Beyond Clean Test Sets: Spurious Correlations in Medical Vision-language Models and the Role of Concept Supervision

Valentina Corbetta[1,2,3,4], Antonio Portaluri[1,5], Muzhen He[1,6], Daniël Boeke[1,2,3], Regina Beets-Tan[1,7,8], Veronica Lachi[4], Elisabeth Wetzer[4], Robert Jenssen[4,8,9], Wilson Silva[1,3⋆], and Kristoffer Wickstrøm[4⋆]

[1] Department of Radiology, The Netherlands Cancer Institute v.corbetta@nki.nl
[2] GROW Research Institute for Oncology and Reproduction, Maastricht University
[3] AI Technology for Life, Department of Information and Computing Sciences, Department of Biology, Utrecht University
[4] Department of Physics and Technology, UiT The Arctic University of Norway
[5] Department of Radiology and Nuclear Medicine, Radboud UMC
[6] Department of Radiology, Shengli Clinical Medical College of Fujian Medical University, Fujian Provincial Hospital,Fuzhou University Affiliated Provincial Hospital
[7] Maastro
[8] Norwegian Computing Center
[9] Pioneer Centre for AI, University of Copenhagen

Abstract. Vision–language models (VLMs) are increasingly used in medical imaging, yet their robustness to spurious correlations remains insufficiently characterized. We propose a controlled evaluation framework that uses synthetic artifacts modelled after common acquisition confounders to parametrically vary correlation strength, and pairs two complementary test protocols — artifact removal and artifact inversion — to isolate whether models rely on clinical features or visual shortcuts. Applying the framework to diabetic retinopathy grading in fundus photography and BI-RADS-based assessment in mammography, we evaluate five architectures spanning a spectrum from no concept supervision to full multi-level image–concept alignment. We find that VLM backbones retain clinical signal when shortcuts are absent, yet actively follow spurious associations when they conflict with pathology, degrading faster than standard visual backbones — a dual encoding that is only exposed when evaluation goes beyond clean test sets. Among concept-based strategies, only architectures that both reshape the feature space toward clinical concepts and shield the classifier from non-clinical signal provide meaningful resilience. The framework is architecture-agnostic and applicable to any vision or multimodal model. To support evaluation in mammography, where the combinatorial richness of the BI-RADS lexicon cannot be feasibly captured by binary concepts alone, we release expert annotations with pixel-wise delineation of findings for 400 images. They can be found, alongside the code, at the following link.

⋆ Equal contribution.

Keywords: Vision-language models · Concept-based XAI · Spurious correlations

1 Introduction

Vision-language Models (VLMs) pre-trained on paired medical images and clinical reports have emerged as powerful backbones for medical image analysis [30]. By aligning visual representations with clinical text through contrastive learning, these models are expected to encode semantically meaningful features grounded in clinical concepts rather than non-clinical visual cues [12,9]. This implicit clinical grounding could, in principle, confer robustness to spurious visual correlations — acquisition artifacts, scanner-specific signatures, or other confounders that frequently contaminate medical imaging datasets [5,20]. However, domain-specific VLMs are typically pre-trained on institutionally homogeneous data [12,9]: while large in volume, these datasets originate from few clinical sites, so both visual features and report language may reflect institutional regularities rather than clinical diversity [30,11]. Recent benchmarks confirm that bias is pervasive across medical Foundation Models (FMs) [15] and that VLM-derived representations can encode demographic and institutional confounders [26,11]. However, these studies evaluate models on existing data where multiple confounders co-occur, making it difficult to isolate the specific role of spurious visual correlations.

One principled approach to reduce reliance on spurious features is to explicitly anchor model representations in clinical concepts. Such concept-based methods have gained considerable traction in medical image analysis, primarily to improve interpretability. Recent work incorporates clinical concepts through multi-level alignment with visual features [2], hierarchical prompting across encoder layers [7], graph-based reasoning over concept interdependencies [29], and label-free concept extraction from VLMs [21]. Beyond interpretability, there is emerging evidence that concept supervision can also improve generalisation: Gao et al. showed that concept-driven logical rules enhance out-of-distribution performance [10], Corbetta et al. demonstrated that regularising in Diabetic Retinopathy (DR) classification [3], and Yan et al. found that Concept Bottleneck Models (CBMs) with Large Language Model (LLM)-derived concepts improve robustness to confounders in chest X-ray classification [25]. However, these studies employ standard Convolutional Neural Networks (CNNs) or general-purpose VLM backbones. Whether concept supervision provides additional robustness when applied to domain-specific medical VLMs — where contrastive pre-training already provides implicit clinical grounding — remains untested.

