R^2-Seg: Training-Free OOD Medical Tumor Segmentation via Anatomical Reasoning and Statistical Rejection

Supplementary Material

1. Method Details

1.1. Planner Outputs and ROI Construction

1.1.1. LLM Planner and Prompt Normalization

As illustrated in Algorithm. 1, the LLM planner Φ transforms a free-form tumor description into structured geometric guidance for ROI construction. Its output consists of anchor organs, ROI parameters, and a concise rationale, each generated through a deterministic template (Appendix 1.1.2). The planner interprets the anatomical site implied by the tumor type, identifies nearby organs that are consistently visible, and determines padding, scale jitter, and optional squaring rules for the ROI. All outputs are formatted as JSON to ensure reproducibility and model-agnostic integration.

Because clinical terminology varies across datasets and institutions, user-provided concepts are normalized before being passed to the segmentor. A dedicated LLM normalizer Π maps arbitrary descriptions from an open prompt space $\mathcal{C}$ onto a fixed vocabulary $\mathcal{C}^\star$ aligned with BiomedParse training. This process relies on a template that restricts outputs to the predefined vocabulary while allowing the LLM to justify its choice. By enforcing a stable set of canonical terms, the pipeline prevents stylistic drift, maintains consistent text conditioning across all R^2-Seg views, and ensures that any variation in predictions arises solely from visual input rather than prompt variability.

Together, the planner and normalizer provide anatomically informed yet text-stable conditioning for downstream segmentation, enabling robust performance across datasets without modifying model parameters.

1.1.2. Prompt Design

The R^2-Seg pipeline relies on a small set of deterministic prompts that govern anatomical planning, concept normalization, and segmentation. These prompts standardize how the planner reasons about anatomy and how the segmentor receives textual instructions, thereby isolating visual variability from linguistic variability and improving cross-dataset robustness.

The planner operates on free-form tumor descriptions using a system prompt that instructs the model to consider anatomical context, identify adjacent organs that are reliably detectable, and generate ROI parameters such as padding, scale jitter, and optional squaring rules. The output follows a fixed JSON schema to ensure consistent behavior across tumor types and imaging modalities. A brief user prompt specifies the tumor type and modality, prompting the planner to derive anchors and ROI rules tailored to the anatomical site.

Algorithm 1 Anatomy-aware Planning and ROI Proposal

Require: Cancer type c, input image I, in-plane spacings (s_x, s_y), padding δ, and scale jitters Γ

$(\mathcal{A}, \mathcal{I}_{\mathrm{ROI}}, r) \leftarrow \Phi(c)$
for each anchor $a \in \mathcal{A}$: $c_a \leftarrow \Pi(a)$ (normalize anchor name) $M_a \leftarrow \mathbb{1}\{f_\theta(I; c_a) \geq \tau_a\}$ (anchor mask)
$M_\cup \leftarrow \bigcup_{a\in\mathcal{A}} M_a$ (union anchor mask)
$B_0 \leftarrow \mathsf{BBox}(M_\cup)$
$(m_x, m_y) \leftarrow (\lceil \delta_x/s_x \rceil\,,\ \lceil \delta_y/s_y \rceil)$
$\widehat{B} \leftarrow \mathsf{Pad}(B_0; m_x, m_y)$
if squaring is requested: $\widehat{B} \leftarrow \mathsf{Square}(\widehat{B})$
return $\{\, B_\gamma = \mathsf{Scale}(\widehat{B}; \gamma) \,\}_{\gamma\in\Gamma}$ and $M_\cup$

To address the diversity of medical terminology, a dedicated normalizer Π maps user-provided concepts to the canonical vocabulary used during BiomedParse training. The normalizer's template restricts output to a predefined vocabulary list while allowing the LLM to provide justification. This step stabilizes the text encoder and ensures consistent conditioning across all TTA crops.

After anchors and normalized concepts are obtained, segmentation prompts are generated through concise templates. Each anchor organ is passed to the segmentor using the directive *"segment the <ORGAN_NAME>"*, a phrasing aligned with the model's training data. Tumor segmentation within each ROI uses the parallel template *"segment the <TUMOR_SITE> tumor"*. These minimal forms avoid unnecessary modifiers, ensuring that prediction differences across crops reflect only visual context changes rather than variations in phrasing.

Overall, the prompt design emphasizes stability, clarity, and consistency. It separates anatomical reasoning from segmentation, maintains uniform textual conditioning across views, and allows R^2-Seg to remain fully training-free while retaining strong cross-domain robustness.

1.1.3. Anchor Masks and Base Bounding Box

For each anchor $a \in \mathcal{A}$, the text-driven segmentor f_θ generates a probability map P_a, which is thresholded at τ_a to obtain a binary mask $M_a = \mathbb{1}\{P_a \geq \tau_a\}$. The union mask $M_\cup = \bigcup_{a\in\mathcal{A}} M_a$ defines an initial region of interest, and its axis-aligned bounding box is denoted by $B_0 = \mathsf{BBox}(M_\cup)$.

