The Metric, Not the Model: Anatomy Confounds Attention Faithfulness in Coronary Calcium

Ashajyothi Bathineni[1*], Siddhant Mishra[4], Marie Zöga Diederichsen[2], Axel Diederichsen[3], Vinay Chakravarthi Gogineni[1], and Victoria Blanes-Vidal[1]

[1] Applied AI and Data Science, The Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark
asba@mmmi.sdu.dk
[2] Department of Radiology, Odense University Hospital, J. B. Winsløws Vej 4, 5000 Odense, Denmark
[3] Department of Cardiology, Odense University Hospital, J. B. Winsløws Vej 4, 5000 Odense, Denmark
[4] Department of Electronics and Instrumentation, Birla Institute of Technology and Science, Pilani, 333031 Rajasthan, India

Abstract. A benchmark score cannot say whether a medical foundation model reads a disease or merely a correlate of it. The common recourse is attention faithfulness, which scores a model's attention against expert annotations and reads agreement as evidence of grounding. On coronary artery calcification (CAC), that agreement conflates patient-specific evidence with the patient-independent anatomical prior. We freeze three ViT-L encoders and train a shared attention-MIL head on the features of each. Their detection intervals overlap, whereas faithfulness ranges from 0.554 to 0.909; read against the chance baseline of 0.5, that range would suggest that only the 3D medical encoder localises the disease. Calcium occurs throughout the thorax. The Agatston score, however, counts only what lies in the coronary arteries, so every annotated lesion sits on the coronary tree, and every volume shares a resampling grid. A patient-independent prior built in the encoder's own instance space therefore reaches 0.937 against the 3D encoder's 0.909, and no encoder beats it. Replacing a patient's attention with the leave-one-out attention prior changes the 3D encoder's faithfulness by +0.001. In this cohort, the attention–annotation agreement conventionally reported as faithfulness is largely explained by where lesions typically occur rather than by additional localisation of an individual patient's lesion. This does not mean the 3D encoder ignores calcium. Grafting a donor lesion into a CAC-negative volume raises its predicted probability by 0.733 and flips 85% of those patients, with a response already visible at 4 voxels, while the 2D encoders show weaker responses. Where lesion location is anatomically constrained, as for CAC, faithfulness should be reported alongside its gain over a patient-independent prior, rather than interpreted against chance alone, so that what attention adds beyond anatomical expectation

*Corresponding author: Ashajyothi Bathineni, asba@mmmi.sdu.dk

is visible, and grounding claims should be complemented by controlled intervention-based evidence.

Keywords: Interpretability · Foundation models · Attention · Evaluation · Cardiac CT.

1 Introduction

Medical foundation models are usually compared by a single benchmark score. However, a score describes a model's output rather than its representation, and cannot distinguish a model that has learned a disease from one that has learned a correlate of it [12]. The distinction matters when a model is deployed on a new scanner fleet or population, where correlates may shift. Attention faithfulness is increasingly used to supply the missing evidence. The model's attention is scored against expert annotations, and agreement is interpreted as evidence that the representation is grounded in the pathology.

We argue that on CAC this interpretation conflates two things, because it depends on a baseline the measure does not state. Scoring attention against annotations with an AUC compares the result to 0.5, the value expected when attention falls on the lesion by chance. That comparison implicitly assumes a spatially uniform null distribution of lesion locations. Calcium occurs throughout the thorax, including the valves, the aorta and the skeleton, but the Agatston score counts only calcium in the four coronary arteries, so every annotated lesion lies on the coronary tree. When volumes are resampled onto a common grid, that tree occupies approximately the same position in every patient. A model that always attends to that region therefore scores highly for localisation without identifying an individual patient's lesion.

Coronary artery calcification (CAC) is a demanding setting for the measure. Lesions occupy a few voxels of a volume whose non-zero signal is dominated by bone, and COCA provides the per-lesion annotations that make the evaluation possible.