We address this question by designing a controlled evaluation framework (Section 2) that injects synthetic confounders of parametrically varying strength and evaluates models under two complementary test protocols. We apply it to two clinical domains, DR (using RetCLIP [9]) and mammography (using Mammo–CLIP [12]), comparing five architectures spanning a spectrum of concept supervision: a standard vision backbone as a non-VLM reference, a VLM baseline, Multi-Task learning with auxiliary concept prediction [23,18], a Post-

Hoc Concept Bottleneck Model (PCBM) [27], and a Multi-level Image-Concept Alignment (MICA) model [2]. We investigate three questions: **Q1 (Balanced generalisation)**: when spurious correlations are absent at test time, can models that trained alongside artifacts still classify based on clinical features? **Q2 (Swapped robustness)**: when artifact–class associations are inverted at test time, do models follow the artifact or the pathology? **Q3 (Concept mitigation)**: across both scenarios, does explicit concept supervision reduce reliance on artifacts beyond what contrastive pre-training alone provides? To answer Q1 and Q2, we design two complementary test protocols, artifact removal (testing on clean images) and artifact inversion (swapping artifact–class associations), while Q3 is addressed by comparing architectures with increasing levels of concept supervision across both protocols. **Contributions:** (1) A controlled evaluation framework for domain-specific medical VLMs under parametrically varying spurious correlations, over two clinical domains and five architectures with increasing levels of concept supervision; (2) expert BI-RADS [19] annotations for 400 mammography images from the EMBED dataset [14], extending its existing labels to the complete lexicon with pixel-wise delineation, capturing the combinatorial richness of mammographic descriptors that cannot be represented as individual binary concepts; (3) evidence that VLM backbones encode clinical and artifactual signal in parallel, retaining clinical features when shortcuts are absent but following artifacts when they conflict with pathology, and that deep concept integration providing both feature–space grounding and classifier shielding yields meaningful robustness.

2 A Framework for Controlled Robustness Evaluation

We design a controlled evaluation framework to isolate the effect of spurious correlations on domain-specific medical VLMs. Inspired by Sun et al. [20], we inject synthetic confounders modelled after common acquisition artifacts into training images at parametrically varying prevalence, and evaluate models under two complementary test protocols that probe whether predictions follow clinical features or visual shortcuts. We compare five architectures that span a spectrum from no concept supervision to full multi-level image–concept alignment.

Synthetic Confounder Design. Each confounder type is deterministically assigned to one class, so that a model exploiting the shortcut must learn which visual cue maps to which label. In both domains, the lowest-severity class remains clean while each remaining class receives a unique perturbation: for fundus images (5-class), four overlays simulate specular reflections, eyelash shadows, uneven illumination, and out-of-focus regions; for mammography (3-class), two overlays simulate grid misplacement and collimator misalignment. The shortcut thus requires distinguishing these class-specific cues from their absence. All perturbations are synthetically generated through parametric image manipulations (e.g., additive bright spots for reflections, directional blurring for out-of-focus regions) and restricted to clinically plausible spatial locations. The overlay percentage $p \in \{0\%, 25\%, 50\%, 75\%, 100\%\}$ controls the fraction of images within

