Pan-FM: A Pan-Organ Foundation Model with Saliency-Guided Masking for Missing Robustness

Qiangqiang Wu[1,2] **Grace McIlvain**[2] **Zhou Yu**[2] **Junhao Wen**[1,2*]

[1]Laboratory of AI and Biomedical Science (LABS)
[2]Columbia University

{qw2487,jw4745}@cumc.columbia.edu, {gm3128,zy2461}@columbia.edu

Abstract

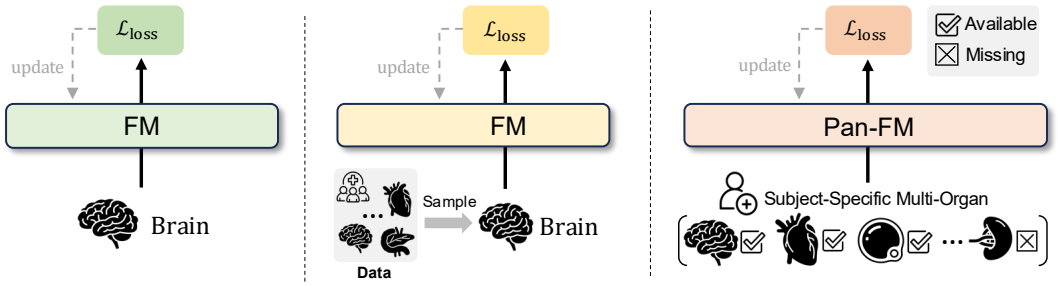
Foundation models (FMs) have shown great promise in medical imaging, but most FMs are trained on unimodal data within isolated domains, such as brain MRI alone. Human aging and disease arise through coordinated biological processes across organs, therefore motivating multimodal FMs that learn whole-body representations. A key challenge, however, is that real-world multimodal biomedical data are often missing not at random, which can reduce power, limit generalizability, and introduce bias. We propose Pan-FM, a pan-organ foundation model pre-trained on imaging from seven organs (*Brain*, *Heart*, *Adipose*, *Liver*, *Kidney*, *Spleen*, and *Pancreas*) under realistic missing-organ scenarios. Pan-FM uses a unified backbone that handles organ missingness during both training and inference, and is pre-trained with masking-based self-distillation. We find that naive multimodal pre-training leads to dominant-organ shortcut learning bias, with the model over-relying on dominant organs such as adipose and heart. To address this, we introduce Saliency-Guided Masking (SGM), which uses the model attention distribution to adaptively mask dominant organs during pre-training, thus encouraging more balanced cross-organ, whole-body learning. Notably, SGM introduces negligible computational overhead and can be seamlessly integrated into existing self-supervised learning frameworks to improve multi-organ representation learning. On the UK Biobank, Pan-FM achieves stronger prediction across 13 disease categories and 14 single disease entities than single-organ and multi-organ baselines, with improved robustness under missing-organ settings. Pan-FM serves as a scalable solution to realistic modality-missingness in multimodal learning in computational, systems and clinical neuroscience, and as a step toward more generalizable whole-body FMs. Codes are available at `https://github.com/LABS-lab/Pan-FM.git`.

1 Introduction

Foundation models (FMs) [2, 8, 32, 42, 45] pre-trained on large-scale biomedical data via self-supervised learning (SSL) have emerged as a powerful paradigm in medical AI, enabling transferable representations across a wide range of downstream clinical tasks. For example, recent efforts have developed FMs for the brain [40] and the heart [38], demonstrating strong performance on organ-specific disease prediction. However, most prior FMs have been trained in isolated domains using imaging data from a single organ system, thereby i) overlooking cross-organ interactions in aging and disease [17–20, 51, 52, 54] and neglecting the recent notion that "No Organ System Is An Island" [49], ii) reducing generalizability under missing-not-at-random settings (e.g., individuals undergoing brain MRI are typically healthier), and iii) limiting training efficiency and power by constraining to a

*Corresponding author

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



(a) Single-Organ FM (b) Independent Multi-Organ FM (c) **Ours**: Joint Multi-Organ FM

Figure 1: (a) Organ-specific FMs [38, 40] trained on data from a single organ system (e.g., brain); (b) Independent multi-organ FMs [45, 60] learned by sampling one organ sample per iteration from a multi-organ dataset, lacking subject-level multi-organ modeling; (c) Our proposed Pan-FM jointly learns cross-organ representations from subject-level multi-organ data, explicitly designed to handle missing-organ scenarios prevalent in biomedical research and clinical settings.

single modality given the data-hungry nature of FMs. This motivates the development of multimodal FMs that integrate multiple organ modalities to more fully capture disease etiology and aging biology.

As shown in Fig. 1, recent multi-organ learning methods [32, 45, 60] process each organ independently, learning multi-organ representations via single-organ training iterations rather than jointly modeling structured inter-organ dependencies. Moreover, they largely overlook missing-organ scenarios (missing-not-at-random), a clinically prevalent setting where imaging data for certain organs is unavailable due to incomplete protocols or acquisition constraints. Existing methods either assume full modality availability or address missingness post-hoc, yielding representations that are neither inherently robust nor biologically meaningful under organ absence. This motivates a unified pan-organ FM that jointly models cross-organ interactions and handles missing organ problems.

To achieve this, we propose Pan-FM, a multi-organ FM that jointly learns cross-organ representations from *in vivo* MRI data across 7 organ systems. We design a unified backbone that supports cross-organ missingness during both training and inference, and pre-train it via masking-based self-distillation. To mitigate dominant-organ shortcut learning bias during the naïve pre-training, where the model over-relies on a subset of organs, we introduce Saliency-Guided Masking (SGM), a simple and effective strategy that adaptively selects organs to mask based on the model's attention distribution. Specifically, SGM identifies organs that the model currently over-relies on and preferentially masks them during pre-training. This self-regulating mechanism progressively balances the model's reliance across more organs, while introducing negligible computational overhead and no architectural changes.

We use UK Biobank [9] as our pre-training data source, which includes MRI-derived organ-specific features across seven organ systems in 41,969 participants, including *Brain*, *Heart*, *Adipose*, *Liver*, *Kidney*, *Spleen*, and *Pancreas*. To demonstrate the pre-training effectiveness of the proposed Pan-FM, we perform downstream classification for organ-specific multi-disease categories (e.g., all diseases related to the central nervous system) and single disease prediction (e.g., Alzheimer's disease). Pan-FM consistently outperforms single-organ and multi-organ baselines, with particularly pronounced advantages on systemic diseases that closely interact with multiple organ systems. Under missing-organ evaluation settings, Pan-FM shows superior robustness compared to multi-organ baselines.

In summary, the main contributions of our work are:

- We present Pan-FM, to the best of our knowledge, the first multi-organ FM for jointly learning cross-organ representations from 7 organ systems in the presence of missing organs.
- We identify dominant-organ shortcut learning bias as a key limitation in multi-organ SSL, and propose Saliency-Guided Masking (SGM), a plug-and-play strategy that facilitates more balanced cross-organ representation learning with negligible overhead and no architectural changes, and integrates seamlessly into existing SSL frameworks.
- We demonstrate that Pan-FM achieves improved performance and missing-organ robustness on downstream disease prediction tasks, including prediction for single disease endpoints and organ-specific disease categories, compared to various baseline approaches.

- We enable a flexible foundation framework that supports both hypothesis-driven and hypothesis-free research in system and clinical neuroscience. In the hypothesis-driven setting, prior knowledge can guide model construction; for example, if Alzheimer's disease is thought to involve the brain-heart-eye axis, Pan-FM can be instantiated on these three organs to test related hypotheses. In the hypothesis-free setting, Pan-FM can be scaled to learn an organ-agnostic, whole-body foundation model with broad clinical generalizability, as demonstrated in the current work.

2 Related Work

Multi-Organ Aging and Disease Analysis. Prior work characterizes multi-organ aging and disease primarily through organ-specific *biological age gaps* (BAGs), defined as the difference between predicted biological age and chronological age derived from imaging or physiological features. Early studies established BAGs across brain and body systems and linked them to disease and mortality [41, 54], with follow-up work extending the framework to MRI-based multi-organ age gaps [20], proteomics [50, 51], metabolomics [17], and the recent *pan-disease* paradigm [19], alongside cross-organ axis studies [18, 31]. Together these efforts outline a roadmap toward a multi-organ *medical digital twin* [52]. Yet they treat BAGs as intermediate variables combined via post-hoc late fusion, which cannot model inter-organ dependencies at the representation level, and assume complete multi-organ acquisitions, misaligned with clinical practice where organ missingness is common. This motivates Pan-FM, which performs multi-organ learning in a unified backbone and remains robust under missing-organ scenarios.

Self-Supervised Learning. Self-supervised learning (SSL) serves as the foundation for FM development. Contrastive learning methods use positive and negative pairs [13, 15, 24, 36, 56]. Negative-free SSL methods rely on architectural asymmetry [3, 4, 14, 23] or covariance regularization [7, 61]. Recent methods use teacher–student momentum updates [10, 34] with self-distillation, and masked image modeling reconstructs masked patches [5, 25, 57, 58, 65]. While recent extensions explore adaptive image-patch masking strategies [27, 28, 44], Pan-FM extends them to the organ level via organ saliency-guided multinomial sampling, which mitigates dominant-organ shortcut learning under missing-organ scenarios. Building on these SSL methods, many biomedical FMs have emerged for brain MRI [6, 30, 37, 40], heart MRI [62], chest imaging [42, 64], retinal imaging [39, 48, 66], pathology [11, 12, 43, 46], ultrasound [26], echocardiography [16], and dermatology [59], together with general-purpose medical foundation models [29, 47, 55, 63]. Despite these advances, most FMs remain organ-specific. While some works [45, 60] extend to multi-organ MRI, they rely on independent organ training and fail to capture cross-organ dependencies. Moreover, current FMs assume complete and homogeneous inputs, which is unrealistic in clinical settings with organ missingness. To the best of our knowledge, Pan-FM is the first multi-organ FM that learns whole-body representations under missing-organ scenarios.

3 Methodology

This section presents Pan-FM, a multi-organ foundation model that jointly learns cross-organ representations across seven organ systems and is designed to handle biologically meaningful missing modalities (miss-not-at-random), a common challenge in system neuroscience and large-scale biobanks where heterogeneous organ data are only partially collected at the population level.

3.1 Architecture

The overall architecture of Pan-FM is shown in Fig. 2. Specifically, Pan-FM treats each organ's MRI features as semantic token patches, and integrates them via a Transformer backbone [22] to facilitate cross-organ representation learning. Modality availability is explicitly encoded and learnt, allowing organ missingness to serve as structured input with biological meaning rather than being treated as discarded missing-at-random noise.

Cross-Attention based Tabular Patch Encoding (CAPE). A key challenge in our multi-organ representation learning setting is the dimensional imbalance across organ systems. Hand-crafted low-dimensional MRI features vary substantially across organs: the *Brain* consists of 122 volumetric regions of interest (ROIs) while the *Pancreas* only has 3 features (i.e., *Brain* > *Heart* > *Adipose* >

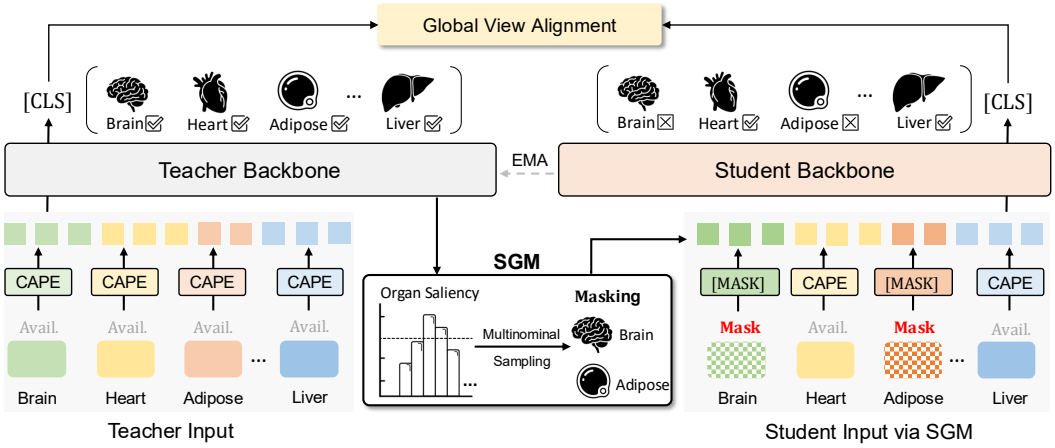


Figure 2: Overview of Pan-FM pre-training with Saliency-Guided Masking (SGM). The teacher backbone receives the full set of available organ tokens as input, while the student backbone receives a masked view constructed by SGM. Both backbones share the same architecture.

Liver = *Kidney* = *Spleen* = *Pancreas*, see Appendix B). To alleviate this, we employ a Cross-Attention Patch Encoder (CAPE) that converts each organ's features into a compact set of semantic tokens with a unified embedding dimension. Formally, for each organ o, organ-specific imaging features $\mathbf{x}_o \in \mathbb{R}^{D_o}$ are projected into feature embeddings $\mathbf{E}_o \in \mathbb{R}^{D_o \times d}$, and aggregated into K_o compact tokens via cross-attention with learnable queries $\mathbf{Q}_o \in \mathbb{R}^{K_o \times d}$:

$$\mathbf{z}_o = \text{softmax}\left(\frac{\mathbf{Q}_o \mathbf{W}_Q (\mathbf{E}_o \mathbf{W}_K)^T}{\sqrt{d_{\text{head}}}}\right) \mathbf{E}_o \mathbf{W}_V \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices and d_{head} is the attention head dimension. Each learnable query acts as a summary slot that attends to the most informative feature combinations within that organ. The resulting tokens $\mathbf{z}_o \in \mathbb{R}^{K_o \times d}$ share a unified dimension d across all organs, enabling direct concatenation with a global `[CLS]` token for cross-organ Transformer interaction.

Missing Organ Learner. A defining characteristic of large-scale biobank data collections is systematic and biologically meaningful whole-body organ incompleteness. In practice, participants frequently have missing imaging data for one or more organ systems due to study design, contraindications, or selection bias, such that missingness itself carries biological meaning and can be explicitly modeled across organ systems. Pan-FM is designed to operate under arbitrary organ subsets during both training and inference. To achieve this, we define an organ availability vector $\mathbf{a} \in \{0, 1\}$, where $a_o = 1$ indicates that organ o is observed. For each organ o, CAPE produces a token set $\mathbf{z}_o \in \mathbb{R}^{K_o \times d}$ if observed, and substitutes a learnable mask token $[\texttt{MASK}]_o$ otherwise:

$$\tilde{\mathbf{z}}_o = a_o \cdot \mathbf{z}_o + (1 - a_o) \cdot [\texttt{MASK}]_o + \mathbf{p}_o \tag{2}$$

where $\mathbf{p}_o \in \mathbb{R}^{K_o \times d}$ is a learnable organ-specific embedding added to all organs regardless of availability, providing the Transformer backbone with explicit position and organ information.

3.2 Naïve Pre-training via Masking-based Self-Distillation

DINOv2 [34] has demonstrated remarkable success in learning transferable visual representations through self-distillation pre-training. Here, we adapted a DINOv2-style self-distillation for multi-organ representation learning to our multi-organ imaging features under real-world organ-missing scenarios. Specifically, we use a teacher model to receive the full set of available organs for each participant, while the student model receives an additional masked view produced by random organ masking, i.e., for each observed organ o, its full token set $\mathbf{z}_o$ is randomly replaced by $[\texttt{MASK}]_o$. Teacher parameters are updated via exponential moving average (EMA).

Pre-training Objective. The original DINOv2 pre-training objective includes three losses: 1) A global alignment loss $\mathcal{L}_{\text{g}}$ encourages the student's `[CLS]` token to match the teacher's full-view

representation; 2) A masked token distillation loss $\mathcal{L}_{\text{m}}$ drives the student to reconstruct missing organ tokens from cross-organ context; 3) A KoLeo regularization term $\mathcal{L}_{\text{k}}$ encourages a uniform `[CLS]` embedding distribution. The overall loss is the combination of all the three losses.

Dominant-Organ Shortcut Learning Bias. We observe an intriguing phenomenon in the naïve DINOv2-style [34] pre-training on our multi-organ data. That is, the model consistently concentrates its attention on a small subset of organs (see Fig. 3), which we term *dominant-organ shortcut learning bias*. Notably, the most dominant organ in our setting is *Adipose*, despite contributing only 16 tokens — neither the largest token budget nor the highest feature dimensionality among the seven organ systems. This phenomenon causes the model to learn shortcut representations that over-rely on *dominant organs*. As a result, as shown in Fig. 3, the model exhibits poor robustness when dominant organs are absent at inference time.

We attribute this to two complementary factors: 1) From a biological perspective, *Adipose*-derived features such as visceral and subcutaneous fat volumes are strong proxies for systemic metabolic state, exhibiting high cross-organ predictive capacity for cardiovascular, hepatic, and endocrine conditions; 2) *Adipose* tokens provide sufficient cross-organ information to effectively construct a representative global `[CLS]` embedding, such that the model can minimize $\mathcal{L}_{\text{global}}$ by attending more to Adipose, bypassing the need to integrate the other organs.

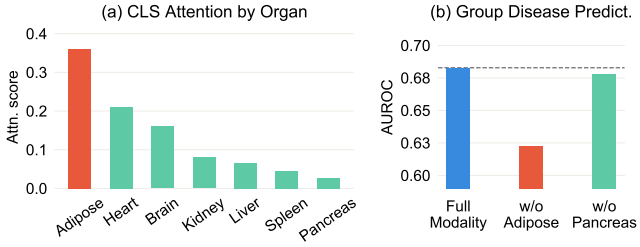


Figure 3: **Dominant-organ shortcut learning bias. (a)** Teacher CLS attention scores across organ systems. Adipose receives disproportionately high attention. **(b)** Mean group disease AUROC obtained by linear probing under organ removal. Removing adipose causes substantial degradation whereas removing pancreas has negligible effect.

3.3 Saliency-Guided Masking

Motivation. A primary goal of Pan-FM is to remain effective in real-world clinical settings, where patients may lack imaging data for one or more organ systems. However, as identified in Section 3.2, the naïve pre-training leads to dominant-organ reliance, where the model disproportionately depends on a small subset of organs for feature extraction. This fundamentally limits clinical applicability, i.e., when several dominant organs are unavailable at inference time, model performance degrades significantly. In addition, this organ dominance can obscure biologically meaningful signals from less dominant organs (e.g., in the hypothesis-free setting for whole-body FMs), limiting the model's utility not only for downstream prediction but also for biological and mechanistic discovery using computational approaches such as genome-wide associations (GWAS) and post-GWAS analyses [20].

To mitigate dominant-organ shortcut learning and move toward a whole-body FM, we propose Saliency-Guided Masking (SGM), which dynamically identifies the most dominant organs and preferentially masks them during pre-training, forcing the student to extract discriminative features from less dominant organs and gradually yielding balanced cross-organ representations (see Appendix Fig. 15).

