Medical AI Encodes a "Feeling of Error": Verifying Cancer Segmentation via Internal Concepts

Mengmeng Ma[1], Yunxiang Peng[2], Tang Li[2], Lu Lin[3], Binsheng Zhao[3], Oguz Akin[3], and Xi Peng[1]

[1] University of Virginia, Charlottesville, VA, USA
[2] University of Delaware, Newark, DE, USA
[3] Memorial Sloan Kettering Cancer Center, New York City, NY, USA
{wkb9cb,naq5rd}@virginia.edu

Abstract. Cancer segmentation models can fail silently, generating plausible but incorrect masks that risk missed findings or unnecessary biopsies. A critical question arises: Do AI models "know" when they are wrong, and if so, can we use the signal to predict their own failures? Humans do have a "Feeling of Error" (FOE): a spontaneous sense of unease that flags a potential error during thinking. We investigate whether cancer segmentation models exhibit an analogous internal signal. Unlike output-level cues (e.g., prediction confidence or uncertainty), which offer no insight into why a failure occurs and suffer from a sensitivity–quality tradeoff where high detection sensitivity could degrade overall segmentation quality. We instead propose to capture the model's FOE from its inner workings. Using mechanistic interpretability tools, specifically Sparse Autoencoders, we decompose internal neural activations into a dictionary of human-interpretable concepts and show that failure cases exhibit a distinct latent signature: fewer active concepts with lower activation magnitudes compared to successful segmentation. By training a classifier on these concept activations, we achieve accurate failure detection along with explanations for the model's mistakes. Experiments on prostate, pancreatic, and brain cancer segmentation demonstrate that our approach outperforms output-based methods in failure detection while preserving segmentation quality. Code is abailable at https://github.com/deep-real/CancerSegFailure.

1 Introduction

The promise of AI in clinical oncology is often undermined by *silent failure*. Even high-performing models can generate anatomically plausible yet incorrect masks that risk missed findings or unnecessary biopsies [22, 27]. Although clinicians remain the last line of defense, manually verifying numerous AI-generated masks under heavy workloads imposes a significant cognitive burden [42]. This raises a critical question: *Do AI models "know" when they are wrong?* If so, can we use this signal to predict the silent failures? Interestingly, humans do, through what psychologist called "Feeling of Error" (FOE) [17, 66]: a spontaneous sensation of cognitive uneasiness arising from conflict detection during thinking, effectively

acting as an internal signal to flag potential error. We ask whether state-of-the-art ViT-based segmentation models [13, 45, 69] exhibit an analogous FOE-like signal. Identifying such signal would turn segmentation from a black-box output into an auditable decision process. By tracing and examining the corresponding FOE signals, we can automatically flag high-risk cases for inspection or provide the justifications clinicians need to trust, refine, or override AI predictions.

A straightforward way to approximate a model's Feeling of Error is through output-level cues [24, 62], such as prediction confidence or uncertainty estimates (Fig. 1). However, this approach faces two fundamental limitations. First, these signals are rarely actionable. They may warn that a mask is risky but offer no insight into why a failure occurs. Second, correcting flagged failure using output-level cues faces a sensitivity–quality tradeoff: Increasing detection sensitivity catches more failures but also flags borderline, clinically acceptable cases, and unnecessary "corrections" can degrade overall segmentation quality. In our experiments, a 20-point gain in failure-detection F1-score via these proxies comes at the cost of a 8–10 point drop in segmentation DSC. This suggests that a model's feeling of error is likely encoded not in what it outputs, but in how it thinks (i.e., its decision-making process).

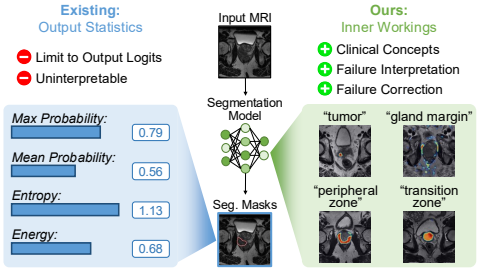

Fig. 1: Two sources for capturing a model's *Feeling of Error* (FOE). Output-level cues (left) are limited to uninterpretable logit statistics, offering no insight into failure causes. We instead extract the model's FOE from its inner workings (right), decomposing internal representations into interpretable concepts that enable failure detection, interpretation, and correction.

To peek under the hood of the model's decision-making, we employ mechanistic interpretability tools [5, 53, 60], specifically Sparse Autoencoders (SAEs) [14, 26, 63]. These tools decompose high-dimensional neural activations into a dictionary of human-interpretable "concepts." Our analysis show that ViT-based cancer segmentation models learn concepts that align with established clinical terms (Fig. 1). Crucially, we observe a distinct latent signature in failure cases: Incorrect segmentations activate fewer internal concepts with lower activation magnitudes compared to successful ones (Fig. 2, Left). The model, in some sense, "knows" it is wrong, even when its outputs say otherwise. This internal concept activation is the machine FOE we can leverage. Translating this signal for reliable failure detection raises two challenges. *First, where does the signal live?* Failures are not monolithic and may not localize to a single layer. A missed tumor might reflect an early-layer visual processing error [15, 38], while misclassifying benign tissue as malignant suggests a deep-layer semantic confusion [54]. *Second, how should the signal be modeled?* Failures may depend on interactions among concepts (and their spatial arrangement), where individually plausible cues become problematic only under certain co-activation patterns.

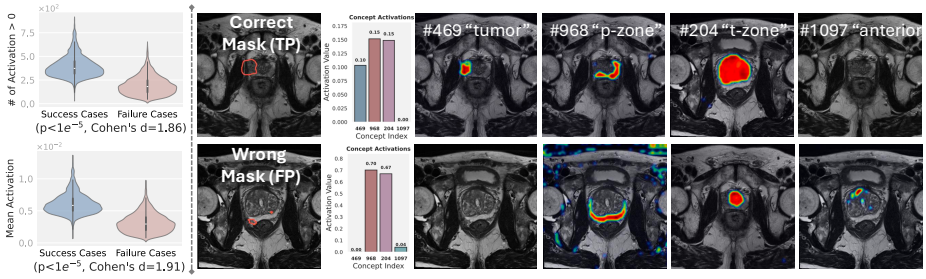

Fig. 2: Distributions of SAE concept activations for success and failure cases on PI-CAI. **Left:** Number of active concepts (activation value > 0) and Mean concept activation value. On average, success cases have more active concepts with stronger activation values than failure cases. Both metrics show statistically significant separation (p-value $< 1e^{-5}$, Cohen's d > 1.8), indicating that concept activation patterns distinguish correct from incorrect segmentation. **Right:** Example success (top) and failure (bottom) cases showing predicted masks and activation patterns for four key concepts. The bar charts display concept activation strengths for each case. Shared concepts #968 ("peripheral zone") and #204 ("transition zone") activate in both success and failure cases, reflecting anatomical context. However, tumor-specific concept #469 ("tumor") shows strong activation only in the success case but minimal in the failure case. *These patterns suggest that the model encodes a "feeling of error" in its internal concept activations.*