When anchors are unreliable or missing, a conservative fallback sets B_0 to the full image frame so that downstream computations remain well-defined.

1.1.4. Padding, Squaring, and Multi-scale Jitter

Let (s_x, s_y) denote the in-plane physical spacings in mm/pixel. The padding vector $\delta = (\delta_x, \delta_y)$ is converted to pixel margins

$$m_x = \left\lceil \frac{\delta_x}{s_x} \right\rceil, \qquad m_y = \left\lceil \frac{\delta_y}{s_y} \right\rceil,$$

and applied to B_0 in image coordinates. If square = true, the padded bounding box is expanded along its shorter side to obtain a square ROI. A family of jittered ROIs $\{B_\gamma\}_{\gamma\in\Gamma}$ is then formed by isotropically scaling the side length of this squared box by each $\gamma \in \Gamma$, while keeping its center fixed.

Each ROI B_γ induces a crop $I|_{B_\gamma}$ that is segmented independently and later restored to the native image canvas during aggregation.

1.2. Segmentation and ROI Fusion

1.2.1. Transform Set and Inverse Alignment

Let $\mathcal{G} = \{g_{\text{id}}, g_{\text{lr}}, g_{\text{tb}}\}$ denote the identity, left–right flip, and top–bottom flip transforms. For each crop $I|_{B_\gamma}$ and each view $g \in \mathcal{G}$, we compute a prediction

$$P^{(\gamma,g)} = f_\theta\big(g(I|_{B_\gamma});\, c_{\text{tumor}}\big).$$

The prediction is then realigned to the native image frame using $\mathsf{Inv}(g)$ and restored to the full canvas based on the known ROI coordinates. This ensures that all predictions, regardless of view or crop, are accumulated in a common spatial reference frame.

1.2.2. Aggregation Across Views and Supports

Let $A \in \{\max, \text{median}, \text{mean}\}$ be an element-wise fusion rule. The fused probability over an ROI is defined as

$$P^{(\gamma)}(x) = A_{g\in\mathcal{G}}\Big[\big(\mathsf{Restore}_{B_\gamma} \circ \mathsf{Inv}(g)\big)\, P^{(\gamma,g)}\Big](x).$$

When restored ROIs overlap, contributions are averaged to avoid bias from differences in ROI coverage. In addition to ROI-level predictions, we compute a full-frame prediction P^{full} (with the same test-time augmentations). To preserve confident detections that may appear only at the global scale or only within a jittered ROI, we adopt a conservative support fusion:

$$P^{\text{final}}(x) = \max\Big(P^{\text{full}}(x),\, \max_{\gamma\in\Gamma} P^{(\gamma)}(x)\Big). \tag{6}$$

This rule retains any high-probability region observed in either support while still exploiting the anatomical focus provided by the ROI crops. Empirically, using $A = \max$ or $A =$ median yields stable results, whereas mean may be preferred when suppressing isolated peaks is desirable.

Algorithm 2 Segmentation and ROI Fusion

Require: Image I, ROIs $\{B_\gamma\}$, transforms $\mathcal{G}$, fusion rule A

$P^{\text{full}} \leftarrow A_{g\in\mathcal{G}}\big[\mathsf{Inv}(g)\, f_\theta(g(I); c_{\text{tumor}})\big]$
for $\gamma \in \Gamma$ **do**
  for $g \in \mathcal{G}$ **do**
    $Q \leftarrow f_\theta\big(g(I|_{B_\gamma});\, c_{\text{tumor}}\big)$
    $Q \leftarrow \mathsf{Inv}(g)(Q)$; $P^{(\gamma,g)} \leftarrow \mathsf{Restore}_{B_\gamma}(Q)$
  end for
  $P^{(\gamma)} \leftarrow A_{g\in\mathcal{G}}[P^{(\gamma,g)}]$ with overlap-count normalization
end for
return $P^{\text{final}} = \max\big(P^{\text{full}},\, \max_\gamma P^{(\gamma)}\big)$

1.2.3. Monotonicity Property

Let $\tau > 0$ be a binarization threshold, and define the super-level set $\mathcal{S}(P,\tau) = \{x : P(x) \geq \tau\}$. By construction of the fusion rule,

$$\begin{aligned}\mathcal{S}\big(P^{\text{final}},\tau\big) &= \mathcal{S}\big(\max(P^{\text{full}}, \max_\gamma P^{(\gamma)}),\tau\big)\\ &\supseteq \mathcal{S}\big(P^{\text{full}},\tau\big) \cup \bigcup_\gamma \mathcal{S}\big(P^{(\gamma)},\tau\big).\end{aligned}$$

Thus no connected component captured by any support is removed when fusing predictions at the same threshold. This property favors recall, while precision is deferred to later statistical screening steps.