Optimization Formulation. We formulate the organ masking strategy as a minimax optimization over the global self-distillation objective. Let $M \subseteq \mathcal{O}_{\text{avail}}$ denote a set of available organs to mask, and $\Delta(\mathcal{M})$ denote the probability simplex over the collection of all valid masking subsets $\mathcal{M} = \{M \subseteq \mathcal{O}_{\text{avail}} : 1 \leq |M| \leq n_{\max}\}$, where $n_{\max} = \min(\lfloor |\mathcal{O}_{\text{avail}}| \cdot r_{\text{mask}} \rfloor, |\mathcal{O}_{\text{avail}}| - 1)$ ensures at least one organ remains visible to the student, and r_{mask} is the masking ratio. We define $\mathcal{O}_{\text{miss}} = \mathcal{O} \setminus \mathcal{O}_{\text{avail}}$ as the set of naturally missing organs. The student and teacher tokens for organ o are constructed as:

$$\tilde{\mathbf{z}}_o^s = \begin{cases} [\texttt{MASK}]_o + \mathbf{p}_o, & o \in \mathcal{O}_{\text{miss}} \cup M, \\ \mathbf{z}_o + \mathbf{p}_o, & \text{otherwise}, \end{cases} \qquad \tilde{\mathbf{z}}_o^t = \begin{cases} [\texttt{MASK}]_o + \mathbf{p}_o, & o \in \mathcal{O}_{\text{miss}}, \\ \mathbf{z}_o + \mathbf{p}_o, & \text{otherwise}, \end{cases} \tag{3}$$

where $[\texttt{MASK}]_o$ is a learnable mask embedding and $\mathbf{p}_o$ is the organ-ID embedding defined in Section 3.2. We train a student backbone $f_S(\cdot;\theta)$ that minimizes the self-distillation loss under the adversarial masking distribution:

$$\min_{\theta} \max_{P(M) \in \Delta(\mathcal{M})} \mathbb{E}_{M \sim P(M)} \left[\mathcal{L}_g \left(f_S\left(\{\tilde{\mathbf{z}}_o^s\}_{o=1}^O; \theta\right), f_T\left(\{\tilde{\mathbf{z}}_o^t\}_{o=1}^O\right) \right) \right] \tag{4}$$

Algorithm 1 Pan-FM PyTorch Pseudocode

```
# f_s, f_t: student/teacher backbones
# m: EMA momentum; n_mask: masking budget
# tau: softmax temperature; lambda: KoLeo loss weight
for x, a in loader: # full-view input, availability
    with no_grad:
        z_t, attn = f_t.forward(x, a) # teacher feed-forward
    # SGM
    S = organ_saliency(attn, organ_spans, a) # Eq.(5)-(6)
    M = torch.Multinomial(softmax(S / tau), n_mask) # Eq.(7)
    x_s = mask_organs(x, M, a, mask_tokens) # student masking view
    z_s = f_s.forward(x_s, a) # student feed-forward w/ masked x_s
    loss = L_g(z_s, z_t) + lambda * L_k(z_s)  # overall loss
    # optimization step
    loss.backward(); optimizer.step()
    f_t.params = m * f_t.params + (1-m) * f_s.params
```

where $\mathcal{L}_g$ is the global alignment loss defined in Section 3.2, $f_T(\cdot)$ is the fixed teacher encoder. The inner maximization seeks the adversarial masking distribution — one that removes the organs most critical to the teacher's global representation, constructing the hard partial view for the student. The outer minimization then optimizes the student to align with the teacher's full-view representation under this adversarial mask, forcing it to leverage complementary information from other available organ systems rather than relying on some dominant organs.

Computational Challenge. Directly solving the inner maximization in (4) requires evaluating the loss over all $|\mathcal{M}|$ masking subsets at every training step, which is computationally intractable. We therefore seek a tractable approximation to the adversarial masking distribution $P^*(M)$.

Attention-Based Approximation. We observe that the optimal inner solution $P^*(M)$ concentrates probability mass on masks that remove the most informative organs — those that contribute most to the teacher's global `[CLS]` representation. We propose to approximate organ importance using the teacher's own attention maps, which are already computed in the forward pass at no additional cost.

Formally, let $\mathbf{a}_l^h \in \mathbb{R}^N$ denote the `[CLS]`-to-patch attention weights from block l and head h, where $N = \sum_{o=1}^{O} K_o$ is the total number of patch tokens. After the teacher's forward pass on the full view, we average across all L blocks and H heads to obtain a per-token attention map, and aggregate over each organ's token span to compute its saliency score:

$$\bar{\mathbf{a}} = \frac{1}{LH} \sum_{l=1}^{L} \sum_{h=1}^{H} \mathbf{a}_l^h \in \mathbb{R}^N, \qquad S_o = \sum_{j \in \text{organ}(o)} [\bar{\mathbf{a}}]_j. \tag{5}$$

The organ-specific saliency scores are then converted into a masking probability distribution via softmax with temperature τ:

$$P_o = \frac{\exp(S_o/\tau)}{\sum_{k \in \mathcal{O}_{\text{avail}}} \exp(S_k/\tau)} \tag{6}$$

where unavailable organs are excluded by setting $S_o = -\infty$ before the softmax. The temperature τ interpolates between two extremes: $\tau \to \infty$ recovers uniform random masking, while $\tau \to 0$ concentrates all probability mass on the single most dominant organ.

Organ Masking. Given the masking probabilities in Eq. (6), the set of organs to mask is drawn without replacement via multinomial sampling:

$$M \sim \text{Multinomial}(\{P_o\}_{o \in \mathcal{O}_{\text{avail}}}, n_{\text{mask}}, \text{replace=False}) \tag{7}$$

where $n_{\text{mask}} \sim \text{Uniform}\{1, \ldots, n_{\max}\}$ follows the same sampling budget as the random masking baseline, ensuring that the only difference between SGM and the uniform random masking baseline lies in the organ masking selection distribution.

The complete pre-training procedure is summarized in Algorithm 1. SGM introduces negligible computational overhead (see Appendix E.1), as organ saliency is derived directly from the teacher's pre-computed attention maps. Moreover, we demonstrate that SGM is a plug-and-play masking strategy that can be easily incorporated into existing SSL frameworks [7, 13, 34, 61] for improving multi-organ learning, as validated in Table 4.

Table 1: **Group-disease evaluation under linear probing (LP) and full fine-tuning (FT).** Mean AUROC across 13 disease groups on the held-out test set. **Bold** and underlined denote the best and second-best within each regime. **Full LP/FT BalAcc results are in Appendix Tables 10 and 12.**

Disease Group	**SimCLR** [13]		**BYOL** [23]		**MoCo-v3** [15]		**VICReg** [7]		**Barlow T.** [61]		**DINOv2** [34]		**Pan-FM**	
	LP	FT	LP	FT	LP	FT	LP	FT	LP	FT	LP	FT	LP	FT
Infectious & Parasitic	0.627	0.629	0.614	0.626	0.648	0.634	0.645	0.642	0.637	0.621	0.657	0.672	**0.668**	**0.694**
Neoplasms	0.628	0.629	0.618	0.622	0.634	0.638	0.625	0.629	0.625	0.637	0.629	0.646	**0.671**	**0.686**
Blood & Immune	0.659	0.660	0.635	0.649	**0.672**	**0.674**	0.635	0.636	0.646	0.650	0.657	0.656	0.666	0.669
Endocrine & Metabolic	0.692	0.703	0.682	0.689	0.708	0.712	0.680	0.694	0.692	0.701	0.700	0.710	**0.713**	**0.718**
Mental & Behavioural	0.673	0.657	0.649	0.645	0.665	0.657	0.649	0.644	0.665	0.656	0.675	0.680	**0.690**	**0.683**
Nervous System	0.667	0.664	0.655	0.655	0.668	0.665	0.645	0.654	0.647	0.653	0.655	0.665	**0.678**	**0.681**
Eye	0.661	0.653	0.611	0.613	0.632	0.631	0.605	0.622	0.603	0.622	0.644	0.661	**0.672**	**0.674**
Circulatory System	0.688	0.688	0.660	0.668	**0.699**	0.702	0.680	0.691	0.686	0.692	0.681	0.702	0.693	**0.704**
Respiratory System	0.646	0.649	0.644	0.651	0.657	0.663	0.644	0.652	0.654	0.660	0.654	0.673	**0.667**	**0.678**
Digestive System	**0.638**	0.633	0.615	0.615	0.632	0.636	0.611	0.617	0.624	0.626	0.619	0.642	0.635	**0.654**
Skin System	0.629	0.623	0.616	0.612	0.630	0.623	0.622	0.621	0.622	0.620	0.621	0.630	**0.671**	**0.666**
Musculoskeletal	0.653	0.650	0.639	0.640	**0.661**	0.664	0.640	0.644	0.643	0.647	0.644	0.667	0.659	**0.669**
Genitourinary	0.636	0.637	0.633	0.631	0.647	0.642	0.644	0.642	0.648	0.650	0.635	0.653	**0.659**	**0.673**
Overall Mean (AUROC)	0.653	0.652	0.636	0.640	0.658	0.657	0.640	0.645	0.646	0.649	0.652	0.666	**0.672**	**0.681**
Overall Mean (BalAcc)	0.622	0.619	0.606	0.607	0.625	0.625	0.612	0.618	0.615	0.618	0.623	0.632	**0.638**	**0.641**

Table 2: **Missing-organ robustness under linear probing (LP) and full fine-tuning (FT).** AUROC comparison of SSL methods across 13 disease groups on the held-out test set. *Standard* uses the full test set with natural missingness; *Full Organs (7)* and *Drop Organ* settings are evaluated on subjects with complete 7-organ imaging data, with random organ dropout repeated over 50 (LP) / 25 (FT) runs to ensure stable and fair assessment. Δ denotes the relative improvement of Pan-FM over the DINOv2 baseline. **BalAcc results and per-regime full tables are in Appendix Tables 11 and 13.**

Method	Standard		Full Organs (7)		Drop 1 Organ		Drop 2 Organs		Drop 3 Organs	
	LP	FT	LP	FT	LP	FT	LP	FT	LP	FT
SimCLR [13]	0.653	0.652	0.659	0.664	0.645	0.652	0.638	0.645	0.625	0.636
BYOL [23]	0.636	0.640	0.647	0.653	0.642	0.645	0.645	0.648	0.628	0.634
MoCo-v3 [15]	0.658	0.657	0.660	0.667	0.653	0.658	0.655	0.662	0.642	0.647
VICReg [7]	0.640	0.645	0.647	0.650	0.640	0.644	0.639	0.642	0.624	0.625
Barlow Twins [61]	0.646	0.649	0.642	0.643	0.632	0.633	0.632	0.630	0.618	0.619
DINOv2 [34]	0.652	0.666	0.683	0.691	0.654	0.681	0.647	0.668	0.617	0.645
Pan-FM (Ours)	**0.672**	**0.681**	**0.694**	**0.704**	**0.681**	**0.696**	**0.672**	**0.687**	**0.654**	**0.675**
Δ (%)	+3.1%	+2.3%	+1.6%	+2.0%	+4.1%	+2.1%	+3.9%	+2.7%	+6.0%	+4.6%

4 Experiments

Dataset. We conduct experiments on the UK Biobank [9], leveraging MRI-derived features from 41,969 participants across 7 organ systems (*Brain*, *Heart*, *Adipose*, *Liver*, *Kidney*, *Spleen*, *Pancreas*). To the best of our knowledge, UK Biobank is the most comprehensive multi-organ imaging resource in the biomedical field. For conceptualizing Pan-FM, we use low-dimensional MRI features extracted by reproducible image preprocessing pipelines, such as brain volumetric ROIs and left ventricular wall thickness based on the American Heart Association (AHA) protocol [33]. Since UK Biobank is a general-population cohort with disease labels derived from clinical history, the goal of this work is not to establish state-of-the-art disease prediction, but to show the technical value of Pan-FM in handling realistic missing-organ scenarios in real-world biobank data and in improving performance over baseline approaches. More dataset details are shown in Appendix A.1.

Data split and evaluation protocol. We hold out 500 cognitively normal (CN) participants and 780 disease-positive participants, identified using ICD codes and clinical history, for evaluation. We study two downstream protocols: (i) Group-disease prediction, covering 13 disease categories as binary classification against CN and evaluated with linear probing (LP) and full-backbone fine-tuning (FT); and (ii) Single-disease prediction, covering 14 specific diseases with smaller positive cohorts, where positives are identified from the full dataset and the same 500 held-out CN serve as negatives. We report 10-run averages for group-disease prediction, and 10-fold cross-validation results for single-disease prediction. Disease definitions, ICD mappings, and split statistics are provided in Appendices A.2 and A.3. Performance is measured by AUROC and balanced accuracy (BalAcc).

Implementation details. We pre-train Pan-FM for 200 epochs with a 12-layer ViT [22] following DINOv2 [34] (batch size 64, AdamW, lr=10^{-4}, hidden dim 384, 8 heads). SGM introduces a single hyperparameter, i.e., the temperature $\tau = 0.25$, with ablations shown in Fig. 6(b). For downstream

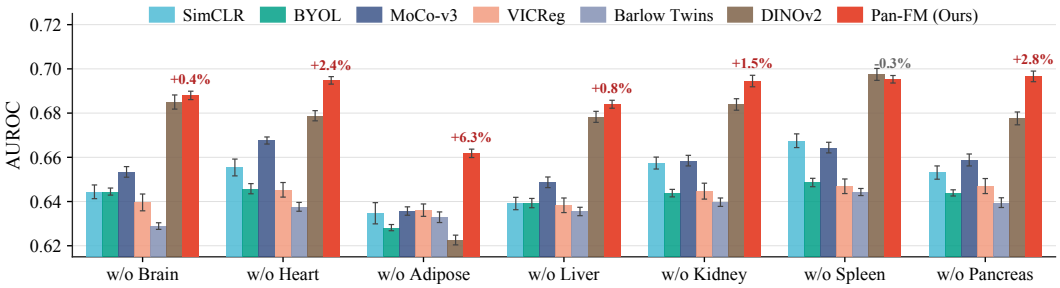


Figure 4: **Single-organ dropout evaluation with linear probing.** Comparison of SSL methods for group-disease prediction across seven single-organ dropout (test-time organ removal) settings. Percentages denote the relative improvement of Pan-FM over the DINOv2 baseline.

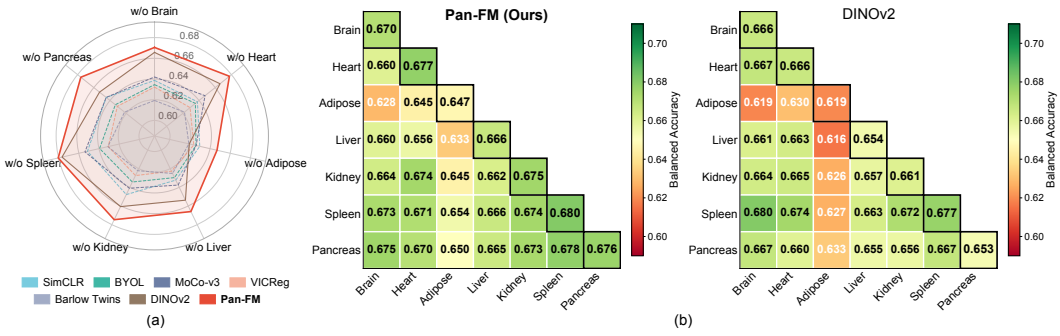


Figure 5: **Robustness under organ dropout with full backbone fine-tuning.** (a) Radar chart of balanced accuracy across seven single-organ dropout settings. (b) Pairwise organ-dropout heatmaps for Pan-FM (left) and DINOv2 (right). Each cell shows the mean balanced accuracy across 13 disease groups when the row and column organs are simultaneously removed; diagonal entries correspond to single-organ removal. **Full radar charts and pairwise heatmaps under both LP and FT regimes (AUROC and BalAcc) are provided in Appendix Figures 7–12.**

evaluation, we attach a linear head to the `[CLS]` token for both linear probing (LP) and full fine-tuning (FT). Additional details are provided in Appendix C.3.

SSL baselines. Since no existing SSL method is tailored to multi-organ representation learning, we compare against representative general-purpose SSL methods [7, 13, 15, 23, 34, 61]. For fairness, all baselines use the same backbone and main pre-training settings as Pan-FM, with method-specific hyperparameters tuned separately. All methods follow identical downstream protocols, with baseline details provided in Appendix C.1.

4.1 Group-Disease Prediction

Overall performance under natural missingness. Table 1 reports group-disease prediction under natural missingness with linear probing (LP) and full-backbone fine-tuning (FT), with detailed per-disease results in Appendix Tables 10 and 12. Pan-FM achieves the best overall performance in both regimes (LP: 0.672/0.638; FT: 0.681/0.641), outperforming DINOv2 and the randomly initialized FT baseline (see Appendix Table 12). Contrastive SSL methods such as SimCLR and MoCo-v3 generally outperform non-contrastive alternatives, suggesting that explicit instance discrimination is beneficial across heterogeneous organ systems. Pan-FM's gains are especially pronounced for systemic conditions such as *Neoplasms* (0.671 vs. 0.629 LP) and *Skin System* (0.671 vs. 0.621 LP), highlighting our Pan-FM's stronger pan-organ representations.

Missing-organ robustness. Table 2 evaluates robustness under natural missingness (*Standard*) and controlled organ dropout on complete-7-organ test subjects. As more organs are removed, all methods degrade, but Pan-FM remains consistently more robust: its AUROC margin over DINOv2 increases from +1.6%/+2.0% under *Full Organs* to +6.0%/+4.6% under *Drop-3* for LP/FT, respectively.

Table 3: **Single-disease evaluation with linear probing.** Comparison of SSL methods for single-disease prediction. Results are reported as AUROC / BalAcc.

Disease	**SimCLR** [13]	**BYOL** [23]	**MoCo-v3** [15]	**VICReg** [7]	**Barlow T.** [61]	**DINOv2** [34]	**Pan-FM**
Alzheimer's Disease	0.663 / 0.739	0.685 / 0.770	0.699 / 0.775	0.667 / 0.748	0.657 / 0.750	0.719 / **0.782**	**0.726** / 0.779
Parkinson's Disease	0.553 / 0.628	0.514 / 0.623	0.569 / 0.642	0.532 / 0.619	0.586 / 0.660	0.598 / 0.652	**0.647 / 0.689**
Multiple Sclerosis	**0.609 / 0.644**	0.598 / 0.638	0.591 / 0.638	0.569 / 0.634	0.535 / 0.607	0.547 / 0.612	0.583 / 0.637
Motor Neuron Disease	0.547 / 0.709	0.576 / 0.741	0.535 / 0.704	0.487 / 0.682	0.500 / 0.702	0.581 / 0.731	**0.594 / 0.742**
Systemic Lupus Eryth.	**0.681** / 0.732	0.614 / 0.695	0.605 / 0.710	0.594 / 0.690	0.560 / 0.682	0.562 / 0.684	0.675 / **0.756**
Sjogren Syndrome	0.659 / 0.711	0.618 / 0.681	0.652 / 0.695	0.629 / 0.679	0.591 / 0.680	0.665 / 0.706	**0.695 / 0.725**
Systemic Sclerosis	0.694 / 0.794	0.620 / 0.735	0.596 / 0.719	0.683 / 0.777	0.671 / 0.758	0.664 / 0.762	**0.719 / 0.804**
Sarcoidosis	0.569 / 0.649	0.553 / 0.645	0.589 / 0.655	0.511 / 0.600	0.482 / 0.596	0.549 / 0.625	**0.590 / 0.656**
Crohn's Disease	0.526 / 0.595	0.520 / 0.601	**0.579 / 0.631**	0.520 / 0.586	0.549 / 0.598	0.503 / 0.586	0.537 / 0.606
Ulcerative Colitis	0.534 / 0.582	0.557 / 0.599	**0.566 / 0.601**	0.527 / 0.581	0.549 / 0.600	0.563 / 0.592	0.552 / 0.592
Primary Biliary Chol.	0.535 / 0.747	0.556 / 0.763	0.622 / 0.793	0.473 / 0.708	0.555 / 0.768	0.636 / 0.795	**0.658 / 0.805**
Type 2 Diabetes	0.777 / 0.733	0.762 / 0.723	**0.808 / 0.754**	0.771 / 0.736	0.758 / 0.719	0.797 / 0.752	0.793 / 0.750
Hypertrophic Cardiom.	0.715 / 0.763	0.729 / 0.793	0.828 / 0.855	0.757 / 0.806	0.726 / 0.776	0.875 / 0.880	**0.893 / 0.900**
Pulmonary Hypertens.	**0.739 / 0.831**	0.571 / 0.747	0.661 / 0.765	0.575 / 0.754	0.580 / 0.763	0.556 / 0.736	0.695 / 0.819
Overall Mean	0.629 / 0.704	0.605 / 0.697	0.636 / 0.710	0.592 / 0.686	0.593 / 0.690	0.630 / 0.707	**0.668 / 0.733**

Table 4: Missing-organ evaluation with linear probing by applying SGM to various SSL methods.

