A 3D Masked-Autoencoder CT Foundation Model under Linear Probing

Gautham Nandakumar[1][0009-0005-6622-5282]

Calico Life Sciences LLC, South San Francisco, CA, USA
gautham@calicolabs.com

Abstract. We present **MAE-ST**, our submission to Task 1 (Linear Probing, LP) **All-Data track** of the CVPR 2026 Challenge on Foundation Models for General CT Image Diagnosis. Our method is forked from the official MAE-ST codebase (`facebookresearch/mae_st`) and adapted for single-channel CT volumes. The central idea is to treat a 3D CT volume as a short *video*: the axial (Z) axis is interpreted as a temporal axis, and a ViT-Large spatiotemporal masked autoencoder (303.6 M parameters) is pre-trained from scratch on 10,819 unlabeled CT volumes drawn from the challenge pre-training corpus. Two design choices drive the representation: a *dynamic* masking ratio sampled uniformly from $[0.6, 0.9]$ on every forward pass, which exposes the encoder to a range of reconstruction difficulties rather than a single fixed one, and a coarse temporal patch size (4 slices) that concentrates capacity on in-plane anatomy. At inference the encoder is *frozen*: each volume is processed by overlapping 16-slice windows, and the layer-normalized mean of the patch tokens is averaged across windows to yield a single 1024-dimensional descriptor; for the eleven region-of-interest (ROI) targets the descriptor is restricted to the organ mask provided by the organizers. On our validation set (15 AMOS abdominal targets) the frozen features reach a **mean AUROC of** 0.678. On the final, held-out *test* set (17 binary targets across three datasets) the same system scores a **balanced accuracy of** 0.600 **and an AUROC of** 0.624 at 6.27 **s per case**, ranking 4th on the All-Data LP leaderboard. We additionally report a recipe-level ablation against an ablated baseline. Code and configurations are available at https://github.com/gautham-calico/cvpr-maest.

Keywords: CT foundation model · Linear probing · Masked autoencoder · Spatiotemporal modeling · Binary classification.

1 Introduction

Computed tomography (CT) is the workhorse of cross-sectional diagnostic imaging, and the past two years have produced a wave of CT *foundation models* that aim to provide a single reusable representation for many downstream diagnostic tasks [13,12,2,8,1]. Despite this rapid progress, two difficulties persist, and they

are precisely the ones this challenge stresses. The first is generalization across anatomy and acquisition: a representation tuned on one body region or one scanner population frequently degrades on another, and the challenge deliberately spans abdomen and chest across multiple source datasets. The second is severe class imbalance, since the downstream targets are binary disease labels whose positive prevalence is often well below 10%, so raw accuracy is uninformative and *balanced accuracy* and AUROC become the metrics of record.

Against this backdrop, we address **Task 1 – Linear Probing (LP)** of the challenge, in which the pre-trained backbone is kept *frozen* and only a lightweight linear classifier is trained on top of its features, with no backbone fine-tuning. This setting isolates the intrinsic discriminative power of a pre-trained representation: how well the features learned during self-supervised pre-training transfer, on their own, to binary CT disease classification. Our submission is built entirely for this Task 1 LP setting.

Prior work approaches CT representation learning from several directions that differ mainly in their source of supervision. *Vision-only* models learn from image content alone. VoCo [13] pre-trains a SwinUNETR with a volume-contrastive objective that predicts the spatial overlap between cropped sub-volumes, pulling anatomically aligned regions together in feature space; CT-FM [12] and the 3D framework of Xu et al. [14] scale contrastive and reconstruction pretext tasks to large unlabeled CT collections; and Curia [5] consolidates many radiology sources into a single multi-task backbone. *Vision-language* models add free-text supervision: Merlin [2], CT-CLIP [8], and Pillar-0 [1] learn a CLIP-style alignment between CT volumes and their radiology reports, which enables zero-shot abnormality detection and stronger supervised transfer, while the hybrid SPECTRE [4] combines self-supervised reconstruction with cross-modal alignment in a volumetric transformer.

