Interpretable Medical Image Diagnosis via VLM-based Concept Alignment and Graph Reasoning

Yilan Hu[1*], Haoyu Xie[1*], Dan Liang[2], Xinhua Wei[2], Hui Zhou[3], Hong Zhou[4], Bingsheng Huang[1]

[1] Medical AI Lab, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China
huangb@szu.edu.cn
[2] Department of Radiology, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China
[3] Department of Radiology, The Affiliated Qingyuan Hospital (Qingyuan People's Hospital), Guangzhou Medical University, Qingyuan, China
[4] Department of Radiology, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang, China

Abstract. Interpretable medical image diagnosis requires predictions grounded in clinically meaningful findings that physicians can inspect and correct. While Concept Bottleneck Models (CBMs) route predictions through predefined imaging signs, they treat concepts as independent scalar scores and struggle to align fine-grained visual features with clinical semantics. We present ConceptAlignE-G, a framework that integrates vision-language model (VLM) based concept alignment with graph reasoning to bridge this gap. Built on BiomedCLIP, a concept-aware perception module grounds each imaging sign in a dedicated visual-semantic embedding space, enabling fine-grained alignment between sign-specific visual features and expert-defined text embeddings. The predicted sign probabilities are then used to construct a dynamic sign-image relational graph, where a GCN incorporates broader image context beyond the concept bottleneck for disease classification. On CT-based appendicitis differentiation and ultrasound-based breast cancer diagnosis, ConceptAlignE-G outperforms both black-box classifiers and existing concept-based methods in diagnostic AUC and sign recognition. Beyond accuracy, it provides per-sign prediction probabilities and interpretable decision weights, and supports clinician-driven intervention at inference time. Code is available at: https://github.com/MedcAILab/med-cag.

Keywords: Medical imaging intelligent diagnosis, Interpretability, Radiological signs, Vision-Language Model, Graph Convolutional Network.

[*] Yilan Hu and Haoyu Xie contributed equally to this work.

1 Introduction

Medical imaging is central to modern clinical diagnosis, and the growing volume of imaging data has placed increasing pressure on radiologists [1]. Deep learning has made substantial inroads in medical image analysis, with models that match or exceed human performance on a range of classification tasks [2,3]. The obstacle to clinical adoption is not accuracy but transparency: without knowing why a model reached a decision, physicians have little basis for trusting or acting on its output, and regulatory frameworks such as the EU AI Act now make interpretability a legal requirement in high-risk applications [4].

Feature attribution methods offer one response to this problem, producing saliency maps that highlight relevant image regions [5]. But a heatmap is not a diagnosis — it tells the physician where the model looked, not what it found. Concept-based methods take a different approach, decomposing predictions into named attributes that correspond to the kind of findings clinicians actually document [6,7]. In radiology, structured reporting systems such as BI-RADS [8] already formalize this sign-based reasoning, making concept-based models a natural fit.

Concept Bottleneck Models (CBMs) [9] implement this idea by inserting a bottleneck layer between the image encoder and the classifier, where each neuron represents a predefined visual concept. Because predictions pass through interpretable concept scores, physicians can inspect individual sign predictions and intervene if they disagree. The practical limitations are well known: CBMs require dense concept annotations, and when the concept set does not cover all diagnostically relevant information, performance suffers because the bottleneck cannot be bypassed [9,10].

Later work has chipped away at these limitations from several directions. Post-hoc CBMs (PCBM) [11] removes the annotation burden by extracting concept vectors from external sources. Concept Embedding Models (CEM) [12] replaces scalar scores with richer embeddings to retain more information per concept. The Concept Complement Bottleneck Model (CCBM) [13] adds a complementary pathway to capture information outside the predefined concept set. More recently, vision-language models (VLMs) have been integrated with concept-based frameworks [14-17], using text descriptions of clinical findings to supervise concept learning — an approach that has shown promise in dermatology [16,18] and chest X-ray interpretation [15,17].

Despite this progress, two challenges remain. First, aligning fine-grained imaging signs with clinical text descriptions is difficult, especially for domain-specific modalities where general-purpose VLMs underperform [19,20]. Second, routing all diagnostic information through a fixed concept set creates a representational bottleneck: when the concept set is incomplete, the model has no mechanism to draw on broader image context [11,12].

