

Sparse Autoencoders for Interpretable Medical Image Representation Learning

Philipp Wesp^{1,2}, Robbie Holland^{1,2}, Vasiliki Sideri-Lampretsa^{3,4}, and Sergios Gatidis^{1,2}

¹ Stanford Center for Artificial Intelligence in Medicine and Imaging, Stanford University, Stanford, CA, USA

² Department of Radiology, Stanford University, Stanford, CA, USA

³ Chair of AI in Healthcare and Medicine, TUM, Munich, Germany

⁴ TUM University Hospital, Munich, Germany

Abstract. Vision foundation models (FMs) achieve state-of-the-art performance in medical imaging. However, they encode information in abstract latent representations that clinicians cannot interrogate or verify. The goal of this study is to investigate Sparse Autoencoders (SAEs) for replacing opaque FM image representations with human-interpretable, sparse features. We train SAEs on embeddings from BiomedParse (biomedical) and DINOv3 (general-purpose) using 909,873 CT and MRI 2D images from the TotalSegmentator dataset. We find that learned sparse features: (a) reconstruct original embeddings with high fidelity (R^2 up to 0.941) and recover up to 87.8 % of downstream performance using only 10 features (99.4 % dimensionality reduction), (b) preserve semantic fidelity in image retrieval tasks, (c) correspond to specific concepts that can be expressed in language using large language model (LLM)-based auto-interpretation, (d) bridge clinical language and abstract latent representations in zero-shot language-driven image retrieval. Our work indicates SAEs are a promising pathway towards interpretable, concept-driven medical vision systems. Code repository: <https://github.com/pwesp/sail>.

Keywords: Sparse Autoencoders · Medical Imaging · Interpretability · Foundation Models · Mechanistic Interpretability

1 Introduction

Vision foundation models (FMs) achieve strong performance in medical imaging tasks such as segmentation, classification, and retrieval, but encode information in abstract, low-dimensional feature representations [14,15]. At the same time, clinical deployment demands interpretability: physicians must justify decisions, detect failure modes, and document reasoning, yet model internals remain inaccessible [13]. This creates a fundamental misalignment between abstract learned representations and the anatomical and clinical concepts that clinicians reason with.

Mechanistic interpretability aims to reverse-engineer model internals into human-understandable components. Sparse Autoencoders (SAEs) [4,7] are a

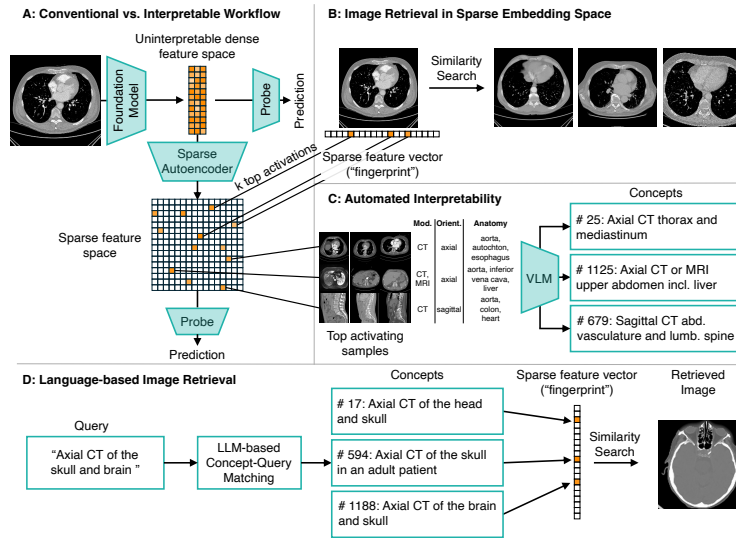


Fig. 1. (A) A Sparse Autoencoder replaces opaque dense FM embeddings with a sparse feature space. (B) Sparse fingerprint retrieval matches images by cosine similarity over k top-activated features. (C) A VLM generates a concept description for each feature from its top-activating images and metadata. (D) An LLM maps a clinical text query to matching feature concepts for zero-shot image retrieval.

