Revisiting 2D Foundation Models for Scalable 3D Medical Image Classification

Supplementary Material

Table of Contents

A. Implementation Details of AnyMC3D

A1. Model Configuration. AnyMC3D can be implemented with any transformer-based 2D foundation model (FM) backbone. We apply LoRA adapters to three components: (1) the patch embedding layer, (2) query, key, and value projection layers in self-attention, and (3) the output projection layer in self-attention. The LoRA rank r is set to 8 and the scaling factor α is set to 16. The task query is initialized as a learnable parameter with values drawn from a truncated normal distribution with standard deviation 0.02. The classification head consists of a single linear layer. The final activation function is sigmoid for multi-label classification tasks and softmax for multi-class classification tasks.

A2. 2D Backbones for 3D Inputs. AnyMC3D processes 3D volumes through slice-wise encoding with 2D FMs (Alg. 1). Each volume is partitioned into 2D slices along the highest-resolution axis, and the slice and batch dimensions are collapsed (reshaped to $(B \cdot S, C, H, W)$) for parallel processing. Single-channel medical slices are replicated three times to match the RGB input format and normalized with ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) to align with the FM's pretraining distribution. We also explored stacking three consecutive slices as a 2.5D input but found it performed comparably to single-slice replication.

A3. Training Details. During training, we employ focal loss $\mathcal{L}_{\text{focal}} = -\alpha(1-p_t)^{\gamma}\log(p_t)$ to address class imbalance. We set γ (the focusing parameter that down-weights easy examples) to 2 and α (the balancing parameter that addresses class imbalance) to 0.25. The batch size is 2 for all tasks except T10, where the batch size is 1 due to the large input dimension. We use a learning rate of 1e-4 with weight decay of 1e-5 for LoRA layers, and a learning rate of 1e-3 with weight decay of 1e-4 for the task query and classification head. The maximum number of training epochs is set to 100, and we select the best checkpoint based on validation AUC. During training, we apply strong data augmentation to both our method and all baselines. Our data augmentation strategy, adapted from nnU-Net [27], is applied directly to 3D images and includes random flipping along all three spatial axes (probability 0.5 each), random rotation ($\pm 30°$ per axis, probability 0.2), random zoom (0.7-1.4×, probability 0.2), random affine translation (± 10 voxels, probability 0.2), Gaussian noise (probability 0.25),

Algorithm 1: Forward Pass of AnyMC3D

Input : 3D volume $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W\times S}$ (or multi-view volumes), pretrained 2D FM f_θ (frozen), optional: return_seg
Parameters: LoRA adapters $\{\psi^{(i)}\}$, task query $\mathbf{q}_t$, classification head g_ω, optional view queries $\{\mathbf{q}^{(i)}\}$, optional 3D decoder D
Output : Classification logits $\mathbf{z} \in \mathbb{R}^{B\times K}$, optional segmentation logits $\mathbf{z}_{\text{seg}}$

Initialize: views $\leftarrow [\,]$
// Process each view $i \in \{1,\ldots,V\}$
for *each view* i **do**
  // Extract and prepare view
  $\mathbf{x}_i \leftarrow \text{ExtractView}_i(\mathbf{X})$
  $\hat{\mathbf{x}}_i \leftarrow \text{Rearrange}(\mathbf{x}_i, (B\cdot S, C, H, W))$
  $\hat{\mathbf{x}}_i \leftarrow \text{Normalize}(\hat{\mathbf{x}}_i)$
  // Slice-wise feature extraction
  $\mathbf{H}^{(i)} \leftarrow \tilde{f}_{\theta,\psi^{(i)}}(\hat{\mathbf{x}}_i)$ (Eq. 4 in the main paper)
  // Query-based attention pooling
  $\mathbf{v}^{(i)} = \text{AttentionPool}(\mathbf{H}^{(i)}, \mathbf{q}^{(i)})$
  views $\leftarrow$ views $\cup \{\mathbf{v}^{(i)}\}$
// Multi-view aggregation
if $V > 1$ **then**
  $\mathbf{V} = \text{Stack}([\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(V)}])$
  $\mathbf{v} = \text{AttentionPool}(\mathbf{V}, \mathbf{q}_t)$
else
  $\mathbf{v} = \mathbf{v}^{(1)}$ // for single view, $\mathbf{q}^{(1)} = \mathbf{q}_t$
// Classification
$\mathbf{z} = g_\omega(\mathbf{v}) \in \mathbb{R}^{B\times K}$
// Optional: segmentation supervision
if *return_seg* **then**
  $\tilde{\mathbf{P}} \leftarrow$ Extract patch tokens from $\mathbf{H}^{(i)}$
  $\mathbf{P} \leftarrow \text{Reshape}(\tilde{\mathbf{P}}, (B, d, S, g_h, g_w))$ (Eq. 7 in the main paper)
  $\mathbf{z}_{\text{seg}} = D(\mathbf{P})$
return $\mathbf{z}$, *optional:* $\mathbf{z}_{seg}$