Method	Standard		Full Organs (7)		Drop 1 Organ		Drop 2 Organs		Drop 3 Organs	
	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc
SimCLR [13]	0.653	0.622	0.659	0.637	0.645	0.629	0.638	0.627	0.625	0.617
SimCLR + **SGM**	**0.671**	**0.635**	**0.682**	**0.656**	**0.663**	**0.639**	**0.656**	**0.639**	**0.648**	**0.635**
VICReg [7]	0.640	0.612	0.647	0.631	0.640	0.626	0.639	0.627	0.624	0.619
VICReg + **SGM**	**0.665**	**0.630**	**0.679**	**0.651**	**0.661**	**0.642**	**0.650**	**0.637**	**0.648**	**0.638**
Barlow Twins [61]	0.646	0.615	0.642	0.624	0.632	0.620	0.632	0.622	0.618	0.613
Barlow Twins + **SGM**	**0.666**	**0.633**	**0.674**	**0.653**	**0.659**	**0.640**	**0.646**	**0.632**	**0.635**	**0.626**
DINOv2 [34]	0.652	0.623	0.683	0.654	0.654	0.642	0.647	0.639	0.617	0.616
DINOv2 + **SGM (Pan-FM)**	**0.672**	**0.638**	**0.694**	**0.670**	**0.681**	**0.662**	**0.672**	**0.655**	**0.654**	**0.638**

Figs. 4 and 5(a) show no pronounced weak single-organ dropout direction for Pan-FM, while Fig. 5(b) localizes DINOv2's failures mainly to *Adipose* and *Pancreas*. Although FT improves severe-missingness performance over LP, the Pan-FM–DINOv2 gap persists in both regimes, indicating that the robustness gain comes from SGM rather than the adaptation protocol.

Single-disease prediction. Table 3 reports linear-probing results on 14 specific diseases. Pan-FM achieves the best overall performance (0.668/0.733), outperforming MoCo-v3 (0.636/0.710) and DINOv2 (0.630/0.707) across most diseases. The gains are most pronounced for multi-organ conditions, including *Parkinson's Disease* (0.647 vs. 0.598) and *Sjogren Syndrome* (0.695 vs. 0.665), indicating that Pan-FM learns system-level representations that improve disease-specific prediction.

4.2 External Evaluation on ADNI

We evaluate our UK Biobank pre-trained Pan-FM on ADNI for Alzheimer's disease versus cognitively normal classification, using brain features only. ADNI differs from UK Biobank in country, scanner fleet, acquisition protocol, and recruitment, and it is disease-enriched rather than a general population cohort. Because ADNI provides only brain imaging, 6 of the 7 organs are absent for our Pan-FM, which is substantially more challenging than the standard organ-drop setting in Table 2.

Pan-FM naturally handles the single-organ input with the masking mechanism. As defined in Sec. 3.1, available organs are represented by their CAPE-encoded feature tokens, whereas unavailable organs are represented by their learned mask tokens together with the corresponding organ embeddings. For ADNI, we simply set the availability vector to $a = (1, 0, 0, 0, 0, 0, 0)$, the six missing organs use their mask tokens, and the backbone and CAPE stay the same as in pre-training. So the brain-only AD classification on ADNI is handled as a standard missing-organ configuration rather than as a special architectural case. For the brain input we use the same 119 MUSE gray matter volumes.

Following [53], we define the AD and CN groups in ADNI based on the diagnosis at the baseline visit and require the diagnosis to remain stable throughout follow-up. We randomly select 100 AD and 100 CN subjects to construct a class-balanced independent held-out test set. The remaining subjects form the development set, on which we perform 5-fold cross-validation. Pan-FM transfers effectively to ADNI despite the domain shift and the absence of six organs. Using brain features alone, it achieves 0.921 ± 0.007 test balanced accuracy, outperforming DINOv2 (0.896 ± 0.012), training from scratch (0.865 ± 0.009), and classical baselines. For a fair comparison, we perform the same learning-rate grid search for all deep models and select hyperparameters using validation performance only. These

Table 5: Validation and test balanced accuracy (BalAcc) on ADNI. Test performance is reported across five folds.

Backbone	Method	Val BalAcc	Test BalAcc (5 folds)	Test BalAcc (mean ± std)
MLP	Raw MUSE features	0.870 ± 0.015	0.890, 0.865, 0.895, 0.915, 0.905	0.894 ± 0.017
SVM	Raw MUSE features	0.879 ± 0.023	0.900, 0.885, 0.920, 0.905, 0.905	0.903 ± 0.011
ViT	Training from scratch	0.885 ± 0.011	0.870, 0.870, 0.875, 0.850, 0.860	0.865 ± 0.009
ViT	DINOv2 pre-trained on UKBB	0.893 ± 0.024	0.890, 0.890, 0.890, 0.920, 0.890	0.896 ± 0.012
ViT	**Pan-FM pre-trained on UKBB**	$\mathbf{0.905 \pm 0.011}$	**0.910, 0.930, 0.915, 0.925, 0.925**	$\mathbf{0.921 \pm 0.007}$

results demonstrate effective single-organ transfer on ADNI without architectural modification and improved robustness under extreme organ missingness.

4.3 Ablation Studies

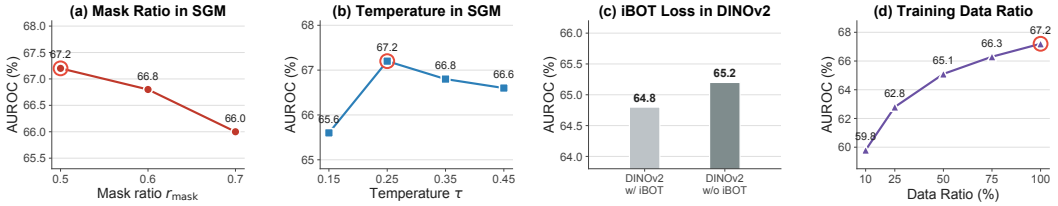


Figure 6: **Ablation studies.** (a) Mask ratio r_{mask} in SGM. (b) Temperature τ in SGM. (c) Effect of the iBOT loss in DINOv2 for multi-organ representation learning. (d) Effect of downstream training data ratio. All results are mean AUROC (%) across 13 disease groups under linear probing.

Mask ratio r_{mask}. Since masking is applied only to participants with at least two available organs (Sec. 3.3), the budget starts from $r_{\text{mask}} = 0.5$. Fig. 6(a) shows performance degrades from 67.2 at 0.5 to 66.0 at 0.7. Unlike patch-level masking on natural images where 0.75 is tractable due to rich local context, full-organ masking removes entire anatomical units and requires a smaller ratio.

Temperature τ. The temperature in Eq. 6 controls the sharpness of the organ-sampling distribution: small τ concentrates probability on a few dominant organs, while large τ collapses to uniform random masking. Fig. 6(b) shows $\tau = 0.25$ achieves the best trade-off.

iBOT loss in DINOv2. The iBOT loss, designed for masked patch reconstruction on natural images, is ill-suited here: full-organ masking leaves insufficient context to reconstruct fine-grained organ-specific features. Fig. 6(c) shows iBOT slightly degrades performance (64.8 vs. 65.2). We therefore remove it from both Pan-FM and DINOv2, reporting the stronger iBOT-free variant throughout.

Downstream training data ratio. Fig. 6(d) shows performance scales with the amount of downstream labeled training data, improving consistently from 59.8 at 10% to 67.2 at full scale with no saturation, indicating that Pan-FM representations continue to benefit from additional supervision.

SGM as a plug-and-play strategy. SGM can be integrated into other multi-view SSL frameworks. As shown in Table 4, applying SGM to SimCLR, VICReg, and Barlow Twins consistently improves both standard and missing-organ performance, demonstrating that SGM generalizes beyond the DINOv2 framework. Detailed per-disease improvements are shown in Appendix Table 14.

5 Conclusion

We introduced Pan-FM, *the first* pan-organ foundation model pre-trained on 7 organs' MRI features under realistic missing-organ scenarios. Critically, we proposed Saliency-Guided Masking (SGM) in Pan-FM, which addresses the shortcut learning bias observed in naive multimodal pre-training, and can be plugged into prior existing multi-view models for superior performance. On the UK Biobank, Pan-FM consistently outperforms single-organ and missing-modality baselines across 13 disease groups and 14 individual diseases, with especially large gains under severe organ missingness. These results establish SGM as an effective remedy to modality shortcut bias and position Pan-FM as a step toward scalable, whole-body FM for systems neuroscience and clinical applications.

6 Acknowledgments

This work was supported by the NIH-funded MULTI Consortium (RF1AG092412; PI: J.W.) and Cure Alzheimer's Fund (PI: J.W.). We want to express our sincere gratitude to the UK Biobank team for their invaluable contribution to advancing clinical research in our field (https://www.ukbiobank.ac.uk/). The MULTI consortium (J.W.; UK Biobank Application Number: 647044) aims to integrate multi-organ imaging and multi-omics data to advance our understanding of human aging and disease mechanisms. We gratefully acknowledge use of the research computing resources of the Empire AI Consortium, Inc, with support from Empire State Development of the State of New York, the Simons Foundation, and the Secunda Family Foundation.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4190–4197, 2020.

[2] Kumar Krishna Agrawal, Longchao Liu, Long Lian, Michael Nercessian, Natalia Harguindeguy, Yufu Wu, Peter Mikhael, Gigin Lin, Lecia V Sequist, Florian Fintelmann, et al. Pillar-0: A new frontier for radiology foundation models. *arXiv preprint arXiv:2511.17803*, 2025.

[3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

[4] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. 2021.

[6] Carlo Alberto Barbano, Matteo Brunello, Benoit Dufumier, Marco Grangetto, Alzheimer's Disease Neuroimaging Initiative, et al. Anatomical foundation models for brain mris. *Pattern Recognition Letters*, 2025.

[7] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. 2022.

[8] Cameron Beeche, Joonghyun Kim, Hamed Tavolinejad, Bingxin Zhao, Rakesh Sharma, Jeffrey Duda, James Gee, Farouk Dako, Anurag Verma, Colleen Morse, et al. A pan-organ vision-language model for generalizable 3d ct representations. *medRxiv*, 2025.

[9] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[11] Richard J. Chen, Chengkuan Chen, Yicong Li, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, 2022.

[12] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[14] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[16] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5):1481–1488, 2024.

[17] MULTI consortium, Filippos Anagnostakis, Sarah Ko, Mehrshad Saadatinia, Jingyue Wang, Christos Davatzikos, and Junhao Wen. Multi-organ metabolome biological age implicates cardiometabolic conditions and mortality risk. *Nature Communications*, 16(1):4871, 2025.

[18] MULTI Consortium, Aleix Boquet-Pujadas, Filippos Anagnostakis, Michael R Duggan, Cassandra M Joynes, Arthur W Toga, Zhijian Yang, Keenan A Walker, Christos Davatzikos, and Junhao Wen. Brain–heart–eye axis revealed by multi-organ imaging genetics and proteomics. *Nature Biomedical Engineering*, pages 1–23, 2025.

[19] MULTI Consortium, Aleix Boquet-Pujadas, Filippos Anagnostakis, Zhijian Yang, Ye Ella Tian, Michael R Duggan, Guray Erus, Dhivya Srinivasan, Cassandra M Joynes, Wenjia Bai, et al. Multi-organ ai endophenotypes chart the heterogeneity of brain, eye and heart pan-disease. *Nature Mental Health*, pages 1–28, 2026.

[20] MULTI Consortium, Huizi Cao, Zhiyuan Song, Michael R Duggan, Guray Erus, Dhivya Srinivasan, Ye Ella Tian, Wenjia Bai, Michael S Rafii, Paul Aisen, et al. Mri-based multi-organ clocks for healthy aging and disease assessment. *Nature Medicine*, 32(1):82–92, 2026.

[21] Jimit Doshi, Guray Erus, Yangming Ou, Susan M Resnick, Ruben C Gur, Raquel E Gur, Theodore D Satterthwaite, Susan Furth, Christos Davatzikos, Alzheimer's Neuroimaging Initiative, et al. Muse: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *Neuroimage*, 127:186–195, 2016.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in neural information processing systems*, pages 21271–21284, 2020.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[26] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis*, 96:103202, 2024.

[27] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318. Springer, 2022.

[28] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022.

[29] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[30] Moona Mazher, Geoff JM Parker, and Daniel C Alexander. Towards generalisable foundation models for brain mri. *arXiv preprint arXiv:2510.23415*, 2025.

[31] Celeste McCracken, Zahra Raisi-Estabragh, Michele Veldsman, Betty Raman, Andrea Dennis, Masud Husain, Thomas E Nichols, Steffen E Petersen, and Stefan Neubauer. Multi-organ imaging demonstrates the heart-brain-liver axis in uk biobank participants. *Nature Communications*, 13(1):7839, 2022.

[32] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.

[33] American Heart Association Writing Group on Myocardial Segmentation, Registration for Cardiac Imaging:, Manuel D Cerqueira, Neil J Weissman, Vasken Dilsizian, Alice K Jacobs, Sanjiv Kaul, Warren K Laskey, Dudley J Pennell, John A Rumberger, Thomas Ryan, et al. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. *Circulation*, 105(4):539–542, 2002.

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[35] Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209, 2010.

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

[37] Daniel Scholz, Ayhan Can Erdur, Viktoria Ehm, Anke Meyer-Baese, Jan C Peeken, Daniel Rueckert, and Benedikt Wiestler. Mm-dinov2: Adapting foundation models for multi-modal medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 320–330. Springer, 2025.

[38] Rohan Shad, Cyril Zakka, Dhamanpreet Kaur, Mrudang Mathur, Robyn Fong, Joseph Cho, Ross Warren Filice, John Mongan, Kimberly Kallianos, Nishith Khandwala, et al. A generalizable deep learning system for cardiac mri. *Nature Biomedical Engineering*, pages 1–16, 2026.

[39] Danli Shi, Weiyi Zhang, Jiancheng Yang, Siyu Huang, Xiaolan Chen, Pusheng Xu, Kai Jin, Shan Lin, Jin Wei, Mayinuer Yusufu, et al. A multimodal visual–language foundation model for computational ophthalmology. *npj digital medicine*, 8(1):381, 2025.

[40] Divyanshu Tak, Biniam A Garomsa, Anna Zapaishchykova, Tafadzwa L Chaunzwa, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Suraj Pai, Sri Vajapeyam, et al. A generalizable foundation model for analysis of human brain mri. *Nature Neuroscience*, pages 1–12, 2026.

[41] Ye Ella Tian, Vanessa Cropley, Andrea B Maier, Nicola T Lautenschlager, Michael Breakspear, and Andrew Zalesky. Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nature medicine*, 29(5):1221–1231, 2023.

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

[43] Eugene Vorontsov, Alican Bozkurt, Adam Casson, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 2024.

[44] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10375–10385, 2023.

[45] Shansong Wang, Mojtaba Safari, Qiang Li, Chih-Wei Chang, Richard LJ Qiu, Justin Roper, David S Yu, and Xiaofeng Yang. Triad: Vision foundation model for 3d magnetic resonance imaging. *arXiv preprint arXiv:2502.14064*, 2025.

[46] Xiyue Wang, Sen Yang, Jun Zhang, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 2022.

[47] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. 2022.

[48] Hao Wei, Bowen Liu, Minqing Zhang, Peilun Shi, and Wu Yuan. Visionclip: An med-aigc based ethical language-image foundation model for generalizable retina image analysis. *arXiv preprint arXiv:2403.10823*, 2024.

[49] Junhao Wen. Multiorgan biological age shows that no organ system is an island. *Nature Aging*, 4(9):1182–1183, 2024.

[50] Junhao Wen. Multi-omics and multi-organ aging clocks digitize human aging. *medRxiv*, 2025. doi: 10.1101/2025.02.06.25321803.

[51] Junhao Wen. Refining the generation, interpretation and application of multi-organ, multi-omics biological aging clocks. *Nature Aging*, 5(9):1897–1913, 2025.

[52] Junhao Wen. Towards a multi-organ, multi-omics medical digital twin. *Nature Biomedical Engineering*, 9(9):1386–1389, 2025.

[53] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63:101694, 2020.

[54] Junhao Wen, Ye Ella Tian, Ioanna Skampardoni, Zhijian Yang, Yuhan Cui, Filippos Anagnostakis, Elizabeth Mamourian, Bingxin Zhao, Arthur W Toga, Andrew Zalesky, et al. The genetic architecture of biological age in nine human organ systems. *Nature aging*, 4(9):1290–1307, 2024.

[55] 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.

[56] Qiangqiang Wu, Jia Wan, and Antoni B Chan. Progressive unsupervised learning for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2993–3002, 2021.

[57] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14561–14571, 2023.

[58] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[59] Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, 2025.

[60] Zhijian Yang, Noel DSouza, Istvan Megyeri, Xiaojian Xu, Amin Honarmandi Shandiz, Farzin Haddadpour, Krisztian Koos, Laszlo Rusko, Emanuele Valeriano, Bharadwaj Swaminathan, et al. Decipher-mr: a vision-language foundation model for 3d mri representations. *npj Digital Medicine*, 2026.

[61] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[62] Chi Zhang, Michael Loecher, Cagan Alkan, Mahmut Yurt, Shreyas S Vasanawala, and Daniel B Ennis. On the foundation model for cardiac mri reconstruction. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 226–235. Springer, 2024.

[63] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024.

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

[65] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. 2021.

[66] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622 (7981):156–163, 2023.

Appendix Contents

A Data Details **16**

A.1 Dataset 16

A.2 Group-Disease Prediction 17

A.3 Single-Disease Prediction 18

B Multi-Organ Input Features **19**

C More Implementation Details **19**

C.1 Self-Supervised Learning (SSL) Baselines 19

C.2 Pre-training Details 21

C.3 Downstream Task 21

D More Results **22**

D.1 Group-Disease Prediction 22

D.2 Quantifying improvements across disease categories 24

D.3 Applying SGM for More SSL Methods 24

D.4 Single Organ Dropout 25

D.5 Pairwise Organ Dropout 25

D.6 Group-Disease Comparison with More Missing Organs 26

D.7 Comparison with Single-Organ Baselines 27

D.8 Comparison with Random (From-Scratch) Baseline 27

E More Ablation Studies **28**

E.1 Computational Overhead 28

E.2 Pre-training Convergence 28

E.3 Saliency Redistribution Under SGM 29

E.4 Organ Saliency Proxy for SGM 30

E.5 Organ Importance Analysis 33

E.6 Effect of Pre-training Loss Components 34

E.7 Downstream Training Data Ratio 34

F Limitations and Future Work **35**

A Data Details

In this section, we detail the data splits for both pre-training and downstream tasks, including group-disease and single-disease prediction.

A.1 Dataset

We conduct experiments on the UK Biobank [9], a large-scale population cohort comprising approximately 500,000 participants recruited between 2006 and 2010. UK Biobank provides the most comprehensive population-scale resources for studying multi-organ biomedical phenotypes.

In this work, we use MRI-derived, organ-specific imaging-derived phenotypes (IDPs) from 41,969 participants, covering seven organ systems: *Brain*, *Heart*, *Adipose*, *Liver*, *Kidney*, *Spleen*, and *Pancreas*.

For conceptualizing our Pan-FM in this work, we use low-dimensional MRI-derived features extracted from standardized image-processing pipelines, including brain volumetric regions of interest and cardiac measurements such as left-ventricular wall thickness following the American Heart Association (AHA) segmentation protocol. This design enables us to systematically study multi-organ representation learning under heterogeneous organ availability, while leaving direct pre-training on raw multi-organ imaging data as an important direction for future work.

Of note, the UK Biobank is a general-population cohort, and disease labels are derived from inpatient records and clinical history rather than disease-specific phenotyping. As a result, absolute predictive performance is expected to be modest. Accordingly, the goal of this section is not to establish state-of-the-art disease prediction, but to demonstrate the technical value of Pan-FM in handling realistic missing-organ scenarios in real-world biobank data and in improving performance over baseline approaches.

A.2 Group-Disease Prediction

Table 6: **Definition of the 13 systemic disease categories via ICD-10.** A participant is assigned a positive label for a category if any of their recorded ICD-10 diagnoses fall within the specified range.

Disease Category	**ICD-10 Range**
Infectious & parasitic diseases	B
Neoplasms	C–D4
Blood & immune system diseases	D50–D89
Endocrine, nutritional & metabolic diseases	E
Mental & behavioural disorders	F
Nervous system diseases	G
Eye diseases	H0–H5
Circulatory system diseases	I
Respiratory system diseases	J
Digestive system diseases	K
Skin system diseases	L
Musculoskeletal system diseases	M
Genitourinary system diseases	N

Disease Category Split. Following [18], disease categories are defined following the International Classification of Diseases, 10th Revision (ICD-10). We exclude Ear diseases (H6–H9) due to its insufficient positive cases, retaining 13 categories for downstream training and evaluation. Specifically, for each UK Biobank participant, we extract hospital inpatient records (field 41270) and self-reported diagnoses and map each ICD-10 code to its corresponding chapter. A participant is assigned a positive label for a given disease category if they have *any* recorded diagnosis whose ICD-10 code falls within the ranges defined in Table 6. Table 6 lists the 13 systemic disease categories used in our group-disease prediction task. These span a broad range of organ systems and pathological mechanisms.

