**78.9**	**75.0**	**97.4**	**86.8**	**87.6**	**84.4**	**86.5**	**64.5**	**67.0**	**68.1**	**82.3**	**8.9**

Table 21: Quantitative evaluation of tooth segmentation on the GEPAR3D [80] test set. FDI notation is used for individual teeth across quadrants (Q1–Q4). Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). The best results are **bolded**. † indicates pre-trained models.

	Upper Right (Q1)								Upper Left (Q2)								Lower Left (Q3)								Lower Right (Q4)								Overall	
Method	18	17	16	15	14	13	12	11	21	22	23	24	25	26	27	28	38	37	36	35	34	33	32	31	41	42	43	44	45	46	47	48	**mDSC**	**mHD95**
VoCo† [84, 85]	63.9	89.6	88.4	86.8	88.6	92.7	89.4	92.3	90.0	86.7	86.3	86.3	89.8	91.8	89.2	65.4	64.3	86.6	85.9	87.4	85.5	78.5	81.2	74.7	77.8	85.8	86.4	84.5	89.1	84.6	83.5	56.2	83.5	6.8
uC 3DU-Net [37]	67.7	89.5	87.7	84.6	83.0	87.3	88.5	92.3	90.7	87.9	87.7	86.5	88.8	91.0	88.2	66.4	64.9	79.0	82.8	86.0	85.9	83.8	87.5	82.8	82.6	89.5	89.0	84.5	87.5	87.1	82.4	53.6	83.8	12.5
SuPreM† [50]	62.7	85.9	87.0	82.1	75.1	81.9	86.8	84.6	87.0	86.1	86.1	83.0	88.4	89.5	85.3	66.8	61.4	79.5	83.7	83.2	76.7	81.0	81.5	76.3	75.8	81.6	88.0	78.6	83.8	85.8	79.2	58.1	80.6	11.0
SegMamba [88]	68.0	87.9	90.2	85.3	85.2	91.7	91.0	93.7	**93.1**	90.9	89.2	86.1	90.0	93.0	85.3	70.6	66.9	78.3	84.5	89.4	89.4	85.4	87.2	86.0	80.3	89.1	92.6	89.1	93.3	84.8	81.4	56.6	85.3	10.6
Mamba-HoME [67]	69.7	90.6	**91.2**	86.5	84.2	**93.9**	91.1	**93.9**	91.4	**92.0**	89.9	86.8	90.9	94.4	90.1	69.6	66.8	84.0	83.5	88.9	87.4	85.8	85.9	81.7	84.4	**89.4**	92.3	89.8	93.5	85.4	84.2	60.0	86.1	6.5
Merlin† [13]	68.2	88.8	89.5	86.0	82.9	91.8	90.9	93.2	89.1	87.2	**90.7**	86.6	**91.6**	94.1	**90.8**	**76.7**	**69.6**	**89.9**	87.5	**90.9**	89.0	**86.7**	85.2	84.3	80.7	87.5	**92.9**	89.7	92.0	89.3	86.9	61.9	86.5	**2.7**
fVLM† [76]	**71.2**	88.2	87.0	86.5	85.7	88.1	87.6	92.6	90.4	86.0	85.3	85.1	90.4	93.8	89.2	**76.7**	68.1	84.1	**88.1**	88.0	85.6	81.5	83.1	76.2	76.6	82.8	89.0	86.8	91.8	83.5	81.1	59.5	84.1	10.1
BTB3D† [33]	71.0	88.0	86.8	86.3	85.5	87.9	87.4	92.4	90.2	85.8	85.1	84.9	90.2	93.6	89.0	76.5	67.9	83.9	87.9	87.8	85.4	81.3	82.9	76.0	76.4	82.5	88.7	86.5	91.5	83.2	80.8	59.2	84.0	9.8
A^{2}MoE†	70.4	**92.2**	89.1	**88.3**	**91.0**	93.5	**90.4**	93.6	91.5	91.7	88.7	**87.4**	91.3	**94.2**	90.7	**76.7**	64.3	87.1	87.5	89.8	**91.1**	86.5	**89.5**	**86.0**	**84.5**	88.5	87.6	**91.2**	**93.8**	**90.0**	**88.3**	**64.3**	**87.7**	8.1

Table 22: Quantitative evaluation of MRISegmenter [98] across all 62 anatomical classes. Labels are sorted strictly by ID. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). The best results are **bolded**. † indicates pre-trained models.

Method	1 Spleen	2 Kid R	3 Kid L	4 Gallb	5 Liver	6 Esoph	7 Stom	8 Aorta	9 IVC	10 P&S Va	11 Pancr	12 Adr R	13 Adr L	14 Lung R	15 Lung L	16 Sm Bow	17 Duod	18 Colon	19 Ilia Art L	20 Ilia Art R	21 Ilia Vn L	22 Ilia Vn R	23 GluMax L	24 GluMax R	25 GluMed L	26 GluMed R	27 Auto L	28 Auto R	29 Ilio L	30 Ilio R	31 Hip L
VoCo [84]	91.0	96.3	88.9	74.4	98.4	84.2	93.0	95.2	92.1	62.0	86.2	59.7	55.1	97.2	96.0	90.9	81.9	89.6	74.9	75.0	59.3	50.8	63.6	57.1	83.6	76.9	97.2	97.1	95.2	95.8	73.7
SuPreM [50]	90.2	96.1	85.7	70.9	98.2	83.1	91.1	95.1	91.5	61.6	85.4	59.1	54.5	97.1	95.7	88.6	80.3	86.8	69.7	72.3	55.2	49.4	50.0	43.2	83.1	75.4	97.0	96.8	94.6	95.6	72.4
uC 3DU-Net [37]	91.0	96.2	87.3	**79.2**	**98.5**	**85.4**	92.9	95.3	92.8	**67.0**	87.9	59.5	57.1	**97.5**	**96.6**	92.0	82.2	91.3	76.1	79.5	66.0	57.5	66.8	58.5	84.5	79.0	**97.5**	97.3	**95.6**	**96.4**	**84.3**
SegMamba [88]	90.7	**96.4**	86.6	73.6	98.4	84.7	92.6	95.2	91.9	61.7	87.1	60.2	58.0	97.1	96.1	91.0	82.6	89.8	75.9	**80.2**	67.1	59.7	63.5	60.4	84.6	78.9	97.3	97.1	95.3	95.8	80.5
Mamba-HoME [67]	91.3	96.3	90.3	76.9	**98.5**	84.7	93.4	95.1	92.3	65.4	88.2	60.2	58.1	97.4	96.5	92.5	83.2	91.7	77.3	79.8	74.3	72.6	74.2	72.8	87.2	84.0	97.3	97.2	95.4	95.9	82.6
Merlin [13]	**91.5**	96.3	**91.7**	78.1	**98.5**	84.7	**93.7**	95.1	92.5	66.8	**88.6**	60.3	58.3	**97.5**	**96.6**	**93.0**	83.4	**92.4**	**77.8**	79.7	**77.0**	**77.4**	**78.1**	**77.3**	**88.1**	**85.9**	97.3	97.2	95.5	96.0	83.4
fVLM [76]	88.3	93.3	86.6	74.7	93.9	80.6	90.2	91.9	89.3	59.5	84.3	56.4	53.9	94.0	93.3	89.1	79.9	88.0	74.5	77.3	70.8	73.7	69.7	69.9	84.5	82.2	94.0	94.0	92.2	93.0	78.6
BTB3D [33]	91.2	96.2	89.5	77.6	98.3	83.5	93.1	94.8	92.2	62.4	87.2	59.3	56.8	97.1	96.2	92.0	82.8	90.9	77.4	**80.2**	73.7	76.6	72.6	72.8	87.4	85.1	97.1	97.1	95.1	95.9	81.5
A^{2}MoE	**91.5**	96.3	91.5	78.6	**98.5**	**85.4**	93.6	**95.5**	**93.1**	64.0	88.1	**61.0**	**59.2**	**97.5**	**96.6**	92.2	**84.2**	91.6	76.8	77.1	66.5	58.2	70.0	59.1	86.8	79.2	**97.5**	**97.4**	95.5	**96.4**	83.4