To tackle the challenges, we propose a two-stage framework. First, we build hierarchical "failure representations" by extracting concepts from early, middle, and deep layers. This gives us a comprehensive "vocabulary" for describing failures from low-level visual issues to high-level diagnostic confusions. Second, to enable both accurate failure detection and interpretation, we apply a lightweight classifier on these internal concepts. This classifier can learn complex interaction patterns (e.g., "hyperintense + irregular texture + weak boundaries = likely failure") while remaining interpretable: each concept's importance score shows exactly how a concept contributes to the failure prediction.

Our contributions: (1) We move beyond output logits to show that a model's internal concept representations can not only detect but also *explain* segmentation failures in clinically meaningful terms. (2) We design a method that extracts failure-predictive concepts across network layers and pinpoints their diagnostic patterns using interpretable classifiers. (3) Empirical verification on prostate and pancreatic cancer shows that our approach achieves superior failure detection performance while providing intuitive explanations and maintaining high segmentation quality.

2 Preliminaries and Findings

In this section, we begin by formally defining cancer segmentation failure, then present two empirical findings that support the existence of a *machine FOE*, the model's internal signal that predicts segmentation failure.

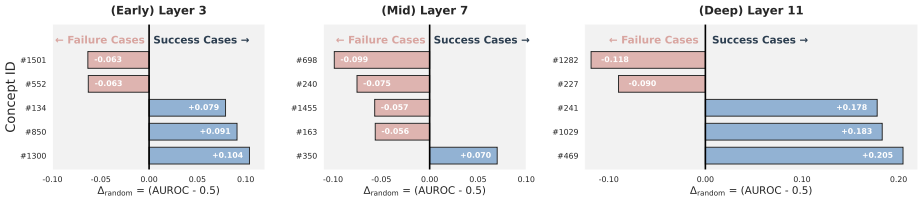

Fig. 3: Predictive power of SAE concepts across layers for segmentation failures. We directly use each concept's activation score and the ground-truth failure labels to compute the AUROC for each concept. Each bar shows AUROC improvement over random chance ($\Delta_{\text{random}} = \text{AUROC} - 0.5$) for the top 5 most discriminative concepts per layer. Blue bars indicate concepts predictive of success cases (success-predictive), while red bars indicate concepts predictive of failure cases (failure-predictive). *Deep layers exhibit stronger predictive power overall, middle layers show more failure-predictive concepts, and early layers have weaker discriminative ability.*

Definition 1 (Cancer Segmentation Failure). *Let $f : \mathcal{X} \to \mathcal{M}$ be a segmentation model mapping medical images to binary masks, where $\mathcal{X}$ is the input image space and $\mathcal{M}$ is the mask space. For an image $x \in \mathcal{X}$, let $\hat{m} = f(x)$ denote the predicted mask and $m^* \in \mathcal{M}$ the ground truth mask. Let $\mathcal{E} : \mathcal{M} \times \mathcal{M} \to [0, 1]$ be a quality metric (e.g., Dice coefficient). A* ***cancer segmentation failure*** *occurs when $\mathcal{E}(\hat{m}, m^*) \leq \tau$ for a predefined threshold $\tau \in [0, 1]$.*

SAE preliminary. SAEs learn interpretable, overcomplete representations by decomposing inputs into sparse combinations of learned features [6,12,18,39, 50]. Given an internal activation $\mathbf{h} \in \mathbb{R}^d$ from a ViT layer; The SAE consists of an encoder $E : \mathbb{R}^d \to \mathbb{R}^D$ (where $D > d$) and decoder $G : \mathbb{R}^D \to \mathbb{R}^d$:

$$\begin{aligned} \mathbf{z} &= E(\mathbf{h}) = \text{ReLU}(\mathbf{W}_{\text{enc}}\mathbf{h} + \mathbf{b}_{\text{enc}}), \\ \hat{\mathbf{h}} &= G(\mathbf{z}) = \mathbf{W}_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{dec}}, \end{aligned} \tag{1}$$

where $\mathbf{W}_{\text{enc}}, \mathbf{W}_{\text{dec}}^\top \in \mathbb{R}^{D \times d}$ and $\mathbf{b}_{\text{enc}}, \mathbf{b}_{\text{dec}} \in \mathbb{R}^D$. The sparse latent $\mathbf{z} \in \mathbb{R}^D$ represents concept activations, where each dimension z_i corresponds to a learned feature. The decoder reconstructs inputs as $\hat{\mathbf{h}} = \sum_{i=1}^{D} z_i \mathbf{w}_i + \mathbf{b}_{\text{dec}}$, where $\mathbf{w}_i$ (the i-th column of $\mathbf{W}_{\text{dec}}$) is the i-th feature direction. SAEs are trained to minimize: $\mathcal{L}(\mathbf{h}) = \|\mathbf{h} - \hat{\mathbf{h}}\|_2^2 + \lambda\|\mathbf{z}\|_1$, balancing reconstruction fidelity with sparsity.

We train SAEs on internal activations of ViT-based segmentation models (e.g., Medical SAM variants [9,45,46]) fine-tuned across multiple cancer datasets. This yields two key findings.

Finding 1: Failure cases exhibit distinct internal concept activation patterns. As shown in Fig. 2, failure cases activate significantly fewer SAE concepts and at substantially lower magnitudes compared to successful cases ($p < 10^{-5}$, Cohen's $d = 1.8$). This large effect size suggests that failures are not merely low-confidence predictions but correspond to a qualitatively different internal state. To further quantify the discriminative power of individual concepts, we compute the AUROC of each concept's activation value against

ground-truth failure labels. As shown in Fig. 3, this reveals two distinct groups: *success-predictive* concepts (high AUROC, strongly active in successful cases) and *failure-predictive* concepts (disproportionately elevated in failure cases). The existence of failure-predictive concepts suggests that failures leave a detectable signature in the latent space.