Disease Category Distribution. We partition the UK Biobank dataset into four splits to support both self-supervised pre-training and downstream disease prediction. As shown in Table 7, the **held-out test set** (N=1,280) is constructed first by randomly sampling 500 cognitively normal (CN) participants together with 780 disease-positive participants, stratified across the 13 ICD-10 categories (60 cases per category). Because a single participant may carry diagnoses spanning multiple categories (i.e., comorbidities), the total positive count aggregated across categories exceeds 780 in the test set, as reflected in Table 7. This test set is strictly excluded from both pre-training and training. The **pre-training set** (N=40,689) comprises all remaining participants with at least one organ imaged, including CN, disease-positive, and unlabeled individuals, thereby maximizing data utilization for self-supervised learning. For downstream evaluation, we further filter the pre-training pool to retain only participants with confirmed CN or disease-positive status, yielding the **training set** (N=36,220), from which 5% is held out as the **validation set** (N=1,907) via stratified sampling on CN status.

Table 7: **Disease category distribution across data splits.** We report the number of positive cases per ICD-10-defined disease category in the downstream training, validation, and held-out test sets.

Disease Category	Train	Val	Test
Infectious & parasitic diseases	1,897	89	137
Neoplasms	9,953	522	340
Blood & immune system diseases	2,532	121	148
Endocrine & metabolic diseases	8,083	410	332
Mental & behavioural disorders	3,243	165	154
Nervous system diseases	4,014	206	184
Eye diseases	1,162	53	90
Circulatory system diseases	13,010	692	455
Respiratory system diseases	6,295	336	281
Digestive system diseases	17,393	910	509
Skin system diseases	4,896	249	213
Musculoskeletal system diseases	11,316	607	383
Genitourinary system diseases	10,694	575	357
Cognitively normal (CN)	5,927	312	500

This design ensures (i) the held-out test set provides an unbiased evaluation benchmark disjoint from all training data, (ii) pre-training leverages the largest possible pool of multi-organ imaging without requiring clean labels, and (iii) downstream supervised evaluation uses only participants with reliable diagnostic annotations.

Table 8: **Organ imaging availability across data splits.** Number of participants with each organ imaged, along with the coverage rate relative to the split total. The majority of participants lack complete 7-organ imaging.

Modality / Metric	Pre-train	Train	Val	Test
Per-organ availability (# participants, % coverage)				
Brain	31,181 (77%)	27,731 (77%)	1,495 (78%)	1,000 (78%)
Heart	33,629 (83%)	29,924 (83%)	1,554 (81%)	1,050 (82%)
Adipose	25,964 (64%)	23,062 (64%)	1,186 (62%)	812 (63%)
Liver	26,166 (64%)	23,206 (64%)	1,210 (63%)	829 (65%)
Kidney	34,119 (84%)	30,346 (84%)	1,590 (83%)	1,087 (85%)
Spleen	26,857 (66%)	23,834 (66%)	1,238 (65%)	835 (65%)
Pancreas	25,964 (64%)	23,062 (64%)	1,186 (62%)	812 (63%)

Organ Availability. A key characteristic of the UK Biobank dataset [9] is that *organ missingness is the norm rather than the exception*: participants are scanned under different acquisition protocols and scheduling constraints, resulting in varying subsets of organs being imaged per individual. Table 8 quantifies this natural missingness across our data splits. On average, each participant has imaging from only 5.0 out of the 7 target organs, and only 33% of participants have complete 7-organ imaging (e.g., 11,828 out of 36,220 in the training set). The remaining 67% exhibit varying degrees of missingness: 37% have 4–6 organs, 23% have 2–3 organs, and 7% have only a single organ imaged. Organ-level coverage further varies substantially across modalities, from 84% (kidney) down to 64% (adipose and pancreas), reflecting protocol-specific acquisition constraints in the UK Biobank imaging study. This consistent pattern across pre-training, training, validation, and test splits establishes missing-organ scenarios as the *default* clinical reality rather than a corner case—motivating foundation models that are trained and evaluated natively under organ-missingness conditions, rather than assuming complete multi-organ acquisitions.

A.3 Single-Disease Prediction

Single-Disease Definitions. In addition to the 13 ICD-10 chapter-level disease groups used for group-disease prediction, we define a complementary set of 14 specific disease entities for fine-grained single-disease prediction. These diseases are selected to span the four major organ systems including

neurological, *autoimmune*, *digestive*, and *cardiometabolic*. Each disease is identified through ICD-10 codes from UK Biobank:

- **Neurological diseases** (*Alzheimer's, Parkinson's, multiple sclerosis, motor neuron disease*): central to aging research and well-established links to multi-organ manifestations (e.g., cardiovascular comorbidity in Alzheimer's, metabolic alterations in Parkinson's).
- **Systemic autoimmune diseases** (*systemic lupus erythematosus, Sjögren syndrome, systemic sclerosis, sarcoidosis*): canonical examples of pathologies intrinsically involving multiple organs (skin, kidney, lung, heart), providing a rigorous test for pan-organ representations.
- **Digestive diseases** (*Crohn's disease, ulcerative colitis, primary biliary cholangitis*): chosen as primarily localized disorders to probe whether Pan-FM remains competitive when the diagnostic signal is concentrated in a narrow anatomical region.
- **Cardiometabolic diseases** (*type 2 diabetes, hypertrophic cardiomyopathy, pulmonary hypertension*): high-prevalence conditions with both organ-specific anchors (heart, pancreas) and broader systemic involvement, complementing the group-level endocrine and circulatory categories.

Evaluation protocol. Because each specific disease has substantially fewer positive casesr, the held-out test set used for group-disease evaluation is too small to provide reliable per-disease estimates. We therefore adopt a different protocol for single-disease prediction. Positive cases for each disease are extracted from the entire dataset based on the ICD-10 codes, while the 500 cognitively normal (CN) participants from the held-out test set serve as shared negatives across all diseases. For each disease, we perform 10-fold cross-validation with linear probing: the frozen Pan-FM (or baseline) encoder produces per-participant representations, and a linear classifier is trained and evaluated fold-by-fold; we report the mean AUROC and balanced accuracy across folds. The encoder itself is never updated during single-disease evaluation, ensuring that the reported performance reflects the transferability of the pre-trained representations rather than task-specific fine-tuning. Because linear probing keeps the encoder frozen, positive cases drawn from the dataset do not constitute label leakage with respect to self-supervised pre-training, as no disease labels are seen during the pre-training stage.

B Multi-Organ Input Features

Multi-organ Input Features. Pan-FM operates on MRI-derived imaging phenotypes (IDPs) rather than raw voxels, ensuring that each organ is represented by a fixed-length vector of clinically validated features. Table 9 summarizes the per-organ feature composition, totaling **228 features** across the 7 organs. Brain features comprise 119 gray matter regional volumes extracted via MUSE [21], covering subcortical structures and cortical regions across both hemispheres. Heart features comprise 80 cardiac IDPs from the UK Biobank cardiac MRI pipeline (field IDs within `24100-24181`), including chamber volumes, ejection fractions, wall thickness, and myocardial strain. Abdominal features (adipose, liver, kidney, spleen, pancreas) are derived from multi-contrast abdominal MRI within UK Biobank Category 149, capturing organ volumes, fat content (PDFF), and iron concentration as appropriate.

We enforce strict *intra-organ completeness*: if any feature within an organ is missing, the entire organ is marked unavailable, i.e., reflecting the clinical reality that organ-level IDPs are either fully acquired or entirely absent. All features are z-score normalized per organ modality using statistics computed on the pre-training set.

C More Implementation Details

C.1 Self-Supervised Learning (SSL) Baselines

We benchmark five representative SSL methods under our multi-organ pre-training protocol, including contrastive-based approaches (SimCLR [13], BYOL [23], MoCo-v3 [15], Barlow Twins [61], and VICReg [7]. All baselines share the same ViT backbone, CAPE tokenizer, and training schedule as the DINOv2 [34] baseline and our Pan-FM. The only differences are the SSL objective and the auxiliary heads required by each method. Below we describe how each method is adapted to the multi-organ setting.

Table 9: Per-organ input features used by Pan-FM and all SSL baselines.

Organ	Dim.	Feature Description (UKBB Field IDs)
Brain	119	Gray matter regional volumes from MUSE: 21 subcortical structures (e.g., hippocampus, amygdala, thalamus, caudate, putamen, cerebellum) and 98 cortical regions across both hemispheres.
Heart	80	UKBB cardiac MRI IDPs: `24100-24181` (excluding `24120, 24123`). Includes LV/RV end-diastolic / end-systolic / stroke volumes, ejection fractions, myocardial mass, AHA-segmented wall thickness (16 segments), and global longitudinal/circumferential strain.
Adipose	16	Body composition / ectopic-fat features: `21085` (VAT), `21090` (pancreas PDFF, used as ectopic-fat proxy), `22403-22406` (anterior/posterior thigh fat-free muscle volumes, L/R), `22408` (ASAT), `22410` (total trunk fat), `22432` (total abdominal adipose tissue index), `22434` (abdominal fat ratio), `22435` (muscle fat infiltration), `23355-23356` (posterior thigh MFI L/R), `24353-24354` (anterior thigh MFI L/R), `40061` (liver fat %, used as ectopic-fat proxy).
Liver	4	`21080` (volume), `21088` (PDFF), `21089` (iron), `40062` (iron-corrected T1, cT1).
Kidney	3	`21081` (left kidney volume), `21162` (right kidney parenchymal volume), `21163` (inter-kidney centroid distance).
Spleen	3	`21083` (volume), `21170` (iron, IDEAL), `21173` (iron, protocol normalised).
Pancreas	3	`21087` (volume), `21090` (PDFF), `21091` (iron).
Total	**228**	

Shared settings across SSL baselines. To ensure fair comparison, all SSL baselines use: (i) the same 12-layer ViT backbone with hidden dimension 384 and 8 attention heads; (ii) the same CAPE tokenizer producing 128 organ-aware tokens per participant (64 brain, 32 heart, 16 adipose, 4 each for liver/kidney/spleen/pancreas); (iii) the same availability-aware input handling, where missing organs are replaced with learnable organ-aware mask tokens; (iv) the AdamW optimizer with a learning rate of $1\mathrm{e}{-4}$, cosine decay, 10 warmup epochs, batch size 64, and 200 pre-training epochs; (v) a two-view training strategy—*the first view* is the full view with only inherent missingness, and *the second view* is a masked view where a subset of available organs is further masked with a mask ratio of 0.5, following Eq. (3).

SimCLR. SimCLR [13] treats the full view and masked view as two augmentations of the same participant, pulls their projected embeddings together, and contrasts them against all other participants in the batch feature space via the NT-Xent loss [13]. Because multi-organ features differ in scale and dimensionality across organs, we apply NT-Xent on CLS-token embeddings rather than raw organ features, ensuring the contrastive signal operates on a unified representation. we use a 2-layer MLP projection head (Linear–BN–ReLU–Linear, hidden/output dimension 256/128) maps each CLS token to an L2-normalized embedding. NT-Xent temperature τ is tuned over $\{0.1, 0.25, 0.5, 0.75, 1.0\}$ and set to 0.5 based on the downstream group-disease prediction performance.

BYOL. BYOL [23] removes negative pairs entirely for self-supervised representation learning, i.e., an online encoder predicts the target encoder's projection of another view of the same participant, where the target encoder is an exponential moving average (EMA) of the online encoder. Collapse is prevented solely by architectural asymmetry—an additional predictor head on the online branch. In the multi-organ setting, this cross-view prediction directly encourages the online encoder to produce consistent representations regardless of which organ subset is visible. Specifically, both online and target branches use a 2-layer MLP projector (Linear–BN–ReLU–Linear–BN, hidden/output dimension 256/256); the online branch additionally has a 2-layer predictor of the same shape. The symmetric loss is $\mathcal{L} = \frac{1}{2}[\text{cosine}(q_1, z_2) + \text{cosine}(q_2, z_1)]$, where q denotes online predictions and z the stop-gradient target projections. The EMA momentum is cosine-scheduled from 0.992 to 1.0.

MoCo-v3. MoCo-v3 [15] combines SimCLR's contrastive objective with BYOL's momentum encoder, without the memory queue used in earlier MoCo versions [24]. In our multi-organ setting,

the query branch (with predictor) encodes both the full view and the masked view of each participant, while the momentum branch (no predictor) produces stop-gradient keys from the same two views. Positives are formed *cross-view* across masking conditions (query of the full view vs. key of the masked view, and vice versa), and negatives are the keys of all other participants in the batch. This design directly aligns with our goal of learning organ-missingness-invariant representations: the query encoder must produce embeddings that remain close to the momentum-stable target across different organ subsets, while staying discriminative against other participants. For the implementation, the query and momentum branches share a 3-layer MLP projector (hidden/output dimension 256/256), and the query branch has an additional 2-layer predictor (hidden/output dimension 256/256). The InfoNCE temperature is tuned over $\{0.1, 0.2, 0.3, 0.5\}$ and set to 0.2. Momentum is cosine-scheduled from 0.992 to 1.0, which is same with the BYOL baseline.

Barlow Twins. Barlow Twins [61] avoids both negative pairs and momentum encoders by regularizing the cross-correlation matrix between two views' projections toward the identity: diagonal entries are pushed to 1 (invariance) and off-diagonals to 0 (decorrelation). In our multi-organ setting, the invariance term drives the CLS embedding to be stable under organ-level masking. For the implementation, a 3-layer expander MLP (hidden 512, output 2048) projects CLS tokens into a high-dimensional redundancy-reduction space. Projections are batch-normalized per feature. The off-diagonal weight λ is set to $5\mathrm{e}{-3}$.

VICReg. VICReg [7] replaces Barlow Twins' cross-correlation with three explicit terms: *invariance* (MSE between view projections), *variance* (hinge loss on per-feature standard deviation to prevent dimensional collapse), and *covariance* (off-diagonal penalty on per-view covariance, which decorrelates features). In our multi-organ setting, the invariance term pulls full-view and masked-view embeddings of the same participant together; the variance term prevents the backbone from collapsing participants with heavy organ missingness to the same embedding; and the covariance term mitigates organ shortcut, analogous to Barlow Twins. Specifically, VICReg shares the same 3-layer expander (hidden 512, output 2048) as Barlow Twins. Loss weights follow the original paper: $\mu = 25.0$, $\nu = 25.0$, $\xi = 1.0$, with variance hinge threshold $\gamma = 1.0$.

Among the SSL baselines, methods with explicit inter-sample discrimination (SimCLR, MoCo-v3, DINOv2) consistently outperform those relying solely on intra-sample invariance or feature regularization (BYOL, Barlow Twins, VICReg) in the multi-organ learning setting. We attribute this gap to the weak view-level augmentation in our setting, i.e., the two views differ only by organ-level masking with no strong appearance transforms, making discrimination-based objectives more robust to shortcut learning than regularization-based ones.

C.2 Pre-training Details

We pre-train Pan-FM for 200 epochs on a single NVIDIA V100 GPU (approximately 8 hours) using a 12-layer ViT backbone [22] with a hidden dim of 384 and 8 attention heads. Optimization uses AdamW with batch size 64 and mixed-precision (FP16) training. The base learning rate is 10^{-4}, linearly warmed up over the first 10 epochs and then cosine-decayed to 10^{-6}. Weight decay is cosine-annealed from 0.04 to 0.4 over the full schedule. The teacher–student architecture uses an 8192-dim projection head with the last layer frozen for the first epoch, an EMA momentum cosine-annealed from 0.992 to 1.0, and a teacher softmax temperature linearly warmed from 0.04 to 0.07 over 10 epochs.

Applying SGM for Various SSL Methods. To verify that SGM is a general-purpose plug-in rather than an architecture-specific trick, we apply it to four representative SSL paradigms: contrastive (SimCLR), feature-decorrelation (VICReg, Barlow Twins), and self-distillation (DINOv2). For a fair comparison, we keep each baseline's original hyper-parameters unchanged and only tune SGM's temperature, setting it to 0.75 for SimCLR, and 0.5 for both VICReg and Barlow Twins.

C.3 Downstream Task

C.3.1 Group-Disease Prediction

Linear Probing. We follow a standard linear probing protocol to evaluate the quality of frozen representations. For each downstream disease, we train an independent binary linear probe (Case

vs. healthy Control) consisting of a single fully-connected layer mapping the frozen CLS token to a binary logit. The backbone is entirely frozen, ensuring that the evaluation purely reflects the quality of the pre-trained representation. Input modalities are normalized per organ using pre-computed training statistics. To address class imbalance between cases and controls, we replace standard BCE with a focal loss ($\alpha = 0.75$, $\gamma = 2.0$) for all compared baselines. Each probe is optimized with AdamW by using a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}) for 20 epochs with batch size 256. All baselines and our Pan-FM are evaluated under this same protocol for fair comparison. Each linear probe is trained with 10 different random seeds, and we report the average performance on the held-out test set.

Full Backbone Fine-tuning. We additionally evaluate each pre-trained backbone under the end-to-end fine-tuning protocol. For each downstream disease, each pre-trained SSL backbone is fine-tuned jointly with a single-layer linear classification head, keeping consistent with the linear probing evaluation. To preserve the pre-trained representations while allowing the task-specific head to adapt freely, we use a smaller learning rate for the backbone (2×10^{-6}) and a larger learning rate for the head (1×10^{-3}). For the Random baseline that uses a randomly initialized backbone, we use a larger learning rate of 1×10^{-3} for backbone training. Training uses AdamW with a linear warmup followed by cosine decay, and the same focal loss ($\alpha = 0.75$, $\gamma = 2.0$) as in linear probing. Each model is fine-tuned for up to 20 epochs with batch size 32, and the best checkpoint is selected by validation loss with early stopping. We report the average over 5 runs with different random seeds on the held-out test set.

Random organ dropout. To assess robustness under incomplete organ availability, we design two complementary missing-organ protocols. Both are conducted on the subset of test subjects with complete 7-organ imaging data to ensure a controlled comparison. For each subject, we randomly drop $k \in \{1, 2, 3\}$ available organs at inference. To ensure stable estimates, for the linear probing setting we train 10 independent linear classifiers (different initialization seeds) and, for each classifier, repeat the random dropout 5 times using fixed seeds $\{0, 1, 2, 3, 4\}$, yielding 50 evaluation runs per drop level. For the fine-tuning setting, due to the longer training time and its empirically smaller run-to-run variance, we instead train 5 independent models and repeat the random dropout with the same 5 seeds (25 evaluation runs). **The same dropout seeds are shared across all compared methods (Pan-FM and SSL baselines), so that every method is evaluated on identical organ-dropout patterns at each run, enabling paired comparisons.** We report the mean performance across all runs. This setting simulates realistic missingness patterns where any subset of organs may be unavailable.

Specific organ dropout. We systematically remove one organ at a time (across all seven organs: brain, heart, adipose, liver, kidney, spleen, pancreas) to measure each model's sensitivity to the absence of a particular organ. This disentangles the contribution of each organ to the learned representation and reveals whether a model relies disproportionately on any single organ. Results are averaged over the same set of classifier seeds as above (10 for linear probing, 5 for fine-tuning).

C.3.2 Single-Disease Prediction

Linear Probing. In contrast to the 13 disease categories, individual diseases contain substantially fewer positive cases, making a single train/test split unreliable for performance evaluation. We therefore adopt 10-fold stratified cross-validation for single-disease evaluation. For each disease, we construct a case-control pool comprising all available cases across all splits and all healthy controls from the held-out test split, which are then partitioned into 10 stratified folds with balanced case/control ratios. Within each fold, a linear probe is trained on the training folds and evaluated on the held-out validation fold, following the same optimization protocol as described above (AdamW, focal loss, cosine annealing and training epochs). We report the mean AUROC and balanced accuracy across the 10 folds.

D More Results

D.1 Group-Disease Prediction

Linear Probing. As shown in Table 10, Pan-FM achieves the best overall performance, with a mean AUROC of 0.672 and BalAcc of 0.638. The gains are especially pronounced on systemic

Table 10: **Group-disease evaluation with linear probing.** Comparison of SSL methods for group-disease prediction on the held-out test set. Values are reported as AUROC / BalAcc.

