SN-AbsTopK: An Unrolled Spatial-Normative Sparse Autoencoder for Medical Foundation Models

Anonymous authors
Paper under double-blind review

Abstract

Sparse autoencoders (SAEs) decompose foundation model representations into monosemantic latents. However, applying independent per-token sparsity across image grids ignores spatial correlations, which causes feature fragmentation in medical imaging. We introduce Spatial-Normative AbsTopK (SN-AbsTopK), an unrolled sparse autoencoder that integrates spatial group support containment and normative reference centering into signed-magnitude dictionary learning. Rather than evaluating tokens independently, SN-AbsTopK operates via a two-stage selection mechanism across spatial token groups. This structure enables effective processing across continuous regions, such as contiguous histology tiles or volumetric MRI slice stacks. In the first stage, the model aggregates pre-activation column energies across all group tokens to rank dictionary atoms and establish a shared group candidate support. In the second stage, each individual token selects its active signed features strictly from within this shared group candidate pool. This guarantees that all token-level active features remain strict subsets of the group candidate set. To isolate disease-specific variations, an exponential moving average reference buffer first centers the input embeddings. This process forces the sparse latent codes to represent residual departures from baseline reference states. Evaluated across histopathology (BCSS with frozen UNI) and radiology (CheXpert and Duke Breast MRI with frozen BiomedCLIP), SN-AbsTopK substantially enhances structural continuity, increasing spatial autocorrelation by 2.9× on BCSS (Moran's 0.244 vs. 0.084) and 3D adjacent-slice support persistence by 60.7% on Duke MRI (Jaccard 0.614 vs. 0.382). Although strict candidate bounds increase reconstruction error and cause feature lock-out (80% dead atoms), raw reconstruction error proves decoupled from clinical decodability. Consequently, SN-AbsTopK yields superior downstream linear probe performance (+0.085 macro-AUROC on CheXpert; +0.033 on Duke MRI).

1 Introduction

Medical foundation models (FMs) have transformed medical imaging by providing general-purpose representations that capture diverse visual patterns across pathology, radiology, and other modalities. Models such as UNI (Chen et al., 2024) and BiomedCLIP (Zhang et al., 2023) map images into high-dimensional embeddings that support classification, retrieval, segmentation, and other downstream tasks. However, these embeddings are dense mixtures of multiple latent factors and are typically consumed as opaque feature vectors, making it difficult to identify meaningful components. For medical imaging, where clinically important findings are often localized to specific anatomical or tissue structures (Chen et al., 2020). An interpretable representation should therefore capture not only which features are present, but also where they occur and how they differ from the dominant structure of the data (Chen et al., 2020; Donteu et al., 2023). Sparse representation learning provides a natural way to obtain such a decomposition.

Sparse autoencoders (SAEs) learn an overcomplete dictionary whose elements can represent different features in the data (Shoga, 2026). The encoder then represents each input using only a small number of these features (Templeton et al., 2026; Gao et al., 2025). Given an input representation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Positioning of SN-AbsTopK relative to existing methods.

Method	Signed	Spatial	Normative	Medical FM
ReLU SAE (Huben et al., 2024)	×	×	×	×
TopK SAE Gao et al. (2025)	×	×	×	×
JumpReLU SAE Rajamanoharan et al. (2024)	×	×	×	×
AbsTopK (Zhu et al., 2026)	✓	×	×	×
SN-AbsTopK (Proposed)	✓	✓	✓	✓

$\boldsymbol{x} \in \mathbb{R}^D$, the underlying sparse-coding problem can be written as

$$\boldsymbol{x} \approx \boldsymbol{D}\boldsymbol{z}, \qquad |\boldsymbol{z}|_0 \leq K, \tag{1}$$

where $\boldsymbol{D}$ is the learned dictionary, $\boldsymbol{z}$ is the sparse code, and K controls the number of active features. This formulation provides a simple and interpretable view of representation learning, in which each input is reconstructed using only a small subset of the learned dictionary elements.

However, standard SAEs do not explicitly model the spatial organization of features or their relationship to a normative reference structure. Recent work has progressively refined sparse support selection, moving from ℓ_1-regularized sparsity toward hard TopK and related activation-based operators (Gao et al., 2025; Rajamanoharan et al., 2024). AbsTopK Zhu et al. (2026) removes the non-negativity constraint by selecting the K features with the largest absolute values from the encoder output. Formally,

$$\boldsymbol{z} = \mathcal{H}_K(\boldsymbol{a}), \tag{2}$$

where $\boldsymbol{a}$ is the encoder output before sparse selection, and $\mathcal{H}_K$ retains the K entries of $\boldsymbol{a}$ with the largest magnitudes while preserving their original signs. This allows a single dictionary feature to represent bidirectional responses along the same learned direction, rather than requiring separate units for positive and negative responses.

Despite these advances, the sparse-coding formulation of existing SAEs does not explicitly account for the spatial structure of medical FM representations Nguyen et al. (2025). The architectural positioning and key differences between existing SAE frameworks and our proposed method are summarized in Table 1. A medical FM typically represents an image as a collection of spatially organized tokens rather than as an unordered set of independent vectors:

$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}, \tag{3}$$

where neighboring tokens correspond to neighboring tissue regions, anatomical structures, or volumetric locations. Standard SAE encoders apply sparse selection independently to each token $\boldsymbol{x}_n$. As a result, neighboring tokens may select different dictionary supports even when they represent the same local structure. Thus, token-wise sparsity does not necessarily produce spatially structured sparsit, a representation may use only a few features per token while still producing fragmented activation patterns across the image Shoga (2026).

A second consideration arises from the structural baseline present in medical imaging modalities. Standard sparse autoencoders reconstruct representations relative to a static bias b, forcing sparse features to encode absolute embedding values Huben et al. (2024).

For medical representations, it is architecturally advantageous to frame sparse coding residually representing activations as departures from a reference baseline state μ updated through an exponential moving average (EMA) (Bricken et al., 2023). This establishes a normative reference prior for residual dictionary learning (Marquand et al., 2019), ensuring that latent features capture relative departures from baseline feature distributions rather than redundant background offsets (Gao et al., 2025).

These observations motivate Spatial-Normative AbsTopK (**SN-AbsTopK**), which extends signed-magnitude AbsTopK with spatial group support constraints and normative reference centering (see Sec. 2). Rather than evaluating token sparsity independently, our module first selects a shared candidate support across neighboring spatial tokens before performing token-wise signed-magnitude selection. This joint support formulation encourages spatial feature identity clustering while preserving local signal variance (Sec. 2.2). We then introduce a normative-reference anchor (Sec. 2.3), which maintains an EMA reference in the FM feature space. This reference serves as the reconstruction baseline, allowing the sparse code to represent deviations from the learned baseline rather than repeatedly encoding common background structure. Finally, we formulate these components using

a restricted absolute Top-k selection mechanism (Sec.2.4), where spatially selected candidate atoms constrain the subsequent token-level sparse selection.

Our main contributions are summarized as follows:

- We reframe sparse dictionary learning in medical FMs from isolated per-token quantization to spatial group support selection. By replacing independent feature selection across tokens with a shared group-level candidate support, our formulation explicitly couples spatially related embeddings across contiguous histology grids and volumetric MRI slice stacks, directly addressing spatial feature fragmentation.
- We introduce a two-stage selection operator that aggregates pre-activation energy within each spatial group to construct a shared candidate support, followed by token-wise signed-magnitude selection, together with an EMA-based normative reference anchor for residual sparse coding.
- We evaluate SN-AbsTopK across histopathology and radiology FM representations using reconstruction, sparsity, spatial-structure, downstream decodability, and controlled component-ablation analyses, characterizing the effects of group support selection and normative centering on learned sparse representations.

