Sparse Local Latents for Explainable Zero-Shot Reasoning in Medical Vision-Language Models

Martin Goetze[1,2⋆], Patrick Wienholt[1,2], Dennis Eschweiler[1,2], Marvin Gazibaric[1,2], Christiane Kuhl[2], Sven Nebelung[1,2], and Daniel Truhn[1,2]

[1] Lab for Artificial Intelligence in Medicine, Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany
[2] Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany
martin.goetze@rwth-aachen.de

Abstract. Vision–language models trained on large-scale medical image–text data achieve strong zero-shot performance but typically rely on global, dense representations that are difficult to interpret and provide little to no visual grounding. We introduce a framework to extract and exploit sparse, local features from a medical vision–language model (MedSigLIP) while preserving text-aligned capabilities. First, we obtain local, text-aligned image embeddings using (1) distillation and (2) token alignment. We analyze these local embeddings using a sparse autoencoder, yielding sparse latents and show that they correspond to recurring, clinically meaningful image patterns. By ranking these latents using MedSigLIP's text embeddings and selectively recombining latents back into the teacher embedding space, we construct new classification probes that focus on interpretable local evidence. We do not fine-tune MedSigLIP and thus retain text-aligned capabilities. Evaluated on VinDR-CXR with external testing on NIH ChestXray, our approach improves mean AUROC on VinDR by up to 0.05 and significantly increases pointing game localization recall by up to 9% on both datasets, across multiple attribution methods. These results demonstrate that sparse latent structure in distilled local features provides an effective mechanism for interpretable and clinically grounded reasoning in medical vision–language models.

Keywords: Medical vision–language models · Sparse representations · Explainable ML.

1 Introduction

Large vision-language models (VLMs) [20,31] have recently emerged as powerful methods for medical image understanding, enabling zero-shot classification, retrieval, and report generation. However, their internal representations are typically global and dense, which limits interpretability and makes it difficult to localize the visual evidence underlying a given prediction [25]. This lack of transparency is a major barrier to clinical acceptance and reliable deployment.

⋆ Corresponding author.

Medical foundation models include both unimodal and multimodal architectures trained on large-scale healthcare data and adapted to downstream tasks through fine-tuning or zero-shot inference [1,31]. MedSigLIP [20] is a CLIP-style [19] model trained with the SigLIP loss [29] on medical images and free-text reports to learn a shared image-text embedding space that supports open-vocabulary zero-shot classification and image-text retrieval. Local attribution methods such as saliency, occlusion, and integrated gradients [22,28,24] are commonly applied to these models, but they operate on dense features that conflate multiple visual concepts and are sensitive to noise [21]. Other gradient based methods include Grad-ECLIP [32]. As a result, their explanations may be unstable and difficult to interpret in practice. Recent work motivated by the linear representation hypothesis [18] suggests that sparse and locally grounded features, learned via sparse autoencoders (SAE) with linear encoders and decoders, can yield more interpretable, robust, and human-aligned representations in both language and vision models [5,17,12,11,7]. In parallel, bag-of-features (BoF) models that treat images as sets of patch embeddings have achieved strong performance by encoding local patches and aggregating them for classification [16,3,4,27].

However, integrating such sparse, local representations into vision-language models while retaining their zero-shot capabilities remains an open challenge and is the focus of this work. In particular, we lack methods that (i) provide local, text-aligned features suitable for text-conditioned attribution, and (ii) expose an interpretable sparse latent structure that can be used to better understand and guide model predictions. In this work, we address this gap by retrieving local, text-aligned features from MedSigLIP and analyzing them with sparse autoencoders to obtain interpretable sparse latents. Our main contributions are:

1. We distill MedSigLIP into a BoF student that encodes image patches, and learn a linear map from teacher tokens to the pooled embedding, yielding two sources of local, text-aligned features that enable text-aligned localization.
2. We train a TopK SAE [8] on these local embeddings and show that the sparse latents correspond to clinically meaningful, localized image patterns, which can be mined and interpreted using MedSigLIP's text embeddings.
3. We construct improved classification probes by recombining the most relevant sparse latents back into the MedSigLIP embedding space, showing that these sparse probes improve both text-aligned classification and localization.