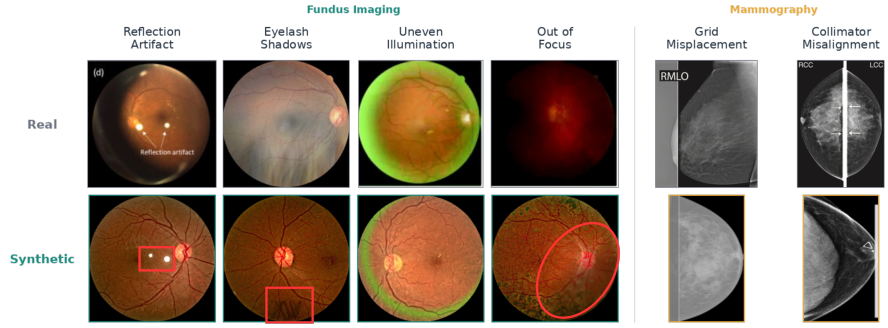

Fig. 1: Synthetic artifacts used to induce spurious correlations, shown alongside real-world examples. Top row: real clinical examples; bottom row: synthetic overlays. Fundus imaging (left): reflection artifact, eyelash shadows, uneven illumination, and out-of-focus region. Mammography (right): grid misplacement and collimator misalignment. Red outlines highlight subtle artifacts. Real fundus examples are adapted from [17,28,24]; real mammography examples are from [1].

Table 1: Models under comparison. All share the same backbone; they differ in encoder initialisation, how clinical concepts are integrated, and the pathway from features to prediction. Only MICA uses the VLM text encoder.

Architecture	Encoder init	Concept role	Prediction pathway
Non-VLM baseline	ImageNet	—	Features → linear head
VLM baseline	VLM	—	Features → linear head
Multi-Task	VLM	Auxiliary loss	Features → linear head (+ concept head)
PCBM [27]	VLM	CAV bottleneck	Features → CAV scores → linear head
MICA [2]	VLM	Multi-level alignment + bottleneck	Features → concept detection → linear head

each affected class that receive their assigned perturbation. At $p = 0\%$, no confounders are present; at $p = 100\%$, every eligible image carries its class-specific overlay. Figure 1 illustrates examples alongside their real-world counterparts.

Evaluation Protocols. Under **artifact removal**, all confounders are removed from the test set, probing whether models learned clinical features despite training alongside shortcuts (Q1). Under **artifact inversion**, the confounder—class associations are cyclically shifted, revealing the extent to which models follow artifacts over pathology (Q2). Comparing models across both protocols at increasing p addresses Q3.

Models Under Comparison. We compare five models that vary along two axes: whether the visual encoder is initialised from domain–specific contrastive pre–training, and the degree of clinical concept supervision during fine–tuning. All models share the same backbone adapted with Low-rank adaptation (LoRA) [13]. Clinical concepts used by concept–based models are defined per dataset (Section 3). These characteristics are summarised in Table 1.

1. Non-VLM baseline. The vision encoder is initialised from ImageNet weights [6] and topped with a linear classification head. No concept information is used. This serves as a reference to isolate the effect of contrastive pre-training. **2. VLM baseline.** The encoder is initialised from domain-specific VLM weights and topped with a linear classification head. No concept information is used. **3. Multi-Task.** The VLM baseline augmented with a parallel concept prediction head. Both heads share the encoder features; concept supervision acts as a regulariser but does not constrain the information pathway. **4. PCBM [27].** For each clinical concept, a Concept Activation Vector (CAV) [16] is derived from a linear Support Vector Machine (SVM) [4] trained on the frozen encoder's feature space. Image features are projected onto CAVs to produce per-concept scores, and a linear classifier maps these to a diagnosis, forming a bottleneck through which all predictive information must pass. **5. MICA [2].** A two-stage framework: Stage 1 aligns visual features with text representations of clinical concepts at three levels: global image-text contrastive alignment, token-level cross-attention, and concept-level CAV supervision. Stage 2 freezes the encoder and trains an explicit concept bottleneck: a concept detection head followed by a linear diagnosis predictor operating solely on predicted concepts.

3 Experiments and Results

Implementation Details. For DR, all architectures use a ViT-B/16 [8] backbone; for mammography, an EfficientNet-b2 [22]. VLM-based models initialise from RetCLIP [9] and Mammo–CLIP [12] weights respectively, while the non-VLM baseline initialises from ImageNet [6]. To isolate the effect of contrastive pre-training, VLM-based models feed the classifier with features projected into the joint image–text embedding space via the VLM projection head, whereas the non-VLM baseline operates directly on vision encoder features. All backbones are adapted with LoRA [13]. For MICA, Stage 1 aligns visual features with textual descriptions of clinical findings: natural-language captions summarising the BI-RADS descriptors (for mammography) or lesion annotations (for fundus) present in each image are fed through a text encoder to produce concept-level representations, against which visual features are aligned at multiple levels. We use the text encoder from the respective domain-specific VLM, preserving the alignment between visual and textual modalities learned during pre-training. Hyperparameter tuning and hardware details are available in the provided repository. For both domains, models are trained on five random train/validation splits with a fixed held-out test set; we report performance averaged across splits.

