"Why Is There a Tumor?": Tell Me the Reason, Show Me the Evidence

Mengmeng Ma [1] **Tang Li** [1] **Yunxiang Peng** [1] **Lu Lin** [2] **Volkan Beylergil** [2] **Binsheng Zhao** [2] **Oguz Akin** [2] **Xi Peng** [1]

Abstract

Medical AI models excel at tumor detection and segmentation. However, their latent representations often lack explicit ties to clinical semantics, producing outputs less trusted in clinical practice. Most of the existing models generate either segmentation masks/labels (localizing where without why) or textual justifications (explaining why without where), failing to ground clinical concepts in spatially localized evidence. To bridge this gap, we propose to develop models that can justify the segmentation or detection using clinically relevant terms and point to visual evidence. We address two core challenges: First, we curate a rationale dataset to tackle the lack of paired images, annotations, and textual rationales for training. The dataset includes 180K image-mask-rationale triples with quality evaluated by expert radiologists. Second, we design rationale-informed optimization that disentangles and localizes fine-grained clinical concepts in a self-supervised manner without requiring pixel-level concept annotations. Experiments across medical benchmarks show our model demonstrates superior performance in segmentation, detection, and beyond. The code is available at https://github.com/deep-real/MedRationale.

1. Introduction

Magnetic Resonance Imaging (MRI) interpretation workflow typically begins with analyzing anatomical and functional sequences to identify deviated visual patterns. These patterns are then mapped to standardized clinical lexicons, ensuring consistency in reporting. Structured reports explicitly link these clinical terms to corresponding visual evidence, guiding clinical action. AI is increasingly integrated into this workflow, assisting radiologists in detecting and segmenting tumors (Ozer et al., 2010; Zhao et al., 2024). Data-driven methods (Ronneberger et al., 2015; Isensee et al., 2021) enable AI models to achieve remarkable accuracy. However, most existing models process images via statistical pattern recognition, mapping pixels to labels through latent representations that lack explicit ties to medical semantics. Consequently, these models often produce outputs that lack clinical interpretability. For instance, they can localize anomalies using segmentation masks without explaining why regions are suspicious. This raises a critical question: *Can AI models justify their prediction using clinical terms and point to visual evidence?*

Methods		Predictions	Prediction Rationales	
			Clinical Concepts	Visual Evidence
Seg. Model	UNet	Seg. Mask	✗	✗
	SAMed			
XML	GradCAM	✗	✗	✓ Category-level
	Concept Bottleneck	Category	✓	✗
MLLM	OpenAI o1	Category	✓	✗
Ours		**Seg. Mask**	✓	✓ **Concept-level**

Figure 1: In contrast to existing methods, our proposed method not only conducts accurate predictions, but can also provide its prediction rationales using clinical concepts supported by valid visual evidence.

There are attempts to enhance models to perform beyond segmentation and detection, yet they often fall short of clinical needs. The large-language models trained on radiology reports (Yang et al., 2024; Li et al., 2024a) can generate textual justifications for their diagnosis, but these explanations often lack spatial grounding, making it difficult for radiologists to verify whether key findings correspond to the correct anatomical regions. Similarly, concept bottleneck models (Gao et al., 2024; Patrício et al., 2023) can justify the prediction using clinical concepts, yet fail at associating the concepts with correct pixels (Margeloiu et al., 2021). Conversely, models focused on visual localization—such as

[1]Department of Computer & Information Sciences, University of Delaware, Newark, DE, USA [2]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. Correspondence to: Xi Peng <xipeng@udel.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

those employing saliency maps (Selvaraju et al., 2020)—can highlight pixels driving predictions, yet they lack explicit clinical semantics, offering no textual rationale to explain why a region appears suspicious. Vision-language models pretrained on image-text pairs enable linking global image features with captions but struggle to disentangle fine-grained concepts, such as distinguishing "capsular invasion" from "benign hyperplasia" within the same gland.

Developing models that justify predictions and point to visual evidence faces two major challenges: (1) Data Challenge: Training such models typically require paired predictions and rationales, but the paired data are largely missing. Most existing public medical benchmarks (Saha et al., 2024; Wang et al., 2017; de Verdier et al., 2024) provide imaging data and lesions or anatomy annotations but lack the corresponding textual rationales. (2) Optimization challenge: Grounding clinical terms to visual evidence requires fine-grained image-concept alignment at the pixel level. Manual labeling of concept-pixels pairs is prohibitively expensive, especially for medical images. This raises the challenge of optimizing the model for concepts localizing without having the ground truth label.

To bridge the data gap, we curate a first-of-its-kind rationale dataset, providing 180K high-quality textual justifications for 1.5K MRI scans. Developed in collaboration with six radiologists over 150+ hours of structured annotation, each rationale captures the detailed reasoning process behind clinical assessments. To ensure reliability, rationales underwent dual-expert validation across key metrics such as factual accuracy and comprehensiveness. Unlike conventional datasets (Adams et al., 2022; Saha et al., 2024) limited to image-mask pairs, our work pioneers image-mask-rationale triples, explicitly linking raw imaging data, expert annotations, and the underlying clinical logic.

To fill the optimization gap, we design rationale-informed optimization, a self-supervised framework that disentangles and localizes clinical concepts without pixel-level annotations. We evaluate the proposed method on multiple medical benchmarks. For segmentation and detection, our method achieves +8.8% DSC and +5.3% AP over the state-of-the-art models. For rationale correctness, our model significantly improves visual evidence localization, achieving 0.403 Relevant Mass Accuracy (RMA ↑) (Arras et al., 2022) compared to the baseline's 0.010 RMA.

2. Problem Formulation

In this section, we provide our definition of rationale and the problem formulation.

Definition 2.1 (Rationale)**.** *We define rationales to be a set of abstract concepts $\{c_k\} \subset \mathcal{C}$ that can sufficiently justify a prediction, supported by valid visual evidence.*

Definition 2.2 (Visual evidence)**.** *Let $x \in \mathcal{X}$ be an input image with ground truth label $y \in \mathcal{Y}$. The valid visual evidence for c_k is the subset of pixels in x that are most relevant to c_k. Let $\Omega \subseteq \mathbb{R}^2$ denote the pixel space, and let $\mathcal{G} : \mathcal{X} \times \mathcal{C} \to 2^\Omega$ map (x, c_k) to $\mathcal{G}(x, c_k) \subset \Omega$. Then, $\mathcal{G}(x, c_k)$ is the* visual evidence *for c_k.*

Taking prostate tumor diagnosis as an example, the rationales could be language descriptions of established clinical concepts $\{c_k\}$ (e.g., hypointense signal intensity, circumscribed margins, or lenticular shape), where each clinical concept c_k should be supported by corresponding image regions $\mathcal{G}(x, c_k)$ to ensure visual validity.

Given a data point (x, y) drawn from distribution $P(\mathcal{X}, \mathcal{Y})$. Let $f \in \mathcal{F}$ be a predictive model, ℓ be a task-specific loss, and $h(\cdot)$ be a method that maps c_k to a region in x depending on f. We propose to solve the following optimization problem to obtain a rationale model:

$$\begin{aligned} &\min_{f \in \mathcal{F}} \mathcal{R}(f) := \mathbb{E}_{(x,y)\sim P(\mathcal{X},\mathcal{Y})}\big[\ell(f(x), y)\big], \\ &\text{s.t. } \; h(x, c_k; f) = \mathcal{G}(x, c_k), \forall c_k \in \mathcal{C}. \end{aligned} \tag{1}$$

Solving the problem is non-trivial, as both the concepts $\{c_k\}$ and $\mathcal{G}(\cdot)$ are not available. In medical image analysis, some efforts have been made to provide textual rationales along with corresponding pixel-level annotations (Tschandl et al., 2018; Daneshjou et al., 2022; Bannur et al., 2024). However, these approaches are typically constrained to small-scale datasets and are impractical for large-scale applications due to the high cost of fine-grained annotations. To address this challenge, we present our approach to automatically obtain the concepts in Section 3 and demonstrate how to train a rationale model without access to $\mathcal{G}(\cdot)$

3. Rationale Dataset Curation

This section describes our strategies to obtain the textual description of "why" and "how" behind experts' reasoning steps, referred to as *rationale data*. Using prostate cancer (PCa) as an example (Sekhoacha et al., 2022), we describe how to: represent domain knowledge in a structured format, source rationale data, convert it into AI-ready annotations using domain knowledge, and ensure data quality. Importantly, our data curation procedures are adaptable to other cancers, such as breast (Cozzi et al., 2024), lung (Mehta et al., 2017), and liver (Chernyak et al., 2018).