leading approach, decomposing polysemantic activations in large language models (LLMs) into monosemantic features [8] that each correspond to a single coherent concept. A recent study applied SAEs to chest radiograph embeddings, demonstrating that a small number of interpretable sparse features can represent clinically relevant visual concepts and support radiology report generation [1]. That study, however, was restricted to a single modality and a single FM architecture with paired text supervision for concept labelling, leaving open whether anatomical structure emerges in self-supervised models across CT, MRI, and diverse anatomical regions. This raises a central question: do anatomical and clinical concepts emerge from self-supervised medical vision training without explicit labels, and can SAEs expose this structure consistently across architecturally distinct foundation models? Because whole-body CT and MRI provide no paired text, we introduce a grounding procedure that scores VLM- and segmentation-derived concept labels against a null baseline to make monosemanticity measurable, together with a configuration-ranking framework that exposes the monosemanticity-performance trade-off.

To this end, we train Matryoshka SAEs [6] with BatchTopK sparsification [5] on frozen embeddings from BiomedParse [21] (biomedical FM) and DINOv3 [18] (general-purpose FM), alongside a random-weight baseline to isolate learned representational structure from architectural effects, across 909,873 CT and MRI 2D images from the TotalSegmentator dataset [19,2] (Fig. 1). We find that sparse

features (a) faithfully reconstruct dense embeddings (R^2 up to 0.941) and recover 87.8% of downstream performance with only 10 features, (b) preserve 97.7% of dense retrieval quality with five-feature fingerprints, (c) correspond to monosemantic concepts verified by an independent large language model judge, and (d) enable zero-shot language-driven image retrieval bridging clinical text and medical image content. These findings indicate that self-supervised vision FMs implicitly encode anatomy-aligned structure that SAEs can expose as language-describable sparse features, a step toward interpretable medical AI aligned with human language.

2 Methods

We train SAEs (the only optimized parameters) on frozen, precomputed embeddings from three vision FMs: BiomedParse [21] (1536-dim, biomedical FM), DINOv3 [18] (1024-dim, general-purpose self-supervised ViT), and an untrained BiomedParse model with randomly initialised weights (1536-dim, random-weight baseline) to isolate learned representational structure from architectural effects.

2.1 Sparse Autoencoder

We adopt the Matryoshka SAE architecture [6] with $L = 4$ nested dictionary levels of increasing size $[D_1, D_2, D_3, D_4]$. A single shared linear encoder projects the input into D_4 pre-activation codes. Level ℓ uses only the first D_ℓ codes as a prefix subset, so that early levels capture coarse structure and later levels refine it progressively. A single shared decoder (encoder weights transposed, columns normalised to unit norm) reconstructs the input at each level by padding smaller-level activations with zeros. During training, we apply BatchTopK sparsification [5]: k features are active per sample on average across the batch, allowing flexible per-sample allocation unlike fixed per-sample TopK. At inference, a JumpReLU [16] threshold, estimated as a running average of the minimum kept activation during training, replaces BatchTopK. The training objective is the mean squared error (MSE) between input and reconstruction, averaged across all L levels, with no auxiliary sparsity or diversity penalties.

Monosemanticity scoring. To compare configurations, we score each feature as $M(f) = C(f)S(f)$. Coherence $C(f) = \binom{n}{2}^{-1} \sum_{i < j} J(O_i, O_j)$ is the mean pairwise Jaccard similarity J over the organ-presence sets O_i of the $n = 10$ top-activating samples of f . Specificity $S(f) = 1 - H(p)/\log K$ is the normalized inverse Shannon entropy of the organ-count distribution p over those samples, where $K = 121$ is the number of organ labels. To distinguish genuine structure from chance co-occurrence, we estimate a null baseline $\bar{C}_{\text{null}}$ (mean of C over 100 random 10-sample subsets), and call a feature above-null when $C(f) > \bar{C}_{\text{null}}$. The configuration-level score M_{config} is the mean $M(f)$ of the top-10 features per configuration.

2.2 Interpretability Evaluation

We evaluate the interpretability of learned sparse features through three complementary demonstrations.