Disease Group	**SimCLR** [13]	**BYOL** [23]	**MoCo-v3** [15]	**VICReg** [7]	**Barlow Twins** [61]	**DINOv2** [34]	**Pan-FM**
Infectious & Parasitic	0.627 / 0.615	0.614 / 0.588	0.648 / 0.617	0.645 / 0.619	0.637 / 0.610	0.657 / 0.627	**0.668 / 0.635**
Neoplasms	0.628 / 0.598	0.618 / 0.586	0.634 / 0.604	0.625 / 0.599	0.625 / 0.601	0.629 / 0.601	**0.671 / 0.643**
Blood & Immune	0.659 / 0.628	0.635 / 0.617	**0.672** / 0.632	0.635 / 0.618	0.646 / 0.624	0.657 / 0.628	0.666 / **0.634**
Endocrine & Metabolic	0.692 / 0.651	0.682 / 0.633	0.708 / 0.662	0.680 / 0.638	0.692 / 0.652	0.700 / 0.657	**0.713 / 0.666**
Mental & Behavioural	0.673 / 0.633	0.649 / 0.612	0.665 / 0.636	0.649 / 0.618	0.665 / 0.631	0.675 / 0.647	**0.690 / 0.648**
Nervous System	0.667 / 0.628	0.655 / 0.617	0.668 / 0.628	0.645 / 0.614	0.647 / 0.608	0.655 / 0.620	**0.678 / 0.643**
Eye	0.661 / 0.635	0.611 / 0.611	0.632 / 0.623	0.605 / 0.590	0.603 / 0.590	0.644 / 0.622	**0.672 / 0.636**
Circulatory System	0.688 / 0.647	0.660 / 0.619	**0.699 / 0.657**	0.680 / 0.635	0.686 / 0.644	0.681 / 0.646	0.693 / 0.648
Respiratory System	0.646 / 0.616	0.644 / 0.604	0.657 / 0.620	0.644 / 0.612	0.654 / 0.618	0.654 / 0.627	**0.667 / 0.641**
Digestive System	**0.638** / 0.603	0.615 / 0.581	0.632 / 0.599	0.611 / 0.590	0.624 / 0.594	0.619 / 0.597	0.635 / **0.606**
Skin System	0.629 / 0.608	0.616 / 0.594	0.630 / 0.614	0.622 / 0.606	0.622 / 0.604	0.621 / 0.599	**0.671 / 0.637**
Musculoskeletal	0.653 / 0.619	0.639 / 0.599	**0.661** / 0.619	0.640 / 0.605	0.643 / 0.603	0.644 / 0.615	0.659 / **0.624**
Genitourinary	0.636 / 0.607	0.633 / 0.610	0.647 / 0.617	0.644 / 0.611	0.648 / 0.618	0.635 / 0.610	**0.659 / 0.628**
Overall Mean	0.653 / 0.622	0.636 / 0.606	0.658 / 0.625	0.640 / 0.612	0.646 / 0.615	0.652 / 0.623	**0.672 / 0.638**

Table 11: **Missing-organ robustness evaluation with linear probing.** Comparison of SSL methods across 13 disease groups on the held-out test set (AUROC / BalAcc). *Full Organs (7)*, *Drop 1 Organ*, *Drop 2 Organs*, and *Drop 3 Organs* settings are evaluated on subjects with complete 7-organ imaging data in the held-out test set, with random organ dropout repeated over 50 runs to ensure stable and comparable assessment. Δ denotes the relative improvement of Pan-Organ over the DINOv2 baseline.

Method	Standard		Full Organs (7)		Drop 1 Organ		Drop 2 Organs		Drop 3 Organs	
	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc
SimCLR [13]	0.653	0.622	0.659	0.637	0.645	0.629	0.638	0.627	0.625	0.617
BYOL [23]	0.636	0.606	0.647	0.634	0.642	0.630	0.645	0.632	0.628	0.621
MoCo-v3 [15]	0.658	0.625	0.660	0.642	0.653	0.636	0.655	0.641	0.642	0.630
VICReg [7]	0.640	0.612	0.647	0.631	0.640	0.626	0.639	0.627	0.624	0.619
Barlow Twins [61]	0.646	0.615	0.642	0.624	0.632	0.620	0.632	0.622	0.618	0.613
DINOv2 [34]	0.652	0.623	0.683	0.654	0.654	0.642	0.647	0.639	0.617	0.616
Pan-FM (Ours)	**0.672**	**0.638**	**0.694**	**0.670**	**0.681**	**0.662**	**0.672**	**0.655**	**0.654**	**0.638**
Δ (%)	+3.1%	+2.4%	+1.6%	+2.4%	+4.1%	+3.1%	+3.9%	+2.5%	+6.0%	+3.6%

conditions such as *Neoplasms*, *Skin System*, and *Metabolic*, where Pan-FM substantially outperforms the DINOv2 baseline, e.g., 0.671 vs. 0.629 on Neoplasms, and 0.671 vs. 0.621 on Skin System. Since Pan-FM and DINOv2 share identical backbones and training configurations, this improvement directly demonstrates the effectiveness of our proposed SGM in capturing systemic pan-organ representations.

Missing Robustness w/ Linear-Probing. Table 11 reports missing-organ robustness under linear probing. *Standard* uses the full test set with natural missingness; the other four settings apply controlled dropout on the complete-7-organ subset. As more organs are dropped, all baselines degrade substantially, while Pan-FM degrades gracefully and consistently leads, with its margin over DINOv2 growing from +1.6% (Full) to +6.0% (Drop-3). Several baselines (BYOL, MoCo-v3) even degrade non-monotonically from Drop-1 to Drop-2, as they rely on a few dominant organs whose redundant counterparts can be dropped without loss (Fig. 4). Pan-FM instead leads in 6/7 single-organ dropout settings, peaking at +6.3% over DINOv2 on *w/o Adipose*, confirming that SGM yields balanced pan-organ representations robust to organ absence.

Full-Backbone Fine-tuning. Table 12 reports group-disease prediction results when the entire backbone is fine-tuned together with the linear head. Pan-FM still achieves the best overall performance (0.681/0.641), outperforming the strongest SSL baseline DINOv2 (0.666/0.632) and substantially surpassing the randomly initialized backbone (0.631/0.610), which demonstrates the effectiveness of multi-organ SSL. In addition, DINOv2 shows a larger relative gain from linear probing to full fine-tuning than other SSL baselines, suggesting that its representations encode transferable features that particularly benefit for fine-tuning. Together with the linear-probing results in Table 10, these results demonstrate that the representations learned by SGM are beneficial both as frozen features and as initialization for end-to-end adaptation.

Missing Robustness w/ Backbone Fine-tuning. Table 13 and Figure 5 assess test-time organ removal after backbone fine-tuning. Pan-FM leads at every missingness level, and its margin over DINOv2 scales with the number of dropped organs (+2.0% at Full Organs, +4.6% at Drop 3). The radar chart in Figure 5(a) indicates no weak direction, and the pairwise heatmap in (b) localizes DINOv2's failure modes to *Adipose* and *Pancreas*, where Pan-FM remains stable. Compared with linear probing, full backbone fine-tuning recovers additional accuracy under severe missingness

Table 12: **Group-disease evaluation with full backbone fine-tuning.** Comparison of SSL methods for group-disease prediction on the held-out test set. Values are reported as AUROC / BalAcc.

Disease Group	**Random**	**SimCLR** [13]	**MoCo-v3** [15]	**VICReg** [7]	**Barlow Twins** [61]	**DINOv2** [34]	**Pan-FM**
Infectious & Parasitic	0.638 / 0.635	0.629 / 0.614	0.634 / 0.610	0.642 / 0.612	0.621 / 0.604	0.672 / 0.630	**0.694 / 0.656**
Neoplasms	0.605 / 0.589	0.629 / 0.601	0.638 / 0.612	0.629 / 0.596	0.637 / 0.608	0.646 / 0.612	**0.686 / 0.646**
Blood & Immune	0.628 / 0.608	0.660 / 0.628	**0.674 / 0.636**	0.636 / 0.609	0.650 / 0.621	0.656 / 0.630	0.669 / 0.632
Endocrine & Metabolic	0.674 / 0.633	0.703 / 0.655	0.712 / 0.669	0.694 / 0.655	0.701 / 0.659	0.710 / 0.661	**0.718 / 0.673**
Mental & Behavioural	0.649 / 0.629	0.657 / 0.613	0.657 / 0.627	0.644 / 0.616	0.656 / 0.633	0.680 / **0.654**	**0.683** / 0.638
Nervous System	0.636 / 0.619	0.664 / 0.624	0.665 / 0.623	0.654 / 0.640	0.653 / 0.618	0.665 / 0.628	**0.681 / 0.652**
Eye	0.615 / 0.610	0.653 / 0.630	0.631 / 0.616	0.622 / 0.610	0.622 / 0.596	0.661 / 0.631	**0.674 / 0.635**
Circulatory System	0.654 / 0.611	0.688 / 0.637	0.702 / **0.663**	0.691 / 0.653	0.692 / 0.647	0.702 / 0.661	**0.704** / 0.656
Respiratory System	0.607 / 0.593	0.649 / 0.612	0.663 / 0.625	0.652 / 0.624	0.660 / 0.624	0.673 / 0.643	**0.678 / 0.645**
Digestive System	0.611 / 0.592	0.633 / 0.601	0.636 / 0.606	0.617 / 0.591	0.626 / 0.595	0.642 / 0.609	**0.654 / 0.610**
Skin System	0.623 / 0.605	0.623 / 0.609	0.623 / 0.605	0.621 / 0.606	0.620 / 0.596	0.630 / 0.602	**0.666 / 0.629**
Musculoskeletal	0.641 / 0.607	0.650 / 0.615	0.664 / 0.624	0.644 / 0.611	0.647 / 0.610	0.667 / 0.622	**0.669 / 0.627**
Genitourinary	0.623 / 0.602	0.637 / 0.605	0.642 / 0.610	0.642 / 0.606	0.650 / 0.623	0.653 / 0.627	**0.673 / 0.634**
Overall Mean	0.631 / 0.610	0.652 / 0.619	0.657 / 0.625	0.645 / 0.618	0.649 / 0.618	0.666 / 0.632	**0.681 / 0.641**

Table 13: **Missing-organ robustness evaluation with full backbone fine-tuning.** Comparison of SSL methods across 13 disease groups on the held-out test set (AUROC / BalAcc). *Full Organs (7)*, *Drop 1 Organ*, and *Drop 3 Organs* settings are evaluated on subjects with complete 7-organ imaging data in the held-out test set, with random organ dropout repeated over 25 runs to ensure stable and comparable assessment. Δ denotes the relative improvement of Pan-FM over the DINOv2 baseline.

Method	Standard		Full Organs (7)		Drop 1 Organ		Drop 2 Organs		Drop 3 Organs	
	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc	AUROC	BalAcc
SimCLR [13]	0.652	0.619	0.664	0.643	0.652	0.635	0.645	0.630	0.636	0.624
BYOL [23]	0.640	0.607	0.653	0.638	0.645	0.632	0.648	0.633	0.634	0.624
MoCo-v3 [15]	0.657	0.625	0.667	0.649	0.658	0.641	0.662	0.645	0.647	0.634
VICReg [7]	0.645	0.618	0.650	0.628	0.644	0.627	0.642	0.629	0.625	0.619
Barlow Twins [61]	0.649	0.618	0.643	0.626	0.633	0.621	0.630	0.622	0.619	0.616
DINOv2 [34]	0.666	0.632	0.691	0.662	0.681	0.660	0.668	0.650	0.645	0.632
Pan-FM (Ours)	**0.681**	**0.641**	**0.704**	**0.675**	**0.696**	**0.672**	**0.687**	**0.662**	**0.675**	**0.650**
Δ (%)	+2.3%	+1.5%	+2.0%	+1.9%	+2.1%	+1.8%	+2.7%	+1.8%	+4.6%	+2.9%

(0.675 vs. 0.654 AUROC at Drop 3), as trainable parameters can reallocate capacity to observed organs; however, the DINOv2–Pan-FM gap persists in both regimes, isolating SGM as the source of robustness rather than the adaptation protocol.

D.2 Quantifying improvements across disease categories

We quantified the improvement using the 13 disease categories as the units of comparison. Table 15 reports the mean paired AUROC difference between Pan-FM and each baseline, with its 95% confidence interval (CI), across the 13 categories. Positive differences favour Pan-FM, and CIs excluding zero indicate statistical significance.

Pan-FM showed significant improvements in 23 of the 24 comparisons. Against MoCo-v3 under three-organ removal, Pan-FM performed better in 9 of the 13 categories, although the CI included zero (mean difference, 0.012; 95% CI, -0.001 to 0.026).

As Pan-FM is built directly on DINOv2, this comparison provides a controlled ablation of SGM. Pan-FM significantly outperformed DINOv2 in every setting, with a larger improvement under three-organ removal than under the standard setting (0.037 versus 0.021 AUROC). Table 4 and Appendix Table 13 further show consistent gains when SGM is incorporated into SimCLR, VICReg and Barlow Twins, supporting its applicability across different self-supervised learning objectives.

D.3 Applying SGM for More SSL Methods

To verify that SGM is a general-purpose plug-in rather than an architecture-specific trick, we apply it to four representative SSL paradigms: contrastive (SimCLR), feature-decorrelation (VICReg, Barlow Twins), and self-distillation (DINOv2). For a fair comparison, we keep each baseline's original hyper-parameters unchanged and only tune SGM's temperature, setting it to 0.75 for SimCLR, and 0.5 for both VICReg and Barlow Twins. As shown in Table 14, SGM consistently improves all four baselines, lifting mean AUROC by +1.8 to +2.5 absolute points and winning 96% (100/104) of head-to-head per-disease comparisons. Notably, SGM benefits are broadly distributed across

Table 14: **Group-disease evaluation with linear probing.** Adding **our proposed SGM** as a plug-in consistently improves four representative SSL methods on the held-out test set across 13 disease groups. Values are AUROC / BalAcc; bold marks the better result within each (base, +SGM) pair.

Disease Group	**SimCLR** [13]		**VICReg** [7]		**Barlow Twins** [61]		**DINOv2** [34]	
	Base	+SGM	Base	+SGM	Base	+SGM	Base	+SGM (Pan-FM)
Infectious & Parasitic	0.627/0.615	**0.638/0.616**	0.645/**0.619**	**0.651**/0.618	0.637/0.610	**0.658/0.648**	0.657/0.627	**0.668/0.635**
Neoplasms	0.628/0.598	**0.647/0.616**	0.625/**0.599**	**0.637**/0.597	0.625/0.601	**0.651/0.612**	0.629/0.601	**0.671/0.643**
Blood & Immune	0.659/0.628	**0.687/0.665**	0.635/0.618	**0.665/0.627**	0.646/0.624	**0.653/0.628**	0.657/0.628	**0.666/0.634**
Endocrine & Metabolic	0.692/0.651	**0.714/0.666**	0.680/0.638	**0.703/0.655**	0.692/0.652	**0.715/0.672**	0.700/0.657	**0.713/0.666**
Mental & Behavioural	**0.673/0.633**	0.672/0.628	0.649/0.618	**0.692/0.645**	0.665/0.631	**0.671/0.640**	0.675/0.647	**0.690/0.648**
Nervous System	0.667/0.628	**0.687/0.650**	0.645/0.614	**0.674/0.642**	0.647/0.608	**0.677/0.641**	0.655/0.620	**0.678/0.643**
Eye	0.661/0.635	**0.664/0.637**	0.605/0.590	**0.658/0.641**	0.603/0.590	**0.672/0.653**	0.644/0.622	**0.672/0.636**
Circulatory System	0.688/0.647	**0.707/0.648**	0.680/0.635	**0.702/0.650**	0.686/0.644	**0.705/0.655**	0.681/0.646	**0.693/0.648**
Respiratory System	0.646/0.616	**0.670/0.634**	0.644/0.612	**0.659/0.628**	0.654/0.618	**0.662/0.621**	0.654/0.627	**0.667/0.641**
Digestive System	0.638/0.603	**0.655/0.617**	0.611/0.590	**0.643/0.605**	0.624/0.594	**0.646/0.612**	0.619/0.597	**0.635/0.606**
Skin System	0.629/0.608	**0.650/0.623**	0.622/0.606	**0.640/0.618**	0.622/0.604	**0.626/0.611**	0.621/0.599	**0.671/0.637**
Musculoskeletal	0.653/0.619	**0.675/0.629**	0.640/0.605	**0.672/0.640**	0.643/0.603	**0.678/0.625**	0.644/0.615	**0.659/0.624**
Genitourinary	0.636/0.607	**0.654/0.623**	0.644/0.611	**0.646/0.618**	0.648/**0.618**	**0.649**/0.615	0.635/0.610	**0.659/0.628**
Overall Mean	0.653/0.622	**0.671/0.635**	0.640/0.612	**0.665/0.630**	0.646/0.615	**0.666/0.633**	0.652/0.623	**0.672/0.638**

Table 15: Mean AUROC improvements of Pan-FM across 13 disease categories, with 95% CIs.

Setting	DINOv2	MoCo-v3	SimCLR	Barlow Twins	VICReg	BYOL
Standard	.021 [.013, .028]	.015 [.004, .025]	.019 [.010, .028]	.027 [.016, .038]	.032 [.023, .041]	.036 [.027, .045]
Drop 1 organ	.027 [.009, .045]	.028 [.014, .042]	.036 [.024, .048]	.049 [.028, .069]	.041 [.021, .062]	.039 [.022, .057]
Drop 2 organs	.025 [.011, .039]	.018 [.007, .028]	.034 [.023, .045]	.040 [.024, .056]	.033 [.018, .049]	.027 [.013, .042]
Drop 3 organs	.037 [.020, .054]	.012 [−.001, .026]	.029 [.016, .042]	.036 [.020, .053]	.030 [.013, .047]	.025 [.009, .042]

disease groups rather than concentrated on a few, i.e., every disease except SimCLR-Mental sees consistent gains. Our final model, Pan-FM (DINOv2 + SGM), attains the best overall performance (0.672 AUROC / 0.638 BalAcc), confirming that SGM compounds with the strongest available backbone. These results demonstrate that SGM is a method-agnostic plug-in that reliably enhances representation quality across both contrastive (SimCLR), feature-decorrelation (VICReg, Barlow Twins), and self-distillation (DINOv2) paradigms.

D.4 Single Organ Dropout

Missing Robustness Evaluation under linear probing. Figure 7 reports AUROC and Balanced Accuracy when each of the seven organs is removed individually at test time, evaluated on subjects with complete 7-organ data. Across all seven removal scenarios, Pan-FM (red curve) consistently dominates the radar plot, forming the outermost envelope on both metrics. The gap between Pan-FM and the strongest competing baseline is most pronounced when removing organs that carry rich diagnostic signal (e.g., *Heart*, *Spleen*, *Kidney*, and *Adipose*), where existing SSL methods drop sharply, while Pan-FM remains stable. This indicates that the cross-organ token interactions learned through SGM enable Pan-FM to recover predictive information from the remaining organs rather than relying on any single modality. In contrast, baselines such as VICReg and Barlow Twins exhibit visibly contracted polygons, reflecting their fragility to organ unavailability, which is a realistic bottleneck in clinical deployment.

Missing Robustness under full backbone fine-tuning. Figure 8 shows that Pan-FM's robustness advantage persists, and is in fact *amplified*, after full backbone fine-tuning. Compared to linear probing, all methods improve in absolute terms once the backbone is fine-tuned, yet the relative ranking remains unchanged: Pan-FM forms the outermost envelope across all seven organ-removal conditions on both AUROC and Balanced Accuracy. The persistent gap suggests that Pan-FM's robustness stems from its pre-trained representations rather than from supervised adaptation, i.e., fine-tuning cannot close the gap for baselines, indicating that SGM injects an inductive bias toward missing-organ generalization that is hard to acquire from labels alone.

Overall, Figs 7 and 8 demonstrate that Pan-FM delivers reliable performance under realistic missing-organ conditions.

D.5 Pairwise Organ Dropout

To see how the models behave when more than one organ is missing at test time, we remove organs two at a time and evaluate the probes on patients who originally have all seven organs available.