Method	32 Hip R	33 Sacr	34 Rib L4	35 Rib L5	36 Rib L6	37 Rib L7	38 Rib L8	39 Rib L9	40 Rib L10	41 Rib L11	42 Rib L12	43 Rib R4	44 Rib R5	45 Rib R6	46 Rib R7	47 Rib R8	48 Rib R9	49 Rib R10	50 Rib R11	51 Rib R12	52 L5	53 L4	54 L3	55 L2	56 L1	57 T12	58 T11	59 T10	60 T9	61 T8	62 T7	**mDSC**	**mHD95**
VoCo [84]	67.2	48.8	15.3	36.8	50.8	55.4	56.1	58.8	65.6	68.7	63.2	13.9	34.3	47.3	57.1	59.7	61.4	62.2	65.9	62.6	66.1	85.1	86.4	86.1	85.3	84.9	84.1	79.6	71.5	55.0	38.5	73.0	7.7
SuPreM [50]	61.2	44.6	12.0	34.7	47.7	49.2	51.7	54.1	63.3	67.5	62.5	14.4	32.9	46.6	52.1	56.3	59.8	62.4	62.2	58.9	60.4	80.4	83.9	86.5	85.4	83.2	78.6	73.6	65.4	50.7	39.5	70.8	8.4
uC 3DU-Net [37]	70.5	49.9	0.0	39.3	57.9	63.3	63.8	67.6	71.3	72.6	68.6	13.6	38.7	57.8	65.4	67.0	67.9	69.7	70.4	65.6	72.9	87.2	86.9	86.7	85.9	87.7	85.8	85.2	77.4	67.6	41.7	76.7	**5.9**
SegMamba [88]	69.6	49.7	15.9	38.8	48.3	53.8	57.1	60.5	69.0	70.0	66.7	18.0	36.0	50.0	60.6	64.6	67.0	69.4	68.1	64.7	66.4	85.6	87.0	87.1	87.3	88.6	88.2	84.9	78.7	69.4	48.9	74.5	6.8
Mamba-HoME [67]	75.7	60.3	4.2	41.3	52.1	58.7	62.0	64.7	70.0	70.9	66.1	19.3	43.2	55.7	63.7	65.8	67.3	68.3	67.9	66.2	74.7	87.1	87.7	87.1	87.3	87.4	87.2	85.6	76.8	68.2	48.1	75.2	6.5
Merlin [13]	**78.0**	**64.4**	0.0	42.3	53.6	60.6	64.0	66.4	70.4	71.2	65.9	19.9	46.0	58.0	65.0	66.2	67.4	67.9	67.8	66.8	77.8	87.6	88.0	87.1	87.3	86.9	86.8	85.9	76.1	67.8	48.0	75.8	6.0
fVLM [76]	70.7	53.0	0.0	42.3	54.1	55.9	58.5	60.9	67.0	67.7	60.4	13.7	36.0	53.7	58.1	61.2	66.7	67.1	66.3	60.1	72.3	85.4	84.8	84.7	83.8	82.8	83.9	81.0	72.5	68.8	45.9	72.0	8.1
BTB3D [33]	73.6	55.9	0.0	45.2	57.0	58.8	61.4	63.8	69.9	70.6	63.3	16.6	38.9	56.6	61.0	64.1	69.6	70.0	69.2	63.0	75.2	88.3	87.7	87.6	86.7	85.7	86.8	83.9	75.4	71.7	48.8	76.9	6.2
A^{2}MoE	75.0	60.0	**22.0**	**48.0**	**58.0**	**64.0**	**68.0**	**72.5**	**75.5**	**76.0**	**72.0**	**24.0**	**51.0**	**60.0**	**68.0**	**70.0**	**71.0**	**75.0**	**73.5**	**70.0**	**80.0**	**90.0**	**90.5**	**90.5**	**90.0**	**91.0**	**91.5**	**90.0**	**85.0**	**78.0**	**55.0**	**79.8**	**5.9**

Table 23: Quantitative segmentation performance on the MVSeg2023 [19] test set. Evaluation metrics include mDSC (↑, %) and mHD95 (↓, mm). Organ abbreviations: PL (posterior leaflet) and AL (anterior leaflet). The best results are **bolded**. † indicates pre-trained models.