Sparse Fingerprint Retrieval. We define a sparse fingerprint as the k most activated features and their values per image, and retrieve similar images by cosine similarity over fingerprints. Retrieval quality is measured as mean cosine similarity to the reference in the dense embedding space, with dense retrieval as the upper bound.

Automated Feature Interpretation. To assess whether individual features encode interpretable and consistent concepts, we greedily select the 5 most dissimilar samples from the top-20 activating images for the top-100 most monosemantic (M score) features using cosine similarity. We then prompt the vision language model (VLM) MedGemma 27B [17] to generate a natural-language concept description from their images and metadata (modality, orientation, anatomy, demographics) [10,3]. A VLM judge (separate MedGemma 27B) then receives the same images and five candidate descriptions and must identify the correct one. The four distractors are sampled uniformly from the concept descriptions of other not-always-active features with more than 10 activations, which excludes trivial and degenerate cases. Generation and judging run as independent inference calls with different prompts, so the task measures generic vision-language matching rather than self-recognition. The rank of the true concept (1 = best, 5 = worst) quantifies interpretability, where chance corresponds to a mean rank of 3.

Language-Driven Image Retrieval. An LLM identifies feature descriptions that match a clinical text query, assembling a sparse fingerprint from their mean activations for cosine retrieval without a reference image. This zero-shot procedure demonstrates that sparse feature concepts can bridge human language and medical image content.

3 Experiments & Results

We evaluate Matryoshka SAEs on BiomedParse and DINOv3 embeddings from the TotalSegmentator dataset: 1,844 scans (1,228 CT, 616 MRI) from 10 institutions, yielding 909,873 2D images with 138 per-image metadata fields spanning anatomy presence, imaging parameters and demographics. Scans from three institutions are withheld entirely as a test set (14.1 % of images), and the remaining scans are split 80/20 into train and validation sets stratified by modality, age group, and sex, yielding 68.6 % and 17.3 % of images respectively. CT volumes use one of 38 study-type-dependent Hounsfield-unit windows (e.g. thorax –200/1400, abdomen 50/1000), and MRI volumes are resampled to isotropic in-plane spacing and clipped to the 1st–99th intensity percentile. Both are rescaled to $[0, 255]$, center-cropped, resized to 1024×1024 , replicated to three channels,

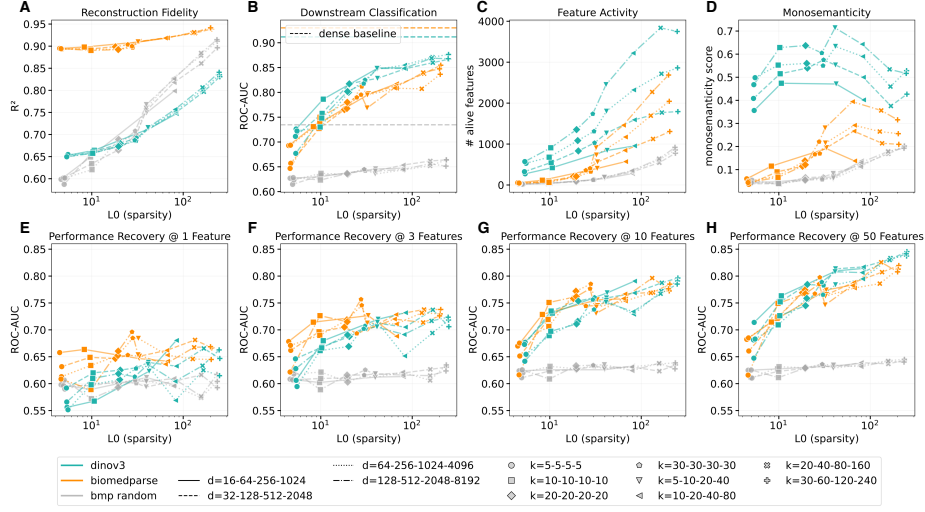


Fig. 2. SAE quality and performance recovery across 32 configurations per FM (DINOv3: blue, BiomedParse: orange, random baseline: grey). (A–D) Reconstruction fidelity (R^2), downstream ROC-AUC, alive features, and monosemanticity score vs. L0 sparsity. (E–H) Performance recovery using only the top- N features ($N = 1, 3, 10, 50$).