We freeze three encoders and train a shared attention-MIL head on the features of each. Their detection intervals overlap, whereas faithfulness ranges widely and identifies the 3D medical encoder as the one that localises what it detects. That inference does not follow, because an attention map that has not seen a patient can rank that patient's calcium, and because such maps order the three encoders exactly as the raw score does. A lesion grafted into a patient who had none, however, moves that same encoder strongly while the others barely respond. Its frozen representation therefore carries information responsive to the calcium-containing signal, and its attention did not reveal this.

The difficulty is not the familiar one that attention weights need not describe the computation, which is what motivates scoring them against annotation in the first place. It is that on CAC the score brought in to settle the question inherits the anatomy, so a high score certifies grounding that has not been shown, and a score at the prior does not deny it. Earlier work reads a failed explanation as

grounds for distrusting the model. Here the graft indicates otherwise, because the encoder the metric favoured is the one whose prediction responds to the inserted calcium, so on this task the metric does not reveal the calcium-sensitive information accessible through the representation.

Frozen encoders from self-supervised and supervised pretraining rival supervised transfer in natural images [10, 22, 26] and in medicine [5, 15, 20, 34], including volumetric ones [35, 27], most recently 3DINO [31] on ∼100k medical volumes. Probing reports what is linearly accessible rather than what a model uses [3, 7], and attention, the more common explanation for vision transformers [1, 11], has long-contested faithfulness to the computation [18], which motivates scoring it against annotation rather than inspecting it qualitatively. That practice has itself been audited, and repeatedly found wanting against baselines that ignore the patient. Adebayo et al. [2] show saliency maps can be insensitive to model randomisation; Arun et al. [4] evaluate eight saliency methods for medical-image localisation and find that the methods often perform no better than simple baseline measures across their trustworthiness criteria. Bigolin Lanfredi et al. [8] register chest radiographs to a common frame and construct a lung-location center-bias baseline, showing that such anatomical bias can substantially affect saliency–gaze similarity metrics. Closest to us, Harvey et al. [14] benchmark attention-MIL on 3D neuroimages with frozen ViTs and find a centre-focused Gaussian that ignores the image entirely outperforms every learned attention method on instance-level localisation, attributing this to the typical neuroanatomic distribution of the lesion. Our contribution is therefore not the observation that patient-independent anatomical priors can outperform learned explanations. It is that we construct a leave-one-out attention prior directly in each model's own instance space, without using the evaluated patient's annotation, so as to quantify the component of attention localisation that exceeds the patient-independent anatomical prior, and that we pair this diagnostic with lesion transplantation to ask a separate question of whether pathology-sensitive information is present in the representation when attention provides little patient-specific gain. We study this distinction for CAC, where pathology is anatomically constrained on a shared coordinate system and diseases with more variable spatial distributions may strike a different balance between the anatomical prior and patient-specific information.

Grafting a donor's lesion into another patient is likewise established machinery [25, 32, 28], as are counterfactual editors that perturb pathology to expose shortcuts [23]. Unlike the insertion-deletion family [24], which perturbs an image under the guidance of the saliency map in order to score that map, we graft annotated pathology into a patient who had none, in order to score the representation. Mechanistic interpretability relates internal computation to human-interpretable structure [21, 9] and has so far been applied mainly outside medical imaging. Coronary calcium scoring is itself well studied as a supervised pipeline [30, 29, 33]. We train no detector, reading frozen encoders out with attention-based multiple-instance learning [16, 19, 13] and using calcium as a probe.

2 Method

2.1 Data

COCA (Stanford; 786 gated cardiac CT with expert per-artery calcium ROIs) is the annotated cohort. ROIs are rasterised by sorting slices on `ImagePosition-Patient` and reading the XML `ImageIndex` as a 0-based index into that order. Matching on `InstanceNumber` instead, as some public code does, puts only 1.8% of ROI pixels on calcium above 130 HU, against 66.9% here. Volumes are HU-windowed from $[130, 800]$ to $[0, 1]$ and resampled to $64 \times 128 \times 128$, so every voxel below the Agatston threshold is zero and the non-zero signal is calcium together with bone. The Agatston score from the ROIs defines presence, meaning any calcium versus none.