	Organ Classes (DSC %) ↑		Overall	
Method	PL	AL	**mDSC**	**mHD95**
VoCo† [84, 85]	83.3	85.3	84.3	5.1
uC 3DU-Net [37]	82.9	85.0	83.9	4.7
SuPreM† [50]	83.2	85.4	84.3	4.9
SegMamba [88]	83.0	84.7	83.8	5.8
Mamba-HoME [67]	83.6	85.3	84.5	4.8
Merlin† [13]	83.3	84.9	84.1	**3.9**
fVLM† [76]	79.0	81.3	80.2	5.0
BTB3D† [33]	83.5	85.7	84.6	4.0
A^{2}MoE†	**84.5**	**86.0**	**85.4**	4.0

E.2 Classification

Table 24: Binary trauma classification on RATIC (0 vs. injury): ROC-AUC values multiplied by 100 and rounded to one decimal place. The best results are **bolded**.

Method	Bowel	Liver	Kidney	Spleen	Mean
VoCo [84]	64.2	67.1	88.1	86.2	76.4
SuPreM [50]	72.6	76.2	90.4	**92.5**	82.9
BTB3D [33]	70.1	74.8	89.4	87.2	80.4
Merlin [13]	79.8	81.2	**94.2**	90.8	86.5
A^2MoE	**82.4**	**84.6**	93.8	92.1	**88.2**

Table 25: Per-grade abdominal trauma classification on RATIC [73]. AUC: area under the receiver operating characteristic curve; Sens: sensitivity; Spec: specificity. All metrics are reported as percentages. The best result for each metric in each row is **bolded**.

		VoCo [84]			SuPreM [50]			BTB3D [33]			Merlin [13]			A^2MoE		
Organ	Grading	AUC	Sens.	Spec.	AUC	Sens.	Spec.	AUC	Sens.	Spec.	AUC	Sens.	Spec.	AUC	Sens.	Spec.
Bowel	0 vs. 1	64.2	61.9	63.1	72.6	71.4	68.4	70.1	66.7	65.8	79.8	**81.0**	74.2	**82.4**	76.2	**79.5**
Liver	0 vs. 1+2	67.1	62.8	64.2	76.2	69.1	72.8	74.8	67.0	71.5	81.2	74.5	77.9	**84.6**	**78.7**	**80.8**
	0 vs. 1	63.8	60.5	64.2	72.4	65.8	72.8	70.9	63.2	71.5	**78.1**	69.7	**77.9**	77.3	**72.4**	74.2
	0 vs. 2	76.8	72.2	64.2	85.4	83.3	72.8	83.2	77.8	71.5	89.1	**88.9**	77.9	**91.8**	83.3	**80.8**
Kidney	0 vs. 1+2	88.1	76.4	82.3	90.4	80.0	**89.2**	89.4	78.2	83.5	**94.2**	83.6	87.1	93.8	**85.5**	88.6
	0 vs. 1	85.4	68.4	**88.6**	87.9	73.5	83.5	86.7	73.5	82.4	90.8	**82.4**	86.3	**92.1**	79.4	87.2
	0 vs. 2	91.8	81.0	84.6	94.2	85.7	87.3	93.2	85.7	85.8	95.7	**95.2**	88.5	**96.8**	90.5	**89.8**
Spleen	0 vs. 1+2	86.2	71.6	79.8	**92.5**	75.2	82.4	87.2	72.5	**87.5**	90.8	79.8	85.8	92.1	**81.7**	87.2
	0 vs. 1	83.1	66.7	78.5	85.2	69.8	81.2	84.2	68.3	79.8	88.7	74.6	84.8	**90.4**	**76.2**	**86.2**
	0 vs. 2	89.8	78.3	81.2	92.1	80.4	83.5	91.2	80.4	**89.0**	93.8	**84.8**	86.2	**94.8**	82.6	87.9

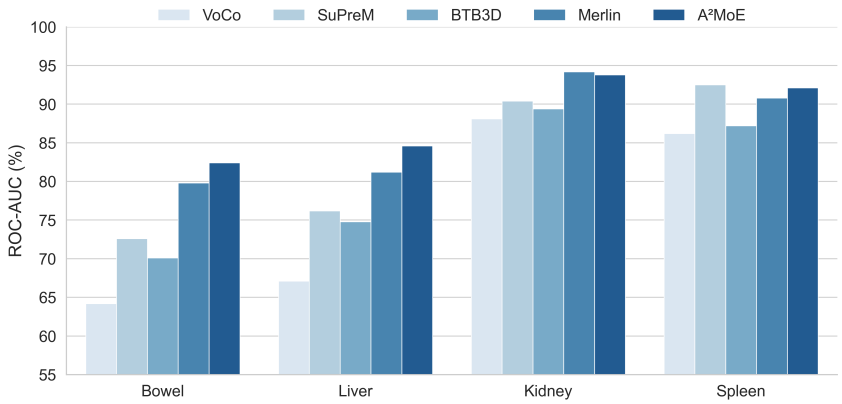

Figure 8: **ROC-AUC scores for the classification tasks** across bowel, liver, kidney, and spleen abdominal injuries on RATIC [73].

E.3 Detection

Table 26: Detection performance on the AbdomenAtlas 3.0 test set. We compare sensitivity (Sens.) for tumor sizes $\leq$2 cm and $>$2 cm, along with specificity (Spec.) in %. The best results are **bolded**.

Model	Pancreatic Tumor (%)			Kidney Tumor (%)			Liver Tumor (%)		
	Sen. ($\leq$2 cm)	Sens. ($>$2 cm)	Spec.	Sens. ($\leq$2 cm)	Sens. ($>$2 cm)	Spec.	Sens. ($\leq$2 cm)	Sens. ($>$2 cm)	Spec.
CT-CHAT	**66.7**	51.9	61.2	31.1	32.8	74.2	5.7	3.2	94.7
CT2Rep	0.0	0.0	92.5	36.5	39.3	70.4	35.8	49.2	70.4
M3D	0.0	7.4	97.2	8.1	16.4	84.1	9.4	12.7	86.0
Merlin	33.3	51.9	71.8	28.4	45.9	86.6	30.2	41.3	**95.9**
RadFM	0.0	0.0	**99.9**	3.7	6.3	**95.6**	3.3	5.7	93.9
RadGPT	**66.7**	81.5	93.2	54.8	93.3	51.8	39.6	**96.8**	64.4
A^{2}MoE (ours)	**66.7**	**83.3**	94.1	**58.2**	**94.1**	55.4	**41.2**	95.2	85.4