How to represent domain knowledge? To systematically capture the why and how behind Prostate Imaging Reporting and Data System (PI-RADS) scoring, we formalize expert reasoning into a PI-RADS Decision Tree (PDT). Developed with radiologists and aligned with PI-RADS v2.1 guidelines, the PDT maps multi-step clinical logic into a traceable hierarchy (Figure 2(a)). Each node represents a reasoning step, ensuring structured, auditable rationale generation.

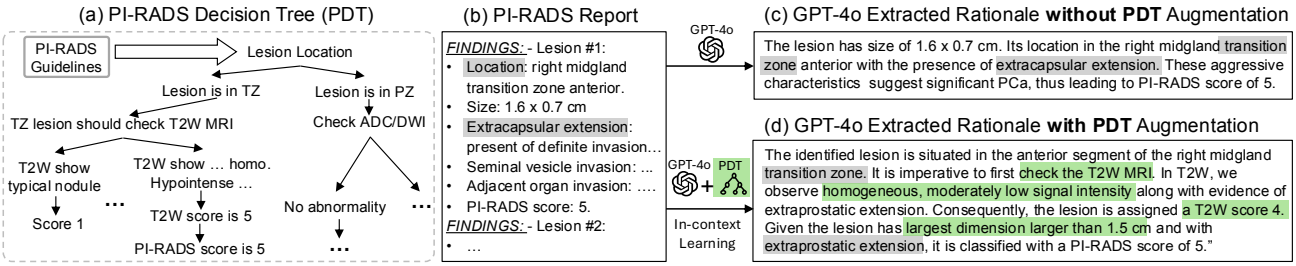

Figure 2: Overview of rationale data curation. **PI-RADS Decision Tree**: the structured representation of domain knowledge. **PI-RADS report**: sources that provide rationale data. Extracted radiologist rationales with and without PDT. Text highlighted in gray represents information existing in the clinical PI-RADS report. Text highlighted in green indicates augmented information by PDT.

Where to source the rationale data? Clinicians routinely document their observations, interpretations, and professional judgments in clinical reports, providing a valuable source to obtain the reasoning process behind medical decisions. For example, reports from the Prostate Imaging Reporting and Data System (PI-RADS) synthesize radiologists' findings, including tumor size, location, MRI signal characteristics, and the likelihood of clinically significant cancer (PI-RADS category; see Figure 2(b)). However, public data repositories (Saha et al., 2024; Adams et al., 2022; Fedorov et al., 2018; Nguyen et al., 2022; Antonelli et al., 2022) typically provide medical imaging scans and annotations but *without accompanying clinical reports*.

To bridge this gap, we collaborated with radiologists to annotate the public data to create MRI and PI-RADS report pairs. We use 1,500 prostate MRI cases from PI-CAI challenge (Saha et al., 2024). Each case includes three MRI sequences: T2W, ADC, and DWI. These sequences have already been annotated with gland, zone, and lesion segmentation masks. For 408 tumor-positive cases, MSK radiologists assess the MRIs and segmentation masks to generate PI-RADS reports.

How to extract rationale data efficiently? The variability inherent in clinical PI-RADS reports could potentially lead to the extraction of incomplete radiologist rationale, often lacking intermediate reasoning steps. We propose to automatically extract detailed rationales from the PI-RADS reports leveraging GPT-4o (Achiam et al., 2023) and PDT. As illustrated in Figure 2(a), we will input 10 example pairs consisting of PI-RADS reports alongside corresponding gold-standard rationales (detailed explanations provided by the radiologist) into GPT-4o for *in-context learning* (Dong et al., 2022). Specifically, GPT-4o will utilize these examples to learn how to apply PDT and augment PI-RADS reports by filling in the missing reasoning steps. Our full prompt is available in Appendix C.

The results in Figures 2(c) and 2(d) show that without PDT augmentation, the extracted rationale missing intermediate reasoning steps. In contrast, the PDT-augmented rationale contains detailed reasoning steps left implicit in the clinical PI-RADS reports. These findings validate the effectiveness of PDT in enabling GPT-4o to generate rationales that remain within factually valid boundaries while effectively capturing the radiologists' detailed reasoning steps. This also aligns with recent broader research findings that confirm GPT-4o's expertise at an expert level in both commonsense reasoning and specialized medical domain knowledge (Bubeck et al., 2023; Liu et al., 2023).

Rationale dataset and statistics. Leveraging the proposed automatic method, we are able to extract rationales from all annotated clinical PI-RADS reports, leading to the first rationale dataset for prostate cancer. The dataset contains 1,500 3D triples (MRI sequence, masks, rationales), decomposed into 36,000 2D triples (MR image, mask, rationale). To further enhance the dataset scale, we generated five GPT-4o paraphrases per rationale, yielding 180,000 total 2D triples.

Can we trust the extracted rationale? We conduct human evaluations of the extracted rationale. The experiment details and metrics are described in Section 5.2. Results show that our rationale data are coherent, consistent, comprehensive, and highly factual accurate.

4. Rationale-Informed Optimization

In this paper, a rationale model should be able to link textual semantics (clinical concepts) with visual patterns (anatomical regions), calling for a shared space where similarity reflects semantic relevance. A straightforward approach is to use the image-text pairs from our rationale dataset and contrastive learning (Radford et al., 2021) to learn a shared vision-language embedding space.

Main objective. Given an image-text-label triple (I, T, M) sampled from $P(\mathcal{I}, \mathcal{T}, \mathcal{M})$, a model $f = (f_{\text{img}}, f_{\text{txt}}, f_{\text{mask}}) \in \mathcal{F}$ that generate image $f_{\text{img}}(I)$ and text $f_{\text{txt}}(T)$ embeddings, as well as segmentation masks $\hat{M} = f_{\text{mask}}(f_{\text{img}}(I))$. Let ℓ_{seg} be the Dice loss (Saha et al.,

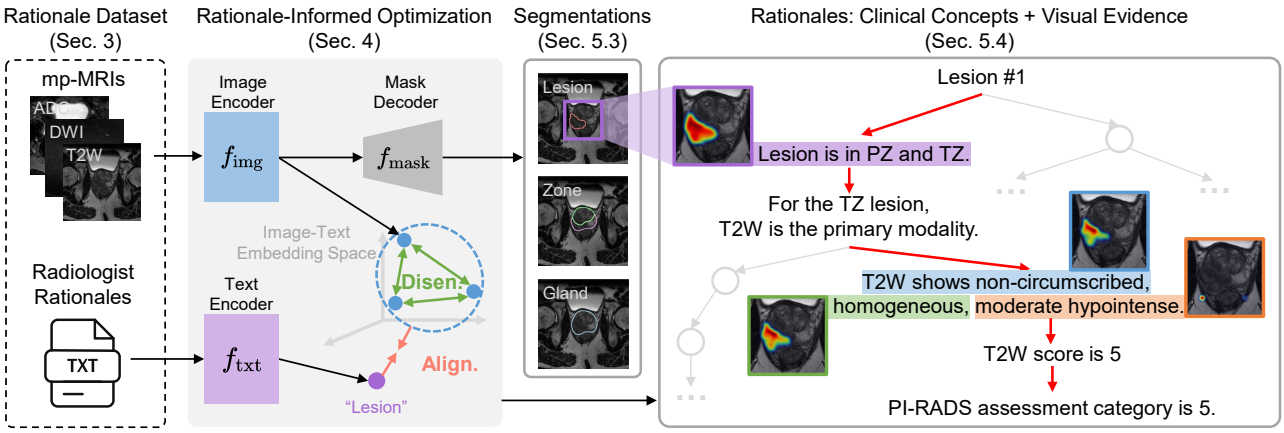

Figure 3: Overview of the proposed method. **Rationale Dataset:** In Section 3, we curate a first-of-its-kind dataset for prostate tumor segmentation that offers textual rationales annotated by expert radiologists. **Rationale-Informed Optimization:** In Section 4, we propose a new optimization method that enables the model to localize the visual evidence of concepts without manual annotations. **Segmentations:** Our model can provide segmentation masks for Lesions, Zones, and Gland. **Rationales:** Our model can provide its prediction rationales using clinical concepts based on valid visual evidence.