1.3. Candidate Extraction and Cross-slice Linking

From the fused probability map P^{final}, a binary mask

$$\mathcal{M} = \mathbb{1}\{P^{\text{final}} \geq \tau_{\text{bin}}\}$$

is obtained and decomposed into 2D connected components $\{C_k\}_{k\in\mathcal{K}}$ using 8-connectivity. Small components are removed, and basic region descriptors (centroid, area, mean probability) are retained for later statistical screening.

1.4. Statistical Screening with MMD and FDR Control

1.4.1. Feature sampling and kernel choice

For each candidate C_k and the control region $M_\cup$, pixel-level features ϕ are sampled uniformly up to a cap of 4k points per set. We use a Gaussian kernel with bandwidth σ determined by the median heuristic on the pooled sample. These implementation details complement the high-level procedure given in the main paper.

1.4.2. Unbiased MMD and permutation testing

We use the unbiased MMD estimator and permutation test exactly as defined in the main text (Sec. 3.2.3); here we detail the sampling caps, bandwidth selection, and smoothed

Algorithm 3 Candidate-level screening with MMD and BH-FDR

Require: Candidates $\{C_k\}$, control mask $M_\cup$, features ϕ, permutations B, FDR level α

1: **for** each k **do**
2: Sample $X \subseteq \phi(I|_{C_k})$ and $Y \subseteq \phi(I|_{M_\cup})$ with caps
3: Set kernel bandwidth σ via median heuristic
4: Compute $\widehat{\mathrm{MMD}}^2_{\mathrm{obs}}(X, Y)$
5: Estimate p_k by B permutations with smoothed counts
6: **end for**
7: Apply BH-FDR at level α and retain significant candidates
8: **return** filtered candidate set

counting for p-values. Let X and Y denote the sampled features from C_k and $M_\cup$. The unbiased MMD estimator is used, and permutation testing is performed by recomputing the statistic on B random shuffles of the pooled sample. The p-value estimate

$$p_k = \frac{\#\{b : \widehat{\mathrm{MMD}}^2_{\mathrm{perm},b} \geq \widehat{\mathrm{MMD}}^2_{\mathrm{obs}}\} + 1}{B + 1}$$

includes the standard smoothing factor to avoid zero-valued estimates. All computations are vectorized and can be executed on GPU for efficiency.

1.4.3. Multiple testing

Across all candidates, the Benjamini–Hochberg procedure at level α determines the subset declared significant. Let $p_{(1)} \leq \cdots \leq p_{(|\mathcal{K}|)}$ be the ordered p-values; the decision index

$$i^* = \max\{i : p_{(i)} \leq \alpha\, i/|\mathcal{K}|\}$$

selects the retained components $\{C_{(1)}, \ldots, C_{(i^*)}\}$. The pipeline remains compatible with alternative corrections when sample sizes are small.

1.5. False-positive Gating for Empty-mask Cases

Let P^{final} denote the fused probability map defined in Eq. (6), where predictions within each view set $\mathcal{G}$ are aggregated by mean, and ROI-vs-full fusion uses the $\max$ rule. We apply a three-level gating strategy complementary to statistical testing:

(L1) Existence gate. Compute the global maximum $p_{\max} = \max_x P^{\mathrm{final}}(x)$, the positive-pixel ratio within the ROI domain

$$\rho = \frac{|\{x \in \Omega_{\mathrm{ROI}} : P^{\mathrm{final}}(x) \geq \tau_{\mathrm{bin}}\}|}{|\Omega_{\mathrm{ROI}}|},$$

and a KS statistic p_{KS} comparing foreground and background probabilities inside Ω_{ROI}. If $p_{\max} < \tau_{\max}$ or $\rho < \tau_\rho$ or $p_{\mathrm{KS}} > \tau_{\mathrm{KS}}$, the case is declared negative and no candidates are extracted.

(L2) Candidate-level gate. For candidates $\{C_k\}$ obtained from P^{final}, require

$$|C_k| \geq A_{\min}, \qquad \overline{P}_k \geq \tau_{\mathrm{mean}}, \qquad \frac{|C_k \cap M_\cup|}{|C_k|} \geq \tau_\cap,$$

where $\overline{P}_k$ is the mean probability over C_k and $M_\cup$ is the organ control region.

(L3) Case-level score. Define $S_k = \overline{P}_k \sqrt{|C_k|}$ and $S^* = \max_k S_k$. If $S^* < \tau_{\mathrm{case}}$, return an all-zero mask. This gate acts as a conservative post-hoc calibration when empty cases dominate the test set.

1.6. Computational Aspects and Approximations

Let $|\mathcal{K}|$ denote the number of candidates, and let m and n be the sampled pixels from each candidate and the control region (both capped). With B permutations, the screening cost scales as $O(|\mathcal{K}|\, B\, (m+n)^2)$, with vectorized distance computations giving a modest constant factor. If necessary, approximations such as random Fourier features or Nyström sampling can reduce the quadratic kernel cost while preserving decision quality and FDR validity. All steps are training-free and do not update model parameters at test time.