2 METHOD

2.1 PROBLEM FORMULATION

Let $\boldsymbol{X} \in \mathbb{R}^{B\times N\times D}$ denote a batch of medical FM representations, where B is the batch size, N is the number of spatial tokens, and D is the embedding dimension. Each token $\boldsymbol{x} \in \mathbb{R}^D$ corresponds to a spatial location in the original image grid. We learn an overcomplete dictionary with $\boldsymbol{W} \in \mathbb{R}^{M\times D}$, where $M \gg D$, a reconstruction bias $b \in \mathbb{R}^D$, and an EMA anchor $\mu \in \mathbb{R}^D$ used by the normative reference 2.3. Each token is first centered against the anchor and bias, then encoded using the dictionary:

$$a = (x - \mu - b)\boldsymbol{W}_{\boldsymbol{e}}^T, \tag{4}$$

producing pre-activations $a \in \mathbb{R}^M$, where as $\boldsymbol{W_e}$ shows encoder. A sparse latent code $\boldsymbol{z} \in \mathbb{R}^M$ is obtained from (a) through the two-stage selection operator $\boldsymbol{\phi}$ (see detailed in Sec. 2.2), using a per-token budget k, a group budget k_g, and $\boldsymbol{Z} \in \mathbb{R}^{B\times N\times M}$. Reconstruction is given by

$$\hat{x} = z\boldsymbol{W_d} + b, \tag{5}$$

where $\boldsymbol{W_d}$ represent decoder, and $b \in \mathbb{R}^D$ is the reconstruction bias. The reconstructed tokens are arranged according to their original spatial locations to form $\hat{\boldsymbol{X}} \in \mathbb{R}^{B\times N\times D}$.

Our objective is to learn a signed sparse code that reconstructs FM representations using a small subset of an overcomplete dictionary. Unlike token-wise sparse coding, SN-AbsTopK incorporates spatial relationships (2.2), and a learned normative reference (2.3) to guide support selection and reconstruction as shown in Figure 1.

2.2 SPATIAL GROUP SPARSITY

Medical images contain strong spatial structure nearby patches describe related tissue or anatomical regions. We use this property to encourage spatially related tokens to activate a similar set of dictionary atoms, while still allowing each token to select its own features. We use a two-stage sparse selection strategy. First, we identify the most important dictionary atoms for each spatial group. Then, each token selects a smaller number of atoms from this shared set.

Let T_g denote the set of tokens belonging to spatial group g, and let $a_n \in \mathbb{R}^M$ be the pre-activation vector of token n, where M is the number of dictionary atoms. We measure the importance of atom j by aggregating its activation magnitude across all tokens in the group. The importance score $s_j^{(g)}$ for atom j across spatial group g is defined as:

$$s_j^{(g)} = \sqrt{\sum_{n\in\mathcal{T}_g} a_{n,j}^2} \tag{6}$$

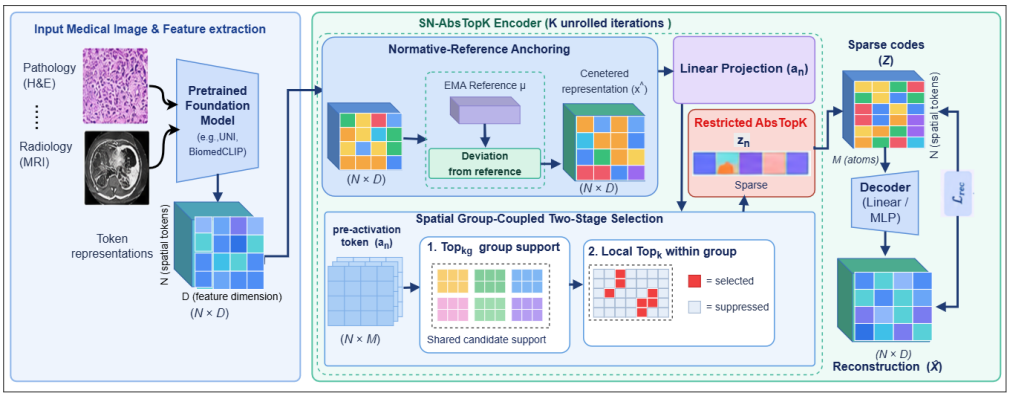


Figure 1: Overview of the proposed Spatial-Normative AbsTopK (SN-AbsTopK) framework.

$$]S_g = \mathrm{TopK}_{k_g}\left(s^{(g)}\right) \subseteq \{1, \ldots, M\}. \tag{7}$$

The resulting set S_g is the shared group candidate support for group g. Thus, all tokens in the same spatial group have access to the same k_g dictionary atoms. Each token then performs a second, more restrictive selection (see Figure 1). Given token $n \in \mathcal{T}_g$, we keep only the atoms in S_g and apply a signed AbsTopK operation with a smaller budget $k \leq k_g$:

$$z_n = \mathrm{AbsTopK}_k\left(a_n \odot \mathbf{1}_{S_{g(n)}}\right), \tag{8}$$

where $\mathbf{1}_{S_{g(n)}}$ is the indicator mask of the shared group candidate support $S_{g(n)}$, and $\odot$ denotes the element-wise product. This enforces block-structured sparsity, restricting each group to a fixed budget of k_g dictionary atoms from which constituent tokens select k features ($k \leq k_g$). In the absence of group structure, tokens form singleton groups ($k_g = k$), and equation 7–equation 8 reduce exactly to per-token $\mathrm{AbsTopK}_k$. The complete procedure for selecting dictionary elements across spatial groups and individual tokens is summarized in Algorithm 1. To further encourage spatially coherent representations across neighboring anatomical structures, we apply a spatial total-variation (TV) regularizer (formulation provided in Appendix 7).

2.3 Normative-Reference Anchoring

In medical imaging, most of an embedding's energy encodes the dominant normal-tissue mode; re-encoding that mode wastes the limited sparse budget. We track a reference μ by an exponential moving average (EMA),

$$\mu \leftarrow m\,\mu + (1-m)\,\overline{x}_{\mathrm{ref}}, \qquad m = 0.99, \tag{9}$$

where x_{ref} is the mean of the preprocessed training batch. The reference is initialized from the first batch mean and updated without gradient propagation during training. At evaluation time, μ is kept fixed. Before encoding, each input is preprocessed using training-set statistics, followed by per-token centering and scaling to unit standard deviation. The encoder and decoder operate relative to the reference:

$$a = (x - \mu - b_d)\,W_e, \qquad \hat{x} = z\,W_d + \mu + b_d. \tag{10}$$

where the encoder has no bias and z is the signed sparse code obtained after group-shared support selection and per-token topk selection (detail provided in Sec. 2.2). Thus, the sparse code represents the residual relative to the reference μ. We also implement a normative-reference anchor that centres the encoder on an EMA estimate of the batch mean, so the sparse code encodes residuals rather than the full input. Atoms that are strongly activated across multiple tokens receive larger scores, and the k_g atoms with the largest scores are selected (see Appendix 7)

Algorithm 1 Two-stage spatial group selection

Require: Pre-activations $\{a_n\}_{n=1}^N$, budgets k_g and $k \le k_g$
1: **for** each spatial group g **do**
2: $\quad s_j^{(g)} \leftarrow \sqrt{\sum_{n \in T_g} a_{n,j}^2}$
3: $\quad S_g \leftarrow \mathrm{TopK}_{k_g}\left(s^{(g)}\right)$
4: **end for**
5: **for** each token n **do**
6: $\quad z_n \leftarrow \mathrm{AbsTopK}_k\left(a_n \odot \mathbf{1}_{S_{g(n)}}\right)$
7: **end for**
8: **return** $\{z_n\}_{n=1}^N$

2.4 Restricted Absolute Top-(K) Selection

Following normative reference anchoring (Sec. 2.3), each token is first centred relative to the reference state:

$$\hat{x}_n = x_n - \mu_{\mathrm{ref}} \qquad a_n = W_{\mathrm{e}}\hat{x}_n + b_{\mathrm{e}}.$$

The spatial group stage then provides a candidate support S_n for each token. Instead of competing over the full dictionary, each token performs sparse selection only within this support. Restricted AbsTopK retains the k largest-magnitude activations in S_n while preserving their original signs:

$$z_{n,m} = \begin{cases} a_{n,m}, & \text{if } m \in \mathrm{TopK}\left(\{|a_{n,j}| \mid j \in \mathcal{S}_n\}\right), \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Thus, spatial group selection determines the candidate atoms, while local Top-k selection determines each token's active subset. The complete restricted-selection operator, support construction, and gradient behavior are provided in Appendix 7, Sec. 7.