Orthogonal to the supervisory signal is the choice of pretext mechanism and how the resulting backbone is adapted. The masked autoencoder [9] hides a large fraction of image patches and reconstructs them with an asymmetric encoder–decoder, learning representations without contrastive pairs or labels; its spatiotemporal extension MAE-ST [7] randomly masks a large fraction of space–time patches so that a video encoder must infer the missing content from the surrounding frames. Once pre-trained, a backbone must be adapted to a downstream decision: linear probing [3] trains only a linear classifier on frozen features and therefore reports their intrinsic quality, whereas LoRA [11] and adapters [10] inject a small number of trainable parameters when the backbone is allowed to move. Across these lines, the field is dominated by contrastive or report-aligned objectives and by CNN/Swin backbones; and while masked autoencoding on native volumetric ViTs is by now well studied for medical images, it is predominantly evaluated by *fine-tuning* for segmentation. The behavior of such a reconstructive ViT under a strict *frozen* linear-probe protocol, with anisotropic through-plane tokenization, is comparatively less charted.

This gap motivates our approach. A CT volume is, geometrically, a stack of axial slices with strong continuity along Z – exactly the structure that video

transformers exploit between frames. We therefore repurpose the spatiotemporal masked autoencoder MAE-ST [7] for CT, with the Z axis playing the role of time. Our implementation is a direct fork of the official MAE-ST codebase [7][1], adapted for single-channel CT volumes. This buys us two things: a masked reconstruction objective in which a large, randomly chosen fraction of space-time patches is dropped, forcing the encoder to infer the occluded anatomy from surrounding 3D context, and a ViT-Large backbone whose patch-token mean is a strong frozen descriptor. Our contributions are:

1. **A CT-adapted spatiotemporal MAE (MAE-ST)** pre-trained *from scratch* on the full 10,819-volume All-Data corpus, with two changes that matter empirically: a *dynamic* mask ratio $\sim \mathcal{U}[0.6, 0.9]$ (a range of reconstruction difficulties within one run) and a coarse temporal patch (4 slices).
2. **A frozen feature-extraction recipe** for arbitrary-depth volumes: overlapping 16-slice windows, layer-normalized token-mean pooling, and mask-guided ROI restriction for the eleven ROI targets – all running at 6.27 s/case, comfortably within the 40 s budget.
3. **An honest empirical study**: validation results (0.678 mean AUROC), the final test ranking (4th on the leaderboard, top six shown; 0.624 AUROC, 0.600 balanced accuracy), and a recipe-level ablation against an ablated baseline.

2 Method

Figure 1 shows the full pipeline: (a) self-supervised spatiotemporal MAE pre-training on unlabeled CT, and (b) frozen feature extraction feeding the organizer-run linear probe.

2.1 Self-supervised pre-training

CT-as-video formulation. A CT volume of shape (D, H, W) is normalized (Sec. 3.2) and a contiguous window of T=16 axial slices is resized to 16×224×224 and treated as a single-channel (C=1) clip. A 3D patch-embedding convolution with spatial kernel 16×16 and *temporal kernel* 4 (stride equal to kernel) produces a token grid of size $(T/4) \times (H/16) \times (W/16) = 4{\times}14{\times}14 = 784$ tokens of dimension 1024, to which a learnable class token and *separable* spatial/temporal positional embeddings are added. Because we adopt MAE-ST's spacetime-*agnostic random* masking, our pretext task is, in effect, identical to randomly masking 3D patches of a volume: whether the third axis is labelled "depth" or "time" does not change the token set or the masking distribution. We therefore make no claim that a video formulation *per se* differs from a native volumetric MAE. What distinguishes our backbone from the volumetric 3D MAEs common in medical imaging is inherited from the video model along two concrete axes: (i) *anisotropic patch embedding* – a 16×16 in-plane patch but a coarser temporal