1.7. Aggregation Rules and Robustness

The ROI fusion rule in (6) ensures that high-confidence signals from either support are preserved, improving recall on challenging cases. The choice of view-aggregation rule A modulates the sharpness of the fused map: $\max$ preserves peaks, median stabilizes against outlier views, and mean suppresses noise but may attenuate small structures. Combined with statistical screening and conservative gating, this design yields a robust balance between sensitivity and specificity under distribution shift.

2. Implementation Details

2.1. LLM Planner

We use Qwen3-0.6B as the planner (temperature 0.2). Since it is **only for anatomical planning**, it requires only basic organ anatomy knowledge, not visual capability or highly precise planning. Even with this 0.6B model, we achieve near-100% planning success.

2.1.1. Inference Latency

Average inference time is ∼**4.1s per image** on an RTX 4090. Breakdown: LLM API (<0.8s), Anchor Seg (0.5s), Tumor Seg + TTA (2s), Statistical Test (0.8s). This is well within acceptable limits for offline radiological screening assistants.

2.2. Parameters

We categorize all hyper-parameters into three coherent categories—*scoring*, *statistical*, and *geometric/gating*. A small unlabeled development split, strictly disjoint from the test data, is used only once to set a few global defaults. All remaining thresholds are derived directly from instance-level prediction statistics at test time; thus, *no supervision* is involved to select parameters.

2.2.1. Scoring thresholds.

The test-time augmentation scheme uses max fusion to retain peak responses while suppressing view-specific noise. Scale jitter is kept moderate ($[0.8, 1.0, 1.2]$). The binarization threshold τ_{bin} is chosen once on the unlabeled development split and thereafter fixed. All other scoring-related quantities operate directly on each image's prediction statistics, ensuring a completely label-free procedure at test time. Empirically, $\tau_{\text{bin}}{=}0.4$ balances recall and precision consistently across organs and imaging planes.

2.2.2. Statistical thresholds.

All statistical decisions rely solely on prediction-derived feature distributions computed inside the ROI. Two-sample testing and FDR control aim to bound false discoveries based on these distributions. Here, the *Type-I error* denotes falsely rejecting the null hypothesis—i.e., misidentifying a normal region as abnormal—when their predicted feature distributions are statistically similar. To control this error in a label-free manner, we adopt the median heuristic for the Gaussian kernel bandwidth, cap each region at 4k samples, and set a single Benjamini–Hochberg level α using only the development split. All statistical thresholds remain fixed at test time and operate exclusively on model-produced statistics.

2.2.3. Geometric and gating thresholds.

Geometric and gating decisions also rely entirely on statistics computed from the model's predictions within each ROI. The existence gate evaluates the positive-pixel ratio r and ROI-max probability m, where reference levels τ_{ratio} and τ_{max} are fixed percentiles estimated from the development split and do not require any supervision. The morphology gate constrains the candidate area and confidence via low-percentile lesion priors (τ_{area}, τ_{mean}), computed from model output statistics. The control-overlap gate enforces anatomical consistency using a modest $\tau_{\text{inter}}{=}0.05$, or an adaptive form $\tau_{\text{inter}} \leftarrow c\rho$ with $c{\approx}0.05$. At the case level, the stability score $S_k = \overline{P}_k \sqrt{|C_k|}$ is compared with a global cutoff $\tau_{\text{case}}{=}2.0$, chosen to suppress false positives without requiring any supervision. ROI padding is defined in millimetres for modality independence; 25 mm corresponds to the 95th percentile of anchor-to-lesion distances in abdominal and pelvic scans.

2.2.4. Default settings and sensitivity.

Unless otherwise specified, we use fixed defaults derived from the development split: aggregation=max, $\tau_{\text{bin}}{=}0.4$, $\alpha{=}0.05$, $\tau_{\text{max}}{=}0.45$, $\tau_{\text{pos}}{=}0.45$, $\tau_{\text{ratio}}{=}2{\times}10^{-4}$, $\tau_{\text{area}}{=}80$ px, $\tau_{\text{mean}}{=}0.5$, $\tau_{\text{inter}}{=}0.05$, $\tau_{\text{case}}{=}2.0$, ROI padding 25 mm, and at most 4k samples per region. Ablation studies show that moderate (±20%) variations around these defaults marginally affect Dice, with τ_{bin} and τ_{case} having the largest influence. In practice, adjusting only these two parameters achieves near-optimal performance while keeping all statistical thresholds and ROI padding fixed.

2.2.5. Computational burden

When finetuning BiomedParse with LoRA, the bakcbone was kept frozen while the pixel-decoder was updated for 5 epochs following its official settings. This process required approximately 20 hours on an NVIDIA A100 GPU (80 GB memory). In contrast, R^2-Seg requires no additional training or finetuning. Its inference was performed on single Nvidia RTX4090 GPU with an average inference time of approximately 4 seconds per image.