2024) and ℓ_{align} be the InfoNCE loss (Oord et al., 2018). We augment Equation 1's objective with the additional objective of learning a shared embedding space:

$$\min_{f\in\mathcal{F}} \mathcal{R}(f) := \mathbb{E}_{(I,T,M)}\left[\ell_{\text{seg}}(f_{\text{mask}}(f_{\text{img}}(I)), M)\right] + \mathbb{E}_{(I,T,M)}\left[\ell_{\text{align}}(f_{\text{img}}(I), f_{\text{txt}}(T))\right]. \tag{2}$$

We conducted training using Equation 2 to learn a shared vision-language space and found that incorporating additional text data improves the model's segmentation and cancer detection performance (see results in Table 1). The model also demonstrates better zero-shot generalization performance compared with image-only models (Table 2). However, optimizing this equation alone can not guarantee accurate localization of the concepts (Figure 7).

Disentanglement constraint. To accurately localize the concept, we propose to disentangle the concept's heatmap. In the shared embedding space, one can link a concept c_k to its most relevant pixels by calculating a heatmap $h(x, c_k; f)$ that compare the similarity between image and text embeddings: $f_{\text{img}}(I) \in \mathbb{R}^{H^{'}\times W^{'}\times D}$, $f_{\text{txt}}(c_k) \in \mathbb{R}^{D}$, $h(x, c_k; f) = \text{upsample}(f_{\text{img}}(x) \cdot f_{\text{txt}}(c_k)^T)$, where $\text{upsampel}(\cdot)$ resizes the heatmap to match the input image resolution. Our insight is that clinically different concepts should highlight different regions in the image. This leads to the following disentanglement constraint:

$$\min_{f\in\mathcal{F}} \mathcal{R}(f) \;\text{ s.t. }\; \mathcal{D}(h(x, c_k; f), h(x, c_k'; f)) \geq \epsilon_1. \tag{3}$$

Where $\mathcal{D}(\cdot,\cdot)$ is a distance metric such as L2 distance. However, naively optimizing the Equation 3 could lead to trivial solutions, where heatmaps of different concepts highlight non-overlapping but random regions in the image.

Localization constraint. To avoid a trivial solution, we introduce additional localizationi constrain. Our idea is that different concepts describing the same anatomical structure should have heatmaps highlighting the same region. Our full objective is as follows:

$$\min_{f\in\mathcal{F}} \mathcal{R}(f) \;\text{ s.t. }\; \mathcal{D}(h(x, c_k; f), h(x, c_k'; f)) \geq \epsilon_1, \quad \mathcal{D}(\sum h(x, c_k; f), M) \leq \epsilon_2 \tag{4}$$

To simplify the optimization, we leverage the KKT condition (Boyd, 2004; Wright, 2006; Qiao & Peng, 2023) and Lagrange multipliers to convert this constraint optimization problem into the unconstrained problem.

Inference. At test time, the model generates rationales by traversing over the PI-RADS decision tree (Figure 2(a)) conditioned on the shared embedding space. We first encode image I into latent embedding $f_{\text{img}}(I)$. For each tree node (e.g., lesion presence, location, margins), we retrieve the rationale with the highest similarity to $f_{\text{img}}(x)$: $\hat{c} = \text{argmax}_{c\in\mathcal{C}_{\text{node}}} \text{mean}(f_{\text{img}}(I) \cdot f_{\text{txt}}(c)^T)$. Note that subsequent nodes condition on prior selections. For example, MRI signal characteristics are inferred only after lesion location is determined, as signal patterns vary anatomically.

5. Experiments

5.1. Datasets, Baselines, Metrics, and Implementation

Datasets. We conduct experiments on our curated rationale dataset and two standard benchmarks (Prostate158 (Adams

et al., 2022) and MSD-Prostate (Antonelli et al., 2022)), to assess the model's segmentation and detection performance, zero-shot generalization capability, and rationale generation. **Rationale Dataset** comprises 1,500 mp-MRI scans with mask annotations at the prostate gland, zone, and lesion levels, sourced from the publicly available PI-CAI dataset (Saha et al., 2024). We enrich the image data with PI-RADS reports and rationale descriptions that capture radiologists' clinical judgment, leading to 180k image-mask-rationale. We split the dataset into training and testing sets using an 80:20 ratio. **Prostate158** (Adams et al., 2022) includes 158 mp-MRIs with T2W, DWI and ADC sequences. **MSD-Prostate** (Antonelli et al., 2022) includes 48 mp-MRIs with T2W, ADC, and DWI sequences. The task is to segment the peripheral and the transitional zones. We conducted training only on rationale dataset and directly evaluated the model on Prostate158 and MSD-Prostate. Detailed descriptions of the datasets are available in Appendix B.

Baselines. We compare with U-Net (Ronneberger et al., 2015), the most referenced baselines in the literature. Additionally, we include comparisons with ITUNet (Kan et al., 2022), Swin UNETR (Hatamizadeh et al., 2021), CSwin U-Net (Li et al., 2023b), and SSL CSwin U-Net (Li et al., 2023b), which are popular 2D and 3D models for prostate cancer segmentation and detection. Finally, we evaluate against foundation models, including SAM-based models (MA-SAM (Chen et al., 2024) and SAMed (Zhang & Liu, 2023)) and large vision-language models such as Biomed-Parse (Zhao et al., 2024).

Evaluation metrics. We use the Dice Similarity Coefficient (DSC) to evaluate the segmentation performance. Following the common practice for tumor detection (Saha et al., 2024), we use Area Under the Receiver Operating Characteristic (AUROC) for patient-level detection and Average Precision (AP) for lesion-level detection.

Implementation details. Our model consists of an image encoder, a text encoder, and a mask decoder (see Fig.3). The image encoder and mask decoder use pre-trained weights from MedSAM (Ma et al., 2024a), while the text encoder uses weights from BiomedCLIP (Zhang et al., 2024). For training, we use Focal Loss (Lin, 2017) with $\alpha = 0.97$ and $\gamma = 2$, and optimize the model using the Adam (Kingma, 2014) with a weight decay of 0.001 and an exponential learning rate scheduler. Additional implementation details on data pre-processing, post-processing, and hyperparameters are provided in Appendix D.

5.2. Human Evaluation of Rationale Dataset Quality

Takeaway: *It is possible to effectively and automatically extract high-quality rationales from clinical reports.*
We present the human evaluation of rationale dataset quality. Two expert radiologists from MSK, each with over 10 years of professional experience, reviewed our extracted textual rationales in terms of ① coherence, ② consistency, ③ comprehensiveness, and ④ factual accuracy. The results in Figure 4 indicate that our rationale data are coherent, consistent, comprehensive, and highly factual accurate.

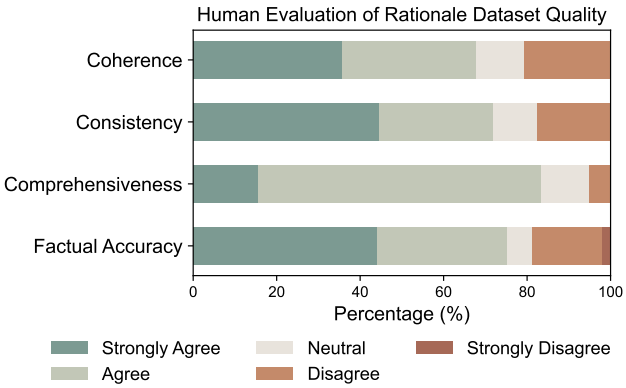

Figure 4: Human evaluation fo rationale quality. Expert radiologists evaluated the textual rationales across four criteria: coherence, consistency, comprehensiveness, and factual accuracy. Each criterion was rated on a scale ranging from Strongly Agree to Strongly Disagree. The rationales received *high ratings across all criteria*, with strong scores in factual accuracy, highlighting the overall good quality of the rationale dataset. Each expert's rating and the experiment settings are available in Appendix A.