While the individual building blocks of our methodology have been studied before, careful combination allows the backbone to stay frozen and retains the shared image-text space. Taken together, our results indicate that sparse local features offer a practical mechanism for improving post-hoc explainability and clinically grounded reasoning in medical VLMs.

2 Methods

2.1 Datasets

We train and evaluate our approach on the VinDR-CXR dataset [15,9] (PhysioNet Credentialed Health Data License), using the official train and test splits.

VinDR includes both diagnosis labels and bounding boxes for localization. We use weighted fusion for bboxes [23]. In addition, we perform external testing on a subset of 5000 samples of the test NIH chest radiograph dataset [26] under the CC0 license. All localization analyses on NIH data are performed on the full set of bounding boxes given in the NIH dataset. We resize all images to 448×448.

2.2 Retrieving local text-aligned features

To obtain local, text-aligned image features suitable for explainability, we consider two approaches. We first train a ResNet18 [10] on patches extracted from chest radiographs to approximate the global image embedding produced by MedSigLIP. During training, we split the images into $H \times H$ patches, feed them into the ResNet18 individually and average the output embeddings. We align the averaged embedding to the teacher embedding using a cosine embedding loss. We note that these embeddings are translation invariant. During inference, we use a sliding window to generate patches for the BoF model (Stride 16×16).

In our second approach, we start from the spatial tokens of the MedSigLIP vision encoder and learn a linear layer to map the mean of the aligned token embeddings to the pooled MedSigLIP image embedding. We effectively align token-level representations with the global embedding. This approach is similar in spirit to PACL [14] but does not require retraining the backbone.

Because MedSigLIP uses a shared image–text embedding space, both types of local embeddings can be directly compared with text prompts to obtain text-conditioned localization maps without task-specific supervision. For a given text embedding, we compute the cosine similarity between the text and local feature.

2.3 Extracting sparse local features

To expose structure in the local feature space, we train a TopK SAE [8] on patch embeddings. Let $X \in \mathbb{R}^{n \times 1152}$ be patch embeddings extracted from the training set. The SAE consists of a linear encoder, a TopK nonlinearity that enforces sparsity, and a linear dictionary. We use an archetypal dictionary [6]. The idea of the archetypal dictionary is that the dictionary is a linear and stochastic combination of representative patch embeddings. We use a tied initialization, meaning $W_{\text{enc}}^{\top} = D$ at initialization [8]. Given an input patch embedding x, the encoder computes a pre-activation $W_{\text{enc}}(x + b_{\text{pre}})$, and the sparse code is obtained by keeping only the top k activations $S = \text{TopK}(W_{\text{enc}}(X + b_{\text{pre}}))$. The decoder reconstructs the input as $SD^{\top}$ and we minimize the squared reconstruction error.

To relate the sparse latents to semantic concepts, we leverage MedSigLIP's text embeddings as probes. For each diagnosis label, we obtain a text embedding P from MedSigLIP, using the label name (with clinically motivated prefixes such as "pleural effusion" or "pulmonary mass" where appropriate). We then compute $D^{\top}P$, which measures how strongly each dictionary atom (and thus each sparse latent) aligns with the text concept. We primarily focus on the latents with the largest positive and negative entries in $D^{\top}P$, which we interpret as the most strongly supportive and opposing local features for that diagnosis.

Table 1. Pointing game scores using patch-text alignment.

	BoF		MedSigLIP	
	VinDR	NIH	VinDR	NIH
Pointing Game Score	0.42	0.36	0.42	0.43

To study how different sparse latents co-activate, we compute the lift between activation events A and B as $\text{lift}(A, B) = \frac{P(A,B)}{P(A)P(B)}$. The lift is 1 for independent events and < 1 for negatively correlated events. We use this measure to quantify whether high- and low-similarity latents for a given diagnosis tend to fire together or represent mutually exclusive visual patterns.