Finding 2: There are SAE concepts that align with established clinical terms. Beyond their discriminative value, we find that SAE concepts correspond to clinically meaningful anatomical structures (Figs. 2 and 4). To validate this, we follow a two-stage protocol. First, we automatically measure spatial overlap between each concept's activated pixels and annotated lesion or anatomical masks across all datasets: concepts with IoU ≥ 0.5 are assigned clinical labels based on the corresponding annotation. We will introduce how to obtain the pixels corresponding to SAE concepts in the next section. Second, we present a representative set of automatically verified concept-label pairs, along with their pixel localizations, to two board-certified radiologists, who independently assess whether each label accurately describes the visualized concept. Both radiologists confirmed agreement with all presented labels. Together, these two findings establish that SAE concepts are both *semantically grounded* and *failure-informative*.

Open problems. While these findings are promising, two critical questions remain. (1) How to effectively identify concepts that are predictive of failures? The SAE concept dictionary can reach 1K–10K entries, making manual inspection or heuristic selection infeasible. (2) Even after discriminative concepts are obtained, how to best use them for failure detection remains an open problem. Segmentation models distribute information across layers; it is unclear which layers' concepts are most informative for detecting failures, or how to effectively aggregate them into a reliable decision signal. In the next section, we address these challenges to build a robust failure detector.

3 Detecting Failures via Internal Concepts

This section details our framework that leverages segmentation models' internal concepts for failure detection. We address two key questions: (1) *How to capture all potential failure signals?* Rather than focusing on a single "best" layer, we capture all possible signals of failure from all layers, obtaining the holistic representation of failures (Sec. 3.1). (2) *How to identify predictive patterns of failure?* Failures may not be linked to a single concept, but to interactions between them. We build a dedicated classifier on top of failure representation to automatically learn the "signature of failure" (Sec. 3.2). Fig. 4 illustrates our complete pipeline.

3.1 How to Capture All Potential Failure Signals?

Cancer segmentation failures could arise from diverse clinical causes that manifest across different representational levels. Low-level image artifacts (motion,

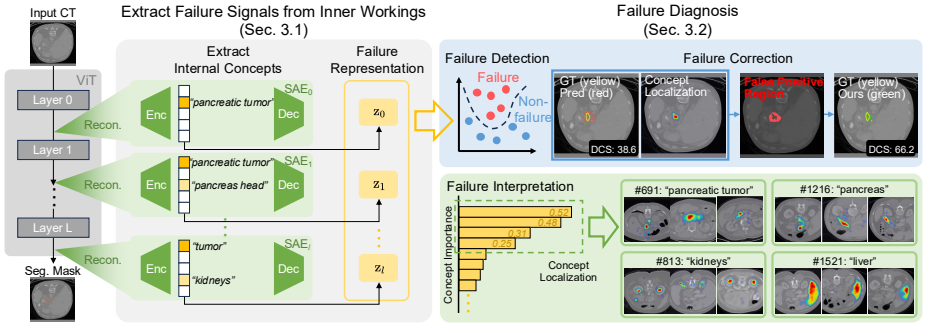

Fig. 4: Overview of the proposed framework. *Failure signal extraction*: SAEs decompose internal representations from multiple layers of a frozen segmentation model into sparse concept vectors, capturing failure-relevant signals across the full representational hierarchy. *Failure diagnosis*: The concatenated concept representation is passed to a lightweight classifier that serves dual purposes: *detection*, which identifies and corrects segmentation failures, and *interpretation*, which ranks concepts by their contribution to the failure prediction, tracing model errors back to interpretable concepts

noise, poor contrast) create visual anomalies, while high-level anatomical complexities (ambiguous boundaries, irregular morphology) cause semantic confusion. Thus, a robust detector should analyze the model's representations from low-level visual features to high-level semantics to holistically characterize all potential failure signals.

Method. To obtain representations of model failures using internal concepts, we train SAEs on latent embedding from early, mid, and deep layers of the segmentation model f. Specifically, we uniformly sampled across the ViT backbone from layers $L = \{1, 3, 5, 7, 9, 11\}$ to ensure coverage of the full representational spectrum and extract their latent embeddings $\mathbf{h}_l(x) \in \mathbb{R}^d$ for each input image x and layer $l \in L$. d is the latent embedding size. For each layer, we train an independent SAE to transform dense embeddings into sparse, interpretable concept vectors. The SAE encoder $E_l : \mathbb{R}^d \to \mathbb{R}^D$ is trained on embeddings from both correct predictions and failures $\{\mathbf{h}_l(x) \mid x \in \mathcal{X}_C \cup \mathcal{X}_F\}$, where D denotes the dictionary size. Each encoder produces sparse concept activations $\mathbf{z}_l(x) = E_l(\mathbf{h}_l(x)) \in \mathbb{R}^D$. We obtain the holistic failure representation by concatenating all sparse concept vectors across layers:

$$\mathbf{c}(x) = \bigoplus_{l \in L} \mathbf{z}_l(x) = \bigoplus_{l \in L} E_l(\mathbf{h}_l(x)) \in \mathbb{R}^{|L| \cdot D}, \tag{2}$$

where $\bigoplus$ denotes concatenation and $|L| = 6$ is the number of selected layers. $\mathbf{c}(x)$ preserves all layer-specific internal concepts, providing a rich signal for failures.

Validation of discriminability. Before proceeding to the next section, we first verify that the learned representation $\mathbf{c}(x)$ captures discriminative information about failures. Fig. 2 visualizes the distribution of failure representations for correct prediction and failures. The two classes exhibit statistically significant separation with minimal overlap (independent t-test, $p < 1e^{-5}$ and effective

size, d > 1.8). This empirical evidence confirms that internal concepts encode discriminative features suitable for failure detection.

3.2 How to Identify Predictive Patterns of Failure?

Given the failure representation $\mathbf{c}(x) \in \mathbb{R}^{|L|\cdot D}$, our goal is to find a decision function $g : \mathbb{R}^{|L|\cdot D} \to [0, 1]$ that satisfies two objectives: (1) *accurate prediction*, where $p_{\text{failure}}(x) = g(\mathbf{c}(x))$ achieves high discriminative for failure prediction, and (2) *interpretability*, where g provides explainable feature importance scores $\mathbf{w} \in \mathbb{R}^{|L|\cdot D}$ that quantify each concept's contribution to failure detection.