2.2 Encoders and read-outs

Three frozen ViT-L encoders span pretraining domain and input dimensionality. They are DINOv3 (2D, natural-image self-supervision), 3DINO (3D, medical self-supervision on ∼100k volumes), and ImageNet-sup (2D supervised control). The 2D encoders embed each axial slice independently and emit 64 per-slice CLS tokens; 3DINO encodes the volume and emits 256 patch tokens on a $4 \times 8 \times 8$ grid. Nothing in the encoders is trained. Patient-level prediction uses a gated attention-MIL head [16] over these instances, so all cross-slice aggregation for the 2D encoders occurs in the head rather than the representation. All arms share an 80/10/10 split (seed 42), where the head trains on the 80%, the epoch is chosen on the validation split, and the test split (N=79, 45 CAC-positive) is scored once. One trained head per encoder and seed supplies every number we report about that encoder. Point estimates are means over five seeds; intervals are 95% bootstrap intervals over test patients, since at N=79 patient sampling dominates seed noise.

2.3 Faithfulness, and the baseline it needs

Each MIL instance corresponds to a region of the volume, and the expert mask records whether that region contains calcium, so we can ask where the attention-MIL head attended. Faithfulness is the per-patient AUC of the attention weights ranking the calcium-bearing instances, averaged over CAC-positive test patients. It is conventionally read against 0.5, the value expected when attention is unrelated to the location of the calcium. We use the term as the medical imaging literature does, although Jacovi and Goldberg [17] distinguish agreement with annotation from faithfulness, reserving the latter for correspondence to the computation. We retain the term for comparability with prior medical-imaging evaluations, but interpret the measured quantity only as attention–annotation agreement unless further evidence links attention to the computation. We therefore separate four things throughout. *Attention–annotation agreement* is the raw

overlap score. *Patient-specific localisation gain* is the component of it that exceeds the patient-independent anatomical prior, which ΔF (Eq. 1) measures. *Representation sensitivity* is the prediction response under lesion intervention (Sec. 2.4). *Grounding* is the stronger interpretation, requiring evidence that links the computation to the pathology. ΔF does not by itself establish representation quality or computational faithfulness, and conversely a ΔF near zero does not imply that the representation lacks pathology information. We show that attention–annotation agreement and representation sensitivity come apart in both directions.

Testing against 0.5 asks whether attention knows anything, when the question is whether it knows anything about the patient being scored. We ask the second question using two baselines, each built in whichever instance space the encoder uses, so that neither depends on the token grid. The leave-one-out prior replaces a patient's attention with the mean attention over the other test patients, giving a map that has not seen them. The population-frequency prior instead ranks instances by empirical calcium frequency across the other patients. It is a label-informed reference for the patient-independent anatomical prior in this cohort and provides a complementary comparison to the leave-one-out attention prior. In contrast, the leave-one-out attention prior uses only the model's own attention, so it can be computed at evaluation time for any saliency method without access to the patient's annotations. Neither prior is used to train, select, or tune anything and both are computed after the encoder and head are frozen, leaving one patient out at a time, so no patient contributes to the prior it is scored against.

The leave-one-out attention prior gives the quantity we argue should be reported alongside the raw score as a patient-independent anatomical reference when anatomy is strongly constrained. Let a_p be the attention the model produces for patient p, let $\bar{a}_{-p}$ be the mean attention over the other patients, and let $F(\cdot)$ be faithfulness. The patient-specific localisation gain is then

$$\Delta F = \mathbb{E}_p\big[F(a_p) - F(\bar{a}_{-p})\big], \tag{1}$$

which measures the patient-specific component of attention–annotation agreement that remains after accounting for the patient-independent anatomical prior. Positive values indicate a patient-specific localisation gain over the patient-independent anatomical prior, whereas values near zero indicate that localisation can be explained almost entirely by the shared anatomy. Computing it costs one further evaluation pass, and it supplies a more appropriate null when lesion locations are anatomically constrained.