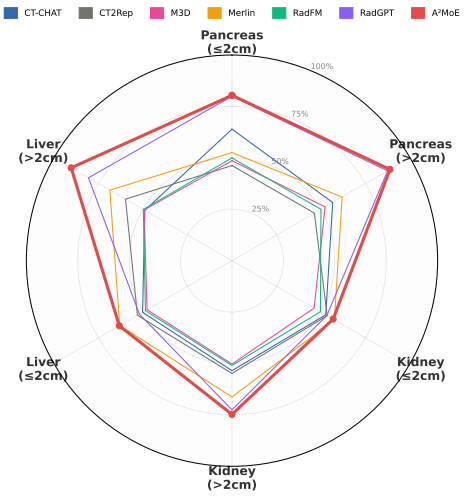

Figure 9: **Detection performance on AbdomenAtlas 3.0** [11]. We compare balanced sensitivity and specificity for tumor sizes $\leq$2 cm vs. $>$2 cm across pancreatic, kidney, and liver tumors.

F Qualitative results

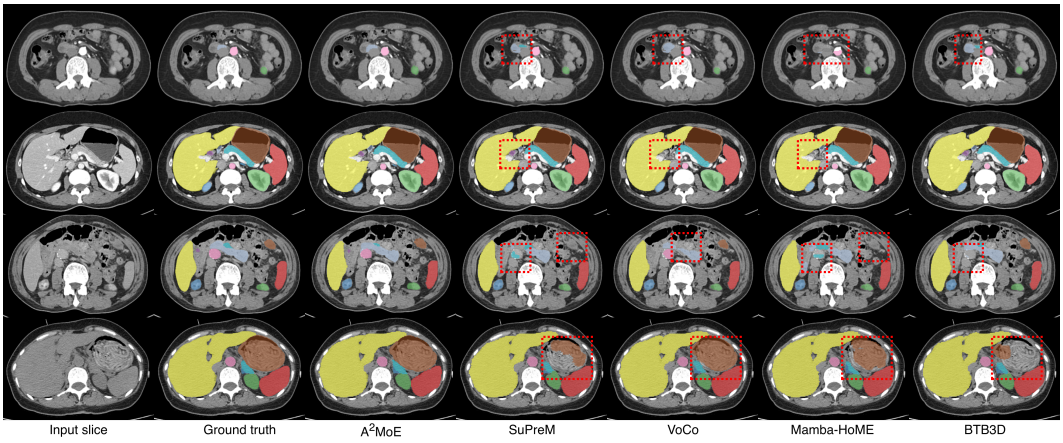

Figure 10: **Qualitative results on the AMOS-CT validation set.** We present the input slice, reference segmentation, and predictions from A^{2}MoE and the four best-performing comparison methods. Key differences are highlighted with red bounding boxes.

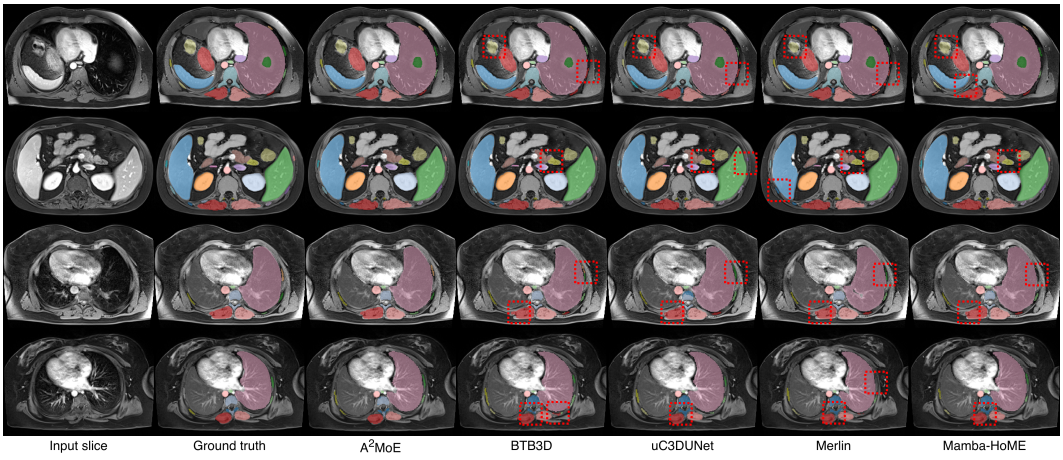

Figure 11: **Qualitative results on the MRISegmenter test set.** We present the input slice, reference segmentation, and predictions from A^{2}MoE and the four best-performing comparison methods. Key differences are highlighted with red bounding boxes.

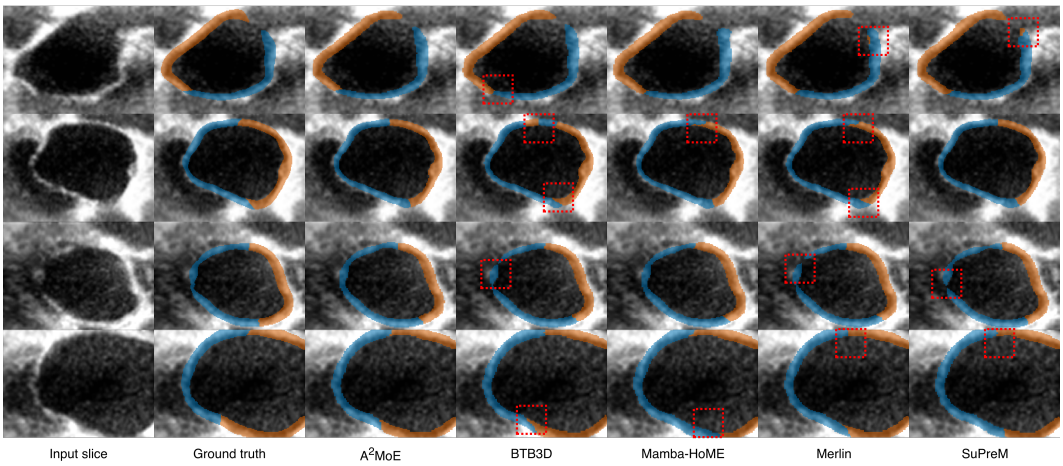

Figure 12: **Qualitative results on the MVSeg test set.** We present the input slice, reference segmentation, and predictions from A^{2}MoE and the four best-performing comparison methods. Key differences are highlighted with red bounding boxes.