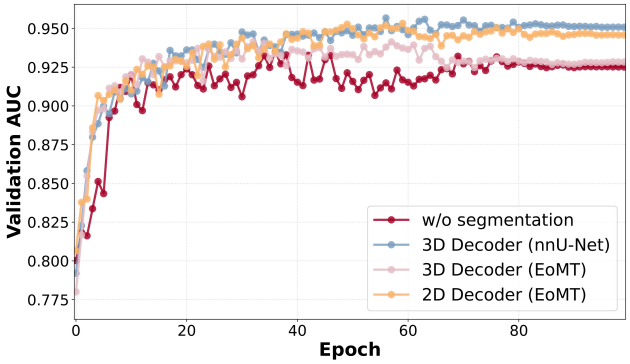

Figure 7. Validation AUC by different segmentation heads.

Gaussian blur (σ=0.5-1.0, probability 0.2), brightness multiplication (0.75-1.25×, probability 0.15), contrast augmentation (probability 0.15), low-resolution simulation (zoom 0.5-1.0×, probability 0.2), and gamma correction (γ=0.7-1.5, probability 0.2-0.3 with/without image inversion).

A4. Choice of Segmentation Decoder Architecture. To incorporate pixel-level supervision, we employ a 3D decoder that upsamples the pseudo-3D token volume (Eq. 7 in the main paper) into a 3D segmentation map. The decoder architecture follows the state-of-the-art 3D medical segmentation framework nnU-Net [27], consisting of consecutive blocks of 3D convolution, leaky ReLU activation (slope $p = 0.01$), and 3D instance normalization. Following [27], we use the combination of Dice loss and cross-entropy loss for training. We also explored alternative decoder designs from EoMT [30], a ViT-based segmentation model, in both 2D and 3D configurations. As shown in Fig. 7, the nnU-Net-based decoder achieves the best performance, followed by the 2D EoMT decoder. These two designs also consistently outperform the baseline, demonstrating that regularizing patch tokens with segmentation effectively improves 3D classification.

B. Dataset and Preprocessing

B1. Preprocessing. In Tab. 3, we present the preprocessing steps, including normalization strategies, reshaped input sizes, data splits, and positive sample prevalence. All compared methods use identical preprocessing per task. For image normalization, we apply different strategies based on imaging modalities. For CT and CECT, we apply task-specific CT windows to highlight relevant pathologies and anatomies, then rescale to [0, 1]. For MRI, we apply z-score normalization to T8 and T10, and apply percentile clipping (0.5th to 99.5th percentiles) followed by rescaling to [0, 1] for T9. All images are resized into a consistent shape before feeding to the network. Following [27], the resized dimension is determined based on the median spacing of each dataset to minimize information loss from downsampling.

B2. Dataset Description.

T1-T4. Abdominal Polytrauma. We use the public RSNA Abdominal Trauma Detection (RATIC) dataset [46], which contains 4,703 contrast-enhanced CT (CECT) scans with annotations for four types of organ injuries: bowel (T1), liver (T2), kidney (T3), and spleen (T4). Two severity grades (Low: AAST [5] grades 1-3; High: AAST grades 4-5) are annotated for liver, kidney, and spleen injuries, while only a single grade is provided for bowel injuries. We treat each organ injury as a binary classification task and train four separate models. For preprocessing, we crop the target organ region from each CECT scan using TotalSegmentator [57]. We exclude cases where the limited field of view results in incomplete organ coverage, as these are unreliable for diagnostic decisions.

T5. Pancreatic Cancer. We use the public PANORAMA challenge dataset [3], which contains 2,238 portal venous phase CECT scans from patients with pancreatic ductal adenocarcinoma (PDAC). We crop the pancreas region using pretrained segmentation models from the challenge baseline [2]. This dataset includes both classification labels for PDAC and segmentation masks for six critical structures: PDAC lesion, veins, arteries, pancreatic duct, common bile duct, and pancreas parenchyma. In our major comparison (Tab. 2 in the main paper), we use only the classification labels to compare 3D classification methods. The segmentation labels are used for auxiliary supervision experiments reported in Sec. 4.2 of the main paper with the method presented in Sec. 3.3 of the main paper.