and ImageNet-normalized. SAEs are optimized with Adam [12] ($\text{lr} = 10^{-4}$, cosine annealing to 10^{-6} , 100 epochs, batch size 2048) across 32 configurations per FM: 4 dictionary size families ([16, 64, 256, 1024] to [128, 512, 2048, 8192]) and 8 sparsity patterns (4 fixed, 4 progressive K). Baselines are a dense embedding upper bound and a random-weight BiomedParse model isolating learned structure from architectural effects.

3.1 SAE Quality

Latent space reconstruction (R^2). Figure 2 shows reconstruction quality, downstream performance, and alive feature counts across all 32 configurations per FM. R^2 ranges from 0.890 to 0.941 for BiomedParse and from 0.649 to 0.841 for DINOv3.

Downstream performance (ROC-AUC). Dense embedding baselines achieve ROC-AUC of 0.907 (BiomedParse) and 0.912 (DINOv3) across anatomical classification tasks. Optimal sparse configurations recover 90.2% and 93.0% of dense performance, respectively. The random-weight baseline reaches only 0.606–0.651 AUC despite a comparable R^2 range (0.587–0.915), confirming that downstream utility reflects learned representational structure, not architectural capacity alone. This dissociation shows that reconstruction fidelity is an insufficient proxy for semantic utility, since a sparse code can faithfully reconstruct a random embedding space while encoding no semantically meaningful structure. Conversely,

Table 1. Top-3 SAE configurations per FM ranked by combined monosemanticity and performance recovery score (32 configurations each). Bold: selected optimal configuration. #Mono/Perf/Comb: monosemanticity/performance/combined rank.

Model	Dict Sizes	Top-K Values	#Mono	#Perf	#Comb
BiomedParse	128, 512, 2048, 8192	20, 40, 80, 160	2	3	1
	128, 512, 2048, 8192	30, 60, 120, 240	3	6	2
	128, 512, 2048, 8192	10, 20, 40, 80	1	12	3
DINOv3	128, 512, 2048, 8192	5, 10, 20, 40	1	11	1
	64, 256, 1024, 4096	5, 10, 20, 40	4	10	2
	128, 512, 2048, 8192	10, 20, 40, 80	2	14	3

DINOv3’s lower R^2 (0.649–0.841) relative to BiomedParse (0.890–0.941) coexists with higher downstream AUC, indicating that task-relevant structure can be preserved under approximate reconstruction.

3.2 SAE Configuration Ranking

Monosemanticity & performance recovery. We quantify the competing properties of monosemanticity and performance recovery [11] across all configurations and select an optimal configuration per model based on a combined score. Figure 2 shows M_{config} and performance recovery. DINOv3 achieves substantially higher monosemanticity (0.356–0.714) than BiomedParse (0.036–0.394), despite BiomedParse’s domain-specific pretraining. The random-weight baseline (0.038–0.202) falls well below both learned models, confirming that monosemanticity reflects learned representational structure rather than architectural capacity. Within the optimal SAEs, the Matryoshka levels realize their intended coarse-to-fine organization, as median feature activation frequency falls from roughly 40,000 (BiomedParse) and 17,000 (DINOv3) images at level 0 to 1,200 and 152 at level 3. The deepest level also holds the majority of coherent (above-null) features, 2,610/3,682 (BiomedParse) and 2,361/3,577 (DINOv3), concentrating organ-specific concepts in deep levels. With $N = 10$ features, BiomedParse recovers 87.8% and DINOv3 recovers 82.4% of dense ROC-AUC, with performance gains diminishing above $N = 10$.

Configuration ranking. Table 1 ranks the top-3 configurations per model by combined monosemanticity and performance recovery score. Progressive Top-K patterns with the largest dictionary family [128, 512, 2048, 8192] dominate the BiomedParse and DINOv3 rankings. BiomedParse’s optimal configuration ($K = [20, 40, 80, 160]$) achieves competitive scores on both dimensions (monosemanticity rank 2, performance rank 3). DINOv3’s optimal ($K = [5, 10, 20, 40]$) leads in monosemanticity (rank 1) but ranks 11th in performance recovery, exemplifying the monosemanticity-performance trade-off inherent to sparser representations.