2.2.6. Normalization Details

We employ **Instance-level Percentile Normalization**. For each ROI, intensities are clipped to the [0.5, 99.5] percentiles and scaled to [0, 1]. This local normalization is robust to scanner-level variations (e.g., HU shifts in CT or signal drift in MRI), unlike fixed-window global normalization.

3. Additional Results

3.1. Quantitative Results

The complete quantitative results on all the categories are shown in Table. 4 and Table. 7.

3.1.1. Comparison on Background Slices

When evaluated on tumor-free slices, R^2-Seg shows a significant improvement in controlling false activations as illustrated in Fig. 7.

Since BiomedParse is trained primarily on in-distribution anatomical structures and rarely outputs empty masks, it produces dense spurious activations on out-of-distribution or background slices, resulting in near-zero Dice scores for several tumor categories, such as bladder, cervix, prostate, and uterus. Although BiomedParse-LoRA partially alleviates this issue by adjusting feature activations through fine-tuning, it often overcompensates, activating irrelevant regions and disrupting structural consistency. By contrast, the hierarchical *false-positive gating* and two-sample *statistical filtering* modules in R^2-Seg adaptively regulate prediction confidence and spatial sparsity during inference, effectively

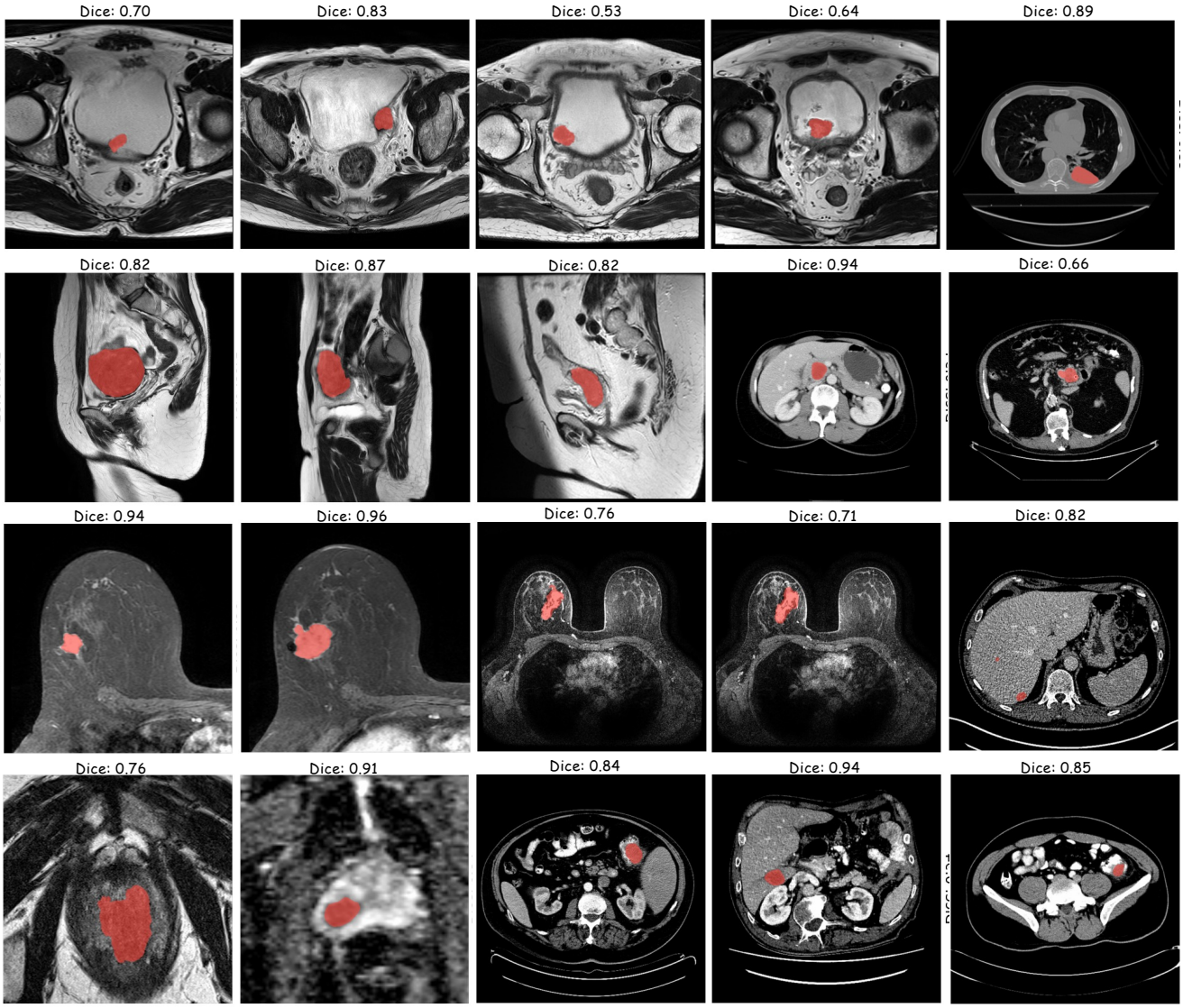

Figure 6. More qualitative results predicted by our R^2-Seg. Our model is capable of recognizing tumors of varying sizes, across different scanning planes and organs, under diverse fields of view.

removing background responses without retraining. As a result, R^2-Seg produces near-empty masks on negative slices while maintaining reliable recall on true lesions, demonstrating its robustness and calibration advantage under severe distribution shifts.