Method. We learn the decision function g in a data-driven manner. The first step is to construct a labeled *failure dataset* $\mathcal{D} = \{(\mathbf{c}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{c}_i = \mathbf{c}(x_i) \in \mathbb{R}^{|L|\cdot D}$ is the failure representation for image x_i, and $y_i \in \{0, 1\}$ indicates correct prediction ($y_i = 0$) or failure ($y_i = 1$). The dataset $\mathcal{D}$ is derived from the segmentation model's training and validation sets, where ground truth annotations enable reliable failure labeling. For both accuracy and interpretability, we model g as a linear or non-linear classifier (e.g., XGBoost [11] or logistic regression). The classifier is trained on the failure dataset, optimizing binary cross-entropy loss:

$$\min_{g} \mathcal{L}(g, \mathcal{D}) = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \log g(\mathbf{c}_i) + (1 - y_i) \log(1 - g(\mathbf{c}_i)) \Big] + \lambda \Omega(g), \tag{3}$$

where $\Omega(g)$ controls model complexity, and $\lambda \geq 0$ is the regularization strength.

Classifier and feature importance. We adopt XGBoost [11] as our primary classifier. Gradient boosting is particularly well-suited for this task as it: naturally handles high-dimensional sparse features, learns non-linear interactions between concepts across layers, and provides feature importance scores that reveal which concepts are most discriminative for failure detection. We also benchmark against other classifiers in the Ablation Study (Tab. 4). Once trained, the classifier provides importance scores for each concept dimension in $\mathbf{c}(x)$, quantifying their contribution to failure prediction. By mapping these scores back to specific SAE concepts and their corresponding layers, we can identify which internal concepts, and at what representational levels, are most indicative of segmentation failures. Fig. 8 visualizes the learned most important concepts.

3.3 Implementation and Discussion

Patch- vs. image-level embedding. We adopt ViTs as our backbone for cancer segmentation. A fundamental design choice when applying SAEs to ViTs is: should we aggregate patch tokens into a single image-level representation, or preserve spatial structure by training SAEs directly on patch-level embeddings? Prior work mainly focuses on the classification task and predominantly uses image-level aggregation [16, 37, 55], where patch embeddings are pooled into a global embedding before SAE training. This is sufficient for classification, as global semantic concepts are adequate for category prediction. However,

segmentation fundamentally differs: it is a *dense prediction task* requiring pixel-wise decisions. Therefore, we train all SAEs directly on patch-level embeddings to preserve spatial granularity.

Concept interpretation via spatial localization. To understand what each concept represents, we leverage the spatial structure preserved in patch-level representations. Given an input image x, the ViT backbone produces spatial patch embeddings $\mathbf{H}_l(x) \in \mathbb{R}^{h \times w \times d}$ at layer l, where h and w denote the spatial grid dimensions. Applying the SAE encoder E_l to each spatial location yields patch-level concept activations:

$$\mathbf{C}_l(x) = E_l(\mathbf{H}_l(x)) \in \mathbb{R}^{h \times w \times D}, \tag{4}$$

where $\mathbf{C}_l(x)[i, j, :] \in \mathbb{R}^D$ represents the concept vector at spatial position (i, j). To visualize where a specific concept k activates in the image, we extract its spatial activation map: $\mathbf{C}_l^{(k)}(x) = \mathbf{C}_l(x)[:, :, k] \in \mathbb{R}^{h \times w}$. This 2D map reveals which image regions mostly activate concept k.

Failure correction. Our framework enables principled *failure correction* as a natural extension of detection. Since the classifier operates at the patch level, we can identify not just whether a prediction is a false positive, but which specific patches drive the wrong decision. We formalize this as a patch retention problem: the corrected segmentation retains patch (i, j) if and only if its local failure score falls below a threshold η:

$$\hat{m}_{\text{corrected}}(i, j) = \hat{m}(i, j) \cdot \mathbf{1}[g(\mathbf{c}(x)_{ij}) < \eta], \tag{5}$$

where $\mathbf{c}(x)_{ij}$ is the concept vector at patch (i, j). This selective correction preserves high-confidence TP regions that output-based methods would discard entirely when flagging a prediction as failed, improving both precision and the clinical utility of the corrected mask.

Relation to broader context. Our framework shares the high-level "concept-then-predict" motivation with concept bottleneck models [35], but differs in two critical ways: (1) it operates post-hoc on any pre-trained segmentation model without architectural modifications, and (2) it discovers concepts unsupervised via SAEs, eliminating the need for manual concept annotations. Additionally, recent medical vision-language models can map visual embeddings to discrete vocabularies aligned with language space [19, 28, 33]. By contrast, our approach does not require internal concepts to be fully aligned with human language. In fact, some of the most predictive failure concepts we discover are non-semantic. They capture internal processing patterns that correlate with failures but are not human-readable. This aligns with recent arguments that AI systems may develop representations outside our existing vocabulary [25].

4 Experiments

4.1 Data, Baselines, Metrics, and Hyperparameters

Datasets. We evaluate our methods on three public cancer datasets: Prostate (PI-CAI [57], Prostae158 [2]) and Pancreatic (PanTS [40]). *PI-CAI* contains

1,500 mp-MRI scans from three Dutch medical centers. We used its provided annotations for prostate lesions and partitioned the dataset into 1,200 training, 60 validation, and 240 testing scans. PI-CAI is one of the largest publicly available MRI datasets for prostate cancer. *Prostate158* includes 158 mp-MRIs with T2W, DWI and ADC sequences. We use all data for zero-shot evaluation. *PanTS* consists of 9,901 CT scans with annotations for the pancreatic tumors, pancreas, and its head, body, and tail. We use the subset PanTS to curate our data. Specifically, we randomly selected 700 CTs and split them into 448 training, 112 validation, and 140 testing scans.

Baselines. We compare against output-based uncertainty quantification methods that do not require additional training: maximum softmax probability (MaxProb) [24], mean softmax probability (MeanProb) [24], entropy [24], energy score [44], and direct probability thresholding. These methods represent the standard practice of using model confidence as a proxy for prediction quality. We do not compare against approaches that incorporate failure detection into the training objective, as such methods require access to and modification of the main network's training pipeline, which is incompatible with the post-hoc deployment scenario this work targets. In clinical practice, segmentation models are typically received as pre-trained, frozen systems from the vendor; our setting and all baselines evaluated here reflect this realistic constraint.

Evaluation metrics. We evaluate segmentation performance using the Dice Similarity Coefficient (DSC, ↑). For failure detection performance, we report the F1 (↑) score as our primary metric. Additionally, following common practice [44, 51, 70], we report the AUROC (↑) to assess how well the method ranks correct versus incorrect segmentations, and the True Positive Rate at 95% False Positive Rate (FPR95, ↓) to measure reliability under strict conditions.