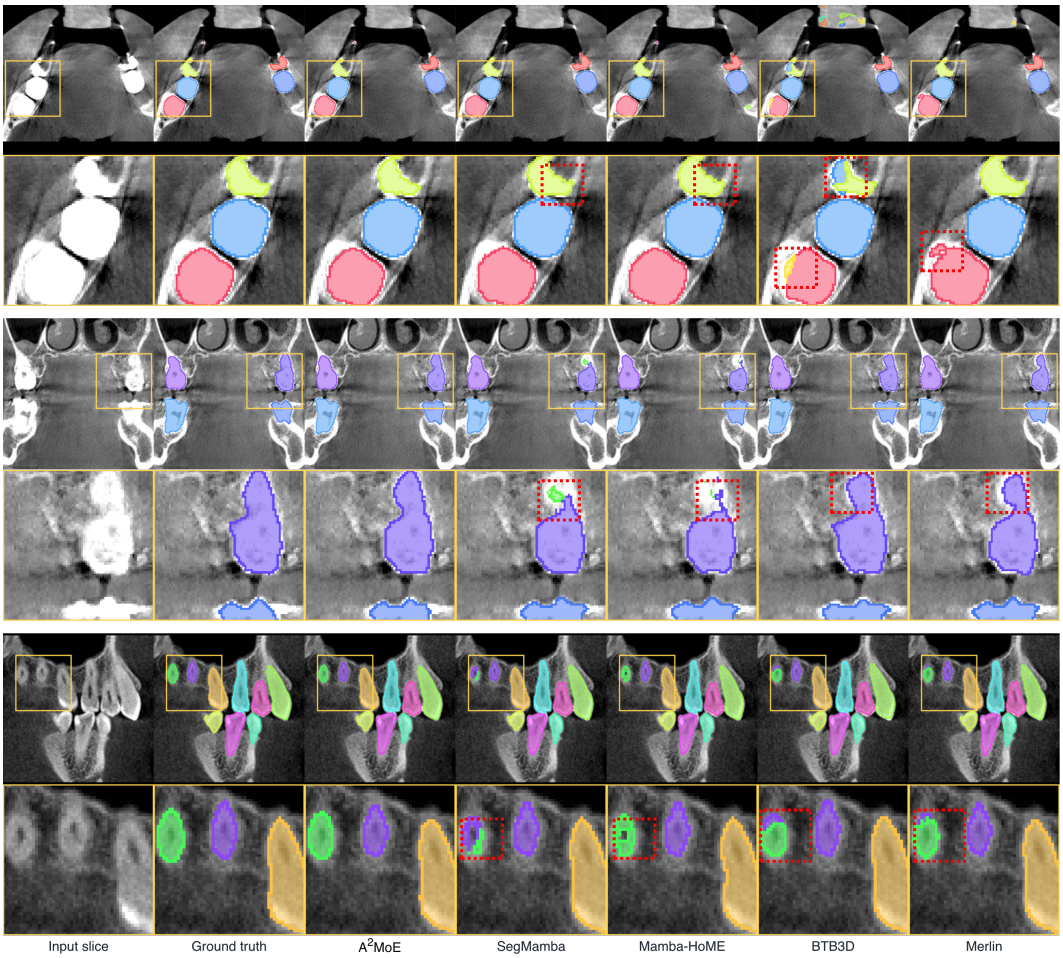

Figure 13: **Qualitative results on the GEPAR3D test set.** We present the 2D input slice, ground truth, A^{2}MoE, and the four next best-performing methods. For each example, the first row shows the original slice, while the second row shows a magnified view of the region. Key differences are highlighted with red bounding boxes.