3.2. Prevalence-sweep under clinically imbalanced evaluation

We also provide the performance under different sampling ratios (prevalence-sweep analysis) as shown in Table. 6, BiomedParse-based models consistently degrade as the imbalance increases, while R^2-Seg remains highly robust and even improves under severe imbalance (1:3, 1:4). This demonstrates that the statistical rejection of R^{2}Seg is effective in excluding false positives, which is critical for real-world medical diagnosis.

3.2.1. Comparison of HD95 Metrics

We have computed the **95% Hausdorff Distance (HD95)**. The HD95 score showed marginal improvement over the original BioMedParse. We attribute this to the unsuitability of HD95 for tumor evaluation, as the metric lacks sensitivity for small, multi-focal lesions.

3.3. Visual Results

More experiment results are shown in Fig. 6 and Fig. 8.

Tumor	Method	Dice	Sensitivity	Specificity	Accuracy	CA
Bladder tumor	BiomedParse	0.069	1.000	0.000	0.976	0.546
	BiomedParse-FT	0.482	0.992	0.254	0.995	0.687
	BiomedParse-LoRA	0.578	0.960	0.456	0.996	0.677
	R^2-Seg (Ours)	0.297	0.335	0.536	0.992	0.762
Breast tumor	BiomedParse	0.144	0.998	0.030	0.952	0.582
	BiomedParse-FT	0.559	0.982	0.411	0.995	0.688
	BiomedParse-LoRA	0.611	0.954	0.520	0.996	0.685
	R^2-Seg (Ours)	0.395	0.412	0.728	0.984	0.762
Cervix tumor	BiomedParse	0.154	1.000	0.000	0.985	0.598
	BiomedParse-FT	0.424	0.991	0.197	0.995	0.699
	BiomedParse-LoRA	0.485	0.949	0.359	0.996	0.686
	R^2-Seg (Ours)	0.355	0.299	0.632	0.993	0.777
Colon tumor	BiomedParse	0.351	1.000	0.042	0.994	0.674
	BiomedParse-FT	0.497	0.987	0.271	0.998	0.683
	BiomedParse-LoRA	0.498	0.970	0.309	0.998	0.675
	R^2-Seg (Ours)	0.479	0.110	0.919	0.997	0.758
Kidney tumor	BiomedParse	0.138	1.000	0.067	0.990	0.553
	BiomedParse-FT	0.567	0.967	0.475	0.998	0.617
	BiomedParse-LoRA	0.537	0.961	0.434	0.998	0.617
	R^2-Seg (Ours)	0.567	0.371	0.778	0.997	0.873
Liver tumor	BiomedParse	0.359	0.998	0.136	0.994	0.647
	BiomedParse-FT	0.529	0.932	0.413	0.998	0.640
	BiomedParse-LoRA	0.556	0.950	0.451	0.998	0.643
	R^2-Seg (Ours)	0.625	0.342	0.866	0.996	0.853
Lung tumor	BiomedParse	0.158	0.977	0.072	0.995	0.562
	BiomedParse-FT	0.393	0.982	0.195	0.998	0.652
	BiomedParse-LoRA	0.460	0.915	0.376	0.998	0.640
	R^2-Seg (Ours)	0.446	0.071	0.884	0.997	0.752
Pancreas tumor	BiomedParse	0.212	1.000	0.078	0.998	0.584
	BiomedParse-FT	0.422	0.955	0.354	0.999	0.576
	BiomedParse-LoRA	0.417	0.980	0.328	0.999	0.584
	R^2-Seg (Ours)	0.739	0.131	0.935	0.999	0.896
Prostate tumor	BiomedParse	0.047	1.000	0.000	0.910	0.552
	BiomedParse-FT	0.377	0.917	0.352	0.991	0.561
	BiomedParse-LoRA	0.428	0.852	0.434	0.992	0.555
	R^2-Seg (Ours)	0.465	0.645	0.587	0.971	0.890
Uterus tumor	BiomedParse	0.190	1.000	0.000	0.977	0.632
	BiomedParse-FT	0.478	0.961	0.273	0.995	0.682
	BiomedParse-LoRA	0.514	0.897	0.394	0.995	0.670
	R^2-Seg (Ours)	0.401	0.424	0.600	0.989	0.806

Table 4. Mean metrics (all classes included) grouped by tumor type and method.