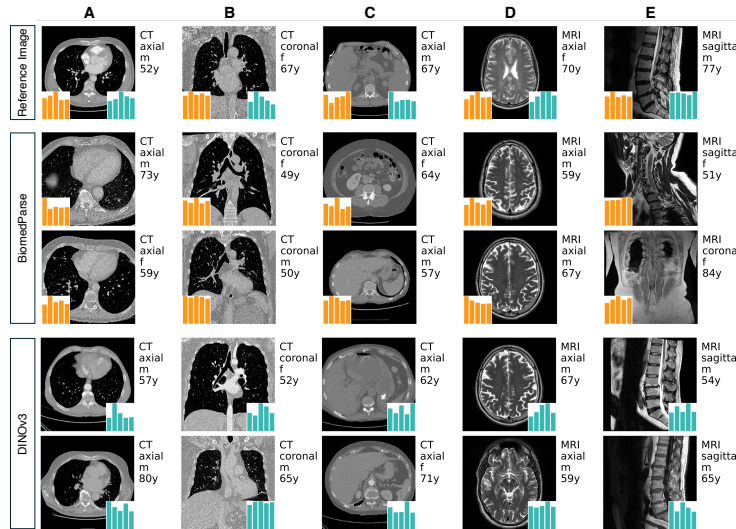


Fig. 3. Sparse fingerprint retrieval at $k = 5$ for five reference cases (A–E) spanning CT and MRI across multiple anatomical regions. Row 1: reference images with BiomedParse (orange) and DINOv3 (blue) fingerprint insets. Rows 2–3: top-2 BiomedParse retrievals. Rows 4–5: top-2 DINOv3 retrievals.

3.3 Sparse Feature Interpretability

For the optimal configurations per FM (Table 1), we evaluate sparse feature interpretability through three demonstrations, excluding the random-weight baseline due to its lack of semantic structure (Sect. 3.1).

Sparse feature-based image retrieval. We evaluate sparse fingerprints, the top- k active features per image, for image retrieval to assess whether sparse features preserve semantic similarity. Retrieval quality is measured as mean cosine similarity of the top-5 retrieved images for 1,000 randomly selected reference images of the test set in the dense embedding space (Table 2). At $k = 5$ features (Fig. 3), BiomedParse achieves 97.7% of dense retrieval quality (0.954 vs. 0.976) and DINOv3 achieves 92.8% (0.831 vs. 0.895). Quality saturates rapidly above $k = 10$ for both models, confirming that semantic content concentrates in a small number of sparse features.

Interpretable sparse feature concepts. We interpret the top-100 interpretable features per model by automated VLM-based concept generation, verified by an independent LLM judge that ranks the true description among five candidates (rank 1 = best, rank 5 = worst). DINOv3 achieves 83/100 rank-1 counts (mean rank 1.26), outperforming BiomedParse (70/100 rank-1 counts, mean rank 1.48). Rank-2 counts are 11/100 and 17/100, respectively (Table 3). Concepts

Table 2. Sparse fingerprint retrieval quality (mean cosine similarity to the reference image in the dense embedding space) as a function of fingerprint size k , averaged over $N = 1,000$ test images. *Dense*: full dense retrieval quality (upper bound).

Model	k=1	k=5	k=10	k=20	Dense
BiomedParse	0.93	0.95	0.96	0.97	0.98
DINOv3	0.75	0.83	0.85	0.86	0.89

Table 3. LLM-as-judge evaluation of automatically generated feature concepts for $N = 100$ features per model. An independent VLM ranks the true concept description among five candidates (1 true + 4 distractors) given the same images. Rank 1 = true concept fits best, Rank 5 = true concept fits worst.

Model	Mean rank	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
BiomedParse	1.48	70	17	9	3	1
DINOv3	1.26	83	11	3	3	0

capture modality, imaging plane, anatomy, and demographics, emerging from self-supervised learning without explicit anatomical labels.