To address both challenges, we propose ConceptAlignE-G, an interpretable diagnostic framework built on BiomedCLIP [21]. A dedicated concept-aware perception module extracts sign-specific features from the input image and aligns them with concept text embeddings in a shared space, enabling fine-grained cross-modal grounding. The resulting sign prediction probabilities are used to construct a dynamic sign-image relational graph, processed by a GCN [22] to fuse image semantics, sign scores, and

concept text knowledge for final disease prediction. We validate the framework on CT-based appendicitis differentiation and ultrasound-based breast cancer diagnosis, where it achieves superior performance over black-box models [23,24] and concept-based baselines [9,12,13] , while supporting traceable sign-level explanations and concept-level expert intervention.

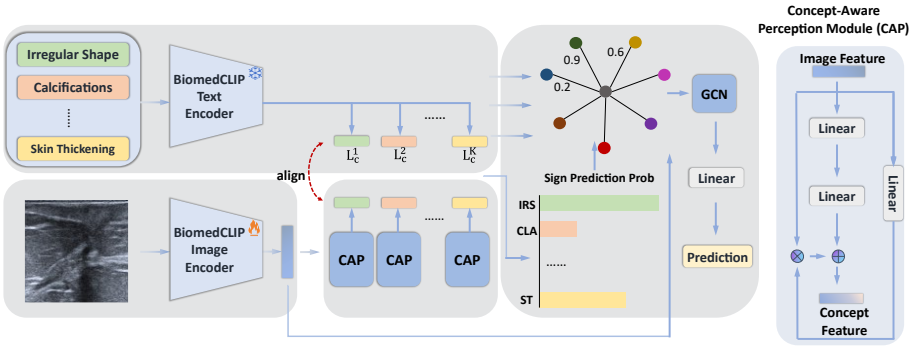

Fig. 1. Overview of the proposed ConceptAlignE-G framework. Imaging sign descriptions are encoded by a frozen text encoder, while the input image is processed by a fine-tuned image encoder. For each sign, a concept-aware perception module (CAP) produces a sign-specific feature that is aligned with the corresponding text embedding to predict sign probabilities, which serve as edge weights in a dynamic sign-image relational graph. A GCN aggregates the graph to obtain the final disease prediction.

2 Methods

2.1 Overview

We propose an interpretable medical image diagnostic framework that bridges the gap between deep feature representations and clinically meaningful imaging signs. As shown in Fig. 1, the framework takes an image $\mathbf{x}$ with disease label $y \in \{0,1\}$ and K binary sign concept labels $\mathbf{c} \in \{0,1\}^K$ as input. The pipeline proceeds in three stages: concept text and image features are first extracted in parallel, then a concept-aware perception module aligns image features with each sign concept in a shared embedding space, and finally a sign-image relational graph is constructed and processed by a GCN to produce the disease prediction. This design allows the model to explicitly reason through imaging signs before reaching a diagnosis, mirroring the workflow of clinical radiologists.

2.2 Concept Text and Image Feature Extraction

Concept text features. Each of the K imaging signs is described by a short, fixed-form sentence. For the appendicitis CT task, descriptions follow the template *"this is a CT image of an appendix with/without \{sign\}"*, while for the breast ultrasound task we

use *"this is an ultrasonic image of a breast tumor with/without \{sign\}"*. These sentences are passed through the frozen text encoder of BiomedCLIP [21], and the resulting representations are projected by a two-layer multi-layer perceptron (MLP) to produce a 512-dimensional embedding $\overrightarrow{T_j} \in \mathbb{R}^{512}$ for each sign j. Freezing the text encoder ensures that these embeddings remain stable throughout training, serving as fixed semantic anchors for cross-modal alignment.

Image features. We use BiomedCLIP's image encoder, a vision transformer-Base [25] followed by a linear projection layer, to extract a 512-dimensional image representation $\vec{I_i} \in \mathbb{R}^{512}$ from each input. The encoder is initialized with BiomedCLIP pretrained weights and fine-tuned end-to-end. Since the encoder only handles 2D inputs, 3D CT volumes are processed slice by slice along the coronal axis and the resulting features are averaged to form a single volume-level representation.

2.3 Concept-Aware Perception and Cross-Modal Alignment

A straightforward approach to computing concept scores is to measure the similarity between the global image feature $\vec{I_i}$ and each concept text embedding. However, this tends to blur the boundaries between different signs, since the same image feature is pulled toward multiple concept anchors simultaneously. Instead, we introduce a separate concept-aware perception module for each of the K signs, which carves out a sign-specific subspace from the image feature before alignment.