2.4 Improved probes from sparse features

Finally, we use the mined sparse latents to construct new classification probes that emphasize interpretable local evidence. For a given diagnosis, we start from the text embedding P and its scores $D^\top P$. We then select the k atoms with the highest scores and the k atoms with the lowest scores, corresponding to the most supportive and most opposing sparse features for that diagnosis.

We create a new probe using $\hat{P} = \text{mask}_k(D^\top P)D$. The masks sets all masked entries to 0. Intuitively, this procedure aims to focus the probe on a small number of interpretable local features that are most relevant to the diagnosis, while suppressing less relevant or noisy directions in the embedding space. We evaluate the probes using AUROC [2] and common explainability localization methods (saliency, occlusion, integrated gradients) [28,24,22] using the pointing game [30]. We use the single highest activation point and count points within bounding boxes. Statistical significance in greater macro pointing game average than baseline is computed using all possible permutations of sign flipping with $\alpha = 0.05$.

3 Results

3.1 Local patches are aligned with teacher text embeddings.

We first evaluate how well the local patch embeddings are aligned with MedSigLIP's text embeddings and how this alignment affects explainability. Qualitatively, we observe strong correspondence between text and image features for both local embedding methods, as seen in Figure 1.

Quantitatively, we compute pointing game scores using raw text embeddings with the local patch embeddings in Table 1. On VinDR-CXR, the BoF matches MedSigLIP (0.42), indicating that distillation preserves the zero-shot localization ability. On NIH, the BoF model underperforms the token alignment (0.36 vs. 0.43), suggesting that the student generalizes less well across datasets.

We also study the effect of patch size on the trade-off between global reconstruction and local localization. As patch size increases, the BoF student's

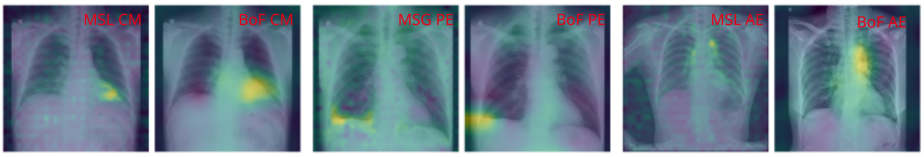

Fig. 1. Similarity maps for cardiomegaly (CM), pleural effusion (PE), and aortic enlargement (AE) using cosine similarity of text embedding with local embeddings. Left: MedSigLIP (MSL), Right: Bag-of-features (BoF) tokens for each pair.

ability to reconstruct the teacher embedding (measured by MSE) improves. However, localization performance (pointing game recall) peaks at a patch size of 64×64, and degrades for larger patches. This suggests that very large patches encode more global information but dilute spatial specificity. We continue with the model trained on patches of size 64×64.

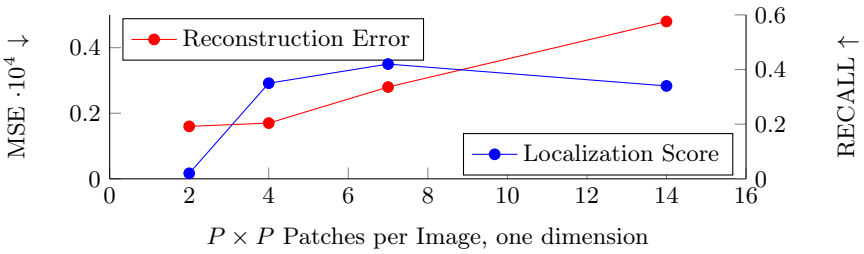

Fig. 2. Reconstruction error and localization score when varying the patches per image.

3.2 Sparse features are prototypes of diagnostic local features.

We analyze the sparse features learned by the TopK autoencoder to assess whether they correspond to diagnostic patterns. We train the SAE on patch embeddings using 2000 sparse features and $k = 10$. Additionally, we use an archetypal dictionary with 5000 K-means centroids [8,6,13].