Datasets and Clinical Concepts. For DR, we use the FGADR Segmentation Set [31]: 1,842 fundus images with five-class severity grading and pixel-level annotations for six lesion types serving as clinical concepts. For mammography, we sample 100 cases (400 images, bilateral CC and MLO) from the EMBED dataset [14] and formulate a three-class task: no findings (BI-RADS 1), probably benign (BI-RADS 2–3), and probably malignant (BI-RADS 4–5). Two radiologists each independently annotated half of the cases following the complete

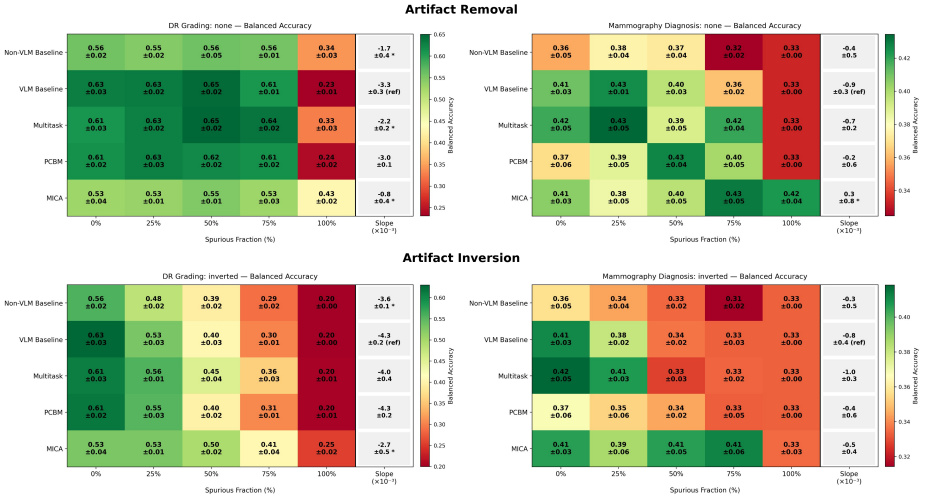

Fig. 2: Balanced accuracy under artifact removal (top) and artifact inversion (bottom) for DR grading (left) and mammography diagnosis (right). Rows within each heatmap: models; columns: spurious fraction p. The rightmost column reports the linear slope ($\times 10^{-3}$) across p; * denotes significantly shallower degradation than the VLM baseline (ref).

BI-RADS lexicon [19]. The combinatorial nature of mammographic findings — where a single mass is characterised by shape, margin, and density — makes it impractical to define individual CAVs for each combination; natural-language encoding through the VLM text encoder provides a scalable alternative. This yields a two-tier concept design: nine binary concepts (five finding-level and four breast density categories) for CAV computation and multi-task supervision, and full BI-RADS descriptors encoded as natural-language captions for MICA's text encoder. Annotations with pixel-wise delineation are publicly released.

Metrics. We report balanced accuracy averaged over five splits. To quantify sensitivity to spurious correlations, we fit a linear slope to each split's balanced accuracy as a function of spurious fraction $p \in \{0, 25, 50, 75, 100\}$ and report the mean slope ($\times 10^{-3}$) across splits. A slope near zero indicates robustness; a large negative slope indicates that the model has learned the spurious association. To test whether an architecture degrades significantly less than the VLM baseline (reference), we apply a one-sided paired permutation test (10,000 permutations) on the per-split slopes ($^* p < 0.05$).