Each module takes $\vec{I_i}$ as input and produces a concept embedding $\overrightarrow{V_i^j} \in \mathbb{R}^{512}$ through a gated residual structure:

$$w_i^j = \sigma\left(W_0^j \vec{I_i} + b_0^j\right) \tag{1}$$

$$\vec{V_i^j} = \mathrm{ReLU}\left(W_2^j \mathrm{ReLU}\left(W_1^j \vec{I_i}\right)\right) + w_i^j \vec{I_i} \tag{2}$$

where $\mathbf{W}_0^j \in \mathbb{R}^{1\times512}$, $\mathbf{W}_1^j \in \mathbb{R}^{128\times512}$, $\mathbf{W}_2^j \in \mathbb{R}^{512\times128}$ are learnable weights, and σ is the sigmoid function. The scalar gate w_i^j controls how much of the original image feature flows through the residual path, letting the module decide how much global context to retain for each concept.

The sign concept score and probability are then computed as:

$$s_i^j = \mathrm{sim}\left(\tilde{\vec{V}}_i^j, \tilde{\vec{T}}_j\right) \cdot \tau, \hat{c}_i^j = \sigma(s_i^j) \tag{3}$$

where $\tilde{\cdot}$ denotes L2 normalization and τ is a learnable temperature parameter. The probability $\hat{c}_i^j \in (0,1)$ reflects how likely sign j is present in image i. All sign predictions are supervised with binary cross-entropy (BCE) loss.

2.4 Sign-Image Graph Reasoning for Disease Classification

Relying solely on concept scores for disease prediction has an inherent limitation: when the predefined sign set does not fully capture all diagnostically relevant information,

the linear decision layer has little room to compensate. In clinical practice, radiologists do not make diagnoses based on a checklist of signs alone — they also draw on the overall appearance of the image and the interplay between different signs. Motivated by this observation, we construct a sign-image relational graph that integrates image semantics, concept text knowledge, and predicted sign probabilities, allowing the model to leverage all three sources of information jointly.

Graph construction. For each image i, we define a star-topology graph $\mathcal{G}_i = (\mathcal{V}, \mathcal{E})$consisting of one central image node $v_i$and Kconcept nodes $\{v_{c_1}, \dots, v_{c_K}\}$, connected by Kdirected edges from concept nodes to the image node. Node features are initialized as:

$$H_i = \left[\vec{I}_i, \vec{T}_1, \dots, \vec{T}_K\right] \in \mathbb{R}^{(K+1)\times 512} \tag{4}$$

The weight of the edge connecting $v_{c_j}$to $v_i$is set to the predicted concept probability $\hat{c}_i^j$, making the graph dynamic: for each input image, edges with higher concept confidence contribute more strongly to the aggregation, naturally focusing the model's attention on the most relevant signs for that case.

Graph convolution and prediction. A single GCN layer [22] aggregates information from all concept nodes into the image node, yielding an updated representation $\vec{f}_l \in \mathbb{R}^{128}$:

$$\vec{f}_l = W_G \frac{1}{k+1}\vec{I}_l + \sum_{j=1}^{k} W_G \frac{\hat{c}_l^j}{\sqrt{2(k+1)}}\vec{T}_j \tag{5}$$

where $\mathbf{W}_G \in \mathbb{R}^{128\times 512}$is the GCN weight matrix. This aggregated feature carries both the image's own semantic content and the domain knowledge encoded in the concept text embeddings, weighted by how strongly each sign is predicted to be present. The final disease prediction is obtained via a linear layer.

Since the edge weights are directly determined by the predicted concept probabilities, correcting a concept prediction at inference time — as a clinician might when reviewing the model's sign-level outputs — propagates through Eq. (5) and changes the final diagnosis accordingly. This makes expert intervention a natural and lightweight operation that requires no retraining.

2.5 Training Objective

The overall loss combines disease classification and sign recognition:

$$\mathcal{L} = \alpha\mathcal{L}_y + \beta\sum_{j=1}^{K}\mathcal{L}_c^j \tag{6}$$

where $\mathcal{L}_y$is the BCE loss for disease prediction, $\mathcal{L}_c^j$is the BCE loss for the j-th sign, and $\alpha = 1, \beta = 0.2$ are set based on preliminary experiments.