2.5 Training Objective

The complete SN-AbsTopK training objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{rec}} \;+\; \mathcal{L}_{\mathrm{aux}} \;+\; \lambda_{\mathrm{tv}}\,\mathcal{L}_{\mathrm{tv}} \;+\; \lambda_{\mathrm{norm}}\,\mathcal{L}_{\mathrm{norm}}, \tag{12}$$

where $\mathcal{L}\mathrm{rec}$ measures reconstruction error and $\mathcal{L}\mathrm{aux}$ prevents dictionary stagnation. $\mathcal{L}\mathrm{tv}$ enforces spatial consistency across neighboring tokens, while $\mathcal{L}\mathrm{norm}$ regularizes normative codes (Sec. 2.2), (Sec. 2.3). No ℓ_1 penalty is used because hard top-k selection directly constrains active code cardinality. This avoids artificial magnitude shrinkage while maintaining the required sparsity level through the selection operators. $\mathcal{L}\mathrm{aux}$ fits reconstruction residuals using broad and fine AbsTopK selections over currently dead atoms. Each auxiliary term is normalized by residual energy, making the dead-feature loss scale-free. An atom is declared dead when its EMA selection frequency falls below threshold or remains unselected for n_{dead} batches. Persistently dead atoms are revived by resampling decoder rows from high-residual input directions. Their corresponding encoder columns are reseeded in parallel, and optimizer moments are cleared. The two-stage spatial mask is binary (see Sec. 2.2), so gradients through $z = a \odot m$ equal one for retained entries. The restricted selection operator changes atom competition during the forward pass but does not modify the loss. We optimize equation 12 with Adam (Kingma & Ba, 2015) and project decoder rows to unit norm after every step.

3 Experimental Setup

3.1 Datasets and Implementation Details

We evaluate SN-AbsTopK on three medical imaging datasets spanning histopathology and radiology. The histopathology experiments use BCSS (Amgad et al., 2019), represented with the frozen UNI encoder (Chen et al., 2024), while the radiology experiments use CheXpert (Irvin et al., 2019) and Duke Breast MRI (Saha et al., 2018), represented with the frozen BiomedCLIP encoder (Zhang et al., 2023). UNI and BiomedCLIP produce 1024- and 512- dimensional patch embeddings, respectively. We use these frozen representations as inputs to train SN-AbsTopK with a dictionary

expansion factor of $\kappa = 16$. Training samples are constructed from spatial groups, with large groups subsampled for computational efficiency. Detailed dataset preprocessing, feature extraction, grouping strategy, optimization settings, and baseline configurations are provided in Appendix 7 $\rightarrow$ Sec. 4,Table 4.

3.2 Evaluation Metrics

Following Gao et al. (2025), we evaluate reconstruction using normalized mean squared error (nMSE) and loss recovered (LR). On held-out test sets with frozen train-split standardization, we report nMSE, explained variance (EV), per-token (ℓ_0) sparsity, and dead-atom fraction. We additionally evaluate modality-specific structure using Moran's (I) for spatial coherence on BCSS, AUROC-5 for clinical information retention on CheXpert, and adjacent-slice Jaccard similarity for support consistency across Duke MRI slices. Finally, we assess downstream utility using linear probing on held-out medical tasks, reporting macro-F1 to measure how well the sparse codes preserve task-relevant information. All results are reported as mean ($\pm$) standard deviation over two random seeds using the final training checkpoint.

4 Results

Table 2 evaluates hard-(k) sparse autoencoders with ($\ell_0 = 32$) across BCSS, CheXpert, and Duke Breast MRI using 4($\times$) dictionary expansion and 20k training steps. SN-AbsTopK achieves the highest structural-coherence score on all three datasets while substantially reducing the number of alive dictionary features, with the strongest reconstruction trade-offs occurring on CheXpert and Duke Breast MRI.

On BCSS, SN-AbsTopK achieves Table 2 (A) a Moran's (I) of 0.244 $\pm$ 0.00, compared with 0.110 for Mat. BatchTopK and 0.108 for TopK + group lasso. This corresponds to increases of 121.8 % and 125.9%, respectively. SN-AbsTopK also reduces the number of alive features from 3,158 for TopK + group lasso and 4,008 for Mat. BatchTopK to 689, corresponding to reductions of 78.2% and 82.8%, respectively. Its reconstruction nMSE is 0.560 $\pm$ 0.03, compared with 0.450 $\pm$ 0.01 for Mat. BatchTopK and 0.540 $\pm$ 0.03 for TopK + group lasso, while EV is 0.436 versus 0.547 and 0.457, respectively. Thus, SN-AbsTopK has the highest Moran's (I) and lowest alive count in the BCSS comparison, with lower reconstruction scores than the reconstruction-focused baselines.

On CheXpert, SN-AbsTopK obtains Table 2 (B) an nMSE of 0.073 $\pm$ 0.00 and EV of 0.930, matching Mat. BatchTopK on EV and remaining close to its nMSE of 0.070 $\pm$ 0.01. SN-AbsTopK achieves the highest AUROC-5 at 0.735 $\pm$ 0.02, compared with 0.734 $\pm$ 0.001 for AbsTopK and 0.730 $\pm$ 0.02 for Mat. BatchTopK. At the same time, the number of alive features is reduced to 388, compared with 1,322 for AbsTopK and 1,589 for Mat. BatchTopK, corresponding to reductions of 70.7% and 75.6%, respectively. Compared with TopK, SN-AbsTopK reduces the alive count from 2,048 to 388 while improving AUROC-5 from 0.723 to 0.735.

On Duke Breast MRI, SN-AbsTopK achieves Table 2 (C) the highest structural metric, with an Adj-Jaccard of 0.614 $\pm$ 0.01, compared with 0.531 $\pm$ 0.02 for Mat. BatchTopK and 0.405 $\pm$ 0.01 for TopK + group lasso. These correspond to relative increases of 15.6% and 51.6%, respectively. SN-AbsTopK also reduces the alive feature count to 387, compared with 1,730 for Mat. BatchTopK and 1,708 for TopK + group lasso, representing reductions of 77.6% and 77.3%. Its reconstruction nMSE is 0.121 $\pm$ 0.01 with EV of 0.883, compared with 0.116 $\pm$ 0.03 and 0.885 for Mat. BatchTopK. SN-AbsTopK therefore attains the highest Adj-Jaccard and substantially fewer alive features while remaining close to the best reconstruction values.

Across the three feature sets, SN-AbsTopK consistently produces the fewest alive dictionary features, 689 on BCSS, 388 on CheXpert, and 387 on Duke Breast MRI. It also achieves the highest reported structural metric on each benchmark: Moran's (I) = 0.244 on BCSS, AUROC-5 =0.735 on CheXpert, and Adj-Jaccard = 0.614 on Duke Breast MRI. The reconstruction metrics show a different ordering: Mat. BatchTopK obtains the lowest nMSE on BCSS 0.450 and Duke Breast MRI 0.116, while SN-AbsTopK obtains the lowest nMSE among the compared methods on neither of these two datasets. On CheXpert, SN-AbsTopK remains close to the minimum nMSE 0.073 versus 0.070 while achieving the highest AUROC-5.

Table 2: **Reconstruction quality and structural coherence of hard-k SAEs ($\ell_0 = 32$).** Evaluated on three medical feature sets ($4\times$ dictionary expansion, 20k steps). Results show mean $\pm$ std over 2 seeds on a held-out test set.