By computing the cosine similarity between dictionary atoms and MedSigLIP text embeddings for each diagnosis, we identify the sparse features with the highest and lowest similarity to a given concept. Qualitatively, these latents act as prototypes: the highest-similarity latents correspond to patches that frequently contain the diagnostic pattern, while the lowest-similarity latents highlight regions that are unlikely to contain that pattern. Figure 3 shows examples of such mutually exclusive feature pairs from the BoF sparse latents.

To quantify co-activation, we compute the lift for the highest- and lowest-similarity features of each diagnosis (Table 2). For most diagnoses, the lift is close to zero, indicating that high- and low-similarity latents almost never fire together

Table 2. Lift of activations of least similar and most similar activations (VinDR) to prompt embedding for Cardiomegaly (CM), pleural effusion (PM), aortic enlargement (AE), pulmonary fibrosis (PF), lung opacity (LO), and pleural thickening (PT).

	BoF						MedSigLIP					
	CM	PE	AE	PF	LO	PT	CM	PE	AE	PF	LO	PT
Lift	0.01	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.53	0.05	0.00

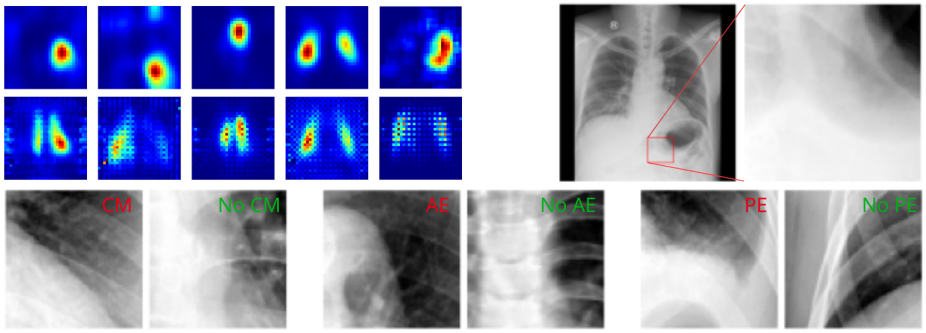

Fig. 3. (Top Left) Highest similarity sparse latents for diagnoses fire in fixed areas, shown are sparse latent fires. Top: BoF, Bottom: MedSigLIP. (Top Right) Mistaken organ in BoF sparse latents. Abdominal feature locally appears like pleural effusion. The patch with the highest activation for the sparse 'pleural effusion' feature is shown. (Bottom) Mutually exclusive sparse features for cardiomegaly (CM), aortic enlargement (AE), and pleural effusion (PE) from the BoF model. Activation for positive label on the left for each pair.

and thus represent distinct visual concepts. We find that we can identify clinically relevant sparse features by computing the cosine between the dictionary entries and text embeddings for a given diagnosis. An exception is pulmonary fibrosis in the MedSigLIP token-based sparse latents, where these features co-activate more often, though still less than expected under independence.

The spatial distribution of sparse latent activations further supports their interpretability. Because chest X-rays have a consistent anatomical layout, we observe that the highest-similarity sparse latents for a given diagnosis tend to fire in anatomically plausible regions (e.g., heart or lungs), as illustrated in Figure 3. In the BoF model some latents occasionally fire in anatomically implausible positions when local appearance resembles another organ (e.g., a pleural effusion-like pattern in the abdomen), as seen in Figure 3, whereas this effect is absent in the MedSigLIP token latents that retain positional context. We also find laterality-specific latents for conditions such as pleural effusion, where separate sparse features predominantly activate on the left or right lung, respectively. These observations indicate that the sparse features capture meaningful, localized, and often anatomically constrained concepts that are useful for explainability.

Table 3. AUC of sparse classification probes using the BoF model and the MedSigLIP aligned features (MSL). We discriminate between the inclusion of most similar $\top$, least similar $\perp$ and a mixture of both in the probes across $k \in \{1, 2, 3\}$. The best method across dataset and inclusion method is bold.