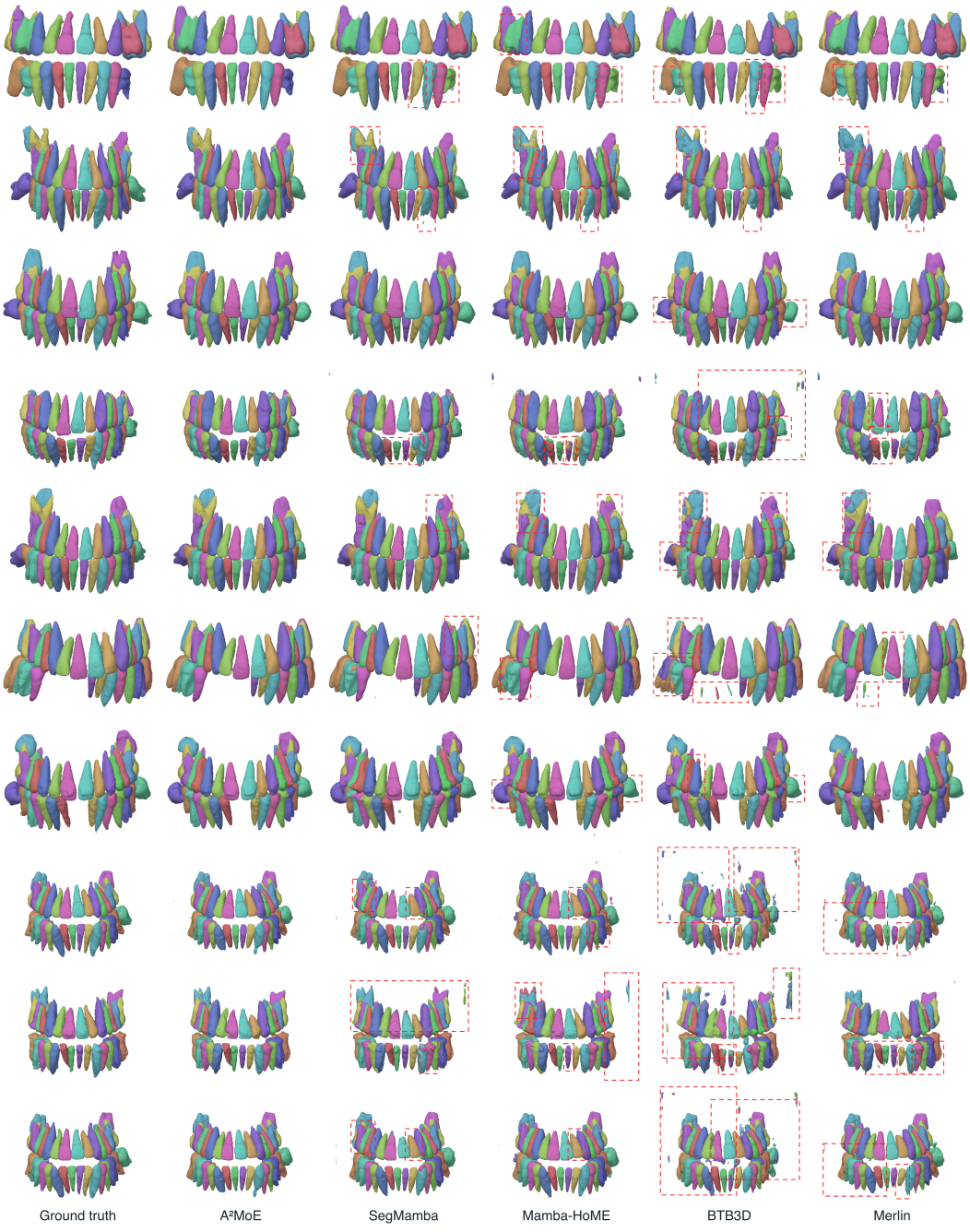

Figure 14: **Qualitative results on the GEPAR3D test set.** We present the 3D reference segmentation, and predictions from A^{2}MoE and the four best-performing comparison methods. Key differences are highlighted with red bounding boxes.

G Statistical significance

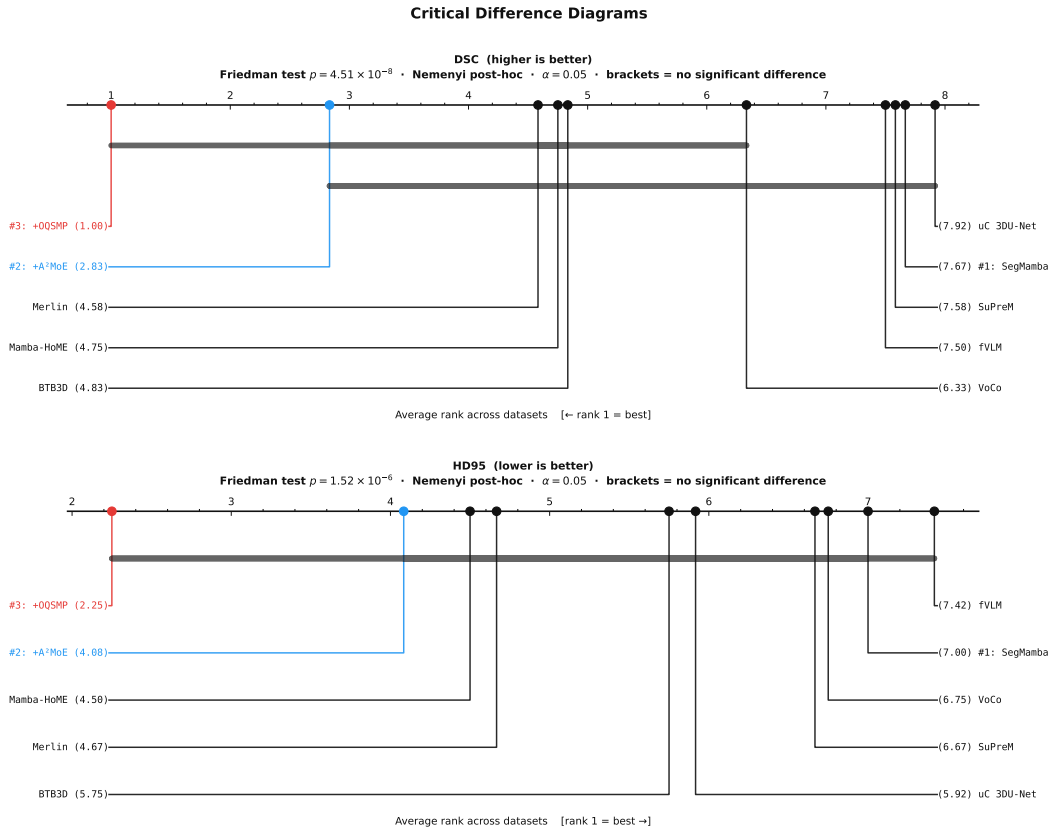

Figure 15: **Critical difference (CD) diagrams comparing ten segmentation methods across six benchmark datasets for DSC (top; higher is better) and HD95 (bottom; lower is better).** Average ranks are computed independently per dataset and then averaged across all six benchmarks; rank 1 denotes the best-performing method. Horizontal brackets connect groups of methods for which no statistically significant difference was detected (Friedman test followed by Nemenyi post-hoc test, α=0.05). Our proposed method A^{2}MoE is evaluated in two configurations: **#2 +A^{2}MoE** applied on top of SegMamba, and **#3 +OQMSP**, which further extends it with organ-query-guided semi-supervised pre-training; #1 SegMamba serves as the baseline. The full model (#3 +OQMSP) achieves the best average rank on both metrics (DSC: 1.00, HD95: 2.25) and is statistically distinguishable from the majority of competing methods ($p = 4.51 \cdot 10^{-8}$ and $p = 1.88 \cdot 10^{-6}$, respectively).

H Generalizability

The strength of a medical foundation model lies not in its performance on a single dataset, but in its ability to transcend the specific conditions of its training [7]. A central design goal of A^{2}MoE is that the inductive biases it encodes – conditional computation routed by anatomical composition and mask-free organ–text alignment via shared queries – should not be specific to any single modality, organ system, or label set. We provide empirical evidence for this claim along four orthogonal axes: (i) *cross-modality transfer* from CT pre-training to MRI and 3D ultrasound; (ii) *cross-anatomy transfer* from abdominal and pelvic supervision to dental cone-beam CT; (iii) *out-of-distribution robustness* on multi-center, multi-vendor, and multi-protocol cohorts that were never seen during pre-training; and (iv) *multi-task transfer* to downstream tasks that the pre-training objective does not directly optimize (organ injury classification on RATIC and tumor detection on AbdomenAtlas 3.0).