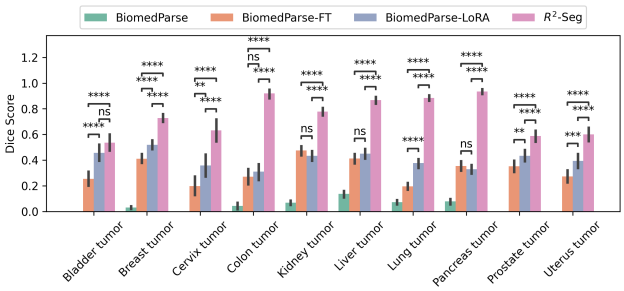

Figure 7. Evaluation of over-segmentation on slices without tumors. Here, p-value is annotated by an asterisk, i.e., ns: $0.05 < p \leq 1$, *: $0.01 < p \leq 0.05$, **: $0.001 < p \leq 0.01$, ***: $0.0001 < p \leq 0.001$, and ****: $p \leq 0.0001$.

Method	**Score**
BiomedParse	0.8917
BiomedParse-LoRA	0.6633
BiomedParse-FT	0.6426
R^2Seg	0.8908

Table 5. We report HD95 for completeness. Its interpretation should be treated with caution in this setting, since many tumors are small or multi-focal and boundary-distance metrics can be less informative than detection-oriented trade-off measures such as FROC.

Ratio	Model	F1-score	Dice	Soft Dice
1:1	BP	0.329	0.362	0.387
	BP-LoRA	0.357	**0.519**	**0.542**
	BP-FT	0.324	0.519	0.541
	R^2Seg	**0.510**	0.510	0.510
1:2	BP	0.219	0.268	0.285
	BP-LoRA	0.238	0.455	0.470
	BP-FT	0.216	0.448	0.464
	R^2Seg	**0.643**	**0.643**	**0.643**
1:3	BP	0.164	0.221	0.234
	BP-LoRA	0.178	0.426	0.438
	BP-FT	0.162	0.425	0.436
	R^2Seg	**0.725**	**0.725**	**0.725**
1:4	BP	0.131	0.193	0.203
	BP-LoRA	0.143	0.405	0.414
	BP-FT	0.128	0.410	0.419
	R^2Seg	**0.765**	**0.765**	**0.765**

Table 6. Model performance under different sampling ratios (tumor:health). BP indicates BioMedParse.

Tumor	Method	Dice	Sensitivity	Specificity	Accuracy	CA
Bladder tumor	R^2-Seg	$0.297_{\pm 0.449}$	0.335	0.536	$0.992_{\pm 0.013}$	$0.762_{\pm 0.245}$
	R^2-Seg wo ST	$0.257_{\pm 0.419}$	0.540	0.419	$0.990_{\pm 0.015}$	$0.770_{\pm 0.241}$
	R^2-Seg wo FPG	$0.099_{\pm 0.265}$	0.923	0.089	$0.989_{\pm 0.015}$	$0.774_{\pm 0.239}$
Breast tumor	R^2-Seg	$0.395_{\pm 0.477}$	0.412	0.729	$0.984_{\pm 0.028}$	$0.762_{\pm 0.246}$
	R^2-Seg wo ST	$0.316_{\pm 0.446}$	0.602	0.544	$0.980_{\pm 0.030}$	$0.767_{\pm 0.243}$
	R^2-Seg wo FPG	$0.152_{\pm 0.325}$	0.934	0.195	$0.978_{\pm 0.030}$	$0.772_{\pm 0.240}$
Cervix tumor	R^2-Seg	$0.355_{\pm 0.465}$	0.299	0.632	$0.993_{\pm 0.008}$	$0.777_{\pm 0.240}$
	R^2-Seg wo ST	$0.337_{\pm 0.446}$	0.496	0.513	$0.993_{\pm 0.007}$	$0.796_{\pm 0.230}$
	R^2-Seg wo FPG	$0.167_{\pm 0.322}$	0.932	0.145	$0.993_{\pm 0.007}$	$0.800_{\pm 0.225}$
Colon tumor	R^2-Seg	$0.479_{\pm 0.494}$	0.110	0.919	$0.997_{\pm 0.006}$	$0.758_{\pm 0.247}$
	R^2-Seg wo ST	$0.460_{\pm 0.490}$	0.174	0.852	$0.997_{\pm 0.007}$	$0.764_{\pm 0.245}$
	R^2-Seg wo FPG	$0.238_{\pm 0.385}$	0.767	0.309	$0.996_{\pm 0.008}$	$0.782_{\pm 0.231}$
Kidney tumor	R^2-Seg	$0.567_{\pm 0.490}$	0.371	0.778	$0.997_{\pm 0.008}$	$0.873_{\pm 0.215}$
	R^2-Seg wo ST	$0.477_{\pm 0.490}$	0.723	0.628	$0.994_{\pm 0.011}$	$0.887_{\pm 0.204}$
	R^2-Seg wo FPG	$0.229_{\pm 0.400}$	0.990	0.248	$0.994_{\pm 0.010}$	$0.904_{\pm 0.186}$
Liver tumor	R^2-Seg	$0.625_{\pm 0.465}$	0.342	0.866	$0.996_{\pm 0.013}$	$0.853_{\pm 0.219}$
	R^2-Seg wo ST	$0.639_{\pm 0.453}$	0.560	0.764	$0.996_{\pm 0.015}$	$0.892_{\pm 0.188}$
	R^2-Seg wo FPG	$0.412_{\pm 0.443}$	0.941	0.291	$0.996_{\pm 0.015}$	$0.928_{\pm 0.140}$
Lung tumor	R^2-Seg	$0.446_{\pm 0.496}$	0.071	0.884	$0.997_{\pm 0.005}$	$0.752_{\pm 0.249}$
	R^2-Seg wo ST	$0.381_{\pm 0.484}$	0.101	0.751	$0.996_{\pm 0.006}$	$0.753_{\pm 0.249}$
	R^2-Seg wo FPG	$0.209_{\pm 0.395}$	0.614	0.379	$0.996_{\pm 0.006}$	$0.758_{\pm 0.246}$
Pancreas tumor	R^2-Seg	$0.739_{\pm 0.435}$	0.131	0.935	$0.999_{\pm 0.003}$	$0.896_{\pm 0.201}$
	R^2-Seg wo ST	$0.697_{\pm 0.454}$	0.237	0.865	$0.999_{\pm 0.003}$	$0.901_{\pm 0.196}$
	R^2-Seg wo FPG	$0.288_{\pm 0.439}$	0.803	0.304	$0.999_{\pm 0.003}$	$0.911_{\pm 0.184}$
Prostate tumor	R^2-Seg	$0.465_{\pm 0.486}$	0.645	0.587	$0.971_{\pm 0.050}$	$0.890_{\pm 0.190}$
	R^2-Seg wo ST	$0.262_{\pm 0.418}$	0.923	0.299	$0.959_{\pm 0.052}$	$0.898_{\pm 0.175}$
	R^2-Seg wo FPG	$0.069_{\pm 0.208}$	0.994	0.033	$0.958_{\pm 0.051}$	$0.900_{\pm 0.171}$
Uterus tumor	R^2-Seg	$0.401_{\pm 0.466}$	0.424	0.600	$0.989_{\pm 0.016}$	$0.806_{\pm 0.234}$
	R^2-Seg wo ST	$0.394_{\pm 0.444}$	0.676	0.482	$0.989_{\pm 0.014}$	$0.836_{\pm 0.215}$
	R^2-Seg wo FPG	$0.184_{\pm 0.324}$	0.955	0.052	$0.988_{\pm 0.014}$	$0.837_{\pm 0.213}$