① *Coherence* refers to rationales that present information in a well-structured, logically smooth manner, ensuring that each sentence contributes to a clear and unified description of tumor segmentation. Figure 4 shows that more than 70% rationales were coherent. ② *Consistency* measures whether the rationale is free from self-contradictions and accurately aligns the lesion's characteristics with the appropriate PI-RADS category. As shown in Figure 4, over 70% of the rationales were rated as consistent, with more than 40% rating as "Strongly Agree". ③ *Comprehensiveness* evaluates rationale including imaging features such as lesion size, location, and MRI findings (e.g., non-circumscribed, moderate hypointensity) and explains how they contribute to the segmentation decision. Figure 4 shows that less than 6% of the rationales were rated as not comprehensive. Noted that the "strongly agree" percentage for comprehensiveness is lower than other metrics. This is primarily due to its stricter requirement of mentioning ALL critical details. However, when combining the "strongly agree" and "agree" categories, comprehensiveness achieves a higher overall percentage than other metrics. This indicates that while some rationales might miss minor details, most successfully include nearly all relevant imaging features. ④ *Factual Accuracy* assesses whether the rationale provides correct information aligned with established clinical standards, e.g.

Table 1: Segmentation (DSC) and cancer detection (AP, AUROC) on the rationale dataset. The reported results are obtained from our implementation and averaged over three independent runs, with † indicating numbers from the original paper. The best scores are highlighted in bold. *Our model outperforms existing baselines in lesion segmentation and cancer detection.*

Method	Gland DSC (↑)	Zone DSC (↑)	Lesion DSC (↑)	Average	AP (↑)	AUROC (↑)	Average
U-Net	0.882±0.005	0.880±0.007	0.504±0.011	0.755	0.518±0.055	0.855±0.028	0.687
ITUNet	0.869 ±0.012	0.858±0.009	0.505±0.006	0.744	0.407±0.033	0.779±0.012	0.593
Swin UNETR	**0.898**±0.020	0.860±0.004	0.431±0.017	0.737	0.513±0.044	0.852±0.020	0.682
CSwin U-Net†	–	–	–	–	0.543±0.042	0.880±0.013	0.712
SSL CSwin U-Net†	–	–	–	–	0.545±0.060	0.888±0.010	0.717
MA-SAM	0.888±0.015	0.813±0.010	0.326±0.018	0.676	0.413±0.058	0.802±0.023	0.608
SAMed	0.807±0.008	0.821±0.006	0.408±0.015	0.679	0.384±0.073	0.807±0.032	0.596
Ours	0.843±0.005	**0.880**±0.007	**0.552**±0.011	**0.758**	**0.574**±0.039	**0.888**±0.009	**0.731**

Table 2: Zero-shot DSC, AP, AUROC on Prostate158 and MSD-prostate. The model, trained on the rationale dataset, is evaluated directly on the two new datasets. Prostate158 collected in a Germany hospital presents notable distribution shifts to the rationale dataset due to differences in demographics or device configurations. The reported results are obtained from our implementation, with † indicating numbers from the original paper. The best performances are highlighted in bold and Δ represents the improvement over the second-best model. Our model consistently *outperforms existing ones in segmenting anatomical structures, delineating lesions, and detecting cancers*, demonstrating strong generalization capability.

	Prostate158						MSD-Prostate
Method	Zone DSC (↑)	Lesion DSC (↑)	Average	AP (↑)	AUROC (↑)	Average	Zone DSC (↑)
U-Net	0.767	0.348	0.558	0.491	0.784	0.638	0.709
ITUNet	0.715	0.383	0.549	0.441	0.760	0.600	0.689
Swin UNETR	0.724	0.362	0.543	0.357	0.731	0.544	0.535
CSwin U-Net†	–	–	–	0.363	0.772	0.568	–
SSL CSwin U-Net†	–	–	–	0.451	0.790	0.621	–
MA-SAM	0.709	0.220	0.465	0.392	0.816	0.604	0.690
SAMed	0.680	0.382	0.531	0.401	0.756	0.579	0.684
BiomedParse	0.356	0.187	0.271	0.002	0.423	0.213	0.461
Ours	**0.787**	**0.426**	**0.607**	**0.565**	**0.835**	**0.700**	**0.717**
Δ	**+0.020**	**+0.043**	**+0.049**	**+0.074**	**+0.019**	**+0.062**	**+0.008**

PI-RADS guidelines. Figure 4 results indicate that fewer than 2% rationales are considered highly inaccurate. Over 75% of the rationales were rated as accurate, with more than 43% receiving a "Strongly Agree" rating.

5.3. Evaluation of Model Prediction

Takeaway: *Learning with additional textual rationales improves models' prediction accuracy and generalization.* We evaluated the segmentation and cancer detection performance across multiple datasets, with results summarized in Tables 1 and 2. We observed that the additional textual rationale data helps to build models that perform better in segmenting anatomical structures and lesions, as well as detecting cancers. From Table 1, our method demonstrates a 12% improvement over SAM-based models and a 1.2% improvement over the classic model in anatomical and lesion segmentation. Additionally, it outperforms the best baseline by 2.0% in cancer detection. Table 2 further highlights the model's robust zero-shot performance, demonstrating consistent superiority on new and unseen datasets. On Prostate158, our model surpasses the best-performing baselines by 8.8% in segmentation and 9.7% in cancer detection. Similarly, in the MSD-Prostate segmentation task, our model outperforms the second-best approach by 1.1%, underscoring its exceptional generalization capabilities. We present qualitative segmentation masks in Figure 5. Our method delivers high-quality segmentation results even under distribution shifts, demonstrating strong potential for real-world clinical applications.

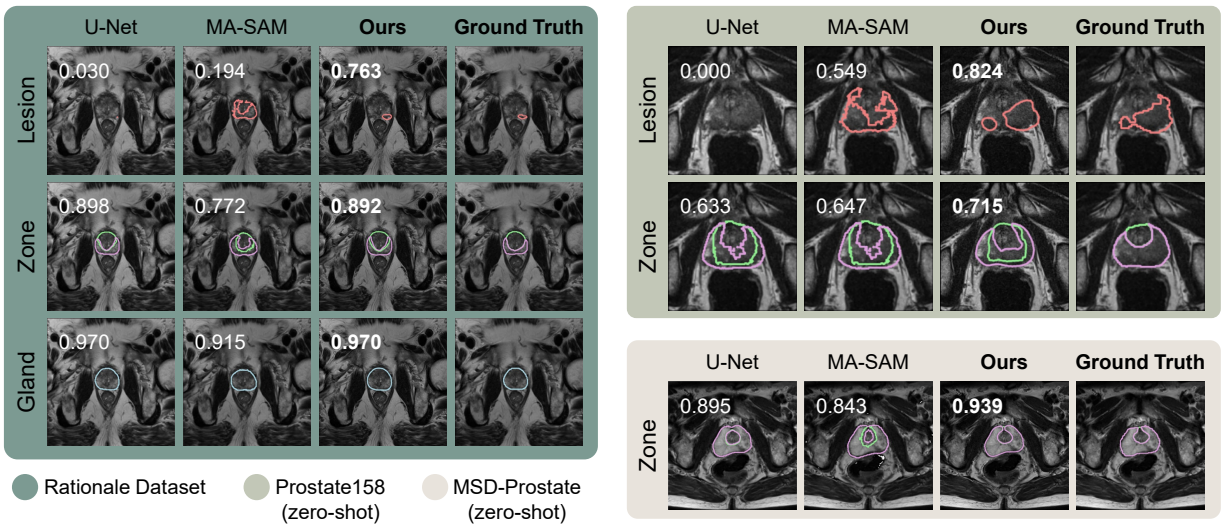

Figure 5: Visualization of segmentation results. All models are trained on our Rationale Dataset's training set and evaluated on its testing set. We further evaluate the zero-shot performance of the models using Prostate158 and MSD-prostate datasets, which are unseen during training. Notably, MRIs in Prostate158 differ significantly from MRIs in training distribution. Despite this, our model achieves superior performance across all datasets, demonstrating strong *generalization capability.*

Table 3: Rationale accuracy on the rationale dataset. We report results for slice- and patient-level tumor classification using the model's rationale. Higher values indicate better performance. We compare with baselines using the model's logits or masks for tumor classification.