3 Experiments

3.1 Datasets and Implementation Details

Datasets. We validate the proposed method on two clinical tasks. For appendicitis differentiation, we retrospectively collected CT images of 1,156 acute appendicitis (AA) patients from three hospitals between January 2018 and March 2022: Qingyuan People's Hospital (Center 1, n=954, complicated:uncomplicated = 328:626), Guangzhou First People's Hospital (Center 2, n=90, 49:41), and the First Affiliated Hospital of University of South China (Center 3, n=112, 36:76). CT scans were acquired using multidetector scanners from different vendors (Siemens SOMATOM Force at Center 1; Philips Brilliance 64 at Centers 2 and 3). The diagnosis of complicated versus uncomplicated AA was established as the gold standard by a senior pathologist with 12 years of experience based on semi-quantitative histopathological assessment. A radiologist with 6 years of abdominal imaging experience manually delineated the inflamed appendix ROI and annotated six CT imaging signs: enlarged ileocecal lymph nodes (EILN), thickening of the cecal wall (TCW), gas in the appendix cavity (GIC), gas outside the appendix cavity (GOC), coproliths in the appendix cavity (CIC), and fluid around the appendix (FLU). The dataset is split into a training set (n=695) and validation set (n=259) from Center 1, and an external test set (n=202) from Centers 2 and 3 to assess generalizability. For breast cancer diagnosis, we use the public BrEaST dataset [26], comprising 252 breast tumor ultrasound images (98 malignant, 154 benign) with seven BI-RADS-based sign annotations marked by radiologists following the BI-RADS lexicon guidelines: irregular shape (IRS), not circumscribed margin (NCM), hypoechoic or heterogeneous echogenicity (HoHE), posterior features (PF), hyperechoic halo (HH), calcifications (CAL), and skin thickening (ST). Five-fold cross-validation is used given the limited dataset size.

Implementation Details. All models are trained with the Adam optimizer under a cosine annealing schedule, with an initial learning rate of 1×10^{-4} (image encoder: 1×10^{-5}) and weight decay of 1×10^{-4} for 200 epochs. Loss coefficients are set to $\alpha = 1$ and $\beta = 0.2$. All experiments are conducted on an NVIDIA A6000 GPU. Statistical differences between receiver operating characteristic (ROC) curves are assessed using the DeLong test.

3.2 Performance Comparison on Two Clinical Tasks

We evaluate ConceptAlignE-G against two groups of baselines: black-box classifiers (ResNet50 [23] and DenseNet121 [24]) and concept-based interpretable methods (CBM [9], CEM [12], and CCBM [13]), all using ResNet50 as the backbone. We also include two ablation variants to isolate the contribution of each component: ConceptAlignE replaces GCN reasoning with a linear layer over concept scores, and BiomedSigLIP further removes the CAP module from ConceptAlignE, directly computing

similarity between the shared image feature and concept text embeddings. Results are reported in Table 1.

Table 1. Comparison of disease diagnosis (AUC/ACC) and sign recognition (mAUC) performance.

Method	Appendicitis (Val)	Appendicitis (Test)	Breast Cancer
ResNet50 [23]	0.735 / —	0.589 / —	0.875 / —
DenseNet121 [24]	0.717 / —	0.656 / —	0.873 / —
CBM [9]	0.745 / 0.690	0.591 / 0.668	0.855 / 0.741
CEM [12]	0.719 / 0.647	0.640 / 0.673	0.861 / 0.774
CCBM [13]	0.725 / 0.632	0.668 / 0.649	0.830 / 0.713
BiomedSigLIP	0.753 / 0.702	0.675 / 0.696	0.823 / 0.758
ConceptAlignE	0.756 / 0.713	0.685 / 0.711	0.890 / 0.786
ConceptAlignE-G	0.771 / 0.770	0.713 / 0.754	0.912 / 0.805

ConceptAlignE-G achieves the best performance across both tasks, with a test area under the ROC curve (AUC) of 0.713 for appendicitis ($p<0.05$ vs. ResNet50, DeLong test) and 0.912 for breast cancer, outperforming all black-box and concept-based baselines. The ablation results further clarify where the gains come from. Removing the CAP module (BiomedSigLIP) leads to a consistent drop in both diagnostic AUC and sign recognition, confirming that per-sign feature extraction is more effective than aligning a shared image feature against multiple concept anchors. Replacing GCN reasoning with a linear layer (ConceptAlignE) also degrades performance, indicating that aggregating image semantics and concept text knowledge through graph convolution provides a more informative representation than operating on concept scores alone.