Table 7. Ablation study: mean $\pm$ std across metrics per tumor and method. ST and FPG refer to statistical tests and false positive gating, respectively.

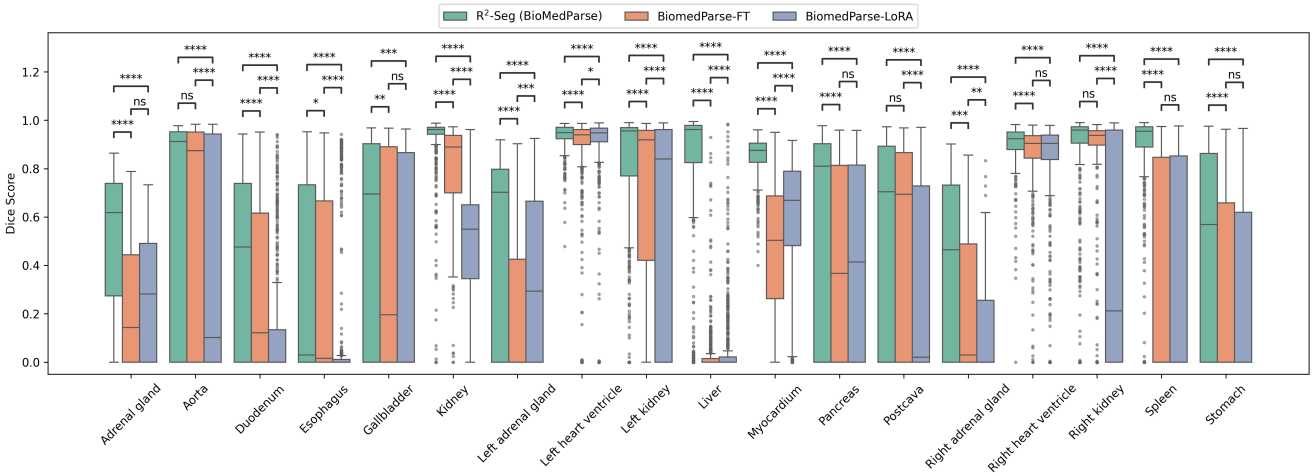

Figure 8. Evaluation of knowledge forgetting on in-distribution MR slices. Statistical tests show that, even finetuned by LoRA, BiomedParse still presents significant performance