	Slice-level			Patient-level		
Method	Precision	Recall	F1	Precision	Recall	F1
Logits	0.563	0.305	0.396	0.508	0.426	0.464
Mask	0.236	0.874	0.372	0.364	0.927	0.523
Ours	0.633	0.513	**0.567**	0.857	0.412	**0.583**

5.4. Evaluation of Model Rationale

Takeaway: *It is feasible to build a rationale model without human annotation at the pixel level.*

We evaluate the accuracy of our model's prediction rationales and its ability to localize the rationales within the images. For comparison, we also tested OpenAI's o1 model for rationale generation and localization.

Rationale and visual evidence. Our model, during inference, generates rationales for a given image by performing image-to-text retrieval from a pool of concepts that describe the tumors in the learned vision-language embedding space. For each retrieved rationale, heatmaps localizing the most relevant pixels are generated by computing the similarity between patch-level embeddings and textual rationale embeddings. *Rationale accuracy* is quantified by verifying whether the generated rationales are tumor-related and *visual evidence accuracy* is measured using Relevant Mass Accuracy (RMA) (Arras et al., 2022). RMA measures the ratio of the sum of heatmap values within the ground truth mask to the total heatmap values, defined as:

$$\mathrm{RMA}(H, M) = \frac{\sum_{(i,j)\in \mathrm{ROI}} H_{i,j} \cdot M_{i,j}}{\sum_{(i,j)} H_{i,j}}, \qquad (5)$$

where H is the heatmap, M is the binary ground truth mask, and ROI represents the spatial indices where $M_{ij} = 1$. Results in Tables 3 and 4 demonstrate that our model significantly outperforms the baseline in both rationale accuracy and localization.

Table 4: Visual evidence accuracy on rationale dataset, Prostate158, and MSD-Prostate. The evaluation measures how accurately the model localizes rationales in the images. Results are compared with a variation of our model, which is optimized without the disentangle and localization constraint (w/o constraint). Our model demonstrates a *significant improvement in rationale localization.*

Method	Rationale Dataset	Prostate158	MSD-Prostate
	Visual evidence (RMA ↑)		
w/o constraint	0.010	0.024	0.012
Ours	**0.403**	**0.330**	**0.497**
	Segmentation (DSC ↑)		
w/o constraint	**0.758**	0.607	0.717
Ours	0.746	0.589	0.687

Comparison with OpenAI's o1. Existing MLLMs, such as OpenAI's o1, are also capable of generating textual justifications for their predictions. To evaluate rationale generation,

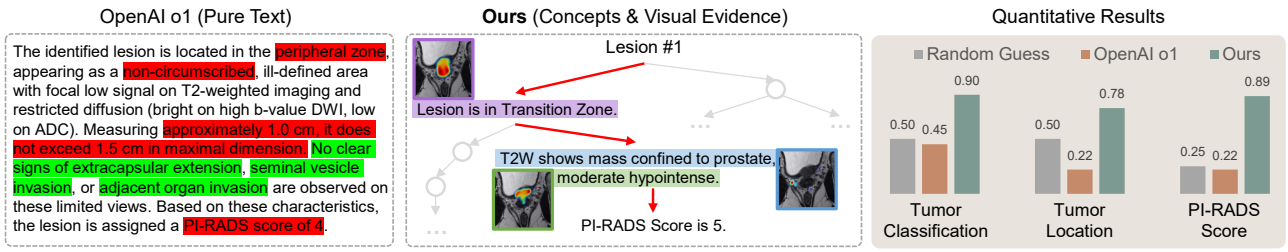

Figure 6: Comparison between OpenAI o1 and our model. Qualitatively, o1 provides only a textual Chain-of-Thought, it might be wrong about critical concepts (highlighted in rad). In contrast, our model offers prediction rationales using clinical concepts supported by valid visual evidence. Quantitatively, o1 performs worse than random guessing in Tumor Classification, Location, and PI-RADS Score prediction, whereas our model achieves significantly higher accuracy.

Table 5: Ablation on training without any constraints (w/o constraint), with disentangle constraint (w/ disen.), and with both constraints (ours) on the rationale dataset.

	disentangle	localize	Segmentation	Visual Evidence
w/o consraint			0.758	0.010
w/ disen.	✓		0.747	0.024
Ours	✓	✓	0.746	**0.403**

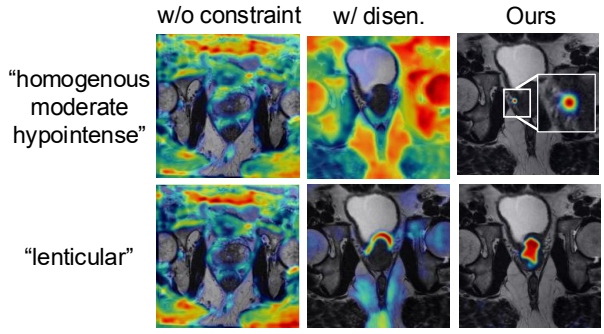

Figure 7: Visualization of visual evidence. *Using the disentanglement constraint alone will lead to trival solutions.*

we tested o1 on 20 selected MRI scans, containing 9 positive and 11 negative MRIs. Figure 6 presents a comparison of the generated rationales. Our model demonstrates superior performance, providing accurate rationales that are consistently supported by valid visual evidence. In contrast, o1 shows lower rationale accuracy and cannot localize the rationales in the image.

5.5. Ablation on the Constraints

Effects of the constraint. We conduct ablation on the contribution of different constraints in terms of segmentation accuracy and rationale localization. Quantitative and qualitative results are provided in Table 5 and Figure 7, respectively. The results shows that both disentangle and localization are indispensable for accurate concept localization. Additionally, we provide analysis of our model's failure case in Appendix E.

6. Related Work

Medical image segmentation. Recent advances in medical image segmentation, such as nnU-Net (Isensee et al., 2021), UNETR (Hatamizadeh et al., 2022), Swin-UNETR (Hatamizadeh et al., 2021), and MedSAM (Ma et al., 2024a), have achieved state-of-the-art performance in lesion and organ segmentation. However, most of these models produce only segmentation masks or labels, offering no insight into why a region is clinically significant. In contrast, our framework not only generates precise segmentation masks but also justifies predictions with clinical terms and points to anatomically localized evidence

Medical image analysis beyond accuracy. Prior efforts to enhance medical AI interpretability fall into three categories: First, *concept-based models* (e.g., CBMs (Koh et al., 2020), ProtoPNet (Chen et al., 2019)) learn the mapping from latent representation to high-level concepts but fail to localize them anatomically (Margeloiu et al., 2021). Second, *explainable methods* (e.g., GradCAM (Selvaraju et al., 2020; Li et al., 2023a)) highlight regions influencing predictions but omit semantic links to clinical terms. Third, *vision-language models* (e.g., BiomedParse (Zhao et al., 2024), multimodal models (Zhang et al., 2024; Ma et al., 2021; Wang et al., 2022; Ma et al., 2022)) align image-text pairs globally but struggle to disentangle fine-grained concepts (Li et al., 2024c;b). Multimodal large language models (e.g., LLaVA-Med (Li et al., 2024a)) trained on clinical content are capable of generating textual rationales but lack spatial grounding or segmentation capabilities. Additionally, while existing datasets (e.g., MIMIC-CXR (Johnson et al., 2019)) provide images and clinical reports, they often lack detailed descriptions of rationales. There are AI models try to provide language justification for their and grounded each sentence to the input image with a bounding box (Bannur et al., 2024; Fallahpour et al., 2025). However, these approaches typically require ground truth bounding boxes for each sentence. Differently, we introduce a rationale dataset, containing paired images, masks, and machine-readable

rationales, and a self-supervised optimization to jointly segment lesions, justify predictions with clinical terms, and localize their visual evidence.