Hyperparameter settings. Our segmentation backbone is MedSAM [45], a widely-adopted foundation model with an image encoder, prompt encoder, and mask decoder. To enable automatic segmentation, we modify the prompt encoder to operate without manual prompts [59]. We fine-tune the model for each cancer segmentation task on its corresponding dataset using Focal Loss [41] ($\alpha = 0.97$, $\gamma = 2$) and the Adam optimizer [34] with a $1e^{-4}$ weight decay and an exponential learning rate scheduler. For SAE training, we follow the BatchTopK [7] approach with learning rate $1e^{-4}$ for 5 epochs. We use dictionary size $D = 1,536$ and sparsity $S = 8$ across all layers.

4.2 Failure Interpretation

Takeaway: *The model's internal concepts provide a mechanistic explanation for segmentation failures and a reliable signal for predicting them.*

Model's internal concepts. To understand what the segmentation model learns, we train SAEs on its latent embeddings. The SAEs disentangle the polysemantic embeddings into monosemantic ones, which can be treated as the internal concepts. As shown in Figs. 1 and 2, these concepts localize to anatomically relevant regions. In prostate cancer, for instance, we identify distinct concepts for the peripheral zone, gland, and tumor, with finer-grained concepts distinguishing

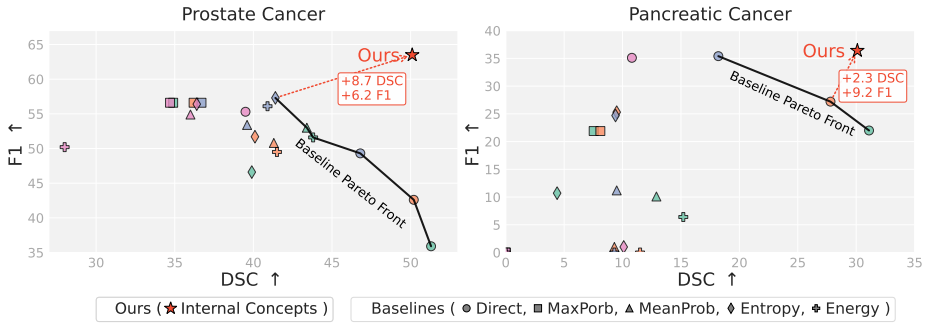

Fig. 5: F1 (↑) vs. DSC (↑) scores on PI-CAI (Prostate Cancer) and PanTS (Pancreatic Cancer). Each marker represents a baseline method varying by threshold [0.5, 0.8]. The solid gray line indicates the Pareto front, showing the best achievable trade-off between segmentation accuracy (DSC) and failure detection quality (F1). *Our method surpasses the baseline Pareto front, achieving both higher DSC and higher F1, demonstrating superior reliability without sacrificing segmentation quality.*

tumor core from edge. We observe that a similar anatomical alignment appears consistently in pancreatic cancer.

Failure analysis. Given these internal concepts, we next analyze how they behave when segmentation fails. We find that the activation of these concepts can distinguish success from failure cases. In Fig. 2, successful ones produce strong activations in tumor concepts. In contrast, failures activate these concepts weakly or not at all, suggesting the model internally "knows" it is uncertain about these areas. This provides a mechanistic explanation for failures: failures often occur when the model misinterprets ambiguous image features, failing to activate the correct internal representations.

4.3 Failure Detection

Takeaway: *Model's internal concepts contain richer signals about failures.*

Our analysis has shown that the internal concept activations are discriminative for segmentation failures. Do internal concepts provide more discriminative information for identifying failures than the model's final output confidence? To answer this question, we train a simple classifier using these internal concept activations to predict segmentation failures and compare it against baselines that use only the model's output (e.g., logits). The results in Fig. 5 and Tab. 1 show that our models are robust and generalizable.

Quantitative performance. We evaluate our method across three cancer segmentation datasets. The results are shown in Fig. 5. Baseline methods exhibit a clear Pareto tradeoff between segmentation quality (DSC) and failure detection performance (F1); Improving one metric requires sacrificing the other. By leveraging internal concepts, our method breaks this tradeoff, simultaneously achieving high performance on both metrics.

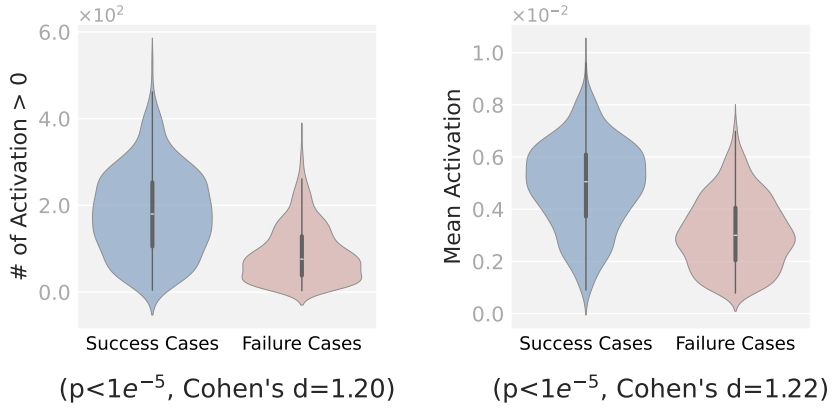

Fig. 6: SAE concept activations for success and failure cases on Prostate158 (unseen data). Both metrics show statistically significant separation ($p < 1e^{-5}$, Cohen's d > 1.2), indicating that *concept activation patterns can generalize across datasets.*

Zero-shot generalization. We evaluate all detectors trained on PI-CAI directly on the unseen Prostate158 dataset without fine-tuning. As shown in Tab. 1, confidence-based baselines, such as MaxProb and Entropy, generalize poorly, with AUROC scores close to random chance. In contrast, our concept-based detector maintains substantially better performance across all three metrics. This indicates that internal concepts are more reliable and generalizable features for failure detection than model confidence, which can be poorly calibrated on unseen data. Fig. 6 further validates generalization: on the unseen Prostate158 dataset, success and failure cases remain separable in concept space, demonstrating that learned failure patterns transfer across domains.

Table 1: Zero-shot failure detection on the unseen Prostate158 dataset using detectors trained on PI-CAI. *Our concept-based detector consistently outperforms confidence-based baselines across all metrics.*

	FPR95 (↓)	AUROC (↑)	AUPR (↑)
MaxProb	100.0	50.0	64.4
MeanProb	100.0	52.6	65.1
Entropy	97.8	53.1	65.6
Energy	100.0	48.0	63.6
Ours	**64.4**	**73.6**	**74.1**

Qualitative analysis of detected failures. Fig. 7 shows representative failures detected by our method: (1) Hallucinated lesions where the model predicts non-existent tumors, and (2) Boundary errors where true lesions are localized but severely over- or under-segmentation. Our approach identifies both complete false positives and boundary-level errors.