2.4 Intervening on the calcium-containing signal

Attention indicates where a read-out looks, not what it requires, so we also edit the input and re-encode. Deletion sets the annotated lesion to zero in a CAC-positive volume. Transplant does the converse, grafting a donor's annotated lesion into a CAC-negative volume at the donor's coordinates. It is the more

Table 1. Patient-level CAC on the COCA test split (N=79; faithfulness over the 45 CAC-positive patients). Frozen encoders, shared attention-MIL read-out. Entries are means over 5 seeds, ± half the 95% bootstrap interval over patients. The leave-one-out prior replaces a patient's own attention with the mean map over the other patients, so ΔF (Eq. 1) is what looking at this patient adds. The population-frequency prior ranks instances by how often calcium occurs across the other patients, and is computed in each encoder's own instance space, so it is directly comparable to the score beside it. Detection intervals overlap and faithfulness appears to rank the encoders, yet no encoder's attention exceeds its own patient-independent prior: ΔF is statistically indistinguishable from zero for 3DINO and negative for the 2D encoders. No encoder exceeds its corresponding population-frequency prior.

	Presence	Faithfulness (AUC of attention ranking calcium)			
Encoder	AUC	$F(a_p)$	$F(\bar{a}_{-p})$	ΔF	population-frequency prior
DINOv3 (2D)	0.773 ± 0.102	0.617 ± 0.058	0.703	-0.086 ± 0.063	**0.787**
ImageNet-sup (2D)	0.752 ± 0.108	0.554 ± 0.056	0.639	-0.085 ± 0.072	**0.787**
3DINO (3D)	0.907 ± 0.073	0.909 ± 0.029	0.908	$+0.001 \pm 0.021$	**0.937**

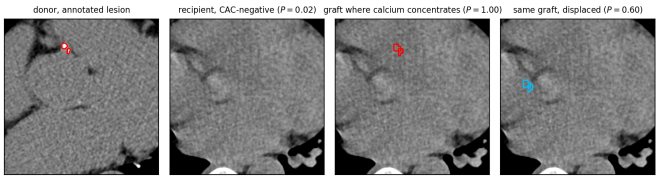

Fig. 1. A transplant and its control. A donor's annotated lesion (left, 48 voxels) is grafted into a CAC-negative recipient where coronary calcium concentrates, and, identically, at a non-coronary cardiac site. 3DINO goes from P=0.02 to 1.00 in the first case but only 0.60 in the second.

informative of the two because it does not need the annotation to be exhaustive (Sec. 3). A graft is accepted only when at least 80% of its voxels fall inside the calcium-prone region, the union of the training positives' lesions dilated by three voxels, which rejects 35 of 136 pairings. That region is a cohort envelope rather than the recipient's own vessel, which a CAC-negative patient does not have annotated. Displacing out of it therefore moves the graft to a non-coronary cardiac site, although this does not guarantee exclusion from every coronary structure. Donors are binned by lesion size and all accepted pairings are retained for analysis.