Language-based image retrieval. As an end-to-end demonstration, an LLM maps a clinical text query to matching sparse feature concepts and assembles a sparse fingerprint for cosine retrieval, requiring no reference image or task-specific training. For the query “Axial CT of the abdomen and retroperitoneum in an elderly patient” (Fig. 4), BiomedParse’s features lacks a modality-pure abdomen feature, selects mixed MRI/CT concepts, and retrieves thoracic images. DINOv3, whose richer feature vocabulary includes three anatomy- and modality-specific abdomen CT concepts, retrieves anatomically correct axial abdominal CT images. This shows that automatically labeled, unsupervised concepts can bridge clinical language and medical image content, capturing anatomy and modality while demographic constraints remain open, complementing text-supervised retrieval such as BiomedCLIP [20] or PubMedCLIP [9] without requiring paired image-text supervision.

4 Conclusion

Sparse features from Matryoshka SAEs faithfully preserve embedding structure, recover strong downstream performance with a handful of features, and enable interpretable retrieval and zero-shot language-driven search, extending prior evidence from chest radiographs to multi-modal volumetric imaging across two architecturally distinct foundation models.

DINOv3, despite no biomedical pretraining focus, consistently produces more monosemantic features and comparable downstream performance, suggesting that representational richness matters more than domain alignment for interpretability. Language-driven retrieval, a single-query proof-of-concept, captures anatomy and modality, with finer-grained constraints such as demographics an open direction. Monosemanticity scoring relies on metadata-derived organ labels and VLM-generated concept descriptions rather than human annotation,

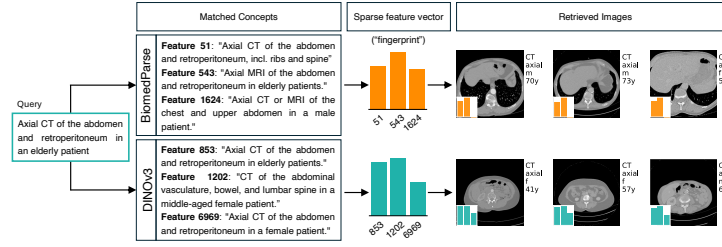


Fig. 4. Zero-shot language-driven retrieval for “Axial CT of the abdomen and retroperitoneum in an elderly patient.” An LLM selects matching feature concepts (left), determining a sparse fingerprint (center) for cosine retrieval (right). BiomedParse selects mixed MRI/CT concepts and retrieves thoracic images. DINOv3 selects CT-specific abdomen features and retrieves correct axial abdominal CT.

providing scalable but proxy-based evidence, with expert evaluation a key next step. The TotalSegmentator dataset covers normal anatomy across 10 institutions and two modalities but excludes pathological cases, and analysis operates at the 2D slice level rather than volumetrically. Because concepts are described from model-specific images, concept quality tracks sparse-feature quality, leaving cross-FM concept similarity, broader clinician-validated query benchmarks, and comparison with text-supervised retrieval as future work.

Overall, sparse autoencoders provide a practical interpretability layer for self-supervised medical vision models, requiring no architectural modification, task-specific labels, or retraining. By bridging abstract FM representations and human-interpretable concepts, sparse autoencoders offer a grounded path toward medical AI systems whose predictions can be inspected, communicated, and trusted in clinical practice.

Acknowledgments. Philipp Wesp is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 553239084.