(A) BCSS (UNI, 1024-d, histopathology)

Method	Dict. ($\times 4$)	nMSE↓	EV↑	Alive	Moran's I↑
TopK (Gao et al., 2025)	4096	0.503 ± 0.00	**0.495**	4096	0.080 ± 0.00
BatchTopK (Bussmann et al., 2024)	4096	0.518 ± 0.01	0.480	4095	0.078 ± 0.00
AbsTopK (Zhu et al., 2026)	4096	0.516 ± 0.01	0.480	4096	0.084 ± 0.01
Mat. BatchTopK (Bussmann et al., 2024)	4096	$\mathbf{0.450 \pm 0.01}$	0.547	4008	0.110 ± 0.00
TopK + group lasso	4096	0.540 ± 0.03	0.457	3158	0.108 ± 0.00
SN-AbsTopK (Ours)	4096	0.560 ± 0.03	0.436	689	$\mathbf{0.244 \pm 0.00}$

(B) CHEXPERT (BiomedCLIP, 512-d, chest X-ray)

Method	Dict.($\times 4$)	nMSE↓	EV↑	Alive	AUROC-5↑
TopK (Gao et al., 2025)	2048	0.128 ± 0.00	0.878	2048	0.723 ± 0.00
BatchTopK (Bussmann et al., 2024)	2048	0.139 ± 0.00	0.867	2048	0.722 ± 0.00
AbsTopK (Zhu et al., 2026)	2048	0.080 ± 0.00	0.921	1322	0.734 ± 0.001
Mat. BatchTopK (Bussmann et al., 2024)	2048	0.074 ± 0.01	0.928	1589	0.730 ± 0.02
TopK + group lasso	2048	0.091 ± 0.00	0.913	1702	0.727 ± 0.01
SN-AbsTopK (Ours)	2048	$\mathbf{0.073 \pm 0.00}$	**0.930**	388	$\mathbf{0.735 \pm 0.02}$

(C) DUKE BREAST MRI (BiomedCLIP, 512-d, breast MRI)

Method	Dict. ($\times 4$)	nMSE↓	EV↑	Alive	Adj-Jaccard↑
TopK (Gao et al., 2025)	2048	0.225 ± 0.01	0.783	2047	0.356 ± 0.03
BatchTopK (Bussmann et al., 2024)	2048	0.236 ± 0.01	0.770	2047	0.358 ± 0.01
AbsTopK (Zhu et al., 2026)	2048	0.171 ± 0.00	0.832	1936	0.382 ± 0.00
Mat. BatchTopK (Bussmann et al., 2024)	2048	0.116 ± 0.03	0.885	1730	0.531 ± 0.02
TopK + group lasso	2048	0.209 ± 0.00	0.798	1708	0.405 ± 0.01
SN-AbsTopK (Ours)	2048	$\mathbf{0.121 \pm 0.01}$	**0.883**	387	$\mathbf{0.614 \pm 0.01}$

Downstream Classification: Tables **??** and **??** evaluate whether the learned representations transfer to clinical prediction tasks. On CheXpert (Table **??**), SN-AbsTopK improves AUROC over AbsTopK for all five clinical findings. The largest gains are observed for Edema (0.669 to 0.789, +0.120), Cardiomegaly (0.657 to 0.774, +0.117), and Pleural Effusion (0.689 to 0.803, +0.114). The macro AUROC increases from 0.632 $\pm$ 0.008 to 0.717 $\pm$ 0.010, an absolute gain of 0.085. On Duke Breast MRI (Table **??**), SN-AbsTopK increases macro AUROC from 0.512 $\pm$ 0.035 to 0.545 $\pm$ 0.025. Improvements are observed for PR (0.346 to 0.498), HER2 (0.630 to 0.672), and Subtype (0.495 to 0.506). ER and Grade decrease from 0.559 to 0.528 and from 0.529 to 0.520, respectively. Overall, SN-AbsTopK yields higher macro AUROC on both downstream benchmarks, with gains of 0.085 on CheXpert and 0.033 on Duke Breast MRI.

Overall, SN-AbsTopK exhibits a favorable trade-off between sparsity, structural coherence, reconstruction, and downstream utility. It activates fewer dictionary features while achieving the strongest structural metrics across all three benchmarks. This structured sparsity comes with higher reconstruction error than Mat. BatchTopK on BCSS and Duke Breast MRI. Despite this trade-off, SN-AbsTopK improves macro-AUROC by $+0.085$ on CheXpert and $+0.033$ on Duke Breast MRI. The qualitative results further show how these structured features respond to specific image patterns across modalities. In Figure 2 BCSS latent activations are predominantly positive, reaching $z \approx +4$ to $+6$ in tumor-enriched and inflammatory regions. In contrast, hypocellular stromal and fatty regions show baseline or negative responses, with $z \leq 0$ in the corresponding regions. This separation in activation magnitude and direction indicates that the latent responds more strongly to cell-dense tissue patterns than to surrounding low-cellularity regions.

Figure 3) shows a clear numerical separation in CheXpert, with pathology-selective latents predominantly activating negatively for correctly identified positive cases. These cases reach $z \leq -15$, whereas missed cases show positive responses from $+16.70$ to $+19.61$, indicating opposite latent responses for correctly detected and missed cases.

For Duke Breast MRI, Figure 7 visualizes latent z_{27} across consecutive 3D axial slices, with the peak response at slice 147 ($z = +2.00$). The top panels track the latent response across the volume, while the bottom grid shows its progression from dark baseline regions ($z \leq 0$) to intense yellow

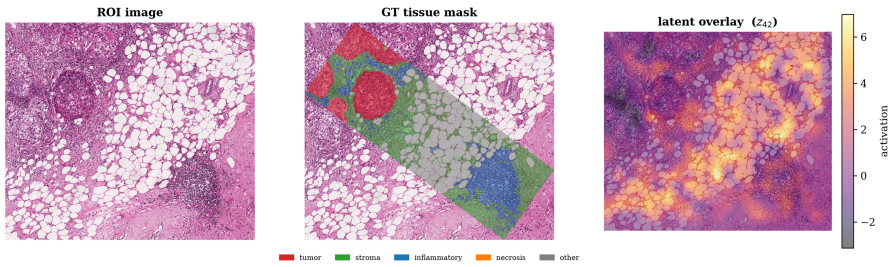


Figure 2: Qualitative visualization of BCSS ROI, showing spatially coherent activation in cell-dense and tumor-enriched regions.

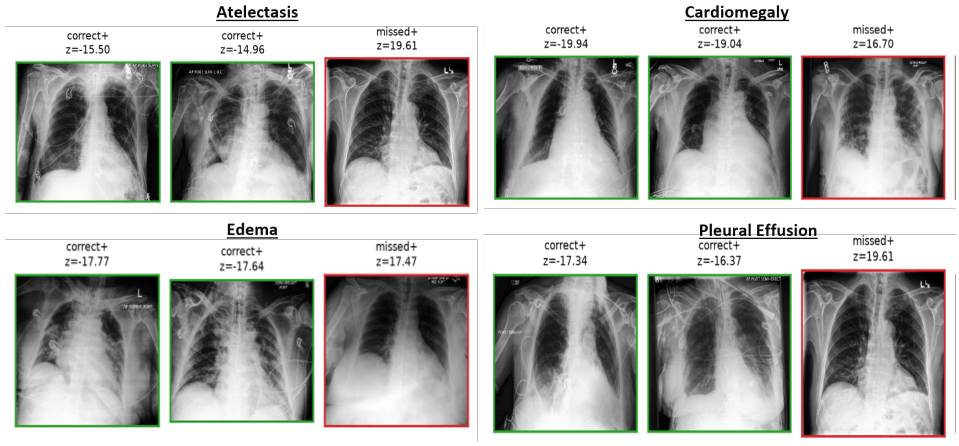


Figure 3: Examples of model feature activations on CheXpert chest X-rays. Green outlines show strong correct detections (strongly negative z-values), while red outlines indicate missed cases.

activations ($z > 1.0$). This numerical transition shows how the latent response changes from inactive regions to localized high-response regions across the MRI volume.

4.1 Ablation

Table 3: Component ablation across BCSS, CheXpert, and Duke-MRI.

	BCSS (UNI)			CheXpert (BiomedCLIP)			Duke MRI (BiomedCLIP)		
Variant	nMSE↓	Dead↓	Moran↑	nMSE↓	Dead↓	AUC-5↑	nMSE↓	Dead↓	Jacc.↑
Full	0.497	0.83	0.245	0.070	0.82	0.724	0.122	0.81	0.613
−group	0.509	0.00	0.078	0.072	0.36	0.730	0.152	**0.06**	0.344
−TV	0.564	0.82	0.236	–	–	–	–	–	–
−norm	0.559	0.83	0.244	0.073	0.82	0.724	0.119	0.81	0.611
−anchor	0.556	0.84	0.245	0.073	0.83	0.727	0.122	0.80	0.619

Table 3 evaluates the contribution of each component using reconstruction, sparsity, and structure-aware metrics across three modalities. Removing group support reduces BCSS Moran's I from 0.245 to 0.078, showing its strong contribution to spatially coordinated latent activation. It also reduces Duke-MRI adjacent-slice Jaccard from 0.613 to 0.344, indicating weaker active-set consistency between neighbouring MRI slices. For CheXpert, removing group support changes AUROC-5 only slightly from 0.724 to 0.730, while nMSE increases from 0.070 to 0.072. Removing TV on BCSS increases nMSE from 0.497 to 0.564 and slightly decreases Moran's I from 0.245 to 0.236. This suggests that TV provides additional local smoothness, while group support primarily controls shared feature selection across spatially related tokens. Removing the normative prior increases BCSS nMSE from 0.497 to 0.559, while Moran's I remains nearly unchanged at 0.244. The corresponding CheXpert nMSE increases from 0.070 to 0.073, with AUROC-5 unchanged at 0.724, showing limited effect on task discrimination. Removing anchor centering produces similar changes, with BCSS nMSE increasing to 0.556 while Moran's I remains unchanged at 0.245.