T6. Lung Nodule Malignancy. We collect a private dataset for classifying malignancy of biopsied high-risk lung nodules. The dataset includes 1,140 subjects from a single imaging site. The cohort is 47.55% male and 52.45% female, with a mean age of 67.30±12.11 years. Each subject has both diagnostic and biopsy CT scans. CT images are acquired on scanners from Siemens (72.25%), GE (14.35%), Philips (12.68%), and Toshiba (0.72%). A radiologist identifies and labels biopsied lung nodules on diagnostic CT images based on biopsy-needle positions in the corresponding biopsy CT images, yielding 1,140 nodules. Biopsy results yield binary labels: 76.32% malignant and 23.68% benign.

T7. Lung Nodule Spiculation. We collect a private dataset for classifying lung nodule spiculation, an important indicator of malignancy. It includes 3,884 CT scans from multiple imaging sites, acquired on scanners from GE (41.04%), Siemens (34.76%), Toshiba/Canon (12.74%), Philips (2.70%), and others (8.76%). Three radiologists independently annotate spiculation for 6–30 mm solid nodules with ground truth defined by majority vote. In total, we obtain 5,668 nodules: 14.04% spiculated and 85.96% non-spiculated.

T8. Bicep Tear. We collect a large MR shoulder dataset comprising 11,828 subjects from two imaging sites. The co-

Table 3. Dataset summary across 12 tasks.

Task	Modality	Normalization	Input size	Total	Train	Val	Test	Pos. Ratio (%)
T1	CECT	[-150, 250]	$288 \times 224 \times 126$	4,679	3,276	467	936	2.23
T2	CECT	[-150, 250]	$288 \times 256 \times 80$	4,701	3,290	471	940	10.09
T3	CECT	[-150, 250]	$288 \times 128 \times 64$	4,677	3,274	469	934	5.99
T4	CECT	[-150, 250]	$160 \times 160 \times 70$	4,695	3,285	471	939	11.59
T5	CECT	[-150, 250]	$432 \times 240 \times 70$	2,238	1,566	224	448	30.20
T6	CT	[-1000, 400]	$224 \times 224 \times 70$	1,140	684	115	341	76.32
T7	CT	[-1000, 400]	$128 \times 128 \times 64$	5,668	3,604	1,032	1,032	14.04
T8	MRI	Z-score	$320 \times 320 \times 28$	12,159	11,143	508	508	38.14
T9	MRI	PercentileClip	$320 \times 320 \times 28$	10,978	10,006	486	486	75.10
T10	MRI	Z-score	$320 \times 320 \times 28$	12,191	11,203	494	494	78.34
T11	CT	A.S.L	$476 \times 476 \times 240$	50,188	45,149	2,000	3,039	N/A
T12	CT	B.T.B	$224 \times 192 \times 40$	29,476	23,657	2,930	2,889	N/A

A.S.L: All, soft tissue, and lung windows ([-1000, 1000], [-150, 250], [-1000, 400]).
B.T.B: Bleeding, tissue, and bone windows.

hort is 55.94% male and 44.06% female, with a mean age of 53.52±15.52 years. Each subject has both axial and sagittal MR scans. All scans are acquired with fat-saturation pulses. The magnetic field strength ranges from 0.7 T to 3.0 T, with a mean of 2.20±0.77 T. Most images (92.15%) are acquired using Siemens scanners, followed by Philips (5.43%), GE (1.72%), and other manufacturers (0.70%). The dataset includes biceps-tendon tear labels at three severity levels: no tear, tendinosis, and tear. Class distributions are 60.36% no tear, 20.64% tendinosis, and 19.01% tear. For evaluation, we report the average of the three one-vs-rest AUC scores corresponding to the three classes.

T9. Bursa Fluid. We assemble an MR shoulder dataset of 10,126 subjects from two imaging sites. The cohort is 55.38% male and 44.62% female, with a mean age of 53.58±15.46 years. Each subject has axial, coronal, and sagittal fat-saturated MR scans. Field strengths range from 0.7 T to 3.0 T (mean 2.23±0.77 T). Scanners are predominantly Siemens (91.57%), with Philips (5.20%), GE (2.39%), and others (0.75%) comprising the remainder. Bursa-fluid labels are provided (no fluid vs. fluid present), with class proportions of 71.27% and 28.73%, respectively.

T10. Labrum Tear. Our MR shoulder dataset includes 11,816 subjects from two imaging sites. The population consists of 56.10% males and 43.90% females, with an average age of 53.37±15.52 years. Each subject undergoes coronal and sagittal fat-saturated MRI. The examinations are performed on scanners operating between 0.7 T and 3.0 T (mean 2.21±0.77 T). Siemens systems produce most scans (92.51%), followed by Philips (5.07%), GE (1.67%), and other vendors (0.75%). Labrum-tear status is annotated, with 67.15% labeled as no tear and 32.85% as tear.