Analysis of concept importance. Since our classifier is trained on using the internal concepts, we can obtain each concept's importance. Fig. 8 visualizes the concept importance scores and concept localization. Crucially, we observe that the most important concepts are tumor-related.

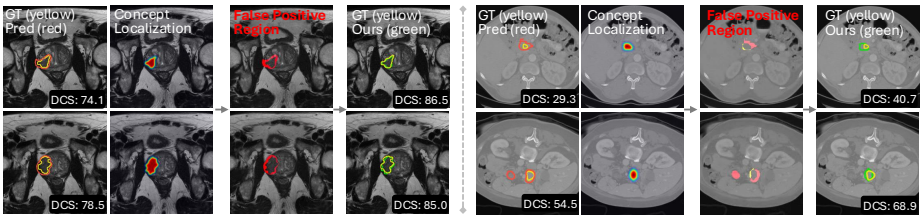

Fig. 7: Visualization of Failure correction on PI-CAI (MRI, left) and PanTS (CT, right). For each example, four columns are shown: (*1st*) initial segmentation prediction (red) and ground truth (yellow); (*2nd*) SAE concept localization; (*3rd*) detected false positive regions (red); (*4th*) final prediction mask (green) after FP removal. DCS scores after correction demonstrate consistent improvement across both datasets. *SAE concepts enable the localization and removal of false-positive regions, consistently improving segmentation accuracy across both MRI and CT datasets.*

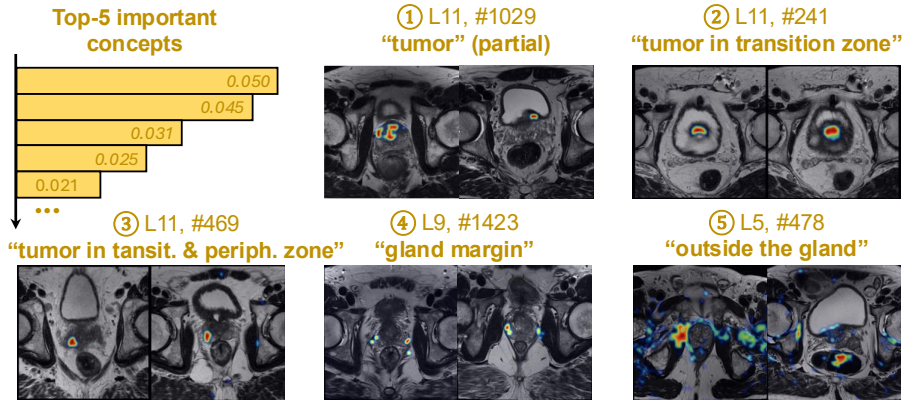

Fig. 8: Top-5 concept importance for failure prediction. Concepts are ranked by importance scores and visualized via localization heatmaps. Tumor-related concepts (1-3) dominate the ranking, capturing partial tumors, transition zone (TZ) tumors, and tumors spanning TZ and peripheral zone (PZ). Anatomical boundary concepts (4-5) identify gland edges and extra-prostatic regions, *revealing that the classifier leverages both tumor characteristics and anatomical context for failure detection.*

4.4 Ablation Study

SAE architecture. We ablate SAE sparsity and dictionary size to determine their impact on failure detection performance. Results are shown in Tab. 2. Optimal performance occurs at sparsity L0 = 8 and dictionary size = 1,536, which are notably lower than typical values for natural images (S $\approx$ 16, dictionary size D $\approx$ 10,000) [14,63]. We hypothesize this reflects the more structured nature of medical imaging modalities compared to natural images.

Early vs. deep layer. We analyze the contribution of concepts from different network depths (Tab. 3). Deep layer concepts achieve the highest failure de-

Table 2: Ablation study on SAE architecture. We vary the dictionary size (D) and sparsity (S) and report the resulting failure detection F1 and segmentation DSC. The optimal configuration (D=1536, S=8) is in bold.

	D=1,536 (768×2)			D=3,072 (768×4)		
Sparsity (S)	4	8	16	4	8	16
DSC (↑)	49.7	**50.1**	49.9	50.1	49.7	49.6
F1 (↑)	60.3	**63.5**	62.5	58.4	57.4	62.6

Table 3: Ablation on concept source depth. We report results using concepts from early, middle, and deep layers. *Combining all three levels (All) achieves the best DSC and F1*, indicating that concepts across network depths provide complementary information.

Layers (L)	Early (1, 3)	Mid (5, 7)	Deep (9, 11)	All
DSC (↑)	47.5	48.8	49.6	**50.1**
F1 (↑)	44.5	50.4	57.6	**63.5**

Table 4: Ablation on the failure detection classifier. We compare performance using our SAE concepts with different classifiers. *XGBoost achieves the best results for both DSC and F1.*

Classifier	XGB	Logistic	Decision Tree	Random Forest
DSC (↑)	**50.1**	47.8	48.5	44.6
F1 (↑)	**63.5**	59.2	58.7	57.0

Table 5: Computation overhead. We compute the per-image latency for our failure detection method. *Our method introduce limited runtime overhead.*

Configuration	Latency (ms)	Δ
ViT model (e.g., MedSAM)	126.3 ± 0.5	—
+ 6 SAE encoders (routine)	130.9 ± 2.7	+3.7%
+ 6 SAE enc+dec (intervention)	134.3 ± 0.3	+6.3%
+ XGBoost (full pipeline)	176.5 ± 13.3	+39.7%

tection performance (F1=57.6) by capturing high-level semantic patterns, while mid-layer concepts excel at segmentation quality (DSC=48.8). Early layer concepts, encoding low-level visual features, contribute moderately to both tasks.

Choice of failure classifier. We tried different classifiers such as logistic regression, random forest, and XGB classifier. The results are shown in Tab. 4. The XGB classifier gives the best results.

Computation overhead. Tab. 5 shows the per-image latency on a single RTX A6000 GPU at 1024^2 input (100 forwards, mean±std). The decoder overhead specifically is ≈ 3.4 ms and is invoked *only* at intervention time, so routine deployment adds only +3.7%. Memory overhead is also small: +60 MB GPU memory (+1.8%) and 54 MB disk storage for all six SAE checkpoints. Relative overhead is stable across batch sizes 1–16 (Δ encoder 3.5–4.3%, Δ enc+dec 6.5–7.4%). Our full failure detection pipeline runs at around 5.66 img/s.