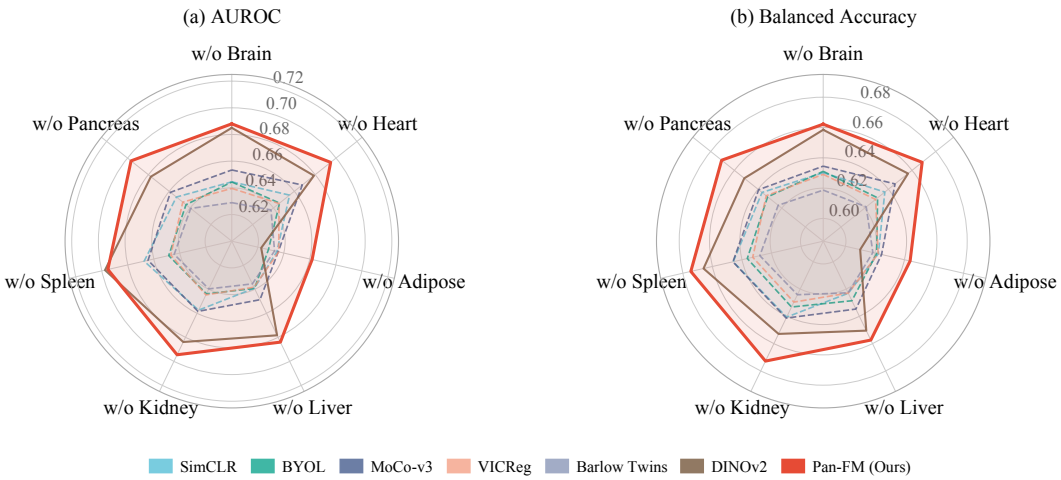


Figure 7: Robustness evaluation under specific organ dropout (test-time organ removal) with **linear probing**.

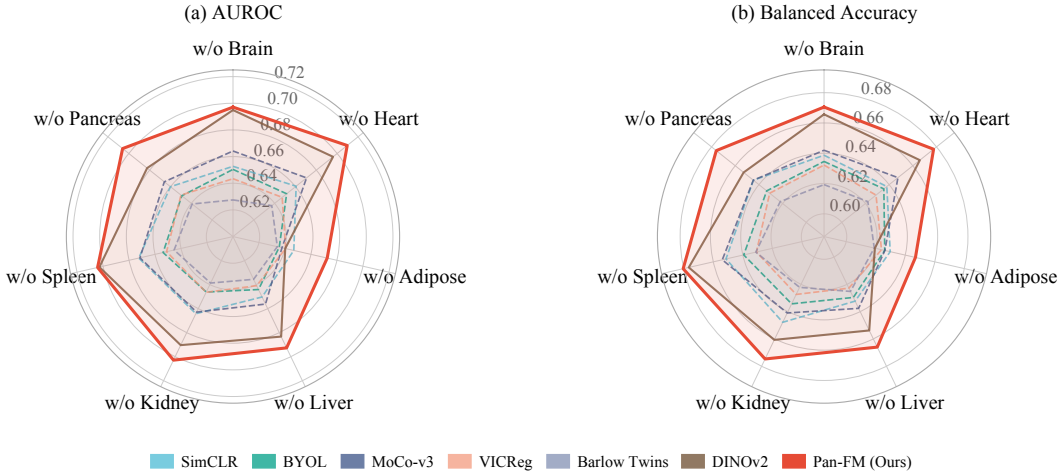


Figure 8: Robustness evaluation under specific organ dropout (test-time organ removal) with **full backbone fine-tuning**.

There are $\binom{7}{2} = 21$ pairs, and together with the 7 single-organ removals they form the lower-triangular heatmaps in Figs 9–12. Specifically, off-diagonal cells correspond to the masked organ pairs, diagonal cells indicate the single-organ dropout, and each cell reports the mean across the 13 diseases. Pan-FM's heatmaps look nearly uniform under both linear probing and full backbone fine-tuning, demonstrating that no pair organs is a critical bottleneck. For DINOv2, it has clear weak spots, i.e., organ drop pairs involving adipose degrade the performance significantly, suggesting that it relies heavily on a few organs and suffers disproportionately when they go missing.

D.6 Group-Disease Comparison with More Missing Organs

To evaluate performance under a more challenging missing-modality regime, we randomly drop three organs at test time and report per-disease AUROC and BalAcc in Tables 16 and 17. Under linear probing, Pan-FM outperforms DINOv2 on 12 of 13 diseases in AUROC, with a mean gain of $+6.00\%$ AUROC and $+3.62\%$ balanced accuracy; the largest improvements appear on Nervous System ($+10.25\%$), Endocrine & Metabolic ($+9.99\%$), and Blood & Immune ($+9.73\%$) diseases, which typically benefit from evidence integrated across multiple organs. The same trend holds under full backbone fine-tuning (mean gains of $+4.61\%$ AUROC and $+2.92\%$ balanced accuracy), showing that the whole-body representations learned by Pan-FM is also beneficial for backbone fine-tuning based downstream adaptation.

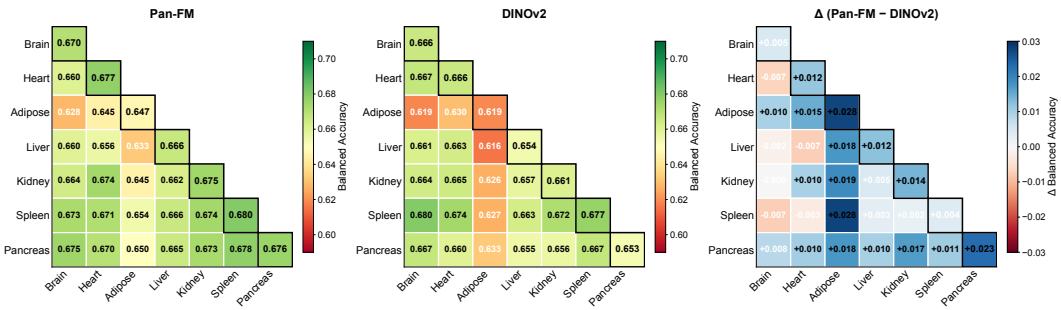


Figure 9: Pairwise organ dropout heatmaps under **full backbone fine-tuning** for our Pan-FM and DINOv2. Each cell shows the mean **balanced accuracy** across 13 group diseases when the corresponding row and column organs are simultaneously removed. Diagonal entries correspond to single-organ removal.

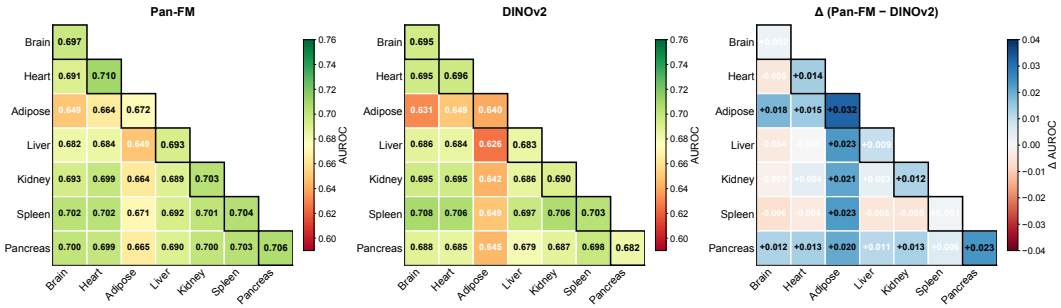


Figure 10: Pairwise organ dropout heatmaps under **full backbone fine-tuning** for our Pan-FM and DINOv2. Each cell shows the mean **AUROC** across 13 group diseases when the corresponding row and column organs are simultaneously removed. Diagonal entries correspond to single-organ removal.

D.7 Comparison with Single-Organ Baselines

We further compare Pan-FM against more classical single-organ baselines, where each organ is trained independently with five predictors (i.e., Lasso, ElasticNet, Linear SVM, Random Forest, and MLP) in a 10-fold cross-validation way. As shown in Tables 18 and 19, Pan-FM achieves the best mean AUROC (0.704) and balanced accuracy (0.674), outperforming all the single-organ baselines. Gains are largest on diseases with multi-system aetiology (Endocrine & Metabolic, Mental & Behavioural, Nervous System, Circulatory System), where no single organ can provide sufficient evidence. These results demonstrate that Pan-FM's advantage arises not from any single organ modality, but from jointly reasoning across all the organs. This yields effective *whole-body representations* and marks a step toward more generalizable whole-body foundation models.

D.8 Comparison with Random (From-Scratch) Baseline

To verify the effectiveness of the proposed Pan-FM pre-training, we compare two identical full ViT multi-organ backbones, fine-tuned end-to-end on the 13-way disease tasks. The main difference is the backbone initialisation: weights from our Pan-FM pre-training, or the standard ViT scheme (truncated-normal). As shown in Fig. 13, Pan-FM outperforms the from-scratch baseline on all 13 disease categories, with a mean gain of $+5.0$ AUROC points. This gap reflects a well-known property of high-capacity Transformers, i.e., end-to-end training from scratch requires data at a scale that downstream clinical cohorts rarely provide. Large-scale multi-organ self-supervision is therefore not merely a favourable initialisation but a practical prerequisite for adapting full ViT backbones in the low-label regime typical of medical applications.

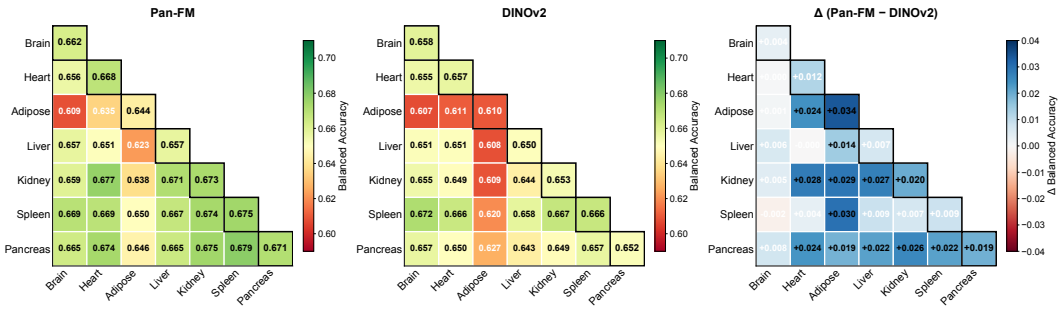


Figure 11: Pairwise organ dropout heatmaps under **linear probing** for our Pan-FM and DINOv2. Each cell shows the mean **balanced accuracy** across 13 group diseases when the corresponding row and column organs are simultaneously removed. Diagonal entries correspond to single-organ removal.

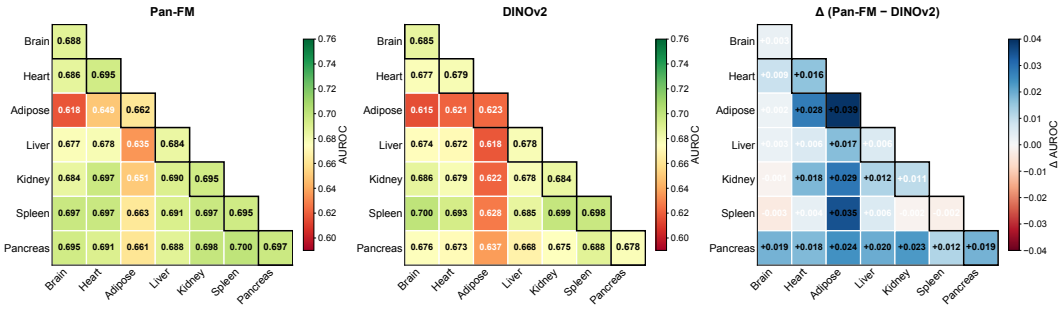


Figure 12: Pairwise organ dropout heatmaps under **linear probing** for our Pan-FM and DINOv2. Each cell shows the mean **AUROC** across 13 group diseases when the corresponding row and column organs are simultaneously removed. Diagonal entries correspond to single-organ removal.

E More Ablation Studies

In this section, we conduct more ablation studies to demonstrate the effectiveness of our proposed Pan-FM.

E.1 Computational Overhead

We compare the per-iteration training cost of our Pan-FM with SGM against the DINOv2 baseline in Table 20. Pan-FM introduces a marginal $+3.5\%$ wall-clock overhead per iteration (82.1 vs 79.3 ms). Notably, the backbone forward FLOPs are identical between the two models (96.4 G), as Pan-FM does not modify the DINOv2 architecture. The entire overhead arises from teacher-side attention extraction and multinomial sampling. Extrapolated to a 200-epoch pretraining run, this leads to only ~ 11 additional minutes of total wall-clock time, which is an essentially negligible cost. Crucially, this overhead is moreover confined to the self-supervised pre-training phase. At inference and during downstream fine-tuning, our model is architecturally identical to the baseline and incurs zero additional cost.

E.2 Pre-training Convergence

To assess how representation quality evolves during self-supervised pre-training, we perform linear probing at every 40 pre-training epochs. At each checkpoint, the backbone is frozen and a linear classifier is trained on top of the [CLS] token using the training set, then evaluated on the full held-out test set. We report the mean AUROC and mean balanced accuracy across the 13 disease categories in Fig. 14. At epoch 40, Pan-FM already outperforms the baseline by a noticeable margin (0.647 vs 0.627 AUROC), indicating that the SGM signal accelerates representation acquisition from early in pre-training. Both methods improve until epoch 80, after which their behaviours diverge: the

Table 16: Group-disease comparison under the Drop-3 Organs setting (linear probing). Results are aggregated over 50 runs. Δ denotes the relative improvement of Pan-FM over the DINOv2 baseline.

Disease	AUROC			Balanced Accuracy		
	DINOv2	Pan-FM (Ours)	Δ (%)	DINOv2	Pan-FM (Ours)	Δ (%)
Infectious & Parasitic	**0.648** $\pm$**0.006**	0.618 $\pm$0.014	-4.57	**0.633** $\pm$**0.009**	0.616 $\pm$0.006	-2.66
Neoplasms	0.584 $\pm$0.006	**0.629** $\pm$**0.006**	$+7.78$	0.592 $\pm$0.004	**0.618** $\pm$**0.006**	$+4.39$
Blood & Immune	0.601 $\pm$0.013	**0.659** $\pm$**0.010**	$+9.73$	0.609 $\pm$0.012	**0.644** $\pm$**0.012**	$+5.76$
Endocrine & Metabolic	0.659 $\pm$0.003	**0.725** $\pm$**0.004**	$+9.99$	0.649 $\pm$0.004	**0.677** $\pm$**0.003**	$+4.32$
Mental & Behavioural	0.634 $\pm$0.010	**0.686** $\pm$**0.012**	$+8.28$	0.634 $\pm$0.008	**0.667** $\pm$**0.011**	$+5.18$
Nervous System	0.606 $\pm$0.006	**0.668** $\pm$**0.011**	$+10.25$	0.609 $\pm$0.005	**0.653** $\pm$**0.009**	$+7.20$
Eye	**0.622** $\pm$**0.009**	0.611 $\pm$0.016	-1.79	0.628 $\pm$0.008	**0.632** $\pm$**0.012**	$+0.60$
Circulatory System	0.630 $\pm$0.003	**0.677** $\pm$**0.005**	$+7.47$	0.623 $\pm$0.003	**0.643** $\pm$**0.005**	$+3.16$
Respiratory System	0.627 $\pm$0.006	**0.657** $\pm$**0.006**	$+4.84$	0.613 $\pm$0.004	**0.638** $\pm$**0.006**	$+4.05$
Digestive System	0.592 $\pm$0.003	**0.624** $\pm$**0.004**	$+5.55$	0.599 $\pm$0.004	**0.613** $\pm$**0.004**	$+2.36$
Skin	0.612 $\pm$0.006	**0.659** $\pm$**0.005**	$+7.65$	0.606 $\pm$0.003	**0.643** $\pm$**0.006**	$+6.14$
Musculoskeletal System	0.602 $\pm$0.003	**0.645** $\pm$**0.006**	$+7.07$	0.604 $\pm$0.003	**0.625** $\pm$**0.006**	$+3.55$
Genitourinary System	0.605 $\pm$0.004	**0.641** $\pm$**0.004**	$+5.92$	0.604 $\pm$0.004	**0.623** $\pm$**0.004**	$+3.22$
Mean	0.617	**0.654**	$+6.00$	0.616	**0.638**	$+3.62$

Table 17: Group-disease comparison under the Drop-3 Organs setting (full backbone fine-tuning). Results are aggregated over 25 runs. Δ denotes the relative improvement of Pan-FM over the DINOv2 baseline.

Disease	AUROC			Balanced Accuracy		
	DINOv2	Pan-FM (Ours)	Δ (%)	DINOv2	Pan-FM (Ours)	Δ (%)
Infectious & Parasitic	0.642 $\pm$0.034	**0.643** $\pm$**0.046**	$+0.23$	0.632 $\pm$0.025	**0.632** $\pm$**0.033**	$+0.03$
Neoplasms	0.619 $\pm$0.019	**0.666** $\pm$**0.014**	$+7.61$	0.605 $\pm$0.009	**0.644** $\pm$**0.010**	$+6.35$
Blood & Immune	0.616 $\pm$0.035	**0.673** $\pm$**0.028**	$+9.33$	0.610 $\pm$0.024	**0.656** $\pm$**0.027**	$+7.52$
Endocrine & Metabolic	0.695 $\pm$0.019	**0.744** $\pm$**0.011**	$+7.11$	0.663 $\pm$0.016	**0.693** $\pm$**0.017**	$+4.51$
Mental & Behavioural	0.654 $\pm$0.030	**0.695** $\pm$**0.016**	$+6.26$	0.639 $\pm$0.011	**0.666** $\pm$**0.016**	$+4.18$
Nervous System	0.629 $\pm$0.019	**0.673** $\pm$**0.016**	$+6.86$	0.618 $\pm$0.017	**0.654** $\pm$**0.010**	$+5.86$
Eye	**0.636** $\pm$**0.014**	0.622 $\pm$0.019	-2.22	0.630 $\pm$0.024	**0.635** $\pm$**0.019**	$+0.75$
Circulatory System	0.675 $\pm$0.019	**0.706** $\pm$**0.005**	$+4.59$	0.652 $\pm$0.021	**0.659** $\pm$**0.005**	$+1.07$
Respiratory System	**0.674** $\pm$**0.023**	0.668 $\pm$0.013	-0.79	**0.654** $\pm$**0.028**	0.637 $\pm$0.009	-2.54
Digestive System	0.628 $\pm$0.021	**0.666** $\pm$**0.020**	$+6.03$	0.613 $\pm$0.018	**0.633** $\pm$**0.017**	$+3.36$
Skin	0.641 $\pm$0.027	**0.660** $\pm$**0.015**	$+3.00$	0.635 $\pm$0.017	**0.637** $\pm$**0.014**	$+0.32$
Musculoskeletal System	0.642 $\pm$0.021	**0.687** $\pm$**0.013**	$+6.93$	0.636 $\pm$0.021	**0.659** $\pm$**0.015**	$+3.58$
Genitourinary System	0.637 $\pm$0.016	**0.671** $\pm$**0.014**	$+5.37$	0.625 $\pm$0.008	**0.647** $\pm$**0.017**	$+3.50$
Mean	0.645	**0.675**	$+4.61$	0.632	**0.650**	$+2.92$

DINOv2 baseline plateaus at ~ 0.652 AUROC and shows essentially no further gain across the remaining 120 epochs, whereas Pan-FM continues to improve and reaches 0.672 AUROC and 0.638 balanced accuracy at epoch 200 ($+2.0$ and $+1.5$ points over the baseline, respectively). The sustained improvement of Pan-FM through the full 200-epoch budget indicates that saliency-guided masking continues to provide useful learning gradient after random masking has saturated, by repeatedly challenging the representation to recover from the loss of the most informative organs.

E.3 Saliency Redistribution Under SGM

The dominant-organ shortcut learning bias (see Sec. 3.2) suggests that, without explicit intervention, a multi-organ self-supervised model tends to allocate most of its representation capacity to only a few highly attended organs, while underusing the remaining organs. We directly examine this effect by tracking per-organ saliency throughout training. Specifically, we define organ saliency as the total [CLS]-token attention mass assigned to all tokens from a given organ in the teacher backbone, computed using the full multi-organ input. We compare Pan-FM with a DINOv2 baseline that uses the same initialization and training schedule but applies uniform random masking.

At step 0 (Fig. 15 c, grey bars), the two models show almost identical saliency patterns. Attention is distributed mainly according to the number of tokens per organ: the brain, with 64 tokens, receives

Table 18: Comparison of AUROC on the held-out test set between more **single-organ baselines** trained with organ-specific data and our proposed pan-organ model across 13 group diseases, evaluated on subjects with complete 7-organ imaging data.