Additional ablations examining support reuse, spatial coherence, and latent statistics are provided in (Appendix 7, → Table 6, Table 7, Table 8).

5 Discussion and Limitations

The central mechanism of SN-AbsTopK is support-level coupling rather than stronger token-level sparsification: Stage 1 aggregates column activation energy across spatially neighboring tokens to construct S_g, while Stage 2 performs exact signed-magnitude ℓ_0 selection within this shared candidate set, yielding $\mathrm{supp}(z^{(n)}) \subseteq S_g$. This constraint converts independently selected token supports into spatially coupled representations, explaining the large coherence gains: removing group sparsity reduces BCSS Moran's I by 68.2% and Duke MRI slice Jaccard by 43.9%, while the full group-constrained setting reaches $I = 0.244$ and Jaccard $= 0.614$. The same mechanism exposes a structural capacity–utilization trade-off: tight $k_f \approx 0,\ k_g = 64$ support bounds produce strong sharing but leave ~ 80–83% of atoms inactive, whereas free selection ($k_f = 24$) or frequency annealing improves utilization and reconstruction at the expense of coherence. Spatial TV provides complementary local code smoothness and reduces reconstruction error, while the EMA normative anchor contributes little on zero-mean standardized features because $\mu \to 0$, indicating that its effect is conditional on the representation geometry rather than intrinsic to the sparse operator. Importantly, the reconstruction penalty on heterogeneous BCSS tiles ($\mathrm{nMSE} = 0.560$ vs. 0.450 for Matrix BatchTopK) does not predict clinical utility: SN-AbsTopK improves CheXpert macro-AUROC by $+0.085$ to 0.717, including gains of $+0.120$, $+0.117$, and $+0.114$ for Edema, Cardiomegaly, and Pleural Effusion, respectively, with a further $+0.033$ gain on Duke MRI.

Limitation: SN-AbsTopK depends on meaningful contiguous groups, with weakly correlated multi-view groups yielding limited coherence gains on CheXpert (view Jaccard ≈ 0.33). Tight shared supports also cause substantial dictionary lock-out ($\sim$ 80–83% inactive atoms), while free selection and frequency annealing trade utilization and reconstruction fidelity against spatial coherence. The EMA anchor represents a dataset centroid rather than clinically curated normality and remains largely inert ($< 2\%$ metric change) under standardized features. Current evaluation covers regular 2D/3D grids, linear probes, spatial autocorrelation, and slice continuity, leaving non-grid topologies and clinician-facing interpretability untested. Future work should explore adaptive group construction, improved selection schedules, clinically grounded references, and extensions to graph-based structures and broader representation evaluations.

6 Conclusion

SN-AbsTopK introduces spatially constrained sparse coding for medical foundation-model representations across histopathology and radiology. Across three datasets, it consistently improves structural coherence, reaching Moran's $I = 0.244$ on BCSS and Adj-Jaccard $= 0.614$ on Duke MRI. On CheXpert, SN-AbsTopK achieves the lowest nMSE (0.073) and highest AUROC-5 (0.735) among evaluated methods. This structural improvement comes with substantial feature sparsity, retaining only 689, 388, and 387 active atoms across BCSS, CheXpert, and Duke MRI. Although reconstruction fidelity decreases on BCSS, the results show that stronger spatial organization can coexist with competitive downstream decodability. Overall, SN-AbsTopK provides a compact framework for learning spatially coherent sparse representations from medical foundation-model features.

Reproducibility Statement

We provide the complete SN-AbsTopK formulation, implementation details, training settings, ablations, and evaluation protocols. Experiments use frozen UNI and BiomedCLIP representations on BCSS, CheXpert, and Duke Breast MRI; detailed implementation and proof details are provided in Appendix 7).

AI USE STATEMENT

We used generative AI tools for language editing, grammar correction, and improving manuscript clarity and readability. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A. Atteya, Mai A. T. Elsebaie, Lamia S. Abo Elnasr, Rokia A. Sakr, Hazem S. E. Salem, Ahmed F. Ismail, Anas M. Saad, et al. Structured crowdsourcing enables convolutional segmentation of histology images. In *Bioinformatics*, 2019. BCSS breast-cancer histopathology dataset.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024. [VERIFY] exact arXiv ID/author order before submission.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. UNI, Mahmood Lab.

Xu Chen, Chunfeng Lian, Li Wang, Hannah Deng, Tianshu Kuang, Steve Fung, Jaime Gateno, Pew-Thian Yap, James J Xia, and Dinggang Shen. Anatomy-regularized representation learning for cross-modality medical image segmentation. *IEEE transactions on medical imaging*, 40(1): 274–285, 2020.

Kerol R Djoumessi Donteu, Indu Ilanchezian, Laura Kühlewein, Hanna Faber, Christian F Baumgartner, Bubacarr Bah, Philipp Berens, and Lisa M Koch. Sparse activations for interpretable disease grading. In *Medical Imaging with Deep Learning*, 2023.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Andre F Marquand, Seyed Mostafa Kia, Mariam Zabihi, Thomas Wolfers, Jan K Buitelaar, and Christian F Beckmann. Conceptualizing mental disorders as deviations from normative functioning. *Molecular psychiatry*, 24(10):1415–1424, 2019.

Duy-Kien Nguyen, Mahmoud Assran, Unnat Jain, Martin R Oswald, Cees G Snoek, and Xinlei Chen. An image is worth more than 16x16 patches: Exploring transformers on individual pixels. In *International Conference on Learning Representations*, volume 2025, pp. 59753–59773, 2025.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024. DeepMind; JumpReLU SAE baseline.

Ashirbani Saha, Michael R. Harowicz, Lars J. Grimm, Che E. Kim, Sujata V. Ghate, Ruth Walsh, and Maciej A. Mazurowski. A machine learning approach to radiogenomics of breast cancer: A study of 922 subjects and 529 dce-mri features. *British Journal of Cancer*, 2018. Duke-Breast-Cancer-MRI dataset (TCIA). [VERIFY] year – project brief cites 2021; the underlying radiogenomics paper is 2018 and the TCIA collection release may be the correct 2021-labelled reference – confirm which is intended before submission.

Waleed R Shoga. Sparse autoencoders for interpreting vision transformers: A systematic review. 2026.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Anonymous Zhu et al. Abstopk: Signed-magnitude hard-sparse selection for sparse autoencoders. *International Conference on Learning Representations (ICLR)*, 2026. [VERIFY] title/author list as given in the project brief ("Zhu et al., ICLR 2026") – confirm against the actual accepted paper before submission.

7 APPENDIX

Let $x \in \mathbb{R}^D$ denote a preprocessed token embedding, $W_e \in \mathbb{R}^{D\times M}$ the encoder, $W_d \in \mathbb{R}^{M\times D}$ the decoder, $b_d \in \mathbb{R}^D$ the decoder bias, and $\mu \in \mathbb{R}^D$ an EMA reference buffer (Sec. 2.1). For a spatial group g containing $|g|$ tokens, k_g denotes the group candidate budget, $k \le k_g$ the per-token budget (Sec. 2.2). The forward pass is defined as:

$$a = (x - \mu - b_d)W_e, \quad z = \Phi(a; g, k_g, k, k_f), \quad \hat{x} = zW_d + \mu + b_d, \tag{13}$$

where Φ executes two-stage selection without encoder bias (Sec. 2.1).