	VinDR						NIH					
	$k=1$		$k=2$		$k=3$		$k=1$		$k=2$		$k=3$	
Latents	BoF	MSL	BoF	MSL	BoF	MSL	BoF	MSL	BoF	MSL	BoF	MSL
$2k\top$	0.79	0.79	**0.80**	0.77	0.79	0.76	0.66	0.64	**0.69**	0.64	**0.69**	0.65
$2k\perp$	0.83	0.73	**0.88**	0.80	**0.88**	0.81	0.62	0.60	0.63	0.60	**0.64**	0.60
$k\top, k\perp$	0.79	0.76	**0.86**	0.80	**0.86**	0.81	0.67	0.65	0.68	0.66	**0.69**	0.66

3.3 Sparse features lead to improved classification and localization.

We evaluate how the probes constructed from the BoF and token-based sparse latents affect zero-shot classification and localization. For each diagnosis and each local feature source (BoF or MedSigLIP tokens), we consider three probe configurations: $2k$ latents with the highest/lowest similarity to the text embedding ($2k\top/2k\perp$), and k highest and k lowest similarity latents ($k\top, k\perp$). We vary $k \in \{1, 2, 3\}$ and use macro average AUROC. For localization, we use $k = 2$.

On VinDR-CXR, the baseline MedSigLIP text probes achieve a mean AUROC of 0.83. Using sparse probes derived from the BoF model improves AUROC across most configurations, with the mixed ($k\top, k\perp$) and bottom configuration ($2k\perp$) reaching up to 0.88 (Table 3). These gains indicate that restricting the classifier to a small number of sparse latents can enhance performance. On NIH, both the sparse probes match the MedSigLIP's AUROC of 0.69, showing that the sparse probes match but do not clearly exceed the baseline on the test dataset.

We also assess localization using the pointing game on both datasets. Table 4 reports pointing game recall for the baseline and for sparse probes, using several attribution methods: integrated gradients (IG), absolute integrated gradients (IG-Abs), occlusion (OC), and saliency, as well as our own similarity maps ('Ours'). On VinDR-CXR, sparse probes based on the BoF model improve IG recall from 0.33 (baseline) to 0.43 and OC recall from 0.40 to 0.44. Similar improvements are observed on NIH, with OC recall increasing from 0.44 to 0.50. These gains are consistent and significant across attribution methods and local feature sources, showing that sparse probes not only maintain classification performance but also yield more localized and plausible explanations.

The two local feature extraction approaches show different strengths. Sparse probes derived from BoF embeddings achieve the highest classification performance on VinDR-CXR. In contrast, sparse probes derived from MedSigLIP tokens provide stronger localization performance on NIH, achieving a pointing-game score of 0.45 compared to 0.38 for BoF similarity maps. Qualitatively, BoF features occasionally activate in anatomically implausible regions because patch representations are translation invariant, whereas token-aligned features retain positional information from the original transformer architecture. These results suggest a trade-off between classification performance and spatial fidelity.

Table 4. Pointing game scores across explainability methods using raw and sparse probes. The 'Ours' column describes the localization scheme using similarity to either BoF or aligned token local features. Significant improvements over baseline underlined.

		VinDR					NIH				
Latent	Probe	IG	IG-Abs	OC	Saliency	Ours	IG	IG-Abs	OC	Saliency	Ours
	Baseline	0.33	0.32	0.40	0.15	-	0.33	0.34	0.44	0.16	-
$2\top, 2\bot$	BoF	**0.43**	**0.42**	0.44	0.27	**0.44**	0.41	0.40	**0.50**	**0.25**	0.38
	MedSigLIP	0.34	0.33	**0.46**	0.17	**0.44**	0.39	0.38	0.48	0.24	**0.45**
$4\bot$	BoF	0.32	0.32	0.35	0.20	0.38	0.32	0.32	0.45	0.17	0.30
	MedSigLIP	0.25	0.25	0.31	0.12	0.37	0.24	0.24	0.38	0.14	0.36
$4\top$	BoF	0.41	0.40	0.44	**0.28**	0.41	**0.42**	**0.41**	0.48	**0.25**	0.38
	MedSigLIP	0.34	0.34	**0.46**	0.19	0.42	0.33	0.33	0.47	0.19	0.32