7. Conclusion, Limitations, and Future Work

Conclusion. We propose to develop models that justify predictions using clinical terms and localize their visual evidence, bridging the semantic-visual gap that hinders clinical trust. Two key challenges are addressed: First, we tackle the lack of prediction and rationale pairs for model training by curating the first rationale dataset pairing 1.5K prostate MRI scans with 180K detailed language rationales. Second, we introduce rationale-informed optimization to disentangle and localize the clinical concepts without pixel-level annotations. Empirical evaluations show that our model delivers accurate rationales, each supported by valid visual evidence.

Limitations and future work. While our current rationale dataset focuses exclusively on prostate cancer, we have developed a scalable pipeline for generating detailed, high-quality textual rationales from structured clinical reports. This foundation enables future expansion to other cancers, such as breast, liver, and lung, where rich sources of clinical reports are available.

Acknowledgment

This work is supported by the National Science Foundation under grant numbers CAREER 2340074, SLES 2416937, III CORE 2412675 and National Institutes of Health under grant number R21CA301093. 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.

Impact Statement

Our work addresses a critical barrier to the clinical adoption of medical AI: the lack of interpretable, auditable reasoning in model predictions. By enabling models to justify segmentation and detection with clinical terms and localize their evidence, we enhance trust and safety in AI-assisted diagnosis. This is particularly impactful in radiology, where misinterpretations can lead to delayed treatment or unnecessary biopsies. Our rationale dataset and self-supervised framework set a new standard for interpretable medical AI, empowering radiologists to audit model logic and accelerating translational research. By aligning AI outputs with clinical workflows, our method could potentially reduce diagnostic errors and radiologist workloads, and improve patient outcomes in resource-constrained settings.

Ethical considerations. While our framework improves transparency, ethical risks may remain. First, while our model's explanations mimic clinical reasoning, they are not a substitute for radiologist judgment; over-reliance on AI could lead to diagnostic complacency. We emphasize that our tool is an assistive, not autonomous, decision-maker. Second, patient privacy is safeguarded by using only de-identified, publicly available data.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. The Medical Segmentation Decathlon. *Nature communications*, 13(1): 4128, 2022.

Arras, L., Osman, A., and Samek, W. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022.

Bannur, S., Bouzid, K., Castro, D. C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Boyd, S. Convex optimization. *Cambridge UP*, 2004.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

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.

Chernyak, V., Fowler, K. J., Kamaya, A., Kielar, A. Z., Elsayes, K. M., Bashir, M. R., Kono, Y., Do, R. K., Mitchell, D. G., Singal, A. G., et al. Liver imaging reporting and data system (li-rads) version 2018: imaging of hepatocellular carcinoma in at-risk patients. *Radiology*, 289(3): 816–830, 2018.

Cozzi, A., Pinker, K., Hidber, A., Zhang, T., Bonomo, L., Lo Gullo, R., Christianson, B., Curti, M., Rizzo, S., Del Grande, F., et al. Bi-rads category assignments by gpt-3.5, gpt-4, and google bard: A multilanguage study. *Radiology*, 311(1):e232133, 2024.

Daneshjou, R., Yuksekgonul, M., Cai, Z. R., Novoa, R., and Zou, J. Y. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems*, 35:18157–18167, 2022.

de Verdier, M. C., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Tahon, N. H., Foltyn-Dumitru, M., Zhang, J., Alafif, M., Baig, S., et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint arXiv:2405.18368*, 2024.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Fallahpour, A., Ma, J., Munim, A., Lyu, H., and Wang, B. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025.

Fedorov, A., Schwier, M., Clunie, D., Herz, C., Pieper, S., Kikinis, R., Tempany, C., and Fennessy, F. An annotated test-retest collection of prostate multiparametric mri. *Scientific data*, 5(1):1–13, 2018.

Gao, Y., Gu, D., Zhou, M., and 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.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., and 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.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Kan, H., Shi, J., Zhao, M., Wang, Z., Han, W., An, H., Wang, Z., and Wang, S. ITUnet: Integration of Transformers and Unet for Organs-at-Risk Segmentation. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2123–2127. IEEE, 2022.

Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Li, T., Qiao, F., Ma, M., and Peng, X. Are data-driven explanations robust against out-of-distribution data? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3821–3831, 2023a.

Li, T., Ma, M., and Peng, X. Beyond accuracy: Ensuring correct prediction with correct rationale. In *Proceedings of the Conference on Neural Information Processing Systems*, 2024b.

Li, T., Ma, M., and Peng, X. Deal: Disentangle and localize concept-level explanations for vlms. In *European Conference on Computer Vision*, pp. 383–401. Springer, 2024c.

Li, Y., Wynne, J., Wang, J., Qiu, R. L., Roper, J., Pan, S., Jani, A. B., Liu, T., Patel, P. R., Mao, H., et al. Cross-Shaped Windows Transformer With Self-Supervised Pretraining for Clinically Significant Prostate Cancer Detection in Bi-Parametric MRI. *arXiv preprint arXiv:2305.00385*, 2023b.

Lin, T. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

Liu, Z., Jiang, H., Zhong, T., Wu, Z., Ma, C., Li, Y., Yu, X., Zhang, Y., Pan, Y., Shu, P., et al. Holistic evaluation of gpt-4v for biomedical imaging. *arXiv preprint arXiv:2312.05256*, 2023.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. Segment anything in medical images. *Nature Communications*, 15(1), January 2024a. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL http://dx.doi.org/10.1038/s41467-024-44824-z.

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., and Peng, X. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Ma, M., Ren, J., Zhao, L., Testuggine, D., and 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.

Ma, M., Li, T., and Peng, X. Beyond the Federation: Topology-aware Federated Learning for Generalization to Unseen Clients. In *Proceedings of the International Conference on Machine Learning*, 2024b.

Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., and Weller, A. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mehta, H. J., Mohammed, T.-L., and Jantz, M. A. The american college of radiology lung imaging reporting and data system: potential drawbacks and need for revision. *Chest*, 151(3):539–543, 2017.

Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T., Dinh, D. H., et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ozer, S., Langer, D. L., Liu, X., Haider, M. A., Van der Kwast, T. H., Evans, A. J., Yang, Y., Wernick, M. N., and Yetik, I. S. Supervised and unsupervised methods for prostate cancer segmentation with multispectral mri. *Medical physics*, 37(4):1873–1883, 2010.

Patrício, C., Neves, J. C., and Teixeira, L. F. Coherent Concept-Based Explanations in Medical Image and Its Application to Skin Lesion Diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3799–3808, 2023.

Qiao, F. and Peng, X. Topology-aware robust optimization for out-of-distribution generalization. In *Proceedings of the International Conference on Learning Representations*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., and Clark, J. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ronneberger, O., Fischer, P., and 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.

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.

Sekhoacha, M., Riet, K., Motloung, P., Gumenku, L., Adegoke, A., and Mashele, S. Prostate cancer review: genetics, diagnosis, treatment options, and alternative approaches. *Molecules*, 27(17):5730, 2022.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Wang, Z., Wu, Z., Agarwal, D., and Sun, J. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods*

in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2022, pp. 3876, 2022.

Wright, S. J. Numerical optimization, 2006.

Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., Kiraly, A., Ahmed, F., Hormozdiari, F., Jaroensri, T., et al. Advancing Multimodal Medical Capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024.

Zhang, K. and Liu, D. Customized Segment Anything Model for Medical Image Segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M. P., Naumann, T., Wang, S., and Poon, H. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640.

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*, pp. 1–11, 2024.