T11. Chest CT Multi-abnormality. We use the public CT-RATE dataset [22] from Istanbul Medipol University, comprising 21,304 unique patients with 25,692 chest CT scans. The cohort ranges in age from 18 to 102 years, with a mean age of 48.8 years. The sex distribution is 41.6% female and 58.4% male. CT scans are acquired using three scanner manufacturers: Philips (61.5%), Siemens (30.1%), and PNMS (8.4%). The number of slices per volume ranges from 100 to 600. Multi-abnormality labels for 18 distinct abnormalities are extracted from the corresponding radiology reports for each CT volume, including medical material, arterial wall calcification, cardiomegaly, pericardial effusion, coronary artery wall calcification, hiatal hernia, lymphadenopathy, emphysema, atelectasis, lung nodule, lung opacity, pulmonary fibrotic sequela, pleural effusion, mosaic attenuation pattern, peribronchial thickening, consolidation, bronchiectasis, and interlobular septal thickening.

T12. Head CT Multi-finding. We curate a large proprietary anonymized dataset of non-contrast head CT (NCCT) volumes for emergency triage, comprising 29,476 studies collected from nine centers across the U.S., Canada, China, and India, under ethics approvals with informed consent waived. Data are drawn from pre-established cohorts or retrospectively selected cases. NCCT scans are acquired using Siemens, GE, and Toshiba scanners. Exclusion criteria include patient age $<$18 years or absence of axial reconstruction. Seventy-five head NCCT findings, including hemorrhagic, vascular, structural, traumatic, mass, and chronic conditions, are extracted from radiology reports using large language models and subsequently verified by board-certified radiologists.

C. Baseline Methods

C1. Implementation. This section describes implementation details for each baseline method. For methods with open-source repositories, we follow the original implemen-

tations and training hyperparameters. For others, we determine optimal hyperparameters through grid search.

3D DenseNet. We use the 3D DenseNet-121 implementation from MONAI[1].

3D ResNet. We use the 3D ResNet-18 implementation from MONAI, as it outperforms other variants (e.g., ResNet-50) in preliminary experiments.

3D ConvNeXt. We extend the 2D ConvNeXt [35] to 3D.

M3T. We follow the official implementation[2].

MST. We follow the official implementation[3] and adopt the best-performing configuration from the paper: DINOv2-pretrained ViT-S as the backbone and a transformer without positional embeddings for slice fusion.

RSNA-Kaggle. We follow [44] and reimplement the model with 2D EfficientNet as the backbone and bidirectional LSTM for slice fusion. Following [6, 21, 44], we stack consecutive slices as different channels to create 2.5D inputs. For fair comparison, we exclude model ensembling and remove the segmentation branch, as not all evaluated tasks include segmentation annotations.

MedicalNet is a 3D medical FM with a ResNet-50 backbone pretrained on large-scale 3D medical datasets [12]. We follow the official implementation[4] and evaluate three finetuning strategies: linear probing, LoRA adaptation, and full finetuning. In Tab. 2 of the main paper, we report full finetuning results as this achieves the best performance (Tab. 4).

VoCo is a 3D medical FM with a Swin-UNETR [23] backbone pretrained on large-scale 3D medical images [59]. We follow the official implementation[5] using the VoComni_B encoder for downstream classification. Similar to MedicalNet, we report full finetuning results as this outperforms other adaptation methods (Tab. 4).

MedImageInsight (MII) is a 2D FM pretrained on large-scale diverse medical images, including 2D modalities (fundus, pathology) and 2D slices from 3D imaging (CT, MRI) [13]. We extract slice embeddings using the publicly available vision encoder[6] as a frozen feature extractor, then aggregate them with our slice fusion method. We attempted full model finetuning but encountered severe overfitting due to the limited 3D training samples relative to the 360M-parameter DaViT backbone. We therefore use the frozen extraction setting, which is also the recommended configuration in the original paper.

MedGemma is a 2D medical FM built by finetuning the Gemma 3 vision encoder (SigLIP-400M) on over 33M medical image-text pairs, including 2D slices from CT and MRI [47]. We use its vision encoder, MedSigLIP[7], as a frozen feature extractor to extract slice embeddings, which are then combined with our fusion method. The MII and MedGemma baselines provide valuable references for evaluating the out-of-the-box quality of 2D FM embeddings when adapted to 3D medical classification tasks through our slice fusion approach.

[1] https://github.com/Project-MONAI/MONAI
[2] https://github.com/KVishnuVardhanR/M3T
[3] https://github.com/mueller-franzes/MST
[4] https://github.com/Tencent/MedicalNet
[5] https://github.com/Luffy03/Large-Scale-Medical
[6] https://huggingface.co/lion-ai/MedImageInsights
[7] https://huggingface.co/google/medsiglip-448

C2. Why Report Full Finetuning for 3D Medical FMs?