Models retain clinical signal when shortcuts are absent (Q1). Under artifact removal (Figure 2, top row), where test images are clean and only the training set contains spurious correlations, all architectures prove remarkably robust up to high corruption levels. In DR grading, the VLM baseline maintains balanced accuracy between 0.63 and 0.65 from p=0% through p=75%, with

degradation only becoming pronounced at p=100% (dropping to 0.23), yielding a slope of -3.3×10^{-3}. The non-VLM baseline follows the same pattern at a lower absolute level (0.56 at p=0%, stable through 75%, dropping to 0.34 at p=100%) but degrades at roughly half the rate (-1.7×10^{-3}, $^{*}p < 0.05$). This indicates that models trained alongside artifacts can still classify based on clinical features when the shortcut is absent at test time. However, the steeper slope of the VLM baseline, despite its higher absolute accuracy, suggests that contrastive pre-training, while providing a stronger starting point, does not protect against shortcut learning. In mammography, where absolute performance is lower across all architectures — consistent with the known difficulty of BI-RADS-based classification with foundation models [11] — the pattern is qualitatively similar: performance remains stable through moderate corruption and begins to decline at p=75%. Slopes are smaller in magnitude across all architectures, reflecting both the narrower performance range and the three-class setting, but the relative ordering between architectures is preserved.

Models follow spurious associations when shortcuts are inverted (Q2). Under artifact inversion (Figure 2, bottom row), where confounder–class associations are cyclically shifted at test time, a different picture emerges. In DR grading, degradation begins earlier and proceeds more steeply than under artifact removal: the VLM baseline already drops from 0.63 to 0.53 between p=0% and p=25%. At p=100%, nearly all architectures converge to 0.20 (chance level for a five-class task), confirming that the inverted artifacts systematically mislead predictions. The VLM baseline's slope steepens to -4.3×10^{-3} (from -3.3 under artifact removal), and again degrades significantly faster than the non-VLM baseline (-4.3 vs. -3.6×10^{-3}, $^{*}p < 0.05$). In mammography, slopes under inversion are comparable to those under removal (e.g., -0.8×10^{-3} for the VLM baseline), reflecting the narrower performance range in this more challenging setting; all models converge to 0.33 at full corruption, confirming the same artifact-following behaviour. Together with Q1, these results suggest that models do not replace clinical features with shortcuts but rather learn both in parallel: when the spurious cue is absent, clinical signal suffices (Q1); when it conflicts with the pathology, the model follows the artifact (Q2). This dual encoding is only revealed by the complementary evaluation protocols and underscores that stable performance on clean test data alone is insufficient to certify robustness.

Only deep concept integration provides meaningful robustness (Q3). Comparing concept-based architectures across both protocols reveals that simply adding concept supervision is insufficient; robustness requires both grounding the encoder's representations in clinical concepts and preventing non-clinical signal from reaching the classifier. In DR under artifact removal, Multi-Task improves over the VLM baseline (-2.2 vs. -3.3×10^{-3}, $^{*}p < 0.05$), but this advantage vanishes under inversion (-4.0 vs. -4.3): the encoder is still fine-tuned on the classification objective and can encode shortcuts alongside concepts. PCBM mirrors the VLM baseline under both protocols (DR : -3.0 / -4.3; mammography: -0.2 / -0.4): although its CAVs are computed offline from the frozen pretrained

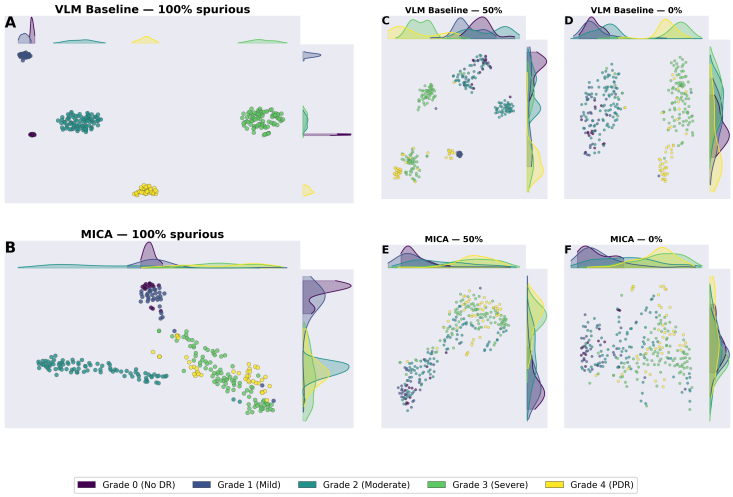

Fig. 3: t-SNE of DR test-set representations for VLM baseline (A,C,D) and MICA (B,E,F) at p=100%, 50%, and 0% spurious fraction in the artifacts removal setup. Colours denote DR severity grades.