A.1 Lemma 1: AbsTopK ℓ_0 Projection Optimality 1

For $u \in \mathbb{R}^M$ and sparsity budget k, the optimization problem

$$\min_z \frac{1}{2}\|u - z\|_2^2 \quad \text{s.t.} \quad \|z\|_0 \le k \tag{14}$$

is solved exactly by $z_i = u_i$; for $i \in \text{TopK}_k(|u|)$ and $z_i = 0$. Thus, signed-magnitude AbsTopK constitutes the exact Euclidean ℓ_0 projection onto k-sparse vectors.

Proof. For any fixed support §⊂ {1, . . . , M} with $|S| \le k$, the objective decouples coordinate-wise, yielding optimal coefficients $z_i = u_i$ for i ∈S and $z_i = 0$ for $i \notin S$. The residual objective value equals $\frac{1}{2}\sum_{i\notin S} u_i^2 = \frac{1}{2}\|u\|_2^2 - \frac{1}{2}\sum_{i\in S} u_i^2$. Minimizing this objective is strictly equivalent to maximizing the retained squared energy $\sum_{i\in S} u_i^2$, which is uniquely maximized by selecting the k largest elements of $|u|$. □

A.2 Lemma 2: Stage-1 Group Energy Ranking 1

Let $U \in \mathbb{R}^{|g|\times M}$ denote pre-activations for tokens $n \in g$. Stage 1 aggregates column activation energies $s_{g,p} = \sqrt{\sum_{n\in g} u_{n,p}^2}$ and selects group candidate support $S_g = \text{TopK}_{k_g}(s^{(g)})$. S_g uniquely solves:

$$\min_{S:|S|\le k_g} \frac{1}{2}\sum_{p\notin S}\sum_{n\in g} u_{n,p}^2 \tag{15}$$

Proof. The total pre-activation group energy $C_g = \sum_{p=1}^{M}\sum_{n\in g} u_{n,p}^2$ is constant w.r.t. S. Minimizing unselected energy reduces to maximizing total retained column energy $\sum_{p\in S} s_{g,p}^2$ subject to $|S| \le k_g$. Because $f(t) = \sqrt{t}$ is strictly monotonic on $[0, \infty)$, ranking columns by $s_{g,p}$ preserves the exact order of $s_{g,p}^2$, proving optimality. □

A.3 Corollary 1: Group Support Containment (supp($z^{(n)}$) ⊆ S_g) 1

supp($z^{(n)}$) ⊆ S_g for all $n \in g$. All tokens within spatial group g draw active latent dictionary atoms strictly from the shared support S_g.

Proof. Stage 2 multiplies pre-activations by the indicator mask 1_{S_g} prior to AbsTopK selection, zeroing all out-of-support entries and guaranteeing supp($z^{(n)}$) ⊆ S_g . □

A.4 Proposition 4: Normative Residual Centering 1

Assume $x = \mu$ and $b_d = 0$. Then $a = (x - \mu - b_d)W_e = 0$, yielding $z = 0$ and $\hat{x} = \mu$. The sparse code z encodes purely residual departures from reference state μ.

A.5 Unrolled Optimization and Feature Lock-out Dynamics 1

Each optimization step executes a single unrolled forward pass (encode → Φ → decode → Adam) [5, 6]. Because hard Stage-1 selection (Lemma 7) discards atoms outside S_g, non-selected encoder columns receive zero gradient updates ($\nabla_{W_e}\mathcal{L} = 0$). Under tight group bounds ($k_f = 0, k_g = 64$), dominant features repeatedly capture S_g, systematically locking out lower-frequency atoms and driving dictionary collapse (80–83% dead atoms).

B Implementation Setup & Dataset Specifications **2**

B.1 Foundation Model Feature Extraction & Standardization 2

Features from frozen foundation-model encoders are standardized per dimension using training-only statistics to zero mean and unit variance. The same statistics are applied to validation and test sets, with strict patient/ROI separation and no train–test overlap.

- **BCSS (Histopathology)**: Extracted from HF `MedOtter/BCSS` TCGA ROIs using frozen **UNI** ($D = 1024$). Non-overlapping 224×224 patches with $< 5\%$ tissue content are discarded, yielding $48,642$ total embeddings. ROI coordinates $(y_{\text{top}}, x_{\text{left}})$ define spatial groups, producing 121 ROI groups ($19/241/1{,}552$ patches/group, min/median/max). Partitioned into $37,921/10,721$ train/test rows, yielding $36,025/1,896/10,721$ SAE train/validation/test rows (see Table 4).
- **CheXpert (Radiology)**: Extracted from CheXpert-v1.0-small frontal view radiographs using frozen **BiomedCLIP** ($D = 512$), yielding $41,066$ embeddings. Grouped by `patient__study__view`, forming $\approx$ 9,765 patient groups ($1/2/74$ images/group). An $80/20$ patient-level split yields $32,957/8,109$ train/test rows, partitioned into $31,309/1,648/8,109$ SAE train/validation/test rows (see Table 4.
- **Duke Breast MRI (Radiology)**: Extracted from 205 usable patient volumes (one post-contrast dynamic phase; central 60% slices retained) using frozen **BiomedCLIP** ($D = 512$), yielding $16,324$ embeddings. Grouped by patient ID across 164 volume groups ($59/76/80$ slices/group). Partitioned into $13,056/3,268$ train/test rows, yielding $12,403/653/3,268$ SAE train/validation/test rows (Table 4)

Table 4: summarizes the exact dataset sizes, split composition, grouping topology, and group-size statistics.

Dataset	Total Rows	SAE Train / Val / Test	# Groups	Group Size (min/med/max)
BCSS	$48,642$	$36,025$ / $1,896$ / $10,721$	121	19 / 241 / 1,552
CheXpert	$41,066$	$31,309$ / $1,648$ / $8,109$	$\approx$ 9,765	1 / 2 / 74
Duke MRI	$16,324$	$12,403$ / 653 / $3,268$	164	59 / 76 / 80

B.2 Hyperparameter Reference and Optimization Settings 2

All SN-AbsTopK models are trained using a dictionary expansion factor of $\kappa = 16$ ($M = \kappa D$, yielding $M = 16{,}384$ for UNI and $M = 8{,}192$ for BiomedCLIP). The per-token local sparsity budget k_{local} is set to 32 for UNI representations (BCSS) and 24 for BiomedCLIP representations (CheXpert and Duke Breast MRI). The group candidate support budget is configured as $k_{\text{group}} = 2k_{\text{local}}$ (64 for UNI, 48 for BiomedCLIP) using ℓ_2 column-energy group aggregation.

Table 5: Consolidated Hyperparameter Configuration and Optimization Reference.

Hyperparameter	Value	Hyperparameter	Value
Dictionary Expansion (κ)	16 ($M = 16 \times D$)	Spatial TV Weight (λ_{tv})	0.03
Local Sparsity Budget (k_{local})	32 (UNI) / 24 (BiomedCLIP)	Normative Loss Weight (λ_{norm})	0.01
Group Support Budget (k_{group})	$2 \times k_{\text{local}}$ (64 / 48)	EMA Anchor Momentum (m)	0.99
Group Aggregation Operator	ℓ_2 Column Activation Energy	TV Adjacency Graph (k, σ)	$k = 4, \sigma = 2.0$
Optimizer / Momentum	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)	Free Selection Budget (k_f)	0 (Default) / 24 (Hybrid)
Learning Rate / Precision	3×10^{-4} / float32	Batch Size / Groups	2048–4096 tokens / 16–24
Gradient Norm Clipping	1.0	Training Steps	15k–20k

Loss regularizers are weighted with $\lambda_{\text{tv}} = 0.03$ for spatial total-variation and $\lambda_{\text{norm}} = 0.01$ for normative code centering. The spatial TV adjacency graph uses $k = 4$ nearest neighbors with Gaussian distance scaling variance $\sigma = 2.0$. The exponential moving average (EMA) reference buffer updates with momentum $m = 0.99$.

Optimization uses Adam (Kingma & Ba, 2015) ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay 0.0) with a constant learning rate of 3×10^{-4}, gradient norm clipping threshold of 1.0, and float32 numerical precision. Batches comprise 16–24 spatial groups (2048–4096 total tokens), with large spatial groups subsampled to at most 1024 tokens per step. Models are trained for 15k–20k steps, with decoder columns normalized to unit ℓ_2 norm after every parameter update (see Table **??**)

C Component Ablations and Utilization Diagnostics **3**

C.1 Full Component Ablation Suite 3

Table 6: BCSS comparison using Moran's I as the coherence metric. Best values are bolded.