In our main comparison (Tab. 2 in the main paper), we report full finetuning results for 3D medical FMs to represent their optimal performance. To justify this choice, we compare three adaptation strategies on T4 (Tab. 4): linear probing, LoRA, and full finetuning. For LoRA, we apply low-rank updates to convolutional layers in MedicalNet's ResNet-50 backbone and to query, key, and value projection layers in VoCo's Swin-UNETR backbone. Full finetuning achieves the best performance (MedicalNet: 0.899, VoCo: 0.919 AUC), significantly outperforming linear probing (0.654 and 0.702 AUC), while LoRA provides a parameter-efficient middle ground. We therefore report full finetuning results to represent 3D medical FMs at their optimal performance.

Table 4. Comparison of different finetuning strategies for 3D medical FMs on T4.

Method	Metric	MedicalNet	VoCo
LP	Trainable Params (M)	0.004	0.002
	AUC	0.654	0.702
LoRA	Trainable Params (M)	1.14	0.07
	AUC	0.889	0.838
Full	Trainable Params (M)	46.16	50.49
	AUC	**0.899**	**0.919**

LP: Linear probing. **Full**: Full finetuning.

C3. Task-Optimized Baselines.

We compare against task-optimized baselines representing state-of-the-art performance for specific tasks, including challenge-winning solutions and specialized FMs tailored for particular clinical applications. This comparison rigorously tests whether *AnyMC3D as a general framework can match or exceed specialized methods without task-specific designs*.

PanDx (T5). PanDx [34] ranked first in the PANORAMA challenge, achieving an AUROC of 0.9263 and AP of 0.7243. The method employs a two-stage coarse-to-fine approach: (1) a low-resolution segmentation model localizes the pancreatic region, and (2) a high-resolution model segments six PDAC-related structures and generates both patient-level likelihood scores and lesion-level detection maps. Both stages use nnU-Net [27] trained on segmentation labels of pancreas-adjacent structures.

CT-CLIP (T11) [22] is a 3D FM trained via contrastive language-image pretraining that aligns CT volumes with report embeddings in a shared latent space. The model employs a 3D vision transformer as the image encoder and a text encoder to extract semantic features from radiology reports. During training, CT-CLIP maximizes cosine similarity between paired CT-report embeddings while minimizing similarity with negative pairs within each batch, enabling zero-shot abnormality detection. We compare against CT-CLIP's ClassFine variant, which finetunes a linear classifier on top of the pretrained frozen encoder, and the supervised baseline CT-Net [17], a fully supervised 3D CNN trained directly on classification labels.

DeepCNTD-Net (T12) [63] is a 3D neuroimaging FM that integrates two independently pretrained, task-specific vision networks through multi-modal fine-tuning with LLM-generated labels. The first network performs hemorrhage subtype segmentation using a 3D Dense U-Net optimized for five subtypes: intraparenchymal, subarachnoid, intraventricular, subdural, and epidural. The second network performs brain anatomy parcellation using a 3D U-Net with a multi-head design for segmenting left-hemisphere, supratentorial vs. infratentorial regions, and remaining brain structures. These pretrained networks are fused into a 3D DenseNet-based FM via feature-level integration, jointly encoding anatomical and pathological features.

C4. Scalability vs. Performance. Fig. 8 illustrates the performance-scalability trade-off on T4. Existing approaches show clear compromises: 3D methods trained from scratch require 11–33M parameters for 0.92–0.93 AUC, while 2D/2.5D transfer learning methods need 23–29M parameters to reach 0.92–0.95 AUC. Fully finetuned 3D medical FMs (MedicalNet, VoCo) achieve 0.89–0.92 AUC with 46-50M parameters, but their parameter-efficient variants sacrifice significant performance (0.65–0.89 AUC) despite using $<$1M parameters. By contrast, AnyMC3D breaks this trade-off, achieving the highest performance (0.957 AUC) with only 1.32M trainable parameters.

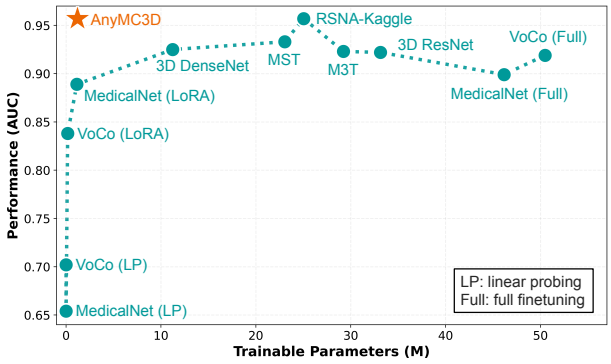

Figure 8. Performance and scalability on T4.