encoder and are themselves uncontaminated, LoRA fine-tuning progressively reshapes the encoder's feature space to encode spurious correlations; when these updated features are projected onto the fixed CAV space, the resulting concept scores inherit the artifact–class association: a fixed bottleneck cannot filter out what is already encoded in the representation. MICA stands apart on both counts: Stage 1 multi-level alignment actively reshapes the encoder's feature space toward clinical concept structure, and Stage 2 freezes this encoder and routes predictions exclusively through detected concepts, so neither the representation nor the classifier are exposed to the artifact–class association during classification training. In DR, its artifact-removal slope (-0.8×10^{-3}, $^*p < 0.05$) is roughly four times less steep than the VLM baseline's, and under inversion it is the only architecture not to collapse to chance (0.25 vs. 0.20 for all others at p=100%). In mammography, MICA is the only model whose artifact-removal slope is positive ($+0.3 \times 10^{-3}$, $^*p < 0.05$). The t-SNE visualisation of DR representations (Figure 3) confirms this at the feature level. As the spurious fraction increases (panels D,C for the VLM baseline; F,E for MICA), the VLM baseline's representations consolidate into artifact-driven clusters (panel A), whereas MICA maintains clinically organised structure even at p=100% (panel B).

4 Conclusion

We presented a controlled evaluation of domain-specific medical VLMs under parametrically varying spurious correlations across two clinical domains. Our results reveal that VLM backbones encode clinical and artifactual signals in

parallel: they retain clinical information when shortcuts are absent, yet follow spurious cues when these conflict with pathology, degrading faster than standard visual backbones. Among concept-based approaches, only MICA provides consistent robustness, as it both grounds representations in clinical concepts and restricts predictions to concept-based information. Overall, robustness in medical VLMs requires counterfactual evaluation protocols and architectural mechanisms that explicitly enforce clinical grounding. The proposed framework is architecture-agnostic and applicable to vision and multimodal models. Future work should compare full fine-tuning, LoRA, and linear probing to clarify how training regime shapes shortcut reliance, validate overlay fidelity through clinician reader studies, and extend the analysis to 3D imaging, multi-organ and multi-site cohorts, and additional cyclic-shift permutations.

Acknowledgments. Research at the Netherlands Cancer Institute is supported by grants from the Dutch Cancer Society and the Dutch Ministry of Health, Welfare and Sport. The authors acknowledge the Research High Performance Computing (RHPC) facility of the Netherlands Cancer Institute (NKI). This research was funded by the René Vogels Stichting and by GROW Research Institute for Oncology and Reproduction. Researchers from UiT are funded by the Research Council of Norway, grant no. 309439 through the Center for Research-based Innovation Visual Intelligence.

Disclosure of Interests. The authors declare no competing interests.

References

1. Ayyala, R.S., Chorlton, M., Behrman, R.H., Kornguth, P.J., Slanetz, P.J.: Digital mammographic artifacts on full-field systems: what are they and how do i fix them? Radiographics **28**(7), 1999–2008 (2008)
2. Bie, Y., Luo, L., Chen, H.: MICA: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 837–845 (2024)
3. Corbetta, V., Dijkstra, F.S., Beets-Tan, R., Kervadec, H., Wickstrøm, K., Silva, W.: In-hoc concept representations to regularise deep learning in medical imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7312–7321 (2025)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
5. DeGrave, A.J., Janizek, J.D., Lee, S.I.: AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Dong, Y., Lin, Y., Yang, X.: Copa: Hierarchical concept prompting and aggregating network for explainable diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 67–76. Springer (2025)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