3 Results and Discussion

Detection performance. Linear probes on the encoders' frozen final-layer CLS tokens predict CAC presence at 0.79 (3DINO), 0.77 (DINOv3) and 0.66 (ImageNet-sup) AUC, so different amounts of CAC information are linearly accessible from their features, and the three are not equivalent representations. Under the shared

Table 2. Grafting a donor's lesion into a CAC-negative volume, by donor lesion size (the median COCA lesion is 17 voxels). Entries are the change in predicted CAC probability, mean over 5 seeds, ± half the 95% bootstrap interval over recipients. In (a) the graft lies where coronary calcium concentrates, and 3DINO responds substantially more strongly than the 2D encoders, particularly at realistic lesion sizes. (b) The identical graft at a non-coronary cardiac site, where valvular or annular calcium would sit and which the Agatston score excludes; only location changes. (c) The difference. 3DINO is location-sensitive: displacing the graft reduces its response at every size, with intervals excluding zero, which indicates the response is not solely attributable to the edit. The reduction is far from complete, and (b) remains most of (a). The 2D encoders show no difference, so their smaller responses in (a) cannot be told apart from the edit.

	change in predicted CAC probability, Δp		
Grafted lesion	DINOv3 (2D nat.)	ImageNet-sup (2D sup.)	3DINO (3D med.)
(a) graft where coronary calcium concentrates			
4 voxels	$+0.017 \pm 0.014$	$+0.027 \pm 0.039$	$+\mathbf{0.229} \pm 0.128$
10 voxels	$+0.043 \pm 0.035$	$+0.006 \pm 0.007$	$+\mathbf{0.406} \pm 0.142$
53 voxels	$+0.042 \pm 0.030$	$+0.084 \pm 0.048$	$+\mathbf{0.706} \pm 0.125$
189 voxels	$+0.182 \pm 0.058$	$+0.271 \pm 0.095$	$+\mathbf{0.733} \pm 0.109$
(b) same graft at a non-coronary cardiac site			
4 voxels	$+0.015 \pm 0.014$	$+0.024 \pm 0.036$	$+0.117 \pm 0.091$
10 voxels	$+0.031 \pm 0.029$	$+0.007 \pm 0.006$	$+0.210 \pm 0.116$
53 voxels	$+0.038 \pm 0.029$	$+0.060 \pm 0.041$	$+0.317 \pm 0.131$
189 voxels	$+0.162 \pm 0.053$	$+0.231 \pm 0.093$	$+0.459 \pm 0.120$
(c) difference, (a) − (b), the part that needs the location			
4 voxels	$+0.002 \pm 0.003$	$+0.003 \pm 0.004$	$+\mathbf{0.112} \pm 0.084$
10 voxels	$+0.012 \pm 0.012$	-0.001 ± 0.005	$+\mathbf{0.196} \pm 0.178$
53 voxels	$+0.004 \pm 0.006$	$+0.023 \pm 0.031$	$+\mathbf{0.389} \pm 0.137$
189 voxels	$+0.020 \pm 0.022$	$+0.040 \pm 0.072$	$+\mathbf{0.275} \pm 0.095$

attention-MIL read-out, however, patient-level detection intervals overlap (0.907, 0.773 and 0.752; Table 1), so detection performance alone cannot determine which representation is better grounded, which is the setting where attention faithfulness is typically invoked.

Attention–annotation agreement ("faithfulness"). Faithfulness ranges from 0.554 for the supervised control, whose interval contains 0.5, through 0.617 for DINOv3, to 0.909 for 3DINO. Read against chance, this ordering suggests that the 3D medical encoder attends to the calcium it reports while the control does not.

Comparison against a patient-independent prior. That reading assumes a spatially uniform null distribution of lesion locations, and the annotated calcium is not uniformly distributed. On the common grid only 55 of 256 patches ever contain any, the twenty most frequent hold 71%, and the median patient contributes 3. Most of the answer to where a patient's calcium lies is therefore available before