D. 1st Place in VLM3D Challenge

To demonstrate AnyMC3D's out-of-the-box generalizability, we participated in the VLM3D challenge for multi-abnormality classification across 18 chest diseases on the CT-RATE dataset [22] (Appendix B2). Without bells and whistles, AnyMC3D achieved first place among 118 participants. For our submission, we use DINOv2 ViT-B as the 2D FM backbone. In this challenge, submissions are ranked based on three metrics: AUC, macro-F1 score, and clinically-weighted relevance gain (CRG) score. While AUC is threshold-agnostic, F1 and CRG require binary predictions at a fixed threshold of 0.5. Since focal loss training typically shifts the optimal operating point below 0.5, we apply model calibration to postprocess AnyMC3D outputs.

Platt Scaling Calibration. We apply Platt scaling [45] to calibrate predictions per class:

$$P_c^{'} = \sigma(z_c^{'}), \quad z_c^{'} = a_c z_c + b_c \tag{1}$$

where $P_c^{'}$ is the calibrated probability for class c, z_c is the raw logit, $z_c^{'}$ is the calibrated logit, σ is the sigmoid function, and a_c, b_c are class-specific learned parameters. For each class, we first identify the raw operating point that maximizes F1 score on the validation set. We then fit a logistic regression model that takes the raw logit z_c as input and the binary label as target, learning a_c and b_c to map the raw operating point to 0.5.

Calibration Strategy Analysis. Fig. 9 compares four calibration strategies: no calibration (No Cali.), optimizing for F1 only, CRG only, or balanced F1+CRG. All calibration strategies preserve AUC at 0.888 while substantially improving F1 and CRG scores. Since CRG weights true positives and false negatives by class prevalence, its optimal operating point differs from F1, which equally weights precision and recall. We optimize for F1 (pink) in our final submission to maximize F1 ranking. However, the balanced F1+CRG strategy may be preferable for real-world deployment where both metrics matter.

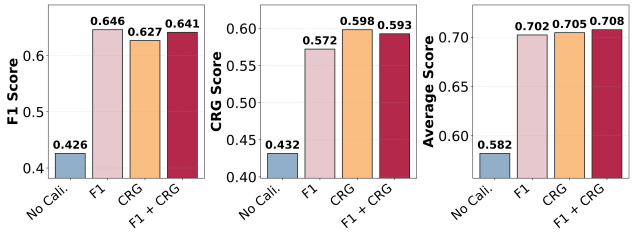

Figure 9. Evaluation metrics under different calibration strategies. Average denotes the mean of AUC, F1, and CRG scores.

Model Ensemble. Our final submission ensembles 10 models by training 10 separate plugins (LoRA adapters, task-query, and classification head) with the shared DINOv2

backbone. Each model uses identical training configurations but different data splits that preserve per-class prevalence according to the global distribution. During inference, we efficiently switch plugins with the loaded FM backbone and average the calibrated logits across all 10 models.

E. Winners of 3D Classification Challenges

We review top-performing solutions from recent 3D medical classification challenges. These winners emerge from highly competitive benchmarks and represent empirically validated strategies that outperformed strong baselines. We summarize key methodological takeaways from three consecutive RSNA-Kaggle challenges below.

E1. RSNA-Kaggle 2022 (Cervical Spine Fracture). The first-place solution [21] uses 2.5D CNN–RNN models for vertebra-level classification. For each vertebra, 15 slices are sampled along the z-axis and concatenated with neighboring slices and segmentation masks to form multi-channel 2D inputs. A 2D backbone (EfficientNet-V2-S [51] or ConvNeXt [35]) encodes each slice, and an LSTM head fuses features across slices for vertebra-level prediction. For patient-level prediction, the model jointly processes all seven vertebrae (7×15 slices total) through the same 2.5D CNN+LSTM architecture, with final predictions obtained via ensemble across backbones and folds.

E2. RSNA-Kaggle 2023 (Abdominal Trauma). The first-place solution [44] adopts a 2.5D slice-fusion approach. Each 96-slice volume is reorganized into 32 triplets of adjacent slices. A 2D CNN backbone (CoaT Lite [62] or EfficientNet-V2-S [51]) encodes each triplet, and a GRU sequence head models inter-slice dependencies. The model is trained with auxiliary segmentation heads and aggregates predictions via max pooling over slice logits.

E3. RSNA-Kaggle 2024 (Lumbar Spine Degenerative Classification). The first-place solution [6] employs a localize-then-classify pipeline. After 3D localization identifies level-wise coordinates, the classifier operates on multi-view crops (2.5D stacks of sagittal and axial slices). A 2D backbone (ConvNeXt-S [35] or EfficientNet-V2-S [51]) encodes each view, with a bidirectional LSTM and attention-based MIL fusing features across slices and views. Auxiliary attention losses regularize training, and ensemble predictions are obtained across backbones and folds.