4 Discussion

We propose a framework that adapts two types of local image representations: BoF distillation and token alignment to enable post-hoc explanations of MedSigLIP while preserving its text-aligned capabilities. By retaining image–text alignment locally, our method allows the model to produce localization maps for arbitrary text prompts within the learned embedding space. By applying a sparse autoencoder to these local embeddings, we exposed a sparse latent structure that can be mined using MedSigLIP's text embeddings to identify a small set of local features most supportive or opposing for each diagnosis. Recombining these sparse latents into probes in the teacher embedding space improved classification on VinDR-CXR and consistently enhanced localization as measured by the pointing game across attribution methods, indicating that focusing predictions on a small number of sparse features benefits performance and explainability.

A central insight of our work is that sparse autoencoders can reveal a structured, clinically meaningful latent space over local features. Our analyses show that high- and low-similarity latents for a diagnosis rarely co-fire, suggesting that they represent distinct visual patterns, and that many latents are constrained to anatomically plausible regions or exhibit laterality (e.g., left vs. right lung). These properties are desirable for explainability because they align the learned latents with stable, human-recognizable features. Our probe construction procedure can also be viewed as a form of concept selection. However, unlike feature-selection approaches using labeled examples to identify features, sparse latent selection is performed only through alignment with text embeddings.

Our study has limitations. First, we test on a single SAE configuration and VLM teacher. We selected MedSigLIP as a publicly available medical contrastive vision-language. The proposed framework is largely model-agnostic and could be applied to alternative medical VLMs that expose compatible image and text embeddings, which we leave to future work. Second, we use simple probes defined directly from MedSigLIP text embeddings of diagnosis labels, with a few hand-

crafted prefixes, and do not perform prompt engineering. Third, our quantitative evaluation relies on the pointing game with bounding box annotations, which provides a coarse measure of localization. We do not perform pixel-level evaluation or user studies with radiologists. We also do not perform statistical analyses on the individual chosen sparse latents corresponding to features and evaluate only on chest X-ray datasets. Finally, although our probes are constructed without labeled supervision, we do not evaluate generalization to previously unseen disease categories limiting the zero-shot evaluation.

Overall, our results support the view that sparse local features provide a promising mechanism for interpretable and clinically grounded zero-shot reasoning in medical vision–language models. By combining distillation, sparse autoencoding, and probe construction, we derive explanations that are more localized, more structured, and more aligned with human concepts. Future work should extend this framework to other modalities, and explore alternative sparse architectures to iterate on local sparse concepts. Code can be found at https://github.com/martingoe/sparse_local_latents.

Acknowledgements & Disclosure of Interests. This research is supported by the Deutsche Forschungsgemeinschaft – DFG (417508432, 517243167, 515639690), the German Federal Ministry of Research, Technology and Space (Transform Liver – 031L0312C, DECIPHER-M, 01KD2420B) and the European Union Research and Innovation Programme (ODELIA – GA 101057091, SAGMA – GA 101222556). The authors have no competing interests in the paper.

References

1. Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., Merhof, D.: Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision (Oct 2023). https://doi.org/10.48550/arXiv.2310.18689
2. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition **30**(7), 1145–1159 (Jul 1997). https://doi.org/10.1016/S0031-3203(96)00142-2
3. Brendel, W., Bethge, M.: Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In: International Conference on Learning Representations (Sep 2018)
4. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. Work Stat Learn Comput Vision, ECCV **Vol. 1** (Jan 2004)
5. Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L.: Sparse Autoencoders Find Highly Interpretable Features in Language Models (Oct 2023). https://doi.org/10.48550/arXiv.2309.08600
6. Fel, T., Lubana, E.S., Prince, J.S., Kowal, M., Boutin, V., Papadimitriou, I., Wang, B., Wattenberg, M., Ba, D., Konkle, T.: Archetypal SAE: Adaptive and Stable Dictionary Learning for Concept Extraction in Large Vision Models (May 2025). https://doi.org/10.48550/arXiv.2502.12892
7. Fel, T., Wang, B., Lepori, M.A., Kowal, M., Lee, A., Balestriero, R., Joseph, S., Lubana, E.S., Konkle, T., Ba, D., Wattenberg, M.: Into the Rabbit Hull: From Task-Relevant Concepts in DINO to Minkowski Geometry (Oct 2025). https://doi.org/10.48550/arXiv.2510.08638