[1] https://github.com/facebookresearch/mae_st

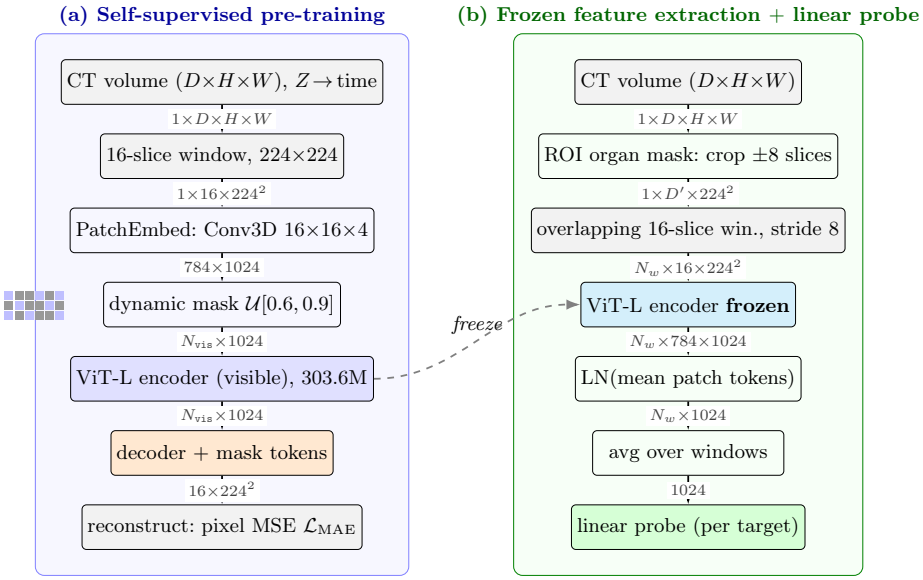

Fig. 1. Pipeline; tensor shapes on each arrow trace the data transformation (channel-first; N_{vis} visible tokens after masking, N_w stride-8 windows, D' cropped depth). **(a) Pre-training:** a CT volume is treated as a video ($Z\rightarrow$time); a random 16-slice window at 224^2 is patch-embedded with a 16×16 spatial / 4-slice temporal kernel into 4×14×14=784 tokens of width 1024, a dynamic fraction (60–90%, grey squares) is masked, the ViT-Large encoder sees only the N_{vis} visible tokens (blue squares), and a lightweight decoder reconstructs the masked voxels under a pixel-MSE loss. **(b) Feature extraction (frozen):** the trained encoder is frozen; a volume is tiled into N_w overlapping 16-slice windows (stride 8), each window yields a layer-normalized patch-token mean ($N_w\times 1024$), and the per-window vectors are averaged into a single 1024-d descriptor on which the organizers train a linear probe. For the eleven ROI targets the volume is first cropped to the provided organ mask (±8 context slices).

patch of 4 slices, tokenizing the through-plane (Z) axis at a different granularity than a cubic 16^3 patch; and (ii) *factorized* (separable) spatial and temporal positional embeddings rather than a single joint 3D embedding. We additionally depart from the prevailing fine-tuned-segmentation protocol by evaluating the encoder strictly *frozen* under linear probing.

Masked reconstruction. Following MAE [9,7], on each forward pass we sample a masking ratio $r \sim \mathcal{U}[0.6, 0.9]$ and randomly drop a fraction r of the 784 tokens. The ViT-Large *encoder* (24 blocks, 1024-d, 16 heads, mlp-ratio 4) processes only the visible tokens. An asymmetric *decoder* (8 blocks, 512-d, 16 heads) receives the encoded visible tokens plus shared mask tokens and reconstructs the masked content; the loss is the mean-squared error between predicted and ground-truth voxels, computed *only over masked patches* (raw-pixel target, no per-patch nor-

malization):

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| \hat{x}_i - x_i \right\|_2^2,$$

where $\mathcal{M}$ indexes masked patches. The temporal prediction dimension is 16, so the decoder reconstructs all 16 input slices.