Takeaways and Motivation. Winning solutions share common design patterns: (1) 2.5D representation via slice sampling or triplet formation, (2) 2D CNN backbones with explicit sequential modeling (LSTM/GRU or attention) for feature fusion, (3) auxiliary heads for training stabilization, and (4) multi-model ensembling for robust predictions. These findings demonstrate that 2D backbones with sequential fusion constitute the most effective approach for 3D medical classification in competitive settings. This motivates us to explore leveraging modern FMs' rich representations within the effective 2D+Fusion paradigm.

F. Detailed Ablation Studies

F1. Impact of Slice Fusion Strategy. We evaluate different strategies for aggregating slice-level features into volume-level predictions on T5 (Tab. 5). Simple pooling operations (average, max, median) require no additional parameters but treat slices independently without modeling inter-slice relationships. Sequential methods like LSTM [6] and Transformer encoder [40] explicitly model slice order but introduce substantial parameters (5.5M and 7.0M, respectively) and show mixed results, with LSTM underperforming (0.903 AUC) despite high parameter cost. Our query-based attention pooling achieves the best performance (0.962 AUC) with minimal parameters (0.001M), demonstrating that effective slice fusion does not require sequential modeling or a large parameter overhead. The learnable query automatically captures relevant cross-slice patterns through attention mechanisms, providing an optimal trade-off between performance and efficiency.

Table 5. Comparison of slice fusion strategies on T5.

Fusion Method	Sequential Modeling	Trainable Params (M)	AUC
Avg. pooling	✗	0	0.958
Max pooling	✗	0	0.946
Median pooling	✗	0	0.944
LSTM	✓	5.5	0.903
Transformer	✓	7.0	0.950
Ours (Query-based)	✗	**0.001**	**0.962**

F2. Impact of Backbone Sizes. We evaluate three DINOv3 backbone sizes: ViT-S (21M), ViT-B (86M), and ViT-L (300M) across three representative tasks (Tab. 6). Results demonstrate that larger backbones consistently yield better performance across all tasks. ViT-L achieves the highest AUC on T3 (0.988), T5 (0.962), and T7 (0.903), outperforming ViT-S by margins of 0.008, 0.029, and 0.022, respectively. The performance gains come at the cost of increased trainable parameters: ViT-L requires 1.20M parameters compared to 0.23M for ViT-S. Notably, all backbone sizes maintain trainable parameters under 1.5M, demonstrating the parameter efficiency of our approach. For practical deployment, we recommend starting with ViT-S for rapid iteration and computational efficiency, then scaling to ViT-B or ViT-L when higher performance is required and computational resources permit.

F3. Impact of DINO Versions. We compare DINOv2 and DINOv3 across four tasks (Tab. 7). Both versions

Table 6. Comparison of DINOv3 backbone sizes across tasks.

Backbone	Trainable Params (M)	AUC		
		T3	T5	T7
ViT-S (21M)	0.23	0.980	0.933	0.881
ViT-B (86M)	0.46	0.975	0.943	0.902
ViT-L (300M)	1.20	**0.988**	**0.962**	**0.903**

achieve comparable performance with negligible differences, and neither consistently outperforms the other. For 3D classification tasks, the choice between DINO versions appears inconsequential. This may be explained by DINOv3's primary improvement over DINOv2: the introduction of a Gram anchoring mechanism [48] that prevents the degradation of local (patch-level) features during long training periods. While this improvement does not translate to better performance for image-level tasks, i.e., classification, it may offer advantages for dense prediction tasks such as segmentation and detection that require fine-grained spatial features. Exploring DINOv3's potential for scalable 3D medical segmentation/detection remains a promising direction for future work, as discussed in our conclusion.

Table 7. Comparison of DINO versions on abdominal trauma classification tasks (T1-T4).

Version	AUC			
	T1	T2	T3	T4
DINOv2	0.956	0.914	0.988	0.957
DINOv3	0.954	0.922	0.984	0.953

G. Attention Heatmaps

We explore multiple explainability methods to generate interpretable heatmaps with our ViT-based framework (Fig. 10). Beyond raw attention maps from the last layer (shown in the main paper), we evaluate: (1) Attention Rollout [1], which aggregates attention weights across all transformer layers to trace information flow; (2) Gradient Attention Rollout, which weights attention maps by gradients of the predicted class to highlight decision-relevant regions; and (3) Gradient Attention Rollout (last layer), which applies gradient weighting only to the final layer.