Setting	Group	nMSE↓	EV↑	Moran's I↑
AbsTopK	no	0.516	0.482	0.084
+EMA	no	0.511	0.486	0.078
+TV	no	0.499	0.498	0.083
+Group ($k_f = 24$)	partial	0.528	0.468	0.101
SN-AbsTopK (full)	yes	0.559	0.436	0.244

Table 7: CheXpert comparison using AUROC as coherence metrics. Best values are bolded.

Setting	Group	nMSE↓	EV↑	AUROC$_f$↑
AbsTopK	no	0.080	0.921	0.632
+EMA	no	0.069	0.933	0.594
+TV	no	0.070	0.934	0.599
+Group ($k_f = 24$)	partial	0.069	0.934	0.720
SN-AbsTopK (full)	yes	0.072	0.936	0.717

Table 8: Duke MRI comparison using Jaccard as the coherence metric. Best values are bolded.

Setting	Group	nMSE↓	EV↑	Jaccard↑
AbsTopK	no	0.1714	0.8317	0.3825
+EMA	no	0.1531	0.8528	0.3366
+TV	no	0.1271	0.8771	0.4299
+Group ($k_f = 24$)	partial	**0.1183**	0.8856	0.5241
SN-AbsTopK (full)	yes	0.1207	0.8833	0.6140

Key observation. Tables 6 and 7 compare step-by-step component additions across the BCSS and CheXpert evaluation datasets. Removing spatial group selection lowers reconstruction error but severely degrades spatial coherence and feature quality across metrics. BCSS Moran's I drops sharply from 0.244 to 0.084, while CheXpert classification AUROC$_f$ decreases from 0.717 to 0.632. Adding total variation without group support improves reconstruction fidelity but completely fails to restore spatial feature coherence. On Duke MRI (Table 8), removing spatial group selection severely degrades slice-to-slice feature coherence, dropping adjacent-slice Jaccard similarity from 0.6140 to 0.3825. Overall, spatial group selection remains the principal driver behind structured feature coherence across all medical image representations.

C.2 Latent-Utilization Diagnostics across Datasets 3

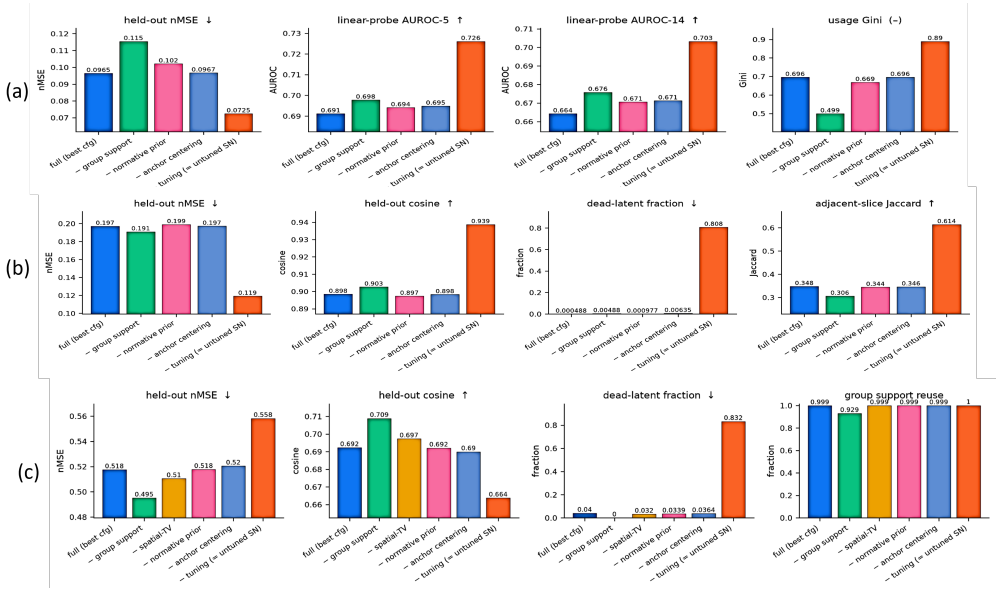


Figure 4: Component ablation across three medical imaging datasets: (a) BCSS (histopathology), (b) CheXpert (chest X-ray), and (c) Duke-MRI (breast MRI). Results compare the full SN-AbsTopK model with individual component ablations, showing their effects on reconstruction, downstream performance, and latent utilization.

Key observation. Figure 4 examines the effect of individual components on reconstruction, downstream clinical decodability, and latent utilization. Removing group support increases held-out nMSE from 0.0965 to 0.115 and reduces usage Gini from 0.696 to 0.499, while the normative-prior and anchor-centering ablations produce smaller changes. The untuned SN variant achieves the

lowest nMSE (0.0725) and highest AUROC-5/14 (0.726/0.703), but its active-latent fraction is only 0.185. These results show that reconstruction quality alone does not capture latent utilization or downstream representation quality.

D Extended Results & Visualizations 4

This section provides additional quantitative and qualitative analyses of latent selectivity and spatial consistency across medical imaging modalities. We evaluate tissue-class selectivity on BCSS, latent coherence across Duke MRI patient subgroups, and active-set persistence across MRI slices.

D.1 Histopathological Tissue-Class Selectivity (BCSS) 4

We evaluate whether learned latent features capture tissue-specific structure in BCSS using one-vs-rest selectivity and top-activating held-out patches. Table 9 reports quantitative selectivity, while Figure 5 provides qualitative examples.

Table 9: BCSS held-out test set, seed 0: best-atom one-vs-rest tissue-class selectivity.

Variant	Tumor	Stroma	Inflam.	Necrosis	Other
SN-AbsTopK-v2	**0.596**	**0.584**	**0.842**	**0.857**	**0.719**
AbsTopK	0.556	0.540	0.556	0.636	0.579
BatchTopK	0.573	0.551	0.553	0.589	0.558

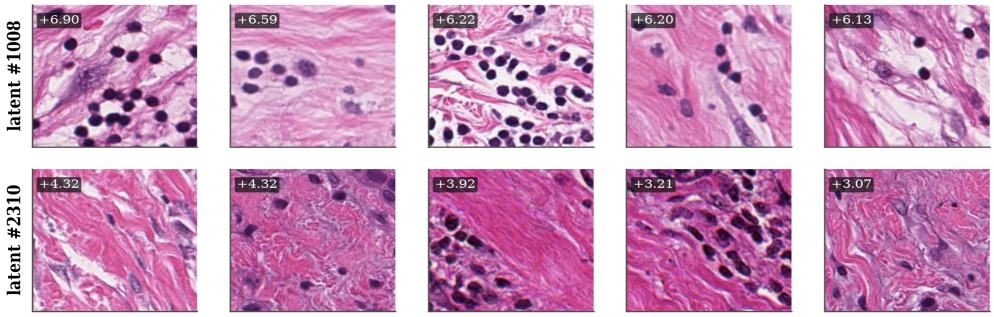


Figure 5: **Top-5 activating held-out BCSS patches for selected latent features.**

Key Observation. SN-AbsTopK achieves higher best-atom selectivity across all five tissue classes on BCSS dataset (see Figure 5). The qualitative examples further show that latent $\#1008$ responds to lymphocyte-rich regions, whereas latent $\#2310$ responds to collagen-rich stromal regions, illustrating distinct tissue-associated activation patterns.

D.2 Radiological Feature Selectivity (CheXpert) 4

We further examine latent responses in chest radiographs using extremal held-out activations. The examples illustrate latent responses to distinct radiological and imaging patterns.

Key Observation. The extremal z-values correspond to distinct radiographic patterns. The top row, with $z = -20.00$ to -18.95, shows relatively clear lung fields. The middle row, with $z = +15.19$ to $+15.10$, contains pleural effusion, blunted costophrenic angles, and support-device artifacts. The bottom row, with $z = -22.07$ to -20.56, shows bilateral consolidation, ground-glass opacities, and diffuse pulmonary edema. These examples show that strong latent activations are associated with recognizable patterns in the radiographs.