5 Related Work

Medical image segmentation for cancer detection. Deep learning has revolutionized cancer segmentation in medical imaging, with U-Net [56] and its variants [30] becoming the de facto standard for tumor delineation across multiple modalities including MRI, CT, and ultrasound. Recent advances have focused on improving segmentation accuracy through architectural innovations such as Vision Transformers (ViT), attention mechanisms, and multi-scale feature fusion. State-of-the-art methods like TransUNet [10], and Swin-UNETR [23] have

achieved remarkable performance on benchmark datasets for prostate cancer, pancreatic cancer, and other malignancies. However, these models still produce clinically significant failures, such as false positives that could lead to unnecessary biopsies and false negatives that miss critical lesions.

Failure detection in medical AI. Existing failure detection methods rely on output-based uncertainty signals: prediction confidence [24], entropy [24], or energy scores [44]. While these approaches can flag potential errors, they provide no insight into failure causes and suffer from poor generalization across datasets [52]. Quality control methods based on image quality assessment [8] or ensemble disagreement [36] similarly lack interpretability. We address this gap by detecting failures leveraging model's internal concepts.

Interpretability methods. Interpretability approaches include gradient-based attribution (GradCAM [58], IG [61]) and concept extraction via Sparse Autoencoders (SAEs) [14, 63]. MedSAE applications include cell analysis [16], disease detection [55], pathology [37], and report generation [1]. Concept-based methods (TCAV [32], CBMs [35,68], medical VLMs [19,33]) require manual annotation, while unsupervised discovery [3,43] targets natural images. Unlike prior work focused on post-hoc explanation, we leverage unsupervised SAE concepts for actionable failure detection in medical segmentation.

6 Conclusion, Limitations, and Future Work

Conclusion. We introduced an interpretable failure detection framework that extracts clinical concepts from segmentation model internals using Sparse Autoencoders. Unlike output-based methods, our approach explains why failures occur by identifying which anatomical concepts drive predictions. Experiments on prostate and pancreatic cancer segmentation show that concept activation patterns effectively distinguish true from false positives, enabling both failure detection and interpretable correction. Critically, our zero-shot transfer results demonstrate that concept-based representations generalize robustly across datasets where confidence-based methods collapse completely.

Limitations and future work. Our method currently captures concept occurrence rather than causal relationships. Understanding how concepts mechanistically interact to produce predictions, rather than merely co-occurring, would enable interventional debugging where clinicians could predict the effect of modifying specific features. Applying causal inference methods such as causal mediation analysis [29, 49, 64] or causal abstraction [20, 21, 67] could reveal these mechanistic pathways. Currently, our SAEs are trained and dedicated to each cancer type, requiring retraining for different anatomies. A promising direction is training a universal SAE across multiple cancers (brain, liver, lung, etc) to discover shared anatomical primitives while capturing cancer-specific concepts. This would enable direct comparison of reasoning processes across cancer types and more efficient clinical deployment. Additionally, extending beyond binary failure classification to identify specific failure modes (boundary errors, shape and size estimation failures) would provide more actionable clinical feedback.

Finally, one promising direction is to extend our framework to medical segmentation with missing modalities [4,31,47,48,65]. By learning modality-specific and shared internal concepts, the model could identify which information is missing and whether the remaining modalities are sufficient for a reliable prediction. These concept-level signals could support failure detection and guide adaptive fusion and modality imputation for robust clinical deployment.

Acknowledgment

This work is supported by the National Science Foundation under grant numbers CAREER 2340074, SLES 2416937, and III CORE 2412675, the National Institutes of Health under grant number R21CA301093, and the Department of Defense under grant number AFOSR FA9550-23-1-0494. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the supporting entities.

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. arXiv preprint (2024)
2. Adams, L.C., Makowski, M.R., Engel, G., Rattunde, M., Busch, F., Asbach, P., Niehues, S.M., Vinayahalingam, S., van Ginneken, B., Litjens, G., et al.: Prostate158-An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. Computers in Biology and Medicine **148**, 105817 (2022)
3. Arefin, M.R., Zhang, Y., Baratin, A., Locatello, F., Rish, I., Liu, D., Kawaguchi, K.: Unsupervised concept discovery mitigates spurious correlations. arXiv preprint arXiv:2402.13368 (2024)
4. Azad, R., Khosravi, N., Dehghanmanshadi, M., Cohen-Adad, J., Merhof, D.: Medical image segmentation on mri images with missing modalities: A review. arXiv preprint arXiv:2203.06217 (2022)
5. Bereska, L., Gavves, E.: Mechanistic interpretability for ai safety–a review. arXiv preprint arXiv:2404.14082 (2024)
6. 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. Transformer Circuits Thread (October 2023), `https://transformer-circuits.pub/2023/monosemantic-features/index.html` (accessed June 30, 2026)
7. Bussmann, B., Leask, P., Nanda, N.: Batchtopk sparse autoencoders. arXiv preprint arXiv:2412.06410 (2024)
8. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3535–3544 (2020)
9. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. Medical Image Analysis **98**, 103310 (2024)

10. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
11. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
12. Chen, Y., Li, T., Peng, X.: World knowledge in the weights: Reading concept circuits of vision transformers. In: Proceedings of the European Conference on Computer Vision (ECCV) (2026)
13. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
14. Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600 (2023)
15. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., et al.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science advances **8**(31), eabq6147 (2022)
16. Dasdelen, M.F., Lim, H., Buck, M., Götze, K.S., Marr, C., Schneider, S.: CytoSAE: Interpretable Cell Embeddings for Hematology . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15973. Springer Nature Switzerland (September 2025)
17. Gangemi, A., Bourgeois-Gironde, S., Mancini, F.: Feelings of error in reasoning—in search of a phenomenon. Thinking & Reasoning **21**(4), 383–396 (2015)
18. Gao, L., la Tour, T.D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J.: Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093 (2024)
19. Gao, Y., Gu, D., Zhou, M., Metaxas, D.: Aligning human knowledge with visual concepts towards explainable medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–56. Springer (2024)
20. Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., et al.: Causal abstraction: A theoretical foundation for mechanistic interpretability. Journal of Machine Learning Research **26**(83), 1–64 (2025)
21. Geiger, A., Lu, H., Icard, T., Potts, C.: Causal abstractions of neural networks. Advances in Neural Information Processing Systems **34**, 9574–9586 (2021)
22. González, C., Gotkowski, K., Fuchs, M., Bucher, A., Dadras, A., Fischbach, R., Kaltenborn, I.J., Mukhopadhyay, A.: Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. Medical image analysis **82**, 102596 (2022)
23. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
24. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
25. Hewitt, J., Geirhos, R., Kim, B.: We can't understand ai using our existing vocabulary. arXiv preprint arXiv:2502.07586 (2025)
26. Hindupur, S.S.R., Lubana, E.S., Fel, T., Ba, D.: Projecting assumptions: The duality between sparse autoencoders and concept geometry. arXiv preprint arXiv:2503.01822 (2025)