the patient is seen. Paired over patients in each encoder's own instance space, the population-frequency prior has a higher mean attention–annotation agreement than each encoder's attention, by +0.028 [+0.002, +0.056] for 3DINO and by +0.170 and +0.233 for the 2D encoders (sign-flip permutation $p = 0.049$, 0.0001 and < 0.0001). The 3DINO margin is narrow, and the population-frequency prior leads in only 51% of individual patients, as expected for a population-level spatial prior that need not match any particular patient's lesion. The leave-one-out attention prior, which replaces a patient's attention with the mean attention map over the other patients, changes 3DINO's faithfulness from 0.909 to 0.908, a gain of $\Delta F = +0.001$ [−0.020, +0.021], while the 2D encoders score 0.086 and 0.085 below it. The claim therefore does not rest on the narrow population-frequency prior margin: ΔF is tightly concentrated around zero, so within this cohort we find no evidence of a patient-specific localisation gain over the patient-independent anatomical prior. This does not depend on the prior being estimated from the evaluation cohort. Re-estimated from the 351 CAC-positive training patients, or from the 38 held-out validation patients, the leave-one-out attention prior gives 3DINO $\Delta F = -0.003$ [−0.025, +0.016] and −0.002 [−0.024, +0.020], and places the 2D encoders further below their priors (−0.152 and −0.220 against the training prior). The leave-one-out attention prior preserves the same encoder ranking as the raw attention score (0.908, 0.703, 0.639), so the ranking is recoverable from maps that never saw the patient scored. 3DINO's attention varies between patients (mean pairwise cosine 0.24 to 0.36) and correlates with the calcium frequency map (Spearman ρ 0.42 to 0.56), so it is neither idle nor constant. It is nonetheless largely explained by a patient-independent anatomical prior concentrated on the coronary tree, and carries little additional information about the individual lesion.

Lesion transplantation. It does not follow that the frozen features carry no calcium information. Grafting a donor's annotated lesion into a CAC-negative volume moves 3DINO in proportion to the graft, from +0.229 at 4 voxels to +0.733 at 189, where it flips 85% of these patients to "present" (Table 2, Fig. 1). Four voxels already flip 28%. The prediction is therefore strongly sensitive to the inserted calcium-containing signal, indicating that calcium-sensitive information is available in the frozen representation. This response occurs despite essentially no patient-specific localisation gain: the head already assigns substantial attention to the region expected under the patient-independent anatomical prior before any graft is introduced. The 2D encoders are substantially less sensitive across sizes, moving by +0.017 and +0.027 where 3DINO moves by +0.229, so the intervention most clearly distinguishes 3DINO from the two 2D encoders, where detection performance could not. We therefore do not generalise the transplantation result equally across all three encoders.

Location control. A graft is an edit, so the response might be a reaction to the edit rather than to the calcium. Repeating the identical paste at a non-coronary cardiac site addresses this, because only the location changes. That site is also the clinically relevant one, since calcium on the aortic valve or mitral

annulus is common and excluded from the Agatston score, so a CAC model should discount it. 3DINO loses part of its response at every size, from +0.112 at 4 voxels to +0.389 at 53 (Table 2), providing evidence that the response is not solely attributable to the edit and that location influences it. It does not discount non-coronary calcium enough, since +0.46 of its +0.73 response to the largest grafts survives displacement, so calcium excluded from the Agatston CAC definition still drives its prediction most of the way up. The 2D encoders lose nothing (+0.002 to +0.020), so a response that does not depend on location cannot be told apart from the edit, which limits what we can say about them.

Calcium outside the coronary arteries also explains why deletion is inconclusive. COCA annotates only the four coronary arteries, so erasing the annotated lesion leaves valvular and annular calcium untouched, and that residue keeps 3DINO's prediction high. It falls by only 0.026 [−0.007, +0.067], with a median of 0.001.

The result does not depend on the graft being a hard paste. Under the LesionMix soft-mask rule [6], which feathers the boundary, 3DINO gives +0.211, +0.261, +0.705 and +0.737 across the four sizes, against +0.229, +0.406, +0.706 and +0.733 for the paste, and the 2D encoders are unchanged.

