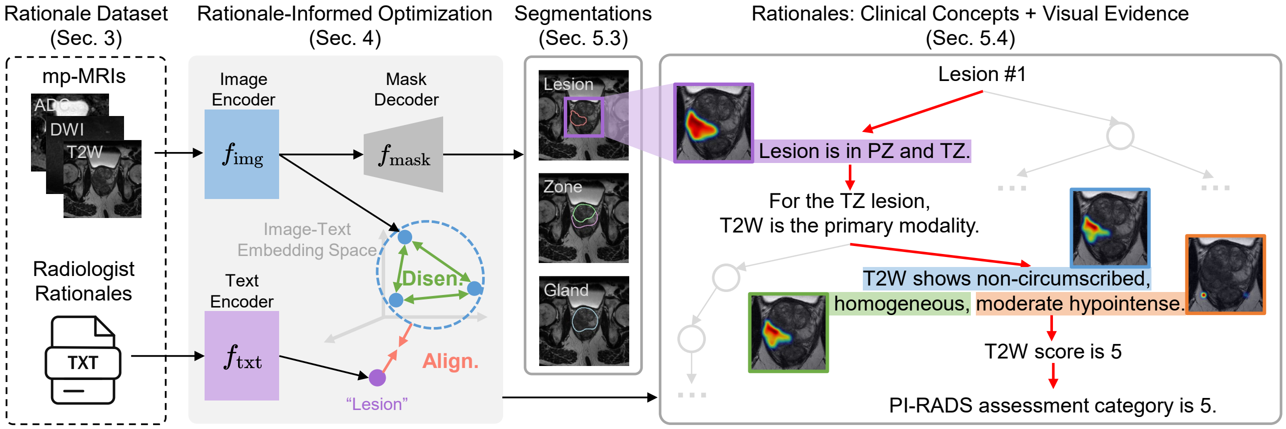
Rationale Dataset (Sec. 3)
mp-MRIs
ADC
DWI
T2W
Radiologist Rationales
TXT
Rationale-Informed Optimization (Sec. 4)
Image Encoder
Mask Decoder
f_{img}
f_{mask}
Image-Text Embedding Space
Disen.
Text Encoder
f_{txt}
Align.
"Lesion"
Segmentations (Sec. 5.3)
Lesion
Zone
Gland
Rationales: Clinical Concepts + Visual Evidence (Sec. 5.4)
Lesion #1
Lesion is in PZ and TZ.
For the TZ lesion, T2W is the primary modality.
T2W shows non-circumscribed, homogeneous, moderate hypointense.
T2W score is 5
PI-RADS assessment category is 5.

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:

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

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:

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

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