Organ	Method	Infect. Paras.	Neo-plasms	Blood Immune	Endoc. Metabol.	Mental Behav.	Nervous System	Eye Disease	Circulat. System	Resp. System	Digest. System	Skin System	Musculo-skeletal	Genito-urinary	Mean
Brain	Lasso	0.534	0.560	0.594	0.650	0.663	0.565	0.647	0.647	0.599	0.581	0.566	0.567	0.561	0.595
	ElasticNet	0.533	0.572	0.602	0.650	0.666	0.565	0.626	0.647	0.603	0.586	0.568	0.567	0.559	0.596
	Linear SVM	0.545	0.565	0.572	0.662	0.646	0.574	0.584	0.576	0.607	0.568	0.576	0.526	0.583	0.583
	RF	0.546	0.592	0.571	0.655	0.682	0.543	0.571	0.669	0.623	0.624	0.603	0.620	0.604	0.608
	MLP	0.515	0.538	0.638	0.657	0.644	0.563	0.629	0.655	0.586	0.605	0.561	0.596	0.564	0.596
Heart	Lasso	0.688	0.630	0.633	0.707	0.656	0.655	0.668	0.684	0.660	0.648	0.664	0.666	0.645	0.662
	ElasticNet	0.688	0.628	0.643	0.706	0.663	0.655	0.651	0.684	0.652	0.648	0.664	0.666	0.642	0.661
	Linear SVM	0.580	0.568	0.601	0.685	0.616	0.663	0.494	0.617	0.678	0.528	0.666	0.544	0.579	0.601
	RF	0.682	0.601	0.602	0.659	0.662	0.627	0.550	0.654	0.629	0.623	0.605	0.634	0.626	0.627
	MLP	0.624	0.609	0.626	0.703	0.644	0.661	0.596	0.670	0.665	0.602	0.611	0.642	0.629	0.637
Adipose	Lasso	0.640	0.665	0.592	0.719	0.644	0.645	0.716	0.679	0.665	0.663	0.660	0.678	0.667	0.664
	ElasticNet	0.653	0.665	0.592	0.723	0.647	0.647	**0.717**	0.680	0.667	0.663	0.659	0.683	0.679	0.667
	Linear SVM	0.587	0.626	0.517	0.722	0.640	0.632	0.430	0.644	0.637	0.645	0.650	0.639	0.580	0.611
	RF	0.663	0.640	0.611	0.701	0.645	0.664	0.638	0.656	0.683	0.657	0.672	0.706	0.670	0.662
	MLP	0.648	0.673	0.581	0.707	0.647	0.700	0.501	0.657	0.683	0.673	**0.692**	0.706	0.688	0.658
Liver	Lasso	0.500	0.555	0.572	0.573	0.550	0.600	0.479	0.576	0.568	0.562	0.577	0.586	0.589	0.561
	ElasticNet	0.500	0.558	0.572	0.573	0.548	0.600	0.446	0.576	0.567	0.558	0.582	0.591	0.590	0.559
	Linear SVM	0.471	0.510	0.485	0.566	0.522	0.521	0.461	0.513	0.561	0.508	0.511	0.529	0.528	0.514
	RF	0.665	0.587	0.593	0.663	0.639	0.594	0.593	0.592	0.627	0.596	0.605	0.598	0.605	0.612
	MLP	0.359	0.612	0.623	0.688	0.647	0.634	0.455	0.658	0.653	0.602	0.609	0.611	0.610	0.597
Kidney	Lasso	0.476	0.505	0.493	0.544	0.500	0.433	0.502	0.531	0.540	0.497	0.521	0.500	0.530	0.506
	ElasticNet	0.476	0.505	0.493	0.546	0.500	0.445	0.452	0.532	0.534	0.503	0.519	0.500	0.530	0.503
	Linear SVM	0.491	0.529	0.513	0.587	0.508	0.538	0.435	0.539	0.561	0.521	0.524	0.511	0.541	0.523
	RF	0.469	0.542	0.490	0.602	0.538	0.580	0.481	0.576	0.568	0.528	0.503	0.521	0.491	0.530
	MLP	0.452	0.587	0.514	0.587	0.520	0.582	0.407	0.593	0.599	0.573	0.588	0.568	0.534	0.546
Spleen	Lasso	0.576	0.521	0.526	0.525	0.533	0.531	0.475	0.515	0.499	0.500	0.528	0.487	0.490	0.516
	ElasticNet	0.576	0.521	0.526	0.525	0.533	0.531	0.443	0.511	0.499	0.500	0.527	0.487	0.490	0.513
	Linear SVM	0.581	0.482	0.525	0.515	0.475	0.456	0.517	0.502	0.483	0.513	0.455	0.466	0.483	0.496
	RF	0.482	0.548	0.548	0.585	0.540	0.542	0.443	0.552	0.520	0.534	0.569	0.507	0.561	0.533
	MLP	0.544	0.526	0.550	0.576	0.533	0.520	0.399	0.527	0.530	0.506	0.555	0.501	0.513	0.522
Pancreas	Lasso	0.578	0.593	0.619	0.618	0.580	0.607	0.637	0.618	0.618	0.628	0.597	0.585	0.608	0.607
	ElasticNet	0.578	0.593	0.619	0.617	0.579	0.617	0.636	0.619	0.621	0.629	0.598	0.581	0.611	0.608
	Linear SVM	0.593	0.599	0.654	0.618	0.533	0.630	0.643	0.613	0.608	0.633	0.606	0.587	0.622	0.611
	RF	0.582	0.562	0.610	0.626	0.592	0.626	0.541	0.597	0.609	0.607	0.578	0.546	0.589	0.590
	MLP	0.550	0.584	0.625	0.616	0.579	0.640	0.603	0.613	0.612	0.591	0.602	0.594	0.621	0.602
Pan-FM (Ours)		**0.691**	**0.702**	**0.681**	**0.772**	**0.735**	**0.737**	0.683	**0.722**	**0.685**	**0.678**	0.671	**0.708**	**0.692**	**0.704**

48% of the total [CLS] attention, whereas each 4-token organ receives only about 3%. This indicates that the model begins training with a strong token-count prior, which provides a natural starting point for shortcut learning. After self-supervised pretraining, the DINOv2 baseline reduces its reliance on the brain ($48\% \rightarrow 15\%$), but the attention budget is largely transferred to another dominant organ, adipose ($13\% \rightarrow 36\%$). Thus, the shortcut bias is not eliminated; it is simply shifted from one organ to another. In contrast, SGM uses the teacher's saliency distribution to guide masking, so organs that become dominant during training are more likely to be masked and therefore discouraged from monopolizing the representation. As a result, adipose saliency is reduced compared with the baseline (32%, -4%), and more attention is redistributed to the small-footprint organs.

At convergence, the four small-footprint organs together receive 27% of the total saliency under SGM, compared with 22% for the baseline and only 13% at initialization, even though they account for just 14% of all tokens. These results provide direct evidence that SGM mitigates the dominant-organ shortcut bias by redistributing representation capacity across organs, rather than merely moving the shortcut from one dominant organ to another.

E.4 Organ Saliency Proxy for SGM

The saliency-guided masking (SGM) module relies on a proxy function $\mathcal{A}(\cdot)$ that converts the teacher's attention maps into per-organ importance scores. Different proxies trade off between local sharpness (later layers focus on task-relevant tokens) and global integration (earlier layers preserve broader context). To validate our design choice, we compare three attention proxies for computing the per-token saliency map $\bar{\mathbf{a}} \in \mathbb{R}^N$:

A1: Last-layer attention. Only the final block's CLS-to-patch attention is used:

$$\bar{\mathbf{a}}^{\mathrm{A1}} = \frac{1}{H} \sum_{h=1}^{H} \mathbf{a}_L^h. \tag{8}$$

This proxy captures the highest-level semantic focus but discards information from intermediate layers.

Table 19: Comparison of balanced accuracy on the held-out test set between more **single-organ baselines** trained with organ-specific data and our proposed pan-organ model across 13 group diseases, evaluated on subjects with complete 7-organ imaging data.

Organ	Method	Infect. Paras.	Neo-plasms	Blood Immune	Endoc. Metabol.	Mental Behav.	Nervous System	Eye Disease	Circulat. System	Resp. System	Digest. System	Skin System	Musculo-skeletal	Genito-urinary	Mean
Brain	Lasso	0.557	0.564	0.603	0.620	0.657	0.572	0.625	0.629	0.606	0.577	0.584	0.588	0.561	0.596
	ElasticNet	0.557	0.573	0.619	0.617	0.656	0.574	0.607	0.626	0.609	0.574	0.579	0.587	0.563	0.595
	Linear SVM	0.553	0.571	0.579	0.628	0.628	0.596	0.610	0.584	0.612	0.584	0.585	0.556	0.592	0.591
	RF	0.566	0.599	0.579	0.644	0.660	0.555	0.586	0.640	0.615	0.614	0.627	0.626	0.629	0.611
	MLP	0.556	0.553	0.629	0.623	0.634	0.582	0.633	0.628	0.601	0.590	0.574	0.600	0.570	0.598
Heart	Lasso	**0.666**	0.623	0.640	0.669	0.644	0.646	0.682	0.666	0.621	0.629	0.647	0.641	0.621	0.646
	ElasticNet	**0.666**	0.623	0.619	0.665	0.650	0.645	0.670	0.666	0.632	0.629	0.647	0.641	0.619	0.644
	Linear SVM	0.603	0.559	0.607	0.649	0.603	0.660	0.552	0.594	0.636	0.548	0.652	0.572	0.593	0.602
	RF	0.664	0.590	0.595	0.654	0.646	0.623	0.573	0.643	0.644	0.614	0.595	0.628	0.602	0.620
	MLP	0.633	0.616	**0.646**	0.662	0.654	0.638	0.611	0.656	**0.656**	0.596	0.612	0.611	0.605	0.630
Adipose	Lasso	0.642	0.644	0.589	0.679	0.646	0.636	**0.722**	0.646	0.636	0.637	0.648	0.668	0.639	0.649
	ElasticNet	0.658	0.644	0.587	0.673	0.657	0.636	0.712	0.647	0.645	0.637	0.661	0.676	0.653	0.653
	Linear SVM	0.591	0.648	0.561	0.665	0.648	0.620	0.529	0.621	0.632	0.630	0.620	0.605	0.581	0.609
	RF	0.659	0.612	0.604	0.662	0.619	0.675	0.639	0.630	0.630	0.630	0.655	0.663	0.672	0.642
	MLP	0.639	0.648	0.588	0.672	0.642	0.670	0.594	0.655	0.648	0.631	**0.673**	0.674	0.662	0.646
Liver	Lasso	0.500	0.578	0.593	0.590	0.584	0.622	0.545	0.597	0.589	0.569	0.586	0.590	0.583	0.579
	ElasticNet	0.500	0.569	0.593	0.590	0.581	0.622	0.513	0.597	0.581	0.566	0.582	0.599	0.580	0.575
	Linear SVM	0.524	0.534	0.518	0.599	0.588	0.540	0.559	0.552	0.583	0.533	0.557	0.544	0.545	0.552
	RF	0.664	0.580	0.603	0.657	0.630	0.619	0.578	0.579	0.635	0.595	0.586	0.602	0.590	0.609
	MLP	0.509	0.609	0.632	0.654	0.637	0.645	0.515	0.629	0.632	0.592	0.603	0.607	0.598	0.605
Kidney	Lasso	0.524	0.536	0.562	0.564	0.500	0.504	0.550	0.549	0.555	0.517	0.552	0.500	0.544	0.535
	ElasticNet	0.524	0.536	0.552	0.561	0.500	0.506	0.547	0.549	0.556	0.525	0.548	0.500	0.544	0.534
	Linear SVM	0.556	0.552	0.552	0.596	0.538	0.597	0.556	0.565	0.563	0.558	0.563	0.534	0.550	0.560
	RF	0.538	0.557	0.543	0.599	0.555	0.573	0.528	0.567	0.586	0.535	0.538	0.541	0.539	0.554
	MLP	0.527	0.592	0.559	0.588	0.552	0.585	0.547	0.595	0.598	0.576	0.589	0.572	0.537	0.571
Spleen	Lasso	0.595	0.549	0.568	0.574	0.564	0.553	0.542	0.551	0.544	0.500	0.556	0.535	0.526	0.551
	ElasticNet	0.595	0.549	0.568	0.573	0.565	0.553	0.515	0.543	0.546	0.500	0.560	0.535	0.526	0.548
	Linear SVM	0.607	0.514	0.561	0.550	0.535	0.519	0.591	0.519	0.529	0.524	0.515	0.516	0.520	0.538
	RF	0.554	0.557	0.574	0.600	0.567	0.560	0.544	0.561	0.544	0.548	0.594	0.529	0.601	0.564
	MLP	0.556	0.553	0.560	0.594	0.567	0.556	0.520	0.547	0.534	0.541	0.565	0.527	0.528	0.550
Pancreas	Lasso	0.580	0.586	0.645	0.596	0.604	0.613	0.639	0.598	0.605	0.614	0.594	0.579	0.608	0.605
	ElasticNet	0.583	0.586	0.645	0.596	0.599	0.616	0.631	0.604	0.599	0.615	0.594	0.579	0.605	0.604
	Linear SVM	0.600	0.595	0.632	0.596	0.548	0.633	0.650	0.602	0.596	0.603	0.606	0.582	0.596	0.603
	RF	0.592	0.581	0.625	0.599	0.599	0.637	0.567	0.576	0.604	0.601	0.572	0.564	0.593	0.593
	MLP	0.562	0.578	0.623	0.595	0.613	0.635	0.617	0.596	0.601	0.588	0.585	0.589	0.618	0.600
Pan-FM (Ours)		0.640	**0.687**	0.645	**0.711**	**0.718**	**0.711**	0.688	**0.685**	0.652	**0.648**	0.628	**0.680**	**0.675**	**0.674**

Table 20: **Computational overhead of saliency-guided masking (SGM).** Measured over 1,000 iterations after 50 warmup iterations on a single V100 GPU with FP16 mixed precision. Both runs use identical model weights and batch sequences. The only difference is the masking strategy.

Metric	DINOv2	Pan-FM	Δ
Time / iter (ms)	79.3 ± 3.3	82.1 ± 0.6	$+3.5\%$
FLOPs (G)	96.4	96.4	0.0%

A2: All-layer average (default). The CLS-to-patch attention is averaged across all L blocks and H heads, which follows the default formulation in (5):

$$\bar{\mathbf{a}}^{\mathrm{A2}} = \frac{1}{LH} \sum_{l=1}^{L} \sum_{h=1}^{H} \mathbf{a}_l^h. \tag{9}$$

This balances semantic focus from later layers with broader contextual signal from earlier ones.

A3: Attention rollout [1]. Following [1], attention rollout recursively multiplies head-averaged attention matrices (with residual connections) across layers to estimate the effective information flow from CLS to each input token:

$$\tilde{\mathbf{A}}_l = 0.5 \cdot \bar{\mathbf{A}}_l + 0.5 \cdot \mathbf{I}, \qquad \mathbf{R} = \prod_{l=1}^{L} \tilde{\mathbf{A}}_l, \qquad \bar{\mathbf{a}}^{\mathrm{A3}} = \mathbf{R}_{[\mathtt{CLS},:]}, \tag{10}$$

where $\bar{\mathbf{A}}_l$ is the head-averaged attention at layer l. Rollout accounts for skip connections and compositional attention paths but can over-smooth saliency in deep networks.

Results. Table 21 compares the three proxies under linear probing across all 13 evaluation settings, averaged over 10 independent runs. To ensure a fair comparison, we performed an independent grid search of the masking temperature $\tau \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$ for each proxy and report results at the optimal value: $\tau = 0.30$ for A1, $\tau = 0.25$ for A2, and $\tau = 0.35$ for A3. A2 (all-layer average) consistently outperforms both alternatives in every setting, with margins that

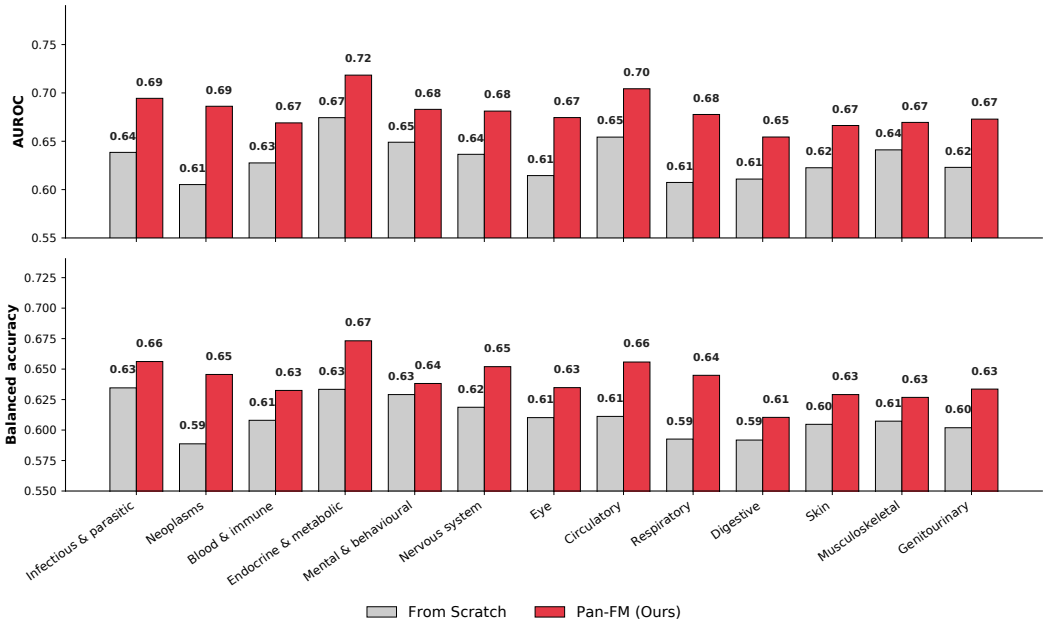


Figure 13: **Per-disease comparison: Pan-FM vs. from-scratch.** Both models use the same full ViT multi-organ backbone and are fine-tuned end-to-end with identical schedules. The from-scratch baseline initialises the backbone with the standard ViT scheme (truncated-normal) while Pan-FM initialises it with our pretrained weights.

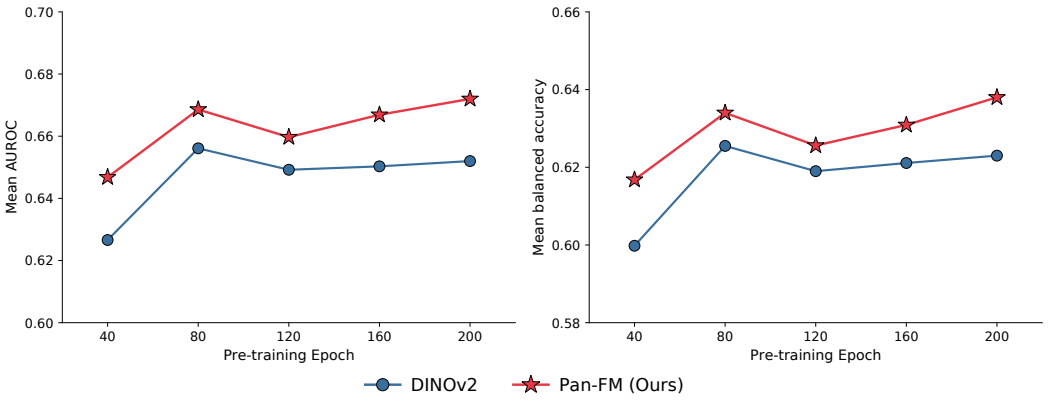


Figure 14: **Linear-probing convergence across pre-training epochs.** For each checkpoint, we report the linear probing performance on the full held-out test set as the mean AUROC (left) and mean balanced accuracy (right) across the 13 categories. Pan-FM with SGM consistently outperforms the DINOv2 baseline throughout the whole pre-training stage.

substantially exceed the run-to-run standard deviations. The gap is most pronounced under severe missingness: at *Drop-3 Organs*, A2 reaches 0.654 AUROC versus 0.622 for A1 and 0.628 for A3 ($+3.2\%$ and $+2.6\%$ respectively), while at *w/o Adipose* the gap reaches $+3.1\%$ over A1 and $+5.6\%$ over A3.