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

References

1. Abdulaal, A., Fry, H., Montaña-Brown, N., Ijishakin, A., Gao, J., Hyland, S., Alexander, D.C., Castro, D.C.: An X-Ray Is Worth 15 Features: Sparse Autoencoders for Interpretable Radiology Report Generation (Oct 2024). <https://doi.org/10.48550/arXiv.2410.03334>
2. Akinci D’Antonoli, T., Berger, L.K., Indrakanti, A.K., Vishwanathan, N., Weiss, J., Jung, M., Berkarda, Z., Rau, A., Reisert, M., Küstner, T., Walter, A., Merkle, E.M., Boll, D.T., Breit, H.C., Nicoli, A.P., Segeroth, M., Cyriac, J., Yang, S., Wasserthal, J.: TotalSegmentator MRI: Robust Sequence-independent Segmentation of Multiple Anatomic Structures in MRI. *Radiology* **314**(2), e241613 (Feb 2025). <https://doi.org/10.1148/radiol.241613>
3. Bills, S.: Language models can explain neurons in language models (Mar 2023)
4. Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N.L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J.E., Hume, T., Carter, S., Henighan, T., Olah, C.: Towards Monosemanticity: Decomposing Language Models With Dictionary Learning (Oct 2023)
5. Bussmann, B., Leask, P., Nanda, N.: BatchTopK Sparse Autoencoders (Dec 2024). <https://doi.org/10.48550/arXiv.2412.06410>
6. Bussmann, B., Nabeshima, N., Karvonen, A., Nanda, N.: Learning Multi-Level Features with Matryoshka Sparse Autoencoders (Mar 2025). <https://doi.org/10.48550/arXiv.2503.17547>
7. 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>
8. Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Wattenberg, M., Olah, C.: Toy Models of Superposition (Sep 2022)
9. Eslami, S., Meinel, C., de Melo, G.: PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain? In: Findings. vol. Findings of the Association for Computational Linguistics: EACL 2023, pp. 1181–1193. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). <https://doi.org/10.18653/v1/2023.findings-eacl.88>
10. Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., Andreas, J.: Natural Language Descriptions of Deep Visual Features (Apr 2022). <https://doi.org/10.48550/arXiv.2201.11114>
11. Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J., Chanin, D., Lau, Y.T., Farrell, E., McDougall, C., Ayonrinde, K., Till, D., Wearden, M., Conmy, A., Marks, S., Nanda, N.: SAEbench: A Comprehensive Benchmark for Sparse Autoencoders in Language Model Interpretability (Jun 2025). <https://doi.org/10.48550/arXiv.2503.09532>
12. Kingma, D.P., Ba, J.L.: Adam: A Method for Stochastic Optimization. In: 3rd International Conference on Learning Representations (ICLR). vol. 3. San Diego, CA, USA (2015)
13. Langlotz, C.P., Allen, B., Erickson, B.J., Kalpathy-Cramer, J., Bigelow, K., Cook, T.S., Flanders, A.E., Lungren, M.P., Mendelson, D.S., Rudie, J.D., Wang, G., Kandarpa, K.: A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology* **291**(3), 781–791 (Jun 2019). <https://doi.org/10.1148/radiol.2019190613>

14. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. *Nature* **616**(7956), 259–265 (Apr 2023). <https://doi.org/10.1038/s41586-023-05881-4>
15. Paschali, M., Chen, Z., Blankemeier, L., Varma, M., Youssef, A., Bluethgen, C., Langlotz, C., Gatidis, S., Chaudhari, A.: Foundation Models in Radiology: What, How, Why, and Why Not. *Radiology* **314**(2), e240597 (Feb 2025). <https://doi.org/10.1148/radiol.240597>
16. Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., Nanda, N.: Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders (Aug 2024). <https://doi.org/10.48550/arXiv.2407.14435>
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S.A., Baby, S.M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.b., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D.F., Kirmizibayrak, C., Pilgrim, R., Golden, D., Yang, L.: MedGemma Technical Report (Jul 2025). <https://doi.org/10.48550/arXiv.2507.05201>
18. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (Aug 2025). <https://doi.org/10.48550/arXiv.2508.10104>
19. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence* **5**(5), e230024 (Sep 2023). <https://doi.org/10.1148/ryai.230024>
20. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (Jan 2025). <https://doi.org/10.48550/arXiv.2303.00915>
21. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moung-Wen, C., Piening, B., Bifulco, C., Wei, M., Poon, H., Wang, S.: BiomedParse: A biomedical foundation model for image parsing of everything everywhere all at once. *Nature Methods* **22**(1), 166–176 (Jan 2025). <https://doi.org/10.1038/s41592-024-02499-w>