What the score does and does not support. 3DINO scores 0.909 yet sits at its prior and is the encoder a graft moves most, so the score is neither sufficient evidence of lesion-specific representation nor, at the prior, evidence against it. A practitioner would select 3DINO and the graft suggests the choice is correct, so the ranking was sound while the mechanism it implied was not. What it ranks is how closely each encoder's average attention follows the coronary tree, which here happens to order them as their detection performance does and, most clearly for 3DINO, as their intervention sensitivity does; the score itself does not establish that this ranking reflects pathology-specific computation. The central dissociation is therefore between attention localisation and representation sensitivity. 3DINO gains essentially nothing over the patient-independent anatomical prior in attention localisation, yet responds strongly and location-dependently to an inserted lesion.

Scope. We claim nothing beyond coronary calcium, having tested one pathology on one cohort, and ΔF removes only the anatomical confound identified here. The effect should be strongest when the annotated pathology occupies a constrained distribution on a shared image coordinate system, as coronary calcium does on a common resampling grid; diseases with more variable or multifocal locations may behave differently. What we offer is therefore a condition, checkable in a dataset before any model is trained, under which attention–annotation agreement should not be read as patient-specific grounding, rather than a claim that ΔF universally replaces conventional faithfulness evaluation.

Limitations. The prior control concerns attention rather than the representation, so we do not claim these encoders fail to localise calcium, only that attention–annotation overlap cannot establish whether they do. The transplantation

experiment does not isolate calcium from all effects of the edited image, and both it and the faithfulness analysis rest on one annotated cohort of 45 CAC-positive test patients, so our intervals quantify uncertainty within this cohort rather than external validity. Our conclusions also concern one attention-MIL read-out, and whether the same anatomy-driven confound appears across other pathologies, datasets and explanation methods is an important direction for further study.

4 Conclusion

On coronary calcium, attention-mask agreement conflates patient-specific evidence with the patient-independent anatomical prior. A map that never saw the patient out-scores every encoder's attention, yet the graft shows that the encoder it favours is strongly sensitive to the inserted calcium-containing signal, despite showing essentially no patient-specific localisation gain. The result concerns interpretation rather than a verdict on attention. Where pathology is anatomically constrained, attention–annotation agreement should be reported together with its gain over a patient-independent anatomical prior, and evidence for grounding should be complemented by controlled interventions that test representation-level sensitivity.

Funding/Support

This work was supported by Innovation Fund Denmark for the DETECT-AI project (Grand Solutions, Grant No. 4351-00005B), and Region of Southern Denmark Research Foundation (2025-0049). The sponsors played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 4190–4197 (2020)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 9505–9515 (2018)
3. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)
4. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M.D., Kalpathy-Cramer, J.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence **3**(6), e200267 (2021). https://doi.org/10.1148/ryai.2021200267
5. Azizi, S., Mustafa, B., et al.: Big self-supervised models advance medical image classification. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3478–3488 (2021)

6. Basaran, B.D., Zhang, W., Qiao, M., Kainz, B., Matthews, P.M., Bai, W.: LesionMix: A lesion-level data augmentation method for medical image segmentation. In: Data Augmentation, Labelling, and Imperfections (DALI), MICCAI Workshop. LNCS, vol. 14379, pp. 73–83. Springer (2024). https://doi.org/10.1007/978-3-031-58171-7˙8
7. Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics **48**(1), 207–219 (2022)
8. Bigolin Lanfredi, R., Arora, A., Drew, T., Schroeder, J.D., Tasdizen, T.: Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays. In: Gaze Meets ML Workshop, NeurIPS (2022), arXiv:2112.11716
9. Bricken, T., Templeton, A., Batson, J., et al.: Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread (2023), https://transformer-circuits.pub/2023/monosemantic-features
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021)
11. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 782–791 (2021)
12. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z
13. Han, Z., Wei, B., Hong, Y., et al.: Accurate screening of covid-19 using attention-based deep 3d multiple instance learning. IEEE Transactions on Medical Imaging **39**(8), 2584–2594 (2020). https://doi.org/10.1109/TMI.2020.2996256
14. Harvey, E., Loevlie, D.J., Satani, A.A., Chen, W., Kent, D.M., Hughes, M.C.: A multi-dataset benchmark of multiple instance learning for 3D neuroimage classification. In: Conference on Health, Inference, and Learning (CHIL). Proceedings of Machine Learning Research, PMLR (2026), arXiv:2604.26807
15. Huang, S.C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digital Medicine **6**(74) (2023). https://doi.org/10.1038/s41746-023-00811-0
16. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning (ICML). PMLR, vol. 80, pp. 2127–2136 (2018)
17. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 4198–4205 (2020)
18. Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of NAACL-HLT. pp. 3543–3556 (2019)
19. Lu, M.Y., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**, 555–570 (2021). https://doi.org/10.1038/s41551-020-00682-w
20. Moor, M., Banerjee, O., et al.: Foundation models for generalist medical artificial intelligence. Nature **616**, 259–265 (2023). https://doi.org/10.1038/s41586-023-05881-4
21. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., Carter, S.: Zoom in: An introduction to circuits. Distill (2020). https://doi.org/10.23915/distill.00024.001