A. Detailed Results of Human Evaluation on Rational Dataset Quality

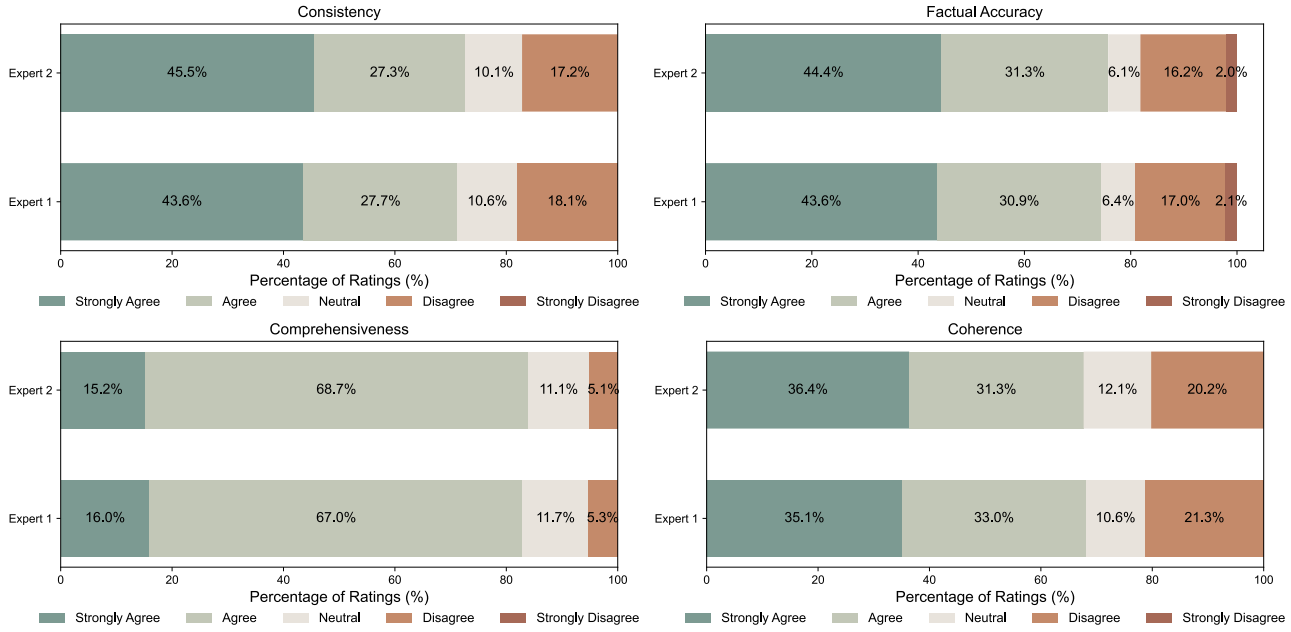

Figure 8: Human evaluation of rationale quality. We randomly sample 100 the extracted rationale descriptions for evaluation. Each radiologist reviewed all 100 cases.

B. Detailed Dataset Description

B.1. PI-CAI Challenge

Rationale dataset is derived from the Public Training and Development Dataset of the PI-CAI challenge (Saha et al., 2024), which comprises 1,500 anonymized prostate mpMRI scans from 1,476 patients. These scans were collected between 2012 and 2021 across three centers in the Netherlands: Radboud University Medical Center, University Medical Center Groningen, and Ziekenhuis Groep Twente. Each patient case was annotated based on histologically-confirmed findings, with Gleason grade group ≥ 2 classified as positive and Gleason grade group ≤ 1 or PI-RADS ≤ 2 classified as negative. For all cases, clinically significant prostate cancer (csPCa) lesions were delineated by one of ten trained investigators or a radiology resident.

Of the 1,500 cases, 1,075 cases contain benign tissue or indolent prostate cancer (PCa), while 425 cases contain csPCa. Among these 425 positive cases, 220 have expert-derived lesion annotations, while the remaining 205 have AI generated lesion annotations. All 1,500 cases have AI generated Gland, Peripheral Zone and Transition Zone annotations. Following the setup provided by the challenge organizers, we combined the human labeled data and AI labeled data and further split them (1,500 cases) into five folds. Then we train and validate all the models with the five folds generated.

B.2. Prostate158

Prostate158 (Adams et al., 2022) includes 158 expert-annotated 3T prostate MRIs with T2-weighted (T2w) and diffusion-weighted imaging (DWI) sequences, including ADC maps, similar to those in PI-CAI. These MRIs were collected at Charité University Hospital in Berlin, Germany, from February 2016 to January 2020. Images were acquired on Siemens VIDA and Skyra 3T scanners (Siemens Healthineers, Erlangen, Germany) following protocols aligned with current guidelines and using B1 shimming.

Two board-certified radiologists, with 6 and 8 years of experience in uro-oncologic imaging, annotated all MR images. They provided pixel-wise segmentations for the central gland (central and transitional zones), peripheral zone, and prostate cancer (PCa) lesions, defined as suspicious areas with a PI-RADS score of ¿= 4. All PCa lesions were segmented on the ADC map and correlated with T2w sequences and high b-value DWI images. We collected anatomical segmentation label from reader

1 on T2w sequences and tumor segmentation label from reader 1 on the ADC map

B.3. MSD-Prostate

MSD-Prostate (Antonelli et al., 2022). This is the prostate subset of Medical Segmentation Decathlon (MSD) dataset. This subset consists of 48 prostate mp-MRI studies comprising T2-weighted and apparent diffusion coefficient (ADC) series, which was acquired at Radboud University Medical Center, Nijmegen Medical Center, Nijmegen, The Netherlands. The data set was further divided into a training set and a testing set. Pixel-wise segmentation masks of the prostate peripheral zone (PZ) and the transition zone (TZ) were provided for the training set. We only use the training set for our zero-shot evaluation.

C. Our Prompts

C.1. The full prompt for rationale generation

Prompt to Extract Rationale From PI-RADS Report

Here is the PI-RADS decision tree that summarizes radiologists' criteria to assign the PI-RADS for prostate cancer tumors:
PI-RADS Decision Tree:

1. **If "Lesion Local" is in the "Peripheral zone":**
   - **Check ADC/DWI MRI:**
     - If "Uniform hyperintense signal intensity (normal)" → DWI score = 1 → PI-RADS score = 1
     - If "Linear or wedge-shaped hypointensity or diffuse mild hypointensity, usually indistinct margin" → DWI score = 2 → PI-RADS score = 2
     - If "Heterogeneous signal intensity or non-circumscribed, rounded, moderate hypointensity" → DWI score = 3:
       - Check DCE MRI:
         - If DCE is "positive" → PI-RADS score = 4
         - If DCE is "negative" → PI-RADS score = 3
     - If "Circumscribed, homogenous moderate hypointense focus/mass confined to the prostate and $<$ 1.5 cm in greatest dimension" → DWI score = 4 → PI-RADS score = 4
     - If "Circumscribed, homogenous moderate hypointense focus/mass confined to prostate and $\geq$ 1.5 cm in greatest dimension, or definite extraprostatic extension/invasive behavior" → DWI score = 5 → PI-RADS score = 5
2. **If "Lesion Local" is in the "Transition zone":**
   - **Check T2W MRI:**
     - If "Normal appearing TZ (rare) or a round, completely encapsulated nodule" → T2W score = 1 → PI-RADS score = 1
     - If "A mostly encapsulated nodule OR a homogeneous circumscribed nodule without encapsulation" → T2W score = 2:
       - Check DWI score:
         - If DWI score $>$ 4 → PI-RADS score = 2
         - If DWI score $\leq$ 4 → PI-RADS score = 3
     - If "Heterogeneous signal intensity with obscured margins" → T2W score = 3:
       - Check DWI score:
         - If DWI score $>$ 5 → PI-RADS score = 3
         - If DWI score $\leq$ 5 → PI-RADS score = 4
     - If "Lenticular or non-circumscribed, homogeneous, moderately hypointense, and $<$ 1.5 cm in greatest dimension" → T2W score = 4 → PI-RADS score = 4

- If "Lenticular or non-circumscribed, homogeneous, moderately hypointense, and $\geq$ 1.5 cm in greatest dimension or definite extraprostatic extension/invasive behavior" $\rightarrow$ T2W score = 5 $\rightarrow$ PI-RADS score = 5

Examples of Radiologist Justifications:

Example 1: PI-RADS Report for Patient 10418

- **PI-RADS Score:** 4
- **Size:** 1.4 x 1.1 cm
- **Location:** Right, anterior, midgland, transition zone

Example 1 Justification: The lesion is located in the transition zone, specifically in the right anterior midgland, with a size of 1.4 x 1.1 cm. The MRI shows a lenticular, non-circumscribed, homogeneous, moderately hypointense lesion, and its size is less than 1.5 cm in the greatest dimension. These characteristics correspond to a T2W score of 4. Since there are no signs of extracapsular extension, seminal vesicle invasion, or adjacent organ invasion, the lesion's PI-RADS score is 4.