Why dynamic masking. A fixed high ratio (e.g. 90%) leaves only ~ 157 visible tokens and makes *every* iteration maximally hard. Sampling $r \sim \mathcal{U}[0.6, 0.9]$ independently each step instead exposes the encoder to a range of reconstruction difficulties – from ~ 627 visible tokens (60%, rich context) to ~ 157 (90%, aggressive inpainting) – within a single training run, rather than committing to one fixed difficulty. The sampling distribution is stationary across training, so this is difficulty *mixing* rather than a scheduled easy-to-hard curriculum. This is consistent with recent evidence that exposing a masked autoencoder to a *range* of masking ratios, rather than a single fixed value, encourages multi-scale feature learning and can surpass even the best-tuned fixed ratio [6]. We expect it to be a major contributor to feature quality; rigorous isolation against the other six changes is left for future work (Sec. 3.4).

Backbone details. ViT-Large, patch 16×16 (spatial) × 4 (temporal), embedding dimension 1024, depth 24, 16 heads, mlp-ratio 4, separable spatio-temporal position embeddings and a class token. The frozen encoder loaded at inference has 303.6 **M parameters** (decoder and mask tokens are stripped from the checkpoint and are not shipped in the container).

2.2 Inference: frozen feature extraction and linear probing

At inference the decoder is discarded and the encoder is frozen; the system maps a full-resolution, arbitrary-depth CT volume to a single descriptor that the organizers' linear probe then classifies.

Global feature from the frozen backbone. Given a single 16-slice window, we run the frozen encoder *without* masking, drop the class token, take the *mean over the* 784 *patch tokens*, and apply the final encoder LayerNorm, yielding a 1024-d vector. We chose mean-pooled patch tokens over the class token because, under a reconstructive objective, the class token is never supervised, whereas patch tokens directly carry localized anatomical evidence.

Windowed aggregation for arbitrary-depth volumes. Clinical CTs have variable depth D (tens to hundreds of slices), but the encoder consumes exactly 16 slices. We therefore slide a 16-slice window along Z with **stride** 8 **(**50% **overlap)**, resize each window in-plane to 224^2 by trilinear interpolation, and *average* the per-window 1024-d descriptors into one volume-level vector. A final window anchored at the last slice guarantees full Z coverage when D is not a multiple of the stride. Overlapping windows make the descriptor robust to where structures fall relative

to window boundaries. Rather than encoding one window at a time, we stack 8 windows into a single GPU batch, which uses the device far more efficiently and speeds up per-case feature extraction.

ROI versus non-ROI targets. Eleven targets are ROI-based (organ-specific): for these the organizers provide a foreground organ mask, and we restrict feature extraction to the masked sub-volume; the remaining targets are treated globally (whole-volume descriptor). For an ROI target we read the provided mask, find the Z-range where the mask is active, expand it by ± 8 context slices, crop the volume to that range, and then apply the same windowed aggregation. This focuses the descriptor on the organ of interest (e.g. adrenal gland for *adrenal_hyperplasia*) while retaining peri-organ context. Cropping also leaves fewer 16-slice windows to pass through the ViT, which lowers per-case compute and runtime. If a mask is empty we fall back to the global descriptor.

Linear-probe head. The container emits only frozen 1024-d features; per challenge protocol, the *organizers* fit the linear probe. For internal model selection we replicate the standard LP recipe: L_2-normalize each descriptor and fit a per-target ℓ_2-regularized logistic regression with *class-balanced* weighting (to counter the severe positive/negative imbalance), sweeping the inverse regularization strength $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$ and selecting by validation AUROC. A linear probe has no capacity to compensate for a weak backbone, which is exactly why it is a faithful probe of feature quality.