3.3 Interpretability Analysis

Concept-level Explanations. Fig. 2 presents representative cases from both tasks. For the appendicitis case (Fig. 2a), the model identifies TCW (80.5%) and CIC (12.8%) as present, while correctly predicting the absence of GOC (0.1%) and FLU (2.1%), leading to a prediction of uncomplicated AA (96.3%). For the breast cancer case (Fig. 2b), high sign probabilities for IRS (99.5%), NCM (80.2%), HoHE (99.6%), and PF (97.3%) drive a malignant prediction (82.7%). The per-sign outputs give physicians a clear account of what the model found and why it reached its conclusion, in terms that map directly onto clinical reporting practice.

Concept Intervention. ConceptAlignE-G supports sign-level intervention at inference time: a clinician who disagrees with a predicted sign probability can override it, and the updated value propagates through the graph to revise the final diagnosis. As shown in Fig. 2c, the model incorrectly predicts the presence of TCW (90.9%) and GIC (99.4%) in an appendicitis case, leading to a false positive for complicated AA (64.1%). After a

clinician corrects both signs to 0%, the prediction is revised to uncomplicated AA (59.3%). Similarly, in the breast cancer case (Fig. 2d), the model fails to detect NCM (38.8%) and HH (1.5%), resulting in a missed malignancy (51.7%). Correcting both signs to 100% successfully flips the prediction to malignant (73.1%). In both cases, the intervention requires no model retraining and takes effect immediately through the graph edge weights.

Linear Decision Weights. Fig. 3 visualizes the weights of the linear decision layer, which reflect each sign's learned contribution to the final prediction. For appendicitis, GIC and TCW carry the strongest positive weights toward complicated AA, while CIC and FLU are negatively weighted. For breast cancer, CAL and PF are the strongest contributors to malignancy. These patterns are broadly consistent with established clinical guidelines [8,27], providing a transparent and inspectable decision logic absent in black-box models.

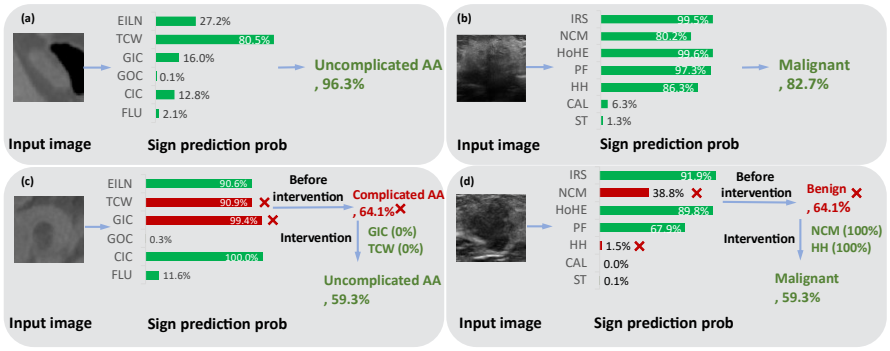

Fig. 2. Interpretability analysis. (a-b) Representative correctly classified cases from the appendicitis and breast cancer tasks, with per-sign prediction probabilities and final diagnostic scores. (c-d) Concept intervention cases: mispredicted signs (highlighted) are corrected by a clinician, and the updated diagnosis is shown alongside the original prediction.

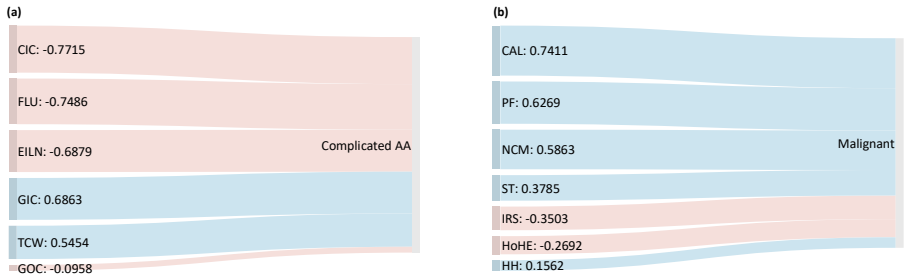

Fig. 3. Learned linear decision weights for both tasks. Positive weights (blue) indicate contribution toward complicated AA (appendicitis) or malignancy (breast cancer); negative weights (red) indicate contribution toward the opposing class. Signs are ranked by absolute weight magnitude.