Example 2: PI-RADS Report for Patient: 10424:

- **PI-RADS Score:** 5
- **Size:** 2.6 x 1.4 cm
- **Location:** Bilateral, posterior, base to apex, peripheral zone

Example 2 Justification: The lesion is located in the peripheral zone, extending bilaterally from the base to the apex. On the MRI, it measures 2.6 x 1.4 cm. The lesion shows circumscribed, homogenous moderate hypointense focus/mass confined to the prostate, greater than 1.5 cm in the greatest dimension, which corresponds to a DWI score of 5. Additionally, there is evidence of extracapsular extension and adjacent organ invasion. Given these characteristics and the lesion's extent, the lesion's overall PI-RADS score is 5.

Task: Could you write a rationale justification for the following patient based on the patient's PI-RADS report and PI-RADS decision tree? Your justification should be short, clear, and follow the logic of the PI-RADS decision tree.

Here goes the new PI-RADS report.

- **PI-RADS Score:** 4
- **Size:** 1.4 x 1.1 cm
- **Location:** Right, anterior, midgland, transition zone

C.2. The full prompt for evaluation with OpenAI o1

Prompt to Evaluation with OpenAI o1

These are T2, DAC, and DWI MRIs for prostate cancer. First, could you tell me if the Image contains tumors? If there is a tumor, where is the location of the tumor? (Peripheral zone or transitional zone?) Then, Is the lesion margin non-circumscribed or circumscribed? Does the lesion's greatest dimension is greater than 1.5 cm? Does the lesion show signs of extracapsular extension, seminal vesicle invasion, or adjacent organ invasion? What's your judgment of the PI-RADS category? Please give me a definitive answer.

D. Additional Implementation Details

Preprocessing. We followed the same preprocessing procedure with the PI-CAI challenge in all three datasets. First, the mpMRI sequence are resampled to a spacing of (3mm, 0.5mm, 0.5mm), then center crop the sequence to (24, 384, 384). To meet the input requirement of our model, all images are bilinearly resized to (24, 1024, 1024), then, z-norm is applied to each slice in the sequence.

Training. To tackle the severe long-tail condition, we sample slices with a significant lesion (spanning ≥ 0.001 of the image) and a nonsignificant lesion (< 0.001) with a probability of 0.5. We use Focal loss with $\alpha = 0.97$ and $\gamma = 2$. We use Adam optimizer with Betas = [0.9,0.999], weight decay = 0.001 and an exponential learning rate schedule with initial lr = 0.0001 and exponent = 0.9. For online image augmentation, we apply random ratation by choosing the rotation angle between [-15, -10, -5, 0, 5, 10, 15] degrees, and randomly flip horizontally and vertically.

For all baseline and our models, we perform 5 fold cross-validation and

Postprocessing. We applied erosion and dilation to the generated lesion mask to eliminate small bubbles. And according to the predicted gland mask, we automatically delete those lesions predicted that have < 0.5 overlap with gland mask. The overlap is defined as:

$$overlap = \frac{|M_g \cap M_l|}{|M_l|}$$

Where M_g and M_l denotes predicted gland mask and lesion mask.

Evaluation metrics. We use Dice score (DSC) for evaluating the segmentation accuracy across all the output channels (Gland, Zone and Lesion). Dice score is defined as

$$\text{Dice} = \frac{2|A \cap B| + \epsilon}{|A| + |B| + \epsilon}$$

Where A,B denotes the predicted mask and ground truth mask respectively. ϵ is a small number that preventing dividing by 0. We choose $\epsilon = 1$. Considering there is a significant amount of images which don't have '1's in the lesion ground truth, making it very easy for the model to get a very high lesion dice score (≈ 0.93 by only output empty mask for lesion), we only report the mean dice score of lesion positive images, meaning images having '1's in their lesion mask. For detection metrics, we adopt the same metrics from the PI-CAI challenge (Saha et al., 2024). The performance of patient-level diagnosis is assessed using the Area Under the Receiver Operating Characteristic (AUROC) metric, while lesion-level detection performance is measured with the Average Precision (AP) metric.

E. Analysis of Failure Case

Figure 9 visualizes the failure cases of our model, highlighting an inconsistency between its segmentation masks and the corresponding rationales. While the model provides both segmentation masks and the reasoning behind its decisions, we observe that these outputs could be inconsistent, particularly in diagnostically transition slices where the tumor is just starting to appear or fade. There are two types of inconsistency. Case 1: mask without rationale.The model predicts tumor segmentation while generating benign rationales. Case 2: rationale without mask. The model output tumor rationales while failing to produce the tumor segmentation. Both cases frequently occurs in transitional slices, where tumor boundaries are ambiguous or tumor presence is subtle. This rationale-mask misalignment could undermine clinical trust of our AI models. We will explore methods to resolve this misalignment in the future work.

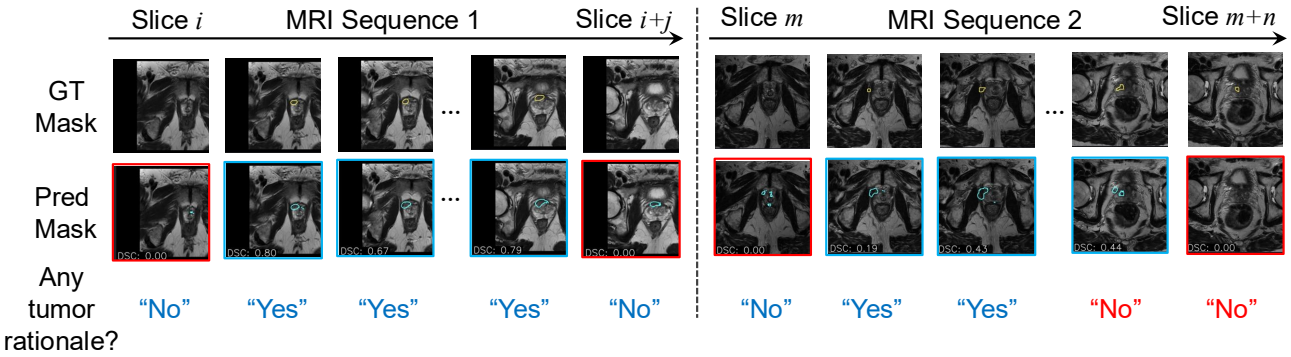

Figure 9: Visualization of failure cases (the inconsistency between segmentation mask and rationale).

F. Future Direction

Expanding the rationale dataset. Although our dataset currently supports robust localized explanations for prostate cancer, a natural next step involves extending this dataset to include other cancer types, such as breast (Cozzi et al., 2024) and lung (Mehta et al., 2017) cancers, where interpretable is equally critical. In addition to expanding data collection within a single institution, future efforts could utilize distributed learning techniques (McMahan et al., 2017; Ma et al., 2024b; Li et al., 2020) to collaboratively train models across multiple hospitals without sharing raw data, implicitly enlarging the effective training dataset.

Building an education tool for junior or attending radiologists. Diagnostic disagreements between expert and junior radiologists—particularly in nuanced cases like PI-RADS scoring—highlight a critical training gap. It is possible to leverage our model's rationale generation ability to build an AI-driven education tool. This tool would enable trainees to analyze cases, submit assessments, and receive feedback contrasting their decisions with expert-backed AI rationales.