1. **Cross-modality transfer (CT $\rightarrow$ MRI/US).** Although A^{2}MoE is pre-trained exclusively on multi-phase CT, the encoder transfers without architectural modification to volumetric MRI (AMOS-MRI [39], MRISegmenter [98]) and 3D transesophageal ultrasound (MVSeg2023 [19]). On AMOS-MRI e observe a gain of 5.6 DSC percentage points over the SegMamba backbone and $+2.2$ over VoCo (Table 1), and on MRISegmenter the full A^{2}MoE attains the best mean DSC across all 62 anatomical classes (Table 22). These improvements are notable because (a) the histogram router was supervised only on TotalSegmentator CT labels and never observed MRI tissue contrast, and (b) OQMSP queries were aligned to English CT-report snippets rather than MRI findings. The fact that the same queries still localize the correct organs on MRI/US supports the interpretation that A^{2}MoE learns a *modality-agnostic* anatomical prior rather than a CT-specific texture model.
2. **Cross-anatomy transfer.** GEPAR3D [80] evaluates fine-grained multi-class dental segmentation in cone-beam CT, an anatomy that is *absent* from the pre-training cohort and the histogram class list (which targets abdominal and pelvic structures). A^{2}MoE nevertheless improves over the SegMamba baseline by $+0.8$ DSC and matches or exceeds modality-matched 3D pre-training baselines, indicating that the encoder retains capacity for transfer to anatomies outside the routing prior.
3. **Out-of-distribution robustness.** The downstream cohorts used in this study cover heterogeneous acquisition conditions: AMOS spans multiple scanners and slice thicknesses; PANORAMA aggregates seven contributing centers from two continents; MRISegmenter spans four imaging phases per patient; RATIC [73] consists of 4,274 trauma studies from 23 institutions across 14 countries; and AbdomenAtlas 3.0 aggregates 17 previously released public CT datasets. None of these datasets contributes to A^{2}MoE pre-training. The consistent improvements reported in Tables 1, 24, 25, and 26 therefore reflect generalization across sites, scanners, and protocols, not in-distribution memorization.
4. **Multi-task transfer.** Although the pre-training loss couples segmentation, presence prediction, and contrastive alignment, A^{2}MoE also generalizes to downstream *classification* (binary trauma 0 versus $\{1, 2\}$ and per-grade 0 versus 1, 0 versus 2 on RATIC; Tables 24–25) and *detection* (per-organ tumor detection split by lesion size on AbdomenAtlas 3.0; Table 26). On RATIC, the same backbone improves binary mean ROC-AUC from 86.5 (Merlin) to 88.2 without organ-specific fine-tuning of the encoder, and on AbdomenAtlas 3.0 it dominates sensitivity for clinically critical ≤ 2 cm lesions across pancreatic, kidney, and liver tumors. Together, these results support the claim that the anatomical prior introduced by A^{2}MoE is broadly useful, rather than a segmentation-specific inductive bias.

I Limitations and future work

1. **Anatomical scope.** The histogram and the per-stage class-to-expert affinity matrix $\mathbf{A}^{(s)}$ are defined over $C-1=14$ abdominal and pelvic classes, with $C=15$ including background. The pre-training labels themselves are *not* raw automatic segmentations: TotalSegmentator outputs and LLM-decomposed report snippets are subsequently revised by experienced radiologists, as detailed in Appendix A. The remaining limitation is therefore one of *coverage* rather than label quality: anatomies outside the 14-class taxonomy are routed using the residual visual context only and do not benefit from the structural prior. Whole-body, head–neck, musculoskeletal, and pediatric imaging are likely to require either a richer class taxonomy or hierarchical routing priors.
2. **Modality and language scope.** Pre-training is performed on adult multi-phase CT paired with English-language reports. Transfer to MRI, ultrasound, and dental CBCT is supported by our results, but A^{2}MoE has not been pre-trained with non-English reports, T2/diffusion-weighted MRI, X-ray, PET, or echocardiographic videos. The mask-free OQMSP architecture is modality-agnostic by construction, so extending pre-training to additional modalities is a natural direction but is not demonstrated here.
3. **Backbone and decoder choice.** We instantiate A^{2}MoE on top of a SegMamba-style 3D encoder [88] with a UNETR-style decoder [35]. The histogram router and OQMSP are designed to be backbone-agnostic and the gating equation makes no assumption about the operator inside each expert (we use bottleneck $1\times1\times1$ convolutions). Verifying that the same gains hold for purely convolutional, transformer-only, or state-space-only backbones is left to future work.
4. **Compute and reproducibility.** Pre-training A^{2}MoE requires non-trivial compute and access to one proprietary CT cohort in addition to the public Merlin [13], and CancerVerse [49] releases. We mitigate the latter by (i) reporting the public-only ablations needed for reproducibility, (ii) providing all hyperparameters and pre-processing details in Appendix C, and (iii) committing to release checkpoints, pre-processed metadata, and training code on acceptance. Reproducing the full proprietary-data results is, however, contingent on data-use agreements that are outside the authors' control.
5. **Clinical readiness.** A^{2}MoE is a representation-learning method evaluated on retrospective benchmarks. It has not been validated prospectively, has not undergone reader studies, and is not approved as a medical device. Its results should not be used to guide clinical decisions.

Future directions. Our findings open several concrete directions: (i) scaling the histogram taxonomy beyond 14 classes via hierarchical or learned anatomical ontologies, with the goal of native whole-body support; (ii) hybrid token/patch routing that recovers fine-grained specialization when crops contain many organs in similar proportions; (iii) longitudinal modeling that pairs the per-organ embeddings of OQMSP across time points to support follow-up and treatment-response reasoning; (iv) extension to non-CT pre-training corpora (MRI, ultrasound, PET/CT) to test whether the same anatomy-activated prior emerges natively in those modalities; (v) replacement of the SegMamba backbone with alternative 3D encoders to confirm backbone independence; and (vi) prospective and human-in-the-loop validation on triage and reporting workflows, paired with calibration and uncertainty estimation, to bridge the gap between retrospective benchmark performance and clinical utility.