Our comparison reveals three key observations. First, all methods successfully identify clinically relevant features, with high activation on pancreatic duct dilatation, a critical secondary sign for PDAC diagnosis. Second, Attention Rollout produces noisier heatmaps with diffuse activations, lacking class-specific guidance. Third, gradient-based methods generate more localized heatmaps by incorporating decision-level information, with the last-layer variant producing the most focused activations on discriminative anatomical structures. Overall, these methods offer complementary visualization options for AnyMC3D.

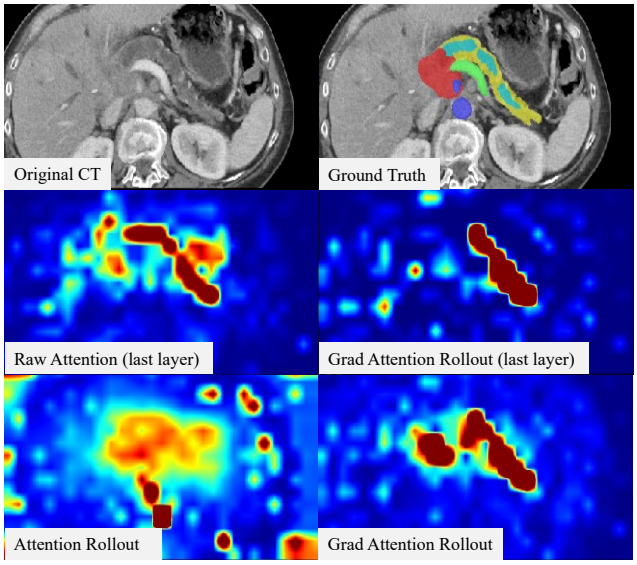

Figure 10. Heatmaps generated by different visualization methods.

H. Additional Evaluation Metrics

H1. Choice of Evaluation Metrics. We primarily use AUROC in the main paper because it evaluates ranking quality independent of classification thresholds, making it robust to class imbalance and enabling fair comparison across tasks with varying positive rates (2.23% to 78.34%). In contrast, accuracy, sensitivity, and specificity are threshold-dependent metrics that require operating point selection based on specific clinical priorities. For example, trauma screening may prioritize high sensitivity to avoid missing injuries, while diagnostic confirmation may require high specificity to reduce unnecessary interventions.

H2. Additional Metrics and Subgroup Analysis. Tab. 8 provides a comprehensive evaluation beyond AUROC, including accuracy, sensitivity, and specificity across trauma grading scenarios using Youden's J statistic for threshold selection. While the main paper reports AUROC for binary classification (injury vs. no injury, i.e., 0 vs. 1+2), we present granular performance by additionally separating low-grade (0 vs. 1) and high-grade (0 vs. 2) scenarios.

Our method demonstrates exceptional performance for severe injuries (0 vs. 2), achieving AUROC values of 0.9934 (Liver), 0.9949 (Kidney), and 0.9864 (Spleen), with perfect sensitivity (1.000) for Liver and Kidney. This indicates the framework reliably identifies high-grade trauma without missing positive cases. For the binary classification task (0 vs. 1+2), performance remains strong with accuracy of 0.862-0.964 and well-balanced sensitivity-

Table 8. Subgroup analysis of trauma organ injury grading.

Task	Grading	Sample Size # Pos | # Neg	AUROC	Accuracy	Sensitivity	Specificity
Bowel (T1)	0 vs. 1	21 | 915	0.9543	0.8996	0.9524	0.8984
Liver (T2)	0 vs. 1+2	94 | 846	0.9219	0.8617	0.8511	0.8629
	0 vs. 1	76 | 846	0.9049	0.8590	0.8158	0.8629
	0 vs. 2	18 | 846	0.9934	0.8657	1.0000	0.8629
Kidney (T3)	0 vs. 1+2	55 | 879	0.9842	0.9636	0.9636	0.9636
	0 vs. 1	34 | 879	0.9775	0.9628	0.9412	0.9636
	0 vs. 2	21 | 879	0.9949	0.9644	1.0000	0.9636
Spleen (T4)	0 vs. 1+2	109 | 829	0.9527	0.9318	0.8807	0.9385
	0 vs. 1	63 | 829	0.9281	0.9316	0.8413	0.9385
	0 vs. 2	46 | 829	0.9864	0.9383	0.9348	0.9385

specificity trade-offs, where Kidney achieves perfectly balanced metrics (0.9636 for all three).

Detection of low-grade injuries (0 vs. 1) proves more challenging, with sensitivity of 0.8158-0.9524, reflecting the inherent difficulty of identifying subtle trauma on CT imaging where findings may be ambiguous even to radiologists. Nevertheless, the method maintains high specificity (0.8629-0.9636) across all scenarios, correctly identifying non-injured cases and minimizing false alarms. These results confirm clinically relevant performance across the full spectrum of trauma severity with an appropriate sensitivity-specificity balance for different diagnostic scenarios.