22. Oquab, M., Darcet, T., Moutakanni, T., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), arXiv:2304.07193
23. Pérez-García, F., Bond-Taylor, S., Sanchez, P.P., van Breugel, B., Castro, D.C., Sharma, H., Salvatelli, V., Wetscherek, M.T.A., Richardson, H., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O., Ilse, M.: RadEdit: Stress-testing biomedical vision models via diffusion image editing. In: European Conference on Computer Vision (ECCV). LNCS, vol. 15102, pp. 358–376. Springer (2024). https://doi.org/10.1007/978-3-031-73254-6˙21
24. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: British Machine Vision Conference (BMVC) (2018)
25. Pezeshk, A., Petrick, N., Chen, W., Sahiner, B.: Seamless lesion insertion for data augmentation in CAD training. IEEE Transactions on Medical Imaging **36**(4), 1005–1015 (2017). https://doi.org/10.1109/TMI.2017.2648229
26. Siméoni, O., Vo, H.V., Oquab, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
27. Tang, Y., Yang, D., et al.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20698–20708 (2022)
28. Tushar, F.I., Momy, U.H., Lo, J.Y., Rubin, G.D.: iTrialSpace: Programmable virtual lesion trials for controlled evaluation of lung CT models. arXiv preprint arXiv:2605.05761 (2026)
29. van Velzen, S.G., et al.: Deep learning for automatic calcium scoring in ct: Validation using multiple cardiac ct and chest ct protocols. Radiology **295**(1), 66–79 (2020). https://doi.org/10.1148/radiol.2020191621
30. Wolterink, J.M., et al.: Automatic coronary artery calcium scoring in cardiac ct angiography using paired convolutional neural networks. Medical Image Analysis **34**, 123–136 (2016). https://doi.org/10.1016/j.media.2016.04.004
31. Xu, T., Hosseini, S., Anderson, C., Rinaldi, A., Krishnan, R.G., Martel, A.L., Goubran, M.: A generalizable 3d framework and model for self-supervised learning in medical imaging. npj Digital Medicine **8**(639) (2025). https://doi.org/10.1038/s41746-025-02035-w
32. Yang, J., Zhang, Y., Liang, Y., Zhang, Y., He, L., He, Z.: TumorCP: A simple but effective object-level data augmentation for tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). LNCS, vol. 12901. Springer (2021). https://doi.org/10.1007/978-3-030-87193-2˙55
33. Zeleznik, R., et al.: Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nature Communications **12**(715) (2021). https://doi.org/10.1038/s41467-021-20966-2
34. Zhou, Y., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023). https://doi.org/10.1038/s41586-023-06555-x
35. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical Image Analysis **67**, 101840 (2021). https://doi.org/10.1016/j.media.2020.101840