8. Gao, L., Dupre la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J.: Scaling and evaluating sparse autoencoders. International Conference on Representation Learning **2025**, 26721–26754 (May 2025)
9. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., et al.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), E215–220 (Jun 2000). https://doi.org/10.1161/01.cir.101.23.e215
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015). https://doi.org/10.48550/arXiv.1512.03385
11. Joseph, S., Suresh, P., Goldfarb, E., et al.: Steering CLIP's vision transformer with sparse autoencoders (Apr 2025). https://doi.org/10.48550/arXiv.2504.08729
12. Lindsey, J., Gurnee, W., Ameisen, E., et al.: On the biology of a large language model. Transformer Circuits Thread (2025)
13. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, vol. 5.1, pp. 281–298. University of California Press (Jan 1967)
14. Mukhoti, J., Lin, T.Y., Poursaeed, O., Wang, R., Shah, A., Torr, P.H., Lim, S.N.: Open vocabulary semantic segmentation with patch aligned contrastive learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19413–19423. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.01860
15. Nguyen, H.Q., Lam, K., Le, L.T., et al.: VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Scientific Data **9**(1), 429 (Jul 2022). https://doi.org/10.1038/s41597-022-01498-w
16. O'Hara, S., Draper, B.A.: Introduction to the Bag of Features Paradigm for Image Classification and Retrieval (Jan 2011). https://doi.org/10.48550/arXiv.1101.3354
17. Pach, M., Karthik, S., Bouniot, Q., Belongie, S., Akata, Z.: Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (Oct 2025)
18. Park, K., Choe, Y.J., Veitch, V.: The linear representation hypothesis and the geometry of large language models. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, vol. 235, pp. 39643–39666. JMLR.org, Vienna, Austria (Jul 2024)
19. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763. PMLR (Jul 2021)
20. Sellergren, A., Kazemzadeh, S., Jaroensri, T., et al.: MedGemma Technical Report (Jul 2025). https://doi.org/10.48550/arXiv.2507.05201
21. Sharkey, L., Chughtai, B., Batson, J., et al.: Open Problems in Mechanistic Interpretability (Jan 2025). https://doi.org/10.48550/arXiv.2501.16496
22. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps (Apr 2014). https://doi.org/10.48550/arXiv.1312.6034
23. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (Mar 2021). https://doi.org/10.1016/j.imavis.2021.104117
24. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. pp. 3319–3328. ICML'17, JMLR.org, Sydney, NSW, Australia (Aug 2017)

25. Wang, G., Zhang, S., Huang, X., Vercauteren, T., Metaxas, D.: Editorial for special issue on explainable and generalizable deep learning methods for medical image computing. Medical Image Analysis **84**, 102727 (Feb 2023). https://doi.org/10.1016/j.media.2022.102727
26. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases. In: Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics, pp. 369–392. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-13969-8_18
27. Wienholt, P., Kuhl, C., Kather, J.N., Nebelung, S., Truhn, D.: MedicalPatchNet: A Patch-Based Self-Explainable AI Architecture for Chest X-ray Classification (Sep 2025). https://doi.org/10.48550/arXiv.2509.07477
28. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks. In: Computer Vision – ECCV 2014. pp. 818–833. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
29. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-Training (Sep 2023). https://doi.org/10.48550/arXiv.2303.15343
30. Zhang, J., Lin, Z., Brandt, J., et al.: Top-Down Neural Attention by Excitation Backprop. In: Computer Vision – ECCV 2016. pp. 543–559. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_33
31. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis **91**, 102996 (Jan 2024). https://doi.org/10.1016/j.media.2023.102996
32. Zhao, C., Wang, K., Hsiao, J.H., Chan, A.B.: Grad-eclip: Gradient-based visual and textual explanations for clip. arXiv preprint arXiv:2502.18816 (2025)