J Impact statement

1. **Contribution.** A^{2}MoE contributes a 3D medical vision–language pre-training framework that combines anatomy-activated mixture-of-experts routing with mask-free organ-query alignment and multi-task supervision. The intended scientific advance is a more structured, scalable use of imaging–report pairs for learning representations that support segmentation, classification, and detection, with explicit modeling of organ-specific evidence rather than purely global image–text matching. The approach is mask-free at inference, transfers across modalities (CT, MRI, ultrasound) and anatomies, and reaches state-of-the-art performance on external benchmarks, providing a reusable building block for future medical foundation models.
2. **Clinical impact.** If validated and integrated thoughtfully, the representations learned by A^{2}MoE could support *assistive* radiology workflows: faster contouring and quality control for segmentation-heavy planning, triage and worklist prioritization when paired with calibrated uncertainty, retrieval of similar prior studies, teaching tools that link images to report language, and research pipelines that reduce manual annotation burden. Because the model exposes an organ-indexed soft-localization map per organ via OQMSP cross-attention, downstream clinical tools can show *which* regions support a given finding, although the interpretability and clinical usefulness of these maps require dedicated validation [22]. These benefits are contingent on rigorous local evaluation, regulatory pathways where applicable, and continued human responsibility for diagnosis and treatment decisions.
3. **Negative impact.** A^{2}MoE inherits limitations common to medical machine-learning systems and can produce confident errors, particularly under distribution shift, in patients with atypical anatomy, or in pediatric or pregnant populations. It may encode and amplify biases present in who is scanned, how findings are documented, and which sites and scanners dominate the training data, and it has not been audited for fairness across demographic, scanner, or protocol strata. Without careful governance, misuse could include over-reliance on automation, privacy risks from mishandling of patient data, deployment to populations underrepresented in training, or use as a stand-alone diagnostic tool. We therefore emphasize that this work is a methodological contribution, not a clinical product, and that any real-world deployment must be preceded by prospective validation, robustness and fairness auditing, clinician oversight, transparent limitation reporting, and full compliance with ethics review, de-identification standards, and data-use agreements.

K Notation

Throughout the paper, we denote the batch size by B and the number of segmentation classes, including background, by C ($C = 15$ in pre-training), with d_e denoting the dimension of the shared vision–text embedding space. An input volume $\mathbf{X} \in \mathbb{R}^{B\times 1\times D\times H\times W}$ is paired with a voxel-wise label map $L \in \{0,\dots,C{-}1\}^{B\times D\times H\times W}$, and the pyramid encoder produces MoE-refined features $\mathbf{X}_s \in \mathbb{R}^{B\times C_s\times D_s\times H_s\times W_s}$ at four stages $s \in \{1,\dots,4\}$, with K_s experts per stage. Let N_s denote the number of projector key/value tokens after any strided convolution. At stage s the A^{2}MoE router combines a global context $\mathbf{v}_s$, a crop-level routing representation $\widetilde{\mathbf{h}} = \mathcal{T}(\mathbf{h})$, where each raw histogram $\mathbf{h} \in \Delta^{C-1}$ is replaced at inference by the stage-specific composition estimate $\hat{\mathbf{h}}_s$ before normalization, and a log-prior built from a learnable affinity matrix $\mathbf{A}^{(s)} \in \mathbb{R}^{C\times K_s}$, producing expert weights $\mathbf{W}_s$ from the joint route $\Psi^{(s)}_{\text{gate}}$, direct route $\Psi^{(s)}_{\text{anat}}$, and structural prior; the stage-specific coefficient $\beta_s > 0$ controls the prior's strength and ϵ stabilizes its logarithm. On top of the encoder, the OQMSP uses a single set of organ queries $\mathbf{Q} \in \mathbb{R}^{(C-1)\times d_e}$ shared across stages, produces per-stage, per-organ embeddings $\hat{\mathbf{E}}_s$, and fuses them with softmax-normalized weights $\boldsymbol{\alpha}$ into $\mathbf{Z}_{\text{vis}} \in \mathbb{R}^{B\times(C-1)\times d_e}$, which is aligned via a learnable log-temperature $\log\tau$ with text embeddings $\mathbf{z}_{\text{txt}} \in \mathbb{S}^{d_e-1}$ drawn from a frozen PubMedBERT through a projection head ϕ_{txt}. The final objective weights are segmentation, anatomy-conditioned routing, organ–text alignment, cross-stage consistency, and query-grounding losses $\lambda_{\text{seg}}, \lambda_{\text{MoE}}, \lambda_{\text{align}}, \lambda_{\text{stage}}, \lambda_{\text{ground}}$.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the anatomy-conditioned routing and mask-free organ-query alignment claims, and the revised wording limits the scope to the evaluated segmentation, classification, detection, and transfer settings.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Appendix I discusses anatomical scope, modality/language scope, backbone choice, compute and reproducibility, and clinical readiness.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper proposes an empirical architecture and pre-training framework; it does not present formal theoretical results or proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The method, losses, preprocessing, model components, downstream tasks, and hardware/software setup are described in the main text and Appendix C.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: The code is available at github.com/gmum/A2MoE.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Training details, preprocessing, optimizer, patch sizes, hardware, and task-specific settings are provided in the main experimental section and Appendix C.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: In Appendix G we provide critical difference (CD) diagrams comparing our proposed method with several state-of-the-art methods across six benchmark datasets for DSC and HD95.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: In Appendix C.1 we provide the hardware and software used to reproduce our results.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The study uses de-identified retrospective medical imaging data and public datasets under their data-use terms; clinical-readiness limitations and responsible-use constraints are discussed in the limitations and impact statement.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Appendix J discusses potential assistive clinical benefits, privacy and deployment risks, fairness concerns, and the need for prospective validation.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [Yes]

    Justification: The paper does not release raw proprietary medical images. Any released checkpoints or metadata will be de-identified and governed by the corresponding data-use agreements and responsible-use terms.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Existing datasets, models, and software packages are cited throughout the paper and appendix; final release materials should include explicit license and access-term summaries for each reused asset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Detailed documentation for the new assets is available in the supplementary material.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The study does not involve crowdsourcing experiments or interaction with human participants; radiologist review concerns retrospective data curation.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: The study uses retrospective de-identified medical data under institutional approvals or data-use agreements. Identifying approval details should be provided in the non-anonymous version or anonymized as required by the submission policy. Ethical approval

for this study was granted by the Bioethics Committee at the Faculty of Medicine, Collegium Medicum, University of Warmia and Mazury in Olsztyn (Approval No. 9/2023, March 30, 2023).

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: We used LLM to extract text snippets from the whole radiology text reports. We described the whole process in Appendix A. Also, we used LLM for editing purposes.

Guidelines:

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