27. Hosny, A., Bitterman, D.S., Guthier, C.V., Qian, J.M., Roberts, H., Perni, S., Saraf, A., Peng, L.C., Pashtan, I., Ye, Z., et al.: Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. The Lancet Digital Health **4**(9), e657–e666 (2022)
28. Hou, J., Liu, S., Bie, Y., Wang, H., Tan, A., Luo, L., Chen, H.: Self-explainable ai for medical image analysis: A survey and new outlooks. arXiv preprint arXiv:2410.02331 (2024)
29. Imai, K., Keele, L., Tingley, D.: A general approach to causal mediation analysis. Psychological methods **15**(4), 309 (2010)
30. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
31. Karimijafarbigloo, S., Azad, R., Kazerouni, A., Ebadollahi, S., Merhof, D.: Mmcformer: Missing modality compensation transformer for brain tumor segmentation. In: Medical imaging with deep learning. pp. 1144–1162. PMLR (2024)
32. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
33. Kim, C., Gadgil, S.U., DeGrave, A.J., Omiye, J.A., Cai, Z.R., Daneshjou, R., Lee, S.I.: Transparent medical image ai via an image–text foundation model grounded in medical literature. Nature medicine **30**(4), 1154–1165 (2024)
34. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
35. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
36. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
37. Le, N.M., Shen, C., Patel, N., Shah, C., Sanghavi, D., Martin, B., Eng, A., Shenker, D., Padigela, H., Biju, R., et al.: Learning biologically relevant features in a pathology foundation model using sparse autoencoders. arXiv preprint arXiv:2407.10785 (2024)
38. Li, M., Zhang, D., He, Q., Liu, H., Zhou, C., Li, M., Zhou, X.: Bridging the semantic gap in medical image segmentation via multi-scale dependency and attention-guided enhancement. Scientific Reports **15**(1), 38064 (2025)
39. Li, T., Chen, Y., Ma, M., Peng, X.: Inside the visual mind: Neuroscience-motivated concept circuits for interpreting and steering vision transformers. In: Proceedings of the International Conference on Machine Learning (ICML) (2026)
40. Li, W., Zhou, X., Chen, Q., Lin, T., Bassi, P.R., Chen, X., Ye, C., Zhu, Z., Ding, K., Li, H., et al.: Pants: The pancreatic tumor segmentation dataset. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
41. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
42. Liu, H., Ding, N., Li, X., Chen, Y., Sun, H., Huang, Y., Liu, C., Ye, P., Jin, Z., Bao, H., et al.: Artificial intelligence and radiologist burnout. JAMA network open **7**(11), e2448714 (2024)

43. Liu, N., Du, Y., Li, S., Tenenbaum, J.B., Torralba, A.: Unsupervised compositional concepts discovery with text-to-image generative models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2095 (2023)
44. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020)
45. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1) (Jan 2024)
46. Ma, M., Li, T., Peng, Y., Lin, L., Beylergil, V., Zhao, B., Akin, O., Peng, X.: "Why Is There a Tumor?": Tell Me the Reason, Show Me the Evidence. In: Forty-second International Conference on Machine Learning (2025)
47. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18177–18186 (2022)
48. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 2302–2310 (2021)
49. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. Advances in neural information processing systems **35**, 17359–17372 (2022)
50. Ng, A.: Sparse autoencoder. CS294A Lecture notes **72**(2011), 1–19 (2011), `https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf` (accessed June 30, 2026)
51. Nguyen, K.X., Li, T., Peng, X.: Interpretable failure detection with human-level concepts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 26326–26334 (2025)
52. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems **32**, 13991–14002 (2019)
53. Peng, Y., Ma, M., Yao, Z., Peng, X.: Inside-out: Measuring generalization in vision transformers through inner workings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
54. Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.F.: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in medicine unlocked **47**, 101504 (2024)
55. Renzulli, R., Lepoutre, C., Cassano, E., Grangetto, M.: Medsae: Dissecting medclip representations with sparse autoencoders. arXiv preprint arXiv:2510.26411 (2025)
56. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
57. Saha, A., Bosma, J.S., Twilt, J.J., van Ginneken, B., Bjartell, A., Padhani, A.R., Bonekamp, D., Villeirs, G., Salomon, G., Giannarini, G., et al.: Artificial Intelligence and Radiologists in Prostate Cancer Detection on MRI (PI-CAI): An International, Paired, Non-Inferiority, Confirmatory Study. The Lancet Oncology (2024)
58. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

59. Shaharabany, T., Dahan, A., Giryes, R., Wolf, L.: Autosam: Adapting sam to medical images by overloading the prompt encoder. arXiv preprint arXiv:2306.06370 (2023)
60. Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al.: Open problems in mechanistic interpretability. arXiv preprint arXiv:2501.16496 (2025)
61. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
62. Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J.D., Dahl, A.B.: Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: International conference on medical image computing and computer-assisted intervention. pp. 824–833. Springer (2020)
63. Thasarathan, H., Forsyth, J., Fel, T., Kowal, M., Derpanis, K.G.: Universal sparse autoencoders: Interpretable cross-model concept alignment. In: International Conference on Machine Learning (2025)
64. Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., Shieber, S.: Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems **33**, 12388–12401 (2020)
65. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15878–15887 (2023)
66. Wessel, J.R.: Error awareness and the error-related negativity: evaluating the first decade of evidence. Frontiers in human neuroscience **6**, 88 (2012)
67. Wu, Z., Geiger, A., Icard, T., Potts, C., Goodman, N.: Interpretability at scale: Identifying causal mechanisms in alpaca. Advances in neural information processing systems **36**, 78205–78226 (2023)
68. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
69. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., et al.: A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nature methods **22**(1), 166–176 (2025)
70. Zhao, Z., Koishekenov, Y., Yang, X., Murray, N., Cancedda, N.: Verifying chain-of-thought reasoning via its computational graph. arXiv preprint arXiv:2510.09312 (2025)