D.3 Volumetric Active-Set Coherence (Duke-MRI) 4

We evaluate spatial consistency of latent selections in Duke-MRI at both patient-subgroup and volumetric levels. Table 10 measures within-subgroup active-set coherence, while Table 11 measures persistence across increasing slice distances.

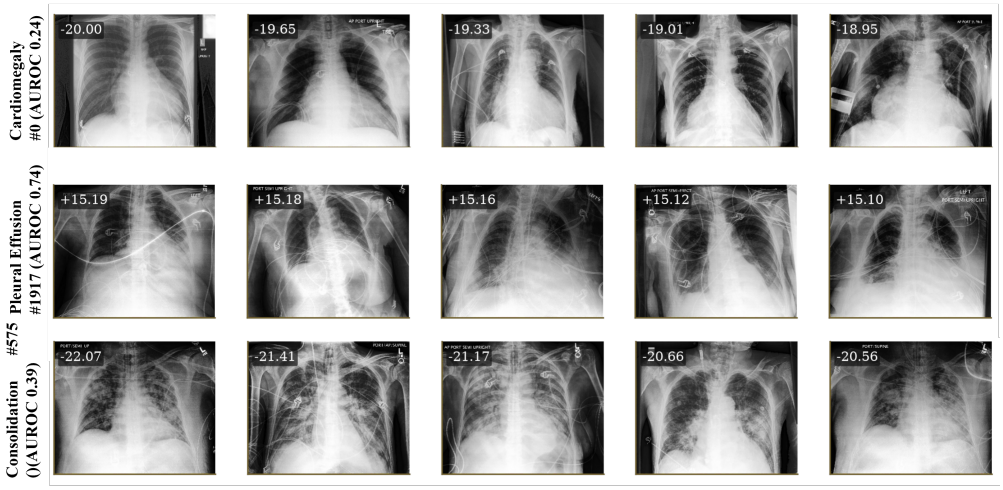


Figure 6: **Extremal held-out CheXpert activations across selected latent directions.** Examples include relatively clear lung fields, pleural effusion and support devices, and bilateral parenchymal abnormalities.

Table 10: Duke-MRI intra-stratum active-set Jaccard coherence across patient subgroups. Values are mean ± std.

Class / Subgroup	**AbsTopK**	**SN-AbsTopK**
ER−	0.379 ± 0.002	**0.610 ± 0.005**
ER+	0.384 ± 0.000	**0.616 ± 0.000**
PR−	0.385 ± 0.002	**0.618 ± 0.002**
PR+	0.381 ± 0.001	**0.612 ± 0.001**
HER2−	0.369 ± 0.001	**0.605 ± 0.001**
HER2+	0.420 ± 0.001	**0.639 ± 0.003**
Grade (low/int)	0.380 ± 0.002	**0.611 ± 0.003**
Grade (high)	0.415 ± 0.002	**0.638 ± 0.003**
Luminal-like	0.372 ± 0.002	**0.607 ± 0.000**
ER/PR+, HER2+	0.416 ± 0.006	**0.634 ± 0.003**
HER2 subtype	0.427 ± 0.007	**0.647 ± 0.005**
Triple-negative	0.358 ± 0.001	**0.599 ± 0.004**
Overall / Macro	0.383 ± 0.001	**0.614 ± 0.002**

Table 11: Duke-MRI held-out test set: active-set Jaccard across slices separated by d. Values are mean ± std. over 40 volumes.

Variant	d=1	d=2	d=5	d=10
SN-AbsTopK	**0.633±0.035**	**0.565±0.032**	**0.497±0.024**	**0.447±0.022**
AbsTopK	0.321±0.060	0.234±0.055	0.165±0.047	0.118±0.044
BatchTopK	0.333±0.055	0.246±0.051	0.167±0.045	0.117±0.041

Key Observation. SN-AbsTopK produces substantially higher active-set overlap within patient subgroups, with macro Jaccard increasing from 0.383 ± 0.001 for AbsTopK to 0.614 ± 0.002 (Table

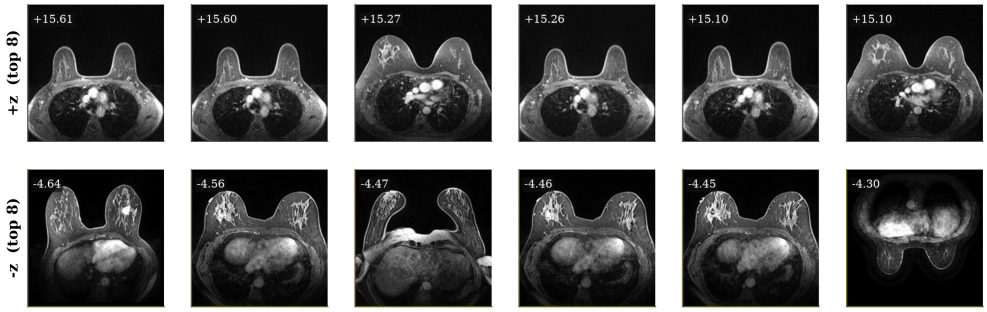


Figure 7: **Extremal held-out Duke-MRI slice activations across latent directions.** Positive responses highlight background parenchymal and fibroglandular patterns, while negative responses include focal enhancing mass-like regions and breast–chest-wall interfaces.

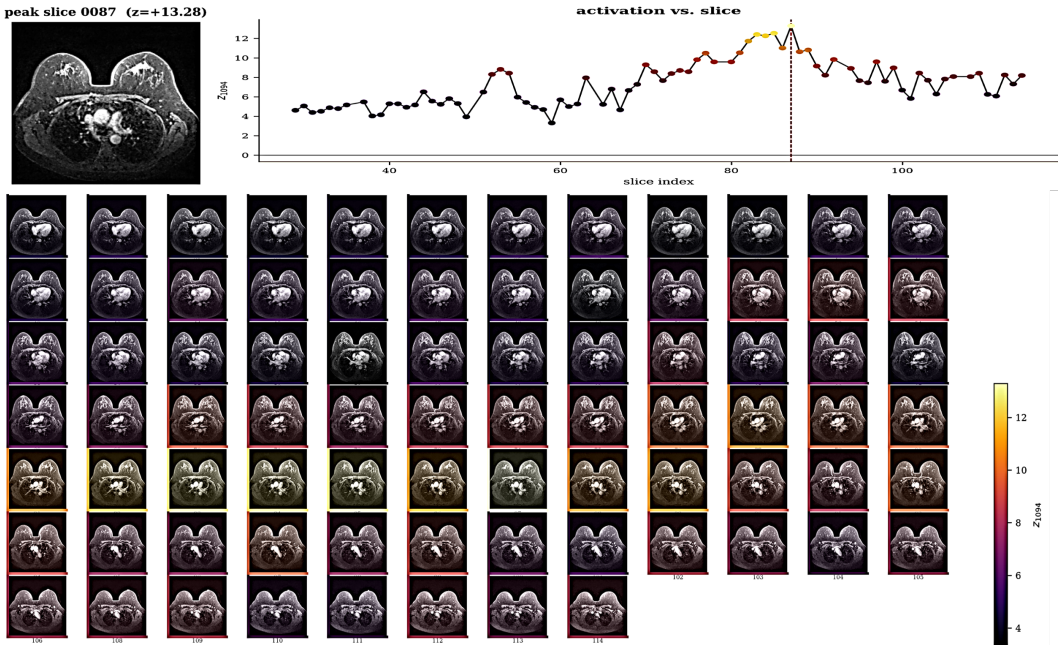


Figure 8: **Volumetric activation profile across axial slices for a learned latent feature (z_{1094}) on Duke-MRI. Top left:** Peak activating axial DCE-MRI slice (Slice 0087) achieving maximum feature magnitude ($z = +13.28$). **Top right:** 1D trajectory plotting latent activation magnitude z_{1094} as a function of axial slice index along the superior-inferior axis, showing a distinct localized peak around slice 87 (dashed line). **Bottom:** Sequential montage of axial breast MRI slices color-coded by activation intensity (ranging from ~ 4 in dark purple to > 12 in bright yellow). The smooth elevation and drop-off in feature activation track the continuous 3D spatial extent of the target anatomical/pathological structure across adjacent MRI slices.