4 Conclusion

We propose an interpretable medical image diagnostic framework that reasons through clinically defined imaging signs before reaching a disease prediction. By combining concept-aware cross-modal alignment with sign-image graph reasoning, the model achieves competitive diagnostic performance on both CT-based appendicitis differentiation and ultrasound-based breast cancer diagnosis, while providing traceable sign-level explanations and supporting expert concept intervention.

Acknowledgments. This study was supported by National Natural Science Foundation of China (NSFC) under grant No. 62371303, and Shenzhen Science and Technology Program under Grant No. JCYJ20250604181405007.

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

References

1. Smith-Bindman, R., Miglioretti, D. L. & Larson, E. B. Rising use of diagnostic medical imaging in a large integrated health system. Health affairs **27**, 1491-1502 (2008)
2. Litjens, G. *et al.* A survey on deep learning in medical image analysis. Medical image analysis **42**, 60-88 (2017)
3. Suganyadevi, S., Seethalakshmi, V. & Balasamy, K. A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval **11**, 19-38 (2022)
4. Act, E. A. I. The eu artificial intelligence act. European Union (2024)
5. Singh, A., Sengupta, S. & Lakshminarayanan, V. Explainable deep learning models in medical image analysis. Journal of imaging **6**, 52 (2020)
6. Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T. & Baralis, E. Concept-based explainable artificial intelligence: A survey. ACM Computing Surveys (2023)
7. Gupta, A. & Narayanan, P. A survey on concept-based approaches for model improvement. arXiv preprint arXiv:2403.14566 (2024)
8. Liberman, L. & Menell, J. H. Breast imaging reporting and data system (BI-RADS). Radiologic Clinics **40**, 409-430 (2002)
9. Koh, P. W. *et al.* in International conference on machine learning, 5338-5348.PMLR, (2020)
10. Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F. & Pan, W. Promises and pitfalls of black-box concept learning models. arXiv preprint arXiv:2106.13314 (2021)
11. Yuksekgonul, M., Wang, M. & Zou, J. Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
12. Espinosa Zarlenga, M. *et al.* Concept embedding models: Beyond the accuracy-explainability trade-off. Advances in neural information processing systems **35**, 21400-21413 (2022)
13. Hongmei, W., Hou, J., He, S., Yang, S. & Chen, H. Concept Complement Bottleneck Model for Interpretable Medical Image Diagnosis. Medical Imaging with Deep Learning (2026)
14. Nie, Y. *et al.* Conceptclip: Towards trustworthy medical ai via concept-enhanced contrastive langauge-image pre-training. arXiv e-prints, arXiv: 2501.15579 (2025)
15. Yan, A. *et al.* Robust and interpretable medical image classifiers via concept bottleneck models. arXiv preprint arXiv:2310.03182 (2023)
16. Kim, C. *et al.* Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature medicine **30**, 1154-1165 (2024)

17. Pellegrini, C. *et al.* in International Conference on Medical Image Computing and Computer-Assisted Intervention, 420-429.Springer, (2023)
18. Bie, Y., Luo, L. & Chen, H. in Proceedings of the AAAI Conference on Artificial Intelligence, 837-845. (2024)
19. Gao, Y. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11194-11204. (2024)
20. Windsor, R., Jamaludin, A., Kadir, T. & Zisserman, A. Vision-language modelling for radiological imaging and reports in the low data regime. arXiv preprint arXiv:2303.17644 (2023)
21. Zhang, S. *et al.* A multimodal biomedical foundation model trained from fifteen million image–text pairs. Nejm Ai **2**, AIoa2400640 (2025)
22. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
23. He, K., Zhang, X., Ren, S. & Sun, J. in Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778. (2016)
24. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. in Proceedings of the IEEE conference on computer vision and pattern recognition, 4700-4708. (2017)
25. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
26. Pawłowska, A. *et al.* Curated benchmark dataset for ultrasound based breast lesion analysis. Scientific Data **11**, 148 (2024)
27. Petroianu, A. Diagnosis of acute appendicitis. International journal of surgery **10**, 115-119 (2012)