3 Experiments

3.1 Dataset and evaluation metrics

The challenge provides a large-scale unlabeled CT pre-training set and a downstream train/validation set of 17 **binary classification targets across three datasets**, spanning abdomen and chest. For Task 1 we enter the **All-Data track**, pre-training from scratch on the full corpus. Our *internal* validation pipeline covers the 15 abdominal (AMOS-derived) targets – 11 ROI and 4 non-ROI; the two additional official targets are chest findings (a COVID-CT target and a lung-nodule-malignancy target) that are absent from our local split, so all internally reported per-disease numbers are *abdominal*.
Classification metrics. *Balanced accuracy* (mean of sensitivity and specificity), robust to imbalance, and *AUROC*. **Efficiency metric:** per-case running time in seconds, including container start-up; for Task 1 this is the average feature-extraction time per target.

3.2 Implementation details

Preprocessing. Each NIfTI volume is loaded and transposed to (D, H, W) with Z first. Hounsfield units are clipped to $[-1024, 3071]$ and linearly mapped

to $[0, 1]$; this single fixed window is used at both pre-training and inference, so train/test preprocessing is identical. During pre-training a random contiguous 16-slice window is drawn (short volumes are tiled to reach 16 slices) and resized in-plane to 224^2 by trilinear interpolation. At inference the deterministic overlapping-window scheme of Sec. 2.2 replaces the random crop. For fast large-scale loading we read NIfTI arrays lazily via memory-mapped `dataobj` access and apply all augmentation on the fly on $(1, T, H, W)$ tensors; no resampling to isotropic spacing is performed, which avoids an expensive global resample on every volume.

Environment settings. Development and runtime environments are summarized in Table 1.

Table 1. Development environments and requirements.

System	Ubuntu 20.04 LTS (kernel 5.4.0)
CPU	Intel Xeon Gold 6130 @ 2.10 GHz (2×16 cores, 64 threads)
RAM (pre-training)	96 GB (Slurm allocation)
RAM (inference)	container limit 32 GB (`-m 32G`)
GPU (pre-training)	3× NVIDIA Titan RTX 24 GB
GPU (inference)	1× NVIDIA GPU (`-gpus device=0`)
CUDA version	12.1
Programming language	Python 3.9
Deep learning framework	PyTorch 2.1.0
Key libraries	nibabel, h5py, timm, iopath, numpy

Training protocols. (1) Data augmentation. During pre-training, augmentations are applied consistently across all 16 slices of a clip: random horizontal flip (p=0.5), vertical flip (p=0.5), random 90/180/270° axial rotation, random intensity scale $\mathcal{U}[0.95, 1.05]$ and shift $\mathcal{U}[-0.05, 0.05]$, and additive Gaussian noise (p=0.3, σ=0.02), with values re-clamped to $[0, 1]$. Inference uses no augmentation (no test-time augmentation). **(2) Sampling.** Each epoch draws one random 16-slice window per volume (random start index), giving stochastic depth coverage across epochs. Repeat-augmentation (×4) replicates each sampled clip into four independently augmented views per iteration, so an optimizer step sees 96 augmented samples spanning the 24 unique clips of the effective batch. **(3) Model selection.** We extract features at candidate checkpoints (epochs 40 and 130) and select by internal validation mean AUROC; epoch 130 was submitted.

Table 2 gives the full pre-training and linear-head protocol. Figure 2 shows the pre-training reconstruction loss and learning-rate schedule: the masked-patch MSE drops sharply within the first few epochs and then decreases smoothly, while the learning rate follows the 40-epoch linear warm-up to 9.4×10^{-5} and the subsequent cosine decay. The submitted checkpoint (epoch 130) sits on the converged plateau.

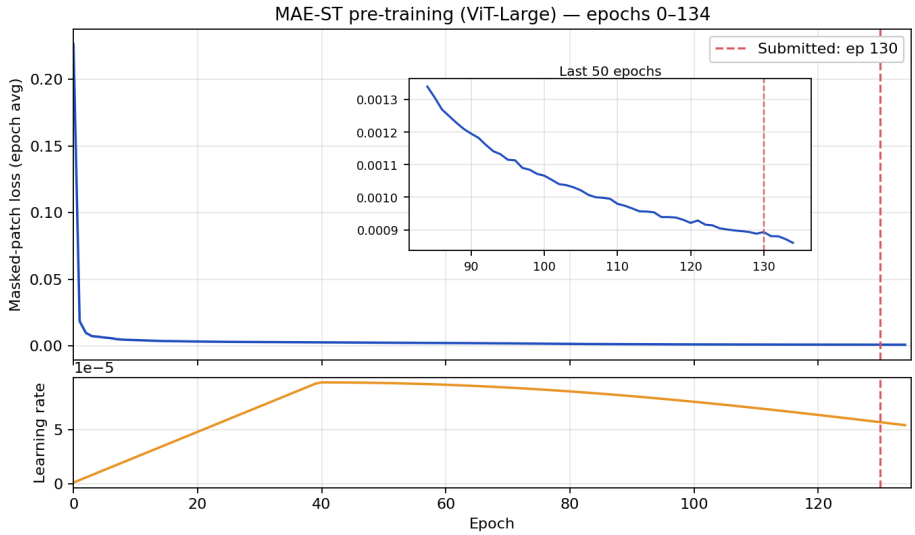

Fig. 2. MAE-ST (ViT-Large) pre-training. **Top:** per-epoch masked-patch reconstruction loss (inset: last 50 epochs, $\sim 9 \times 10^{-4}$); the dashed line marks the submitted epoch-130 checkpoint. **Bottom:** the per-epoch learning rate – linear warm-up over 40 epochs to 9.4×10^{-5} followed by cosine decay (linear-scaling rule, effective batch 24). Training ran to epoch 133 of 250 scheduled and was stopped on the converged plateau; axes are cropped to the trained range.

Table 2. Training protocols for Task 1 – Linear Probing (pre-training and linear-head training).

Pre-trained backbone (if any)	None – MAE-ST pre-trained from scratch (All-Data)
Backbone tuning strategy	Frozen
Batch size (pre-training)	24 effective (4/GPU $\times$ 2 accum $\times$ 3 GPU)
Patch / volume size	input $16 \times 224 \times 224$; patch $16 \times 16 \times 4$
Total pre-training epochs	250 scheduled; epoch 130 submitted
Linear-head epochs	Logistic regression, C-sweep (6 values), no SGD epochs
Optimizer	AdamW ($\beta = 0.9, 0.95$), weight decay 0.05
Initial learning rate (lr)	9.4×10^{-5} (blr $10^{-3} \times 24/256$)
Lr decay schedule	Cosine to 0, 40-epoch linear warmup
Pre-training time	$\approx$ 65.8 hours (1 node, 3$\times$ Titan RTX)
Linear-head training time	$<$ 1 min/target (logistic regression)
Loss function	Masked-patch pixel MSE
Number of model parameters	303.6 M (frozen encoder)

3.3 Quantitative results on the validation set

Validation performance. Under the frozen-LP protocol, MAE-ST attains a validation **mean AUROC of** 0.678 (mean balanced accuracy 0.638) over the 15 AMOS abdominal targets. This is higher than the 17-target *test* AUROC (0.624, Sec. 3.5) because the official test set additionally averages in two chest findings (COVID, lung-nodule malignancy) that are absent from our internal validation split.

Per-disease validation detail. Table 3 reports per-target AUROC and balanced accuracy on the class-balanced AMOS-15 validation set (epoch-130 submitted checkpoint). The per-target dynamic range is large (0.587 → 0.841 AUROC), which the mean hides: the strongest targets are organ-scale, high-contrast findings (splenomegaly, fatty liver), while the weakest are focal/cystic findings (liver cyst, liver calcifications, gallstone). Aggregated by region of interest, ROI and non-ROI targets are nearly tied (0.685 vs. 0.660 mean AUROC).

Table 3. Task 1 LP **All-Data**, **validation set**. Per-disease AUROC and balanced accuracy on the class-balanced AMOS-15 internal validation split (MAE-ST, epoch-130 submitted checkpoint; 303.6 M params). Targets sorted by AUROC; n_{val} is the number of balanced validation cases.

Target	Type	AUROC	BalAcc	n_{val}
splenomegaly	ROI	**0.841**	0.729	70
fatty_liver	ROI	0.825	0.679	78
cholecystitis	ROI	0.727	0.706	102
adrenal_hyperplasia	ROI	0.713	0.686	70
colorectal_cancer	non-ROI	0.695	0.673	52
hydronephrosis	ROI	0.692	0.703	64
lymphadenopathy	non-ROI	0.677	0.625	72
kidney_stone	ROI	0.676	0.633	158
ascites	non-ROI	0.655	0.647	68
renal_cyst	ROI	0.639	0.605	266
liver_lesion	ROI	0.620	0.522	134
atherosclerosis	non-ROI	0.612	0.603	68
gallstone	ROI	0.607	0.612	116
liver_calcifications	ROI	0.603	0.569	72
liver_cyst	ROI	0.587	0.578	244
All AMOS-15		**0.678**	**0.638**	15
ROI (11)		0.685	0.638	11
non-ROI (4)		0.660	0.637	4

3.4 Ablation study

Submitted recipe vs. baseline. Table 4 compares the submitted MAE-ST recipe against an ablated baseline that differs in nine training choices simultaneously

(dynamic vs. fixed 90% mask, raw-pixel vs. per-patch-normalized reconstruction target, coarse vs. fine temporal patch, decoder depth, repeated augmentation, base learning rate, gradient clipping, warmup/training length, and an expanded augmentation pipeline). The full recipe gains +0.051 AUROC (0.627 → 0.678), improving 12 of 15 targets. The largest per-target gains are on organ-scale findings: splenomegaly 0.663 → 0.841 (+0.178), liver calcifications 0.440 → 0.603 (+0.163, rescued from below chance), and atherosclerosis 0.493→0.612 (+0.119, raised above chance). Because the two configurations differ on nine axes, we cannot isolate the contribution of any single change from this comparison; per-axis attribution – in particular for dynamic masking and the raw-pixel loss target, which we expect to be major drivers – is left for follow-up work.

Table 4. Ablation: submitted MAE-ST recipe vs. an ablated baseline (AMOS-15 internal validation, mean AUROC). The two configurations differ in nine training choices listed below; per-design-choice attribution requires further runs (Sec. 3.6).

Configuration	Mean AUROC
Ablated baseline	0.627
MAE-ST submitted (ep130)	**0.678**
Differences (baseline → submitted): mask $0.9 \to \mathcal{U}[0.6, 0.9]$; loss target per-patch-normalized → raw pixels; t_p $2 \to 4$; decoder depth $4 \to 8$; repeat-aug $1 \to 4$; blr $1.5\times10^{-4} \to 10^{-3}$; clip-grad $0.02 \to 1.0$; warmup $10 \to 40$ ep; total $100 \to 250$ ep; augmentation pipeline expanded.	

Aggregation and ROI cropping (design rationale). Two further design choices were guided by internal observations; we summarize the rationale here. *(i) Multi-window overlap.* Overlapping (50%) windows with mean pooling were more stable than non-overlapping windows, particularly for findings that can fall between window boundaries. *(ii) ROI cropping.* For the eleven ROI targets, mask-guided cropping (organ range ±8 slices) concentrates the descriptor on the relevant organ and reduces runtime by shrinking the window list. We did not run a head-to-head comparison against whole-volume features, so we make no discrimination claim; ROI and non-ROI targets in fact reach comparable validation AUROC (0.685 vs. 0.660).

3.5 Results on the final testing set

Table 5 reports the official final-test ranking for the Task 1 LP All-Data track. Our team (**maest**) places 4th on the published leaderboard (top six shown) with 0.600 balanced accuracy and 0.624 AUROC at 6.27 s/case – the second-fastest runtime among the top four and roughly 5× faster than the slowest valid entry. The leading cluster (fomofo / raidium / cemrg) sits at 0.705–0.714 AUROC; we attribute most of the gap to cross-region (chest) generalization (Sec. 3.6).

Table 5. Official final-test ranking, Task 1 LP **All-Data** track (17 targets). Our submission is **maest**. Runtime is seconds per case including container start-up.

Rank	Team	Balanced Acc	AUROC	Runtime (s)
1	fomofo	0.654	0.714	9.414
1	raidium	0.639	0.707	4.393
3	cemrg	0.643	0.705	4.467
4	**maest (ours)**	**0.600**	**0.624**	**6.273**
5	xupy	0.566	0.599	10.860
6	sdfmu	0.160	0.170	33.830

3.6 Limitations and future work

Our encoder is pre-trained on a predominantly abdominal corpus and never sees the two chest targets in local validation; closing the gap to the leaderboard leaders (~ 0.09 mean AUROC) likely requires chest-CT pre-training data and region-balanced sampling. A single mean-pooled descriptor is a strong but blunt summary; attention pooling or multi-layer feature concatenation may recover focal-lesion signal lost to averaging. We resize every volume to 224×224 in-plane and treat the Z axis as time, discarding native pixel spacing in all three axes; the model therefore never sees true physical scale. Spacing-aware positional encodings or isotropic resampling that preserves the original x, y, and z resolution are natural extensions. Internal results are single-seed on a small validation split; multi-seed pre-training and bootstrap confidence intervals would tighten the claims. Finally, our ablation contrasts the submitted recipe with a single all-changes-flipped baseline; per-axis attribution requires one-change-at-a-time runs that did not fit the challenge timeline.

4 Conclusion

We framed a 3D CT volume as a short video and pre-trained a spatiotemporal masked autoencoder (MAE-ST, ViT-Large, 303.6 M parameters) from scratch on 10,819 unlabeled CT volumes for the All-Data linear-probing track. A dynamic masking ratio and a coarse temporal patch yielded a frozen 1024-d descriptor. Via overlapping-window aggregation and mask-guided ROI cropping, this descriptor achieved 0.678 mean AUROC on our abdominal validation set and 0.600 balanced accuracy / 0.624 AUROC at 6.27 s/case on the final test set, ranking 4th on the published All-Data LP leaderboard (top six shown). The main lesson is methodological: a purely reconstructive ViT with explicit through-plane modeling and a simple frozen-feature recipe is a competitive, fast, and reproducible CT foundation-model baseline, whose principal remaining weakness — cross-region (chest) generalization — chest-CT pre-training should directly address.

Acknowledgements. We thank the data owners for making the medical images publicly available and Codabench [15] for hosting the challenge platform.

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

References

1. Agrawal, K.K., et al.: Pillar-0: A new frontier for radiology foundation models. arXiv preprint arXiv:2511.17803 (2025) 1, 2
2. Blankemeier, L., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square (2024) 1, 2
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2
4. Claessens, C., et al.: Scaling self-supervised and cross-modal pretraining for volumetric CT transformers. arXiv preprint arXiv:2511.17209 (2025) 2
5. Dancette, C., et al.: Curia: A multi-modal foundation model for radiology. arXiv preprint arXiv:2509.06830 (2025) 2
6. Dong, M., Wang, L., Kluger, Y.: Understanding and enhancing mask-based pretraining towards universal representations. In: Advances in Neural Information Processing Systems (NeurIPS) (2025) 5
7. Feichtenhofer, C., Fan, H., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 2, 3, 4
8. Hamamci, I.E., et al.: A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. CoRR (2024) 1, 2
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2, 4
10. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning (ICML) (2019) 2
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022) 2
12. Pai, S., et al.: Vision foundation models for computed tomography. arXiv preprint arXiv:2501.09001 (2025) 1, 2
13. Wu, L., Zhuang, J., Chen, H.: 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 (CVPR) (2024) 1, 2
14. Xu, T., et al.: A generalizable 3d framework and model for self-supervised learning in medical imaging. npj Digital Medicine **8**(1), 639 (2025) 2
15. Xu, Z., et al.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 11