We attribute A1's underperformance to its narrow reliance on final-layer attention, where the teacher's [CLS] focus is already biased toward dominant organs that the model has learned to over-rely on—precisely the shortcut that SGM aims to disrupt. A3 partially mitigates this through layer-wise integration, but the iterative matrix multiplication in attention rollout [1] amplifies low-magnitude noise across our 12-layer ViT, leading to overly diffuse organ saliency estimates and even worse

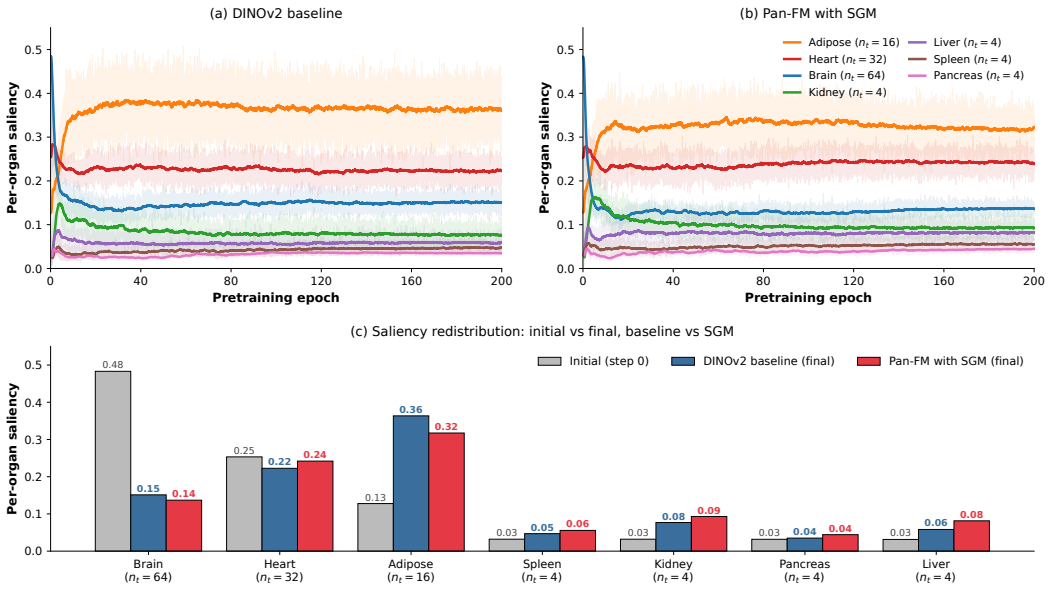


Figure 15: **The dominant-organ shortcut bias and its mitigation by SGM.** Per-organ saliency is defined as the [CLS]-token attention mass aggregated over each organ's tokens in the teacher backbone, computed on the full multi-organ input without any masking. The two runs share identical initialisation, optimisation schedule, and training data, and differ only in whether SGM is applied to the student. **(a, b)** Saliency trajectories across 200 pretraining epochs for the DINOv2 baseline and Pan-FM, respectively; faint lines show raw values and bold lines show exponential-moving-average-smoothed trajectories. **(c)** Per-organ saliency at initialisation (step 0, grey) compared to the final state of the baseline (blue) and Pan-FM (red). At initialisation, saliency is approximately proportional to each organ's token-count footprint (n_t). Self-supervised pretraining alone displaces brain but *relocates* the shortcut to a new dominant organ (adipose, 36%), while SGM mitigates this bias by redistributing representation capacity across organs, rather than merely moving the shortcut from one dominant organ to another.

performance than A1 on several settings (e.g., *w/o Adipose*). A2's simple averaging strikes the best balance: it aggregates signal across all layers without compounding noise, providing both stability (low variance across runs) and discriminative organ-level saliency.

E.5 Organ Importance Analysis

Fig. 16 reveals several interesting patterns. First, *adipose* features are near-universal contributors. Removing them degrades performance on the majority of disease categories, with the largest drops on nervous system ($\Delta = +5.84$), skin ($+5.58$), musculoskeletal ($+5.32$), and genitourinary ($+4.76$) diseases. This is consistent with the role of body composition as a systemic marker of metabolic and inflammatory state. Second, brain features are highly disease-specific. Their contribution concentrates on mental & behavioural ($+5.06$) and, to a smaller extent, neoplasms ($+1.50$) and endocrine & metabolic ($+1.18$) disorders. Third, the remaining organs show clear disease-specific specialisation. Heart contributes most to eye ($+1.85$) and infectious & parasitic diseases ($+1.90$), reflecting the cardiovascular footprint of systemic infection and microvascular eye conditions; liver is the strongest contributor for infectious & parasitic ($+2.50$) and a consistent secondary signal across endocrine, neoplastic, and metabolic categories ($+1.16$ to $+1.47$), in line with its central role in immune and metabolic homeostasis; kidney plays a notable role in blood & immune disease ($+1.42$) and mental & behavioural disorders ($+1.03$); spleen and pancreas contribute meaningfully to endocrine & metabolic ($+0.64$, $+0.21$) and mental & behavioural ($+1.15$) diseases.Overall, these results indicate that our pre-trained representation learns disease-specific reliance on specific organs rather than treating modalities uniformly.

Table 21: **Comparison of saliency proxies for SGM under linear probing.** Mean AUROC across 13 disease groups on the held-out test set, averaged over 10 independent runs (mean ± std). *Standard* uses the full test set; *Full Organs (7)* and *Drop k Organ(s)* are evaluated on the complete-7-organ subset; *w/o X* removes only organ X at test time. Best result per row in **bold**.

Setting	A1: Last-layer	A3: Rollout [1]	**A2: All-layer avg. (Default)**
Standard	0.668 ± 0.002	0.664 ± 0.002	$\mathbf{0.672 \pm 0.002}$
Full Organs (7)	0.685 ± 0.003	0.673 ± 0.002	$\mathbf{0.694 \pm 0.002}$
Drop 1 Organ	0.653 ± 0.002	0.653 ± 0.001	$\mathbf{0.681 \pm 0.002}$
Drop 2 Organs	0.649 ± 0.002	0.647 ± 0.001	$\mathbf{0.672 \pm 0.002}$
Drop 3 Organs	0.622 ± 0.004	0.628 ± 0.002	$\mathbf{0.654 \pm 0.002}$
Drop 4 Organs	0.602 ± 0.002	0.601 ± 0.002	$\mathbf{0.622 \pm 0.002}$
w/o Brain	0.663 ± 0.003	0.670 ± 0.003	$\mathbf{0.688 \pm 0.002}$
w/o Heart	0.685 ± 0.004	0.674 ± 0.003	$\mathbf{0.695 \pm 0.002}$
w/o Adipose	0.631 ± 0.003	0.606 ± 0.003	$\mathbf{0.662 \pm 0.002}$
w/o Liver	0.682 ± 0.002	0.670 ± 0.002	$\mathbf{0.684 \pm 0.002}$
w/o Kidney	0.684 ± 0.003	0.671 ± 0.002	$\mathbf{0.695 \pm 0.003}$
w/o Spleen	0.686 ± 0.003	0.680 ± 0.002	$\mathbf{0.695 \pm 0.002}$
w/o Pancreas	0.676 ± 0.003	0.672 ± 0.002	$\mathbf{0.697 \pm 0.002}$

E.6 Effect of Pre-training Loss Components

The full Pan-FM training objective combines two loss components: the DINO consistency loss $\mathcal{L}_{\text{DINO}}$ on the [CLS] token and the KoLeo regularization loss $\mathcal{L}_{\text{KoLeo}}$ [34]. To isolate the contribution of each, we re-train Pan-FM under two reduced objectives and compare against the full model: 1) DINO only: $\mathcal{L} = \mathcal{L}_{\text{DINO}}$, 2) KoLeo only: $\mathcal{L} = \mathcal{L}_{\text{KoLeo}}$, and 3) Full (default): $\mathcal{L} = \mathcal{L}_{\text{DINO}} + \lambda\mathcal{L}_{\text{KoLeo}}$ All other configurations (backbone, optimizer, SGM, masking ratio, training epochs) are held fixed.

Results. Table 22 reports linear-probing AUROC across all 13 evaluation settings. Three observations emerge. **(i)** *KoLeo alone fails to learn meaningful representations*, achieving only 0.551 mean AUROC, essentially at chance level on hardest settings (e.g., 0.521 at Drop-4). This is because KoLeo provides no supervisory target and merely pushes features apart on the unit sphere. **(ii)** *DINO alone learns a competent representation* (0.624 mean AUROC), confirming that the teacher–student consistency objective is the primary driver of feature quality, but it lags the full model by 5.4 points on average. **(iii)** *Adding KoLeo as a regularizer to DINO substantially improves performance* across every setting, with gains ranging from $+4.1\%$ (Standard) to $+5.3\%$ (Drop-3) and reaching $+7.0\%$ on *w/o Heart*. We attribute this to KoLeo's role as an anti-collapse regularizer: by encouraging uniform feature spread on the embedding sphere, it prevents the model from collapsing onto a few dominant organs and complements DINO's invariance objective. The two losses are complementary: DINO encourages discriminative representations, while KoLeo promotes a well-distributed feature space; together, they yield stronger representations than either loss alone.

E.7 Downstream Training Data Ratio

Table 23 reports linear-probing performance across five label fractions (5%-75%) under both standard and missing-organ evaluation protocols. For training with various data ratio, we construct training sets by sampling both *disease cases* and *cognitively normal (CN)* subjects according to predefined label fractions. A linear classifier is trained with these sampled data on frozen backbone features. To ensure robustness, each experimental variant is repeated for 5 runs with different random seeds, and the average performance is reported. Evaluation is performed on a fixed held-out test set for fair comparison across all methods and data regimes. Pan-FM achieves consistent performance gains over DINOv2 across all downstream training data ratios under the linear probing setting. The improvements are particularly pronounced in the low-label regime, suggesting superior data efficiency and stronger representation quality. As the amount of labeled data increases, Pan-FM continues to maintain its advantage, demonstrating that its learned representations generalize well across both scarce and abundant supervision scenarios.

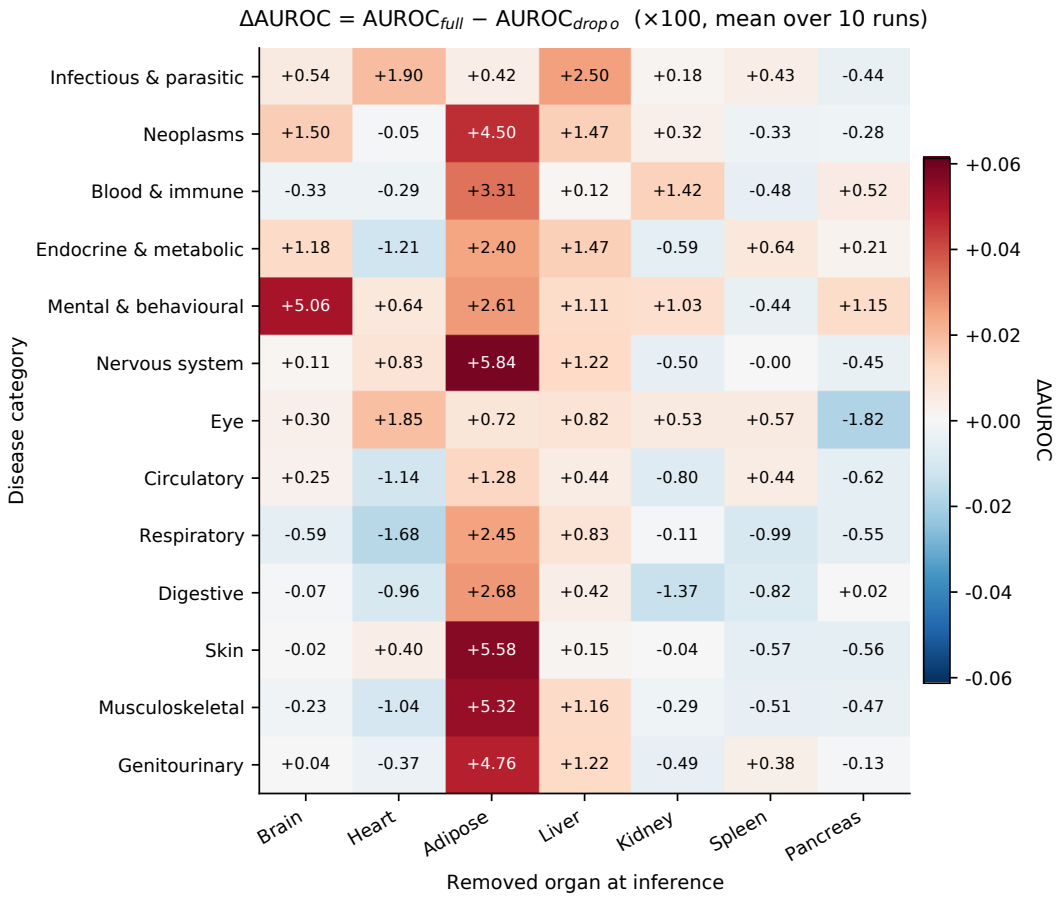


Figure 16: **Leave-one-organ-out ablation**. Each cell reports $\Delta\text{AUROC} \times 100 = 100 \times (\text{AUROC}_{\text{full}} - \text{AUROC}_{\text{drop } o})$, averaged over 10 linear-probe training runs on top of the frozen pre-trained backbone. Probes are trained on the full training set and evaluated on the 7-organ-complete test subset. Positive values (red) indicate organ o contributes to a specific disease d.

F Limitations and Future Work

Pan-FM is the first foundation model to jointly learn cross-organ representations across seven organ systems under realistic missing-organ scenarios, but the present work has several limitations that we acknowledge and view as directions for future research. First, our experiments are designed primarily as a *conceptualization* of multi-organ representation learning rather than as a clinical-grade prediction system: we operate on low-dimensional, pre-extracted MRI-derived imaging phenotypes (IDPs) instead of raw voxel data, which constrains the representational capacity of the backbone and leaves direct pre-training on raw multi-organ imaging as a natural next step. Second, the UK Biobank is a general-population cohort whose disease labels are derived from inpatient records and ICD codes rather than disease-specific phenotyping, so absolute predictive performance is inherently modest. Our goal is to demonstrate the effectiveness of Pan-FM in handling realistic missing-organ scenarios and in improving over baseline approaches, rather than to establish state-of-the-art disease prediction. Validating Pan-FM on disease-specific cohorts (e.g., ADNI [35] for neurodegeneration) is therefore an important next step. Finally, the UK Biobank is predominantly white-British and middle-aged-to-elderly. Generalization to under-represented populations and other healthcare systems requires further validation.

Table 22: **Effect of loss components on Pan-FM under linear probing.** Mean AUROC across 13 disease groups on the held-out test set. *Standard* uses the full test set; *Full Organs (7)* and *Drop k Organ(s)* are evaluated on the complete-7-organ subset; *w/o X* removes only organ X at test time. Best result per row in **bold**.

Setting	KoLeo only	DINO only	**DINO + KoLeo (Default)**
Standard	0.561	0.624	**0.672**
Full Organs (7)	0.543	0.636	**0.694**
Drop 1 Organ	0.564	0.628	**0.681**
Drop 2 Organs	0.538	0.616	**0.672**
Drop 3 Organs	0.525	0.603	**0.654**
Drop 4 Organs	0.522	0.606	**0.622**
w/o Brain	0.559	0.611	**0.688**
w/o Heart	0.546	0.626	**0.695**
w/o Adipose	0.527	0.614	**0.662**
w/o Liver	0.529	0.637	**0.684**
w/o Kidney	0.579	0.642	**0.695**
w/o Spleen	0.568	0.636	**0.695**
w/o Pancreas	0.597	0.637	**0.697**
Mean	0.551	0.624	**0.678**

Table 23: Comprehensive performance comparison (Mean AUROC) across five downstream training data ratios (5%, 10%, 25%, 50%, 75%) over 5 runs under linear probing. Bold values indicate the better performance between **DINOv2** and **Pan-FM** under each setting.

Evaluation Setting	5% Ratio		10% Ratio		25% Ratio		50% Ratio		75% Ratio	
	DINOv2	Pan-FM	DINOv2	Pan-FM	DINOv2	Pan-FM	DINOv2	Pan-FM	DINOv2	Pan-FM
Standard Protocols										
Standard	0.564	**0.580**	0.581	**0.598**	0.614	**0.628**	0.635	**0.651**	0.647	**0.663**
Full Organs (7)	0.586	**0.606**	0.615	**0.629**	0.649	**0.657**	0.668	**0.679**	0.682	**0.688**
Random Organ Dropout										
Drop 1 Organ	0.573	**0.587**	0.588	**0.620**	0.608	**0.644**	0.640	**0.659**	0.655	**0.677**
Drop 2 Organs	0.578	**0.581**	0.594	**0.606**	0.614	**0.633**	0.638	**0.653**	0.647	**0.667**
Drop 3 Organs	0.568	**0.576**	0.579	**0.595**	0.593	**0.622**	0.609	**0.637**	0.617	**0.648**
Drop 4 Organs	0.561	**0.564**	0.561	**0.582**	0.574	**0.600**	0.595	**0.606**	0.597	**0.618**
Specific Organ Dropout										
w/o Brain	0.585	**0.585**	0.592	**0.629**	0.628	**0.651**	0.665	**0.668**	0.684	**0.683**
w/o Heart	0.592	**0.606**	0.602	**0.632**	0.632	**0.663**	0.659	**0.680**	0.684	**0.691**
w/o Adipose	0.571	**0.576**	0.585	**0.611**	0.607	**0.614**	0.620	**0.645**	0.626	**0.660**
w/o Liver	0.586	**0.605**	0.608	**0.631**	0.632	**0.653**	0.665	**0.679**	0.679	**0.690**
w/o Kidney	0.586	**0.599**	0.613	**0.631**	0.641	**0.661**	0.673	**0.690**	0.688	**0.698**
w/o Spleen	0.590	**0.597**	0.607	**0.632**	0.639	**0.664**	0.679	**0.690**	0.697	**0.699**
w/o Pancreas	0.583	**0.600**	0.604	**0.634**	0.631	**0.665**	0.661	**0.691**	**0.677**	0.676

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: We present Pan-FM, to the best of our knowledge, the first multi-organ FM for jointly learning cross-organ representations from 7 organ systems in the presence of missing organs. We identify dominant-organ shortcut learning bias as a key limitation in multi-organ SSL for missing robustness, and propose Saliency-Guided Masking (SGM), a plug-and-play strategy that facilitates more balanced cross-organ representation learning with negligible overhead and no architectural changes, and integrates seamlessly into existing SSL frameworks.

   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: The limitation and future work are discussed in Appendix Section F.

   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: [No]

   Justification: This paper does not contain formal theoretical results. The proposed Saliency-Guided Masking (SGM) is an empirical methodology motivated by qualitative analysis of dominate-organ learning bias. We do not claim any theorem.

   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: All the implementation details are included in Section 4 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: [No]

   Justification: We release the training and evaluation code. Raw UK Biobank data cannot be redistributed by us due to UK Biobank's Material Transfer Agreement. We provide the full list of UKBB field IDs and brain ROI definitions in Appendix B so that, given approved UK Biobank access, the input feature space is exactly reconstructible.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All the details are provided in Section 4 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: We report downstream evaluation results averaged over multiple runs to account for variability.

   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: All experiments are conducted on a single NVIDIA V100 GPU. A full 200-epoch Pan-FM pre-training run takes approximately 8 hours. Running details are in Appendices C and E.1.

   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: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Broader impacts are discussed in Appendix **??**.

    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: [N/A]

    Justification: The model is a foundation model for biomedical representation learning, not a generative model or scraped-data model. It does not generate text, images, or other content with high misuse risk.

    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: All third-party assets used in this work are properly cited.

13. **New assets**

    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.

14. **New assets**

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

Answer: [Yes]

Justification: We will release pre-trained Pan-FM checkpoints, training/evaluation code, and a documentation README specifying input feature schemas (UKBB field IDs, MUSE ROI list, normalization statistics) upon publication. The released code includes scripts for re-running all reported experiments given UK Biobank access.

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.

15. **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: We do not perform crowdsourcing or recruit any new human subjects.

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.

16. **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: This research uses UK Biobank, which has ethical approval from the North West Multi-Centre Research Ethics Committee (REC reference 11/NW/0382), , and all participants provided written informed consent prior to data collection. The present study used only de-identified data made available through UK Biobank. No participant-level identifiable information was accessed.

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.

17. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core methodology. LLMs were used only for editing and language polishing of the manuscript text.

    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.