9. Du, J., Guo, J., Zhang, W., Yang, S., Liu, H., Li, H., Wang, N.: Ret-CLIP: A retinal image foundation model pre-trained with clinical diagnostic reports. In: International conference on medical image computing and computer-assisted intervention. pp. 709–719. Springer (2024)
10. Gao, Y., Zhou, H., Gao, Z., Wang, B., Gao, S., Wang, S., Zhuang, X.: Learning concept-driven logical rules for interpretable and generalizable medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 291–300. Springer (2025)
11. Germani, E., Selin-Türk, I., Zeineddine, F., Mourad, C., Albarqouni, S.: Bias and generalizability of foundation models across datasets in breast mammography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–34. Springer (2025)
12. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-CLIP: A vision language foundation model to enhance data efficiency and robustness in mammography. In: International conference on medical image computing and computer-assisted intervention. pp. 632–642. Springer (2024)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
14. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilies, G., Smith, G., et al.: The emory breast imaging dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiology: Artificial Intelligence **5**(1) (2023)
15. Jin, R., Xu, Z., Zhong, Y., Yao, Q., QI, D., Zhou, S.K., Li, X.: FairMedFM: fairness benchmarking for medical imaging foundation models. Advances in Neural Information Processing Systems **37**, 111318–111357 (2024)
16. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: ICML. pp. 2668–2677. PMLR (2018)
17. Lin, J., Yu, L., Weng, Q., Zheng, X.: Retinal image quality assessment for diabetic retinopathy screening: A survey. Multimedia Tools and Applications **79**(23), 16173–16199 (2020)
18. Playout, C., Duval, R., Cheriet, F.: A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images. In: International conference on medical image computing and computer-assisted intervention. pp. 101–108. Springer (2018)
19. Spak, D.A., Plaxco, J., Santiago, L., Dryden, M., Dogan, B.: BI-RADS® fifth edition: A summary of changes. Diagnostic and interventional imaging **98**(3), 179–190 (2017)
20. Sun, S., Koch, L.M., Baumgartner, C.F.: Right for the wrong reason: Can interpretable ML techniques detect spurious correlations? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 425–434. Springer (2023)
21. Sun, S., Tessier, L., Meeuwsen, F., Grisi, C., van Midden, D., Litjens, G., Baumgartner, C.F.: Label-free concept based multiple instance learning for gigapixel histopathology. arXiv preprint arXiv:2501.02922 (2025)
22. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. pp. 6105–6114. PMLR (2019)
23. Tardy, M., Mateus, D.: Leveraging multi-task learning to cope with poor and missing labels of mammograms. Frontiers in Radiology **Volume 1 - 2021** (2022)

24. Toslak, D., Liu, C., Alam, M.N., Yao, X.: Near-infrared light-guided miniaturized indirect ophthalmoscopy for nonmydriatic wide-field fundus photography. Optics letters **43**(11), 2551–2554 (2018)
25. Yan, A., Wang, Y., Zhong, Y., He, Z., Karypis, P., Wang, Z., Dong, C., Gentili, A., Hsu, C.N., Shang, J., et al.: Robust and interpretable medical image classifiers via concept bottleneck models. arXiv preprint arXiv:2310.03182 (2023)
26. Yang, Y., Liu, Y., Liu, X., Gulhane, A., Mastrodicasa, D., Wu, W., Wang, E.J., Sahani, D., Patel, S.: Demographic bias of expert-level vision-language foundation models in medical imaging. Science Advances **11**(13) (2025)
27. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
28. Zhang, F., Xu, X., Xiao, Z., Wu, J., Geng, L., Wang, W., Liu, Y.: Automated quality classification of colour fundus images based on a modified residual dense block network. Signal, Image and Video Processing **14**(1), 215–223 (2020)
29. Zhao, L., Chen, C., Pu, B., Qi, X., Peng, F., Wang, C., Li, K., Tan, G.: Concept-induced graph perception model for interpretable diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 215–225. Springer (2025)
30. Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., Shen, D.: CLIP in medical imaging: A survey. Medical Image Analysis **102**, 103551 (2025). https://doi.org/https://doi.org/10.1016/j.media.2025.103551
31. Zhou, Y., Wang, B., Huang, L., Cui, S., Shao, L.: A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. IEEE Transactions on Medical Imaging **40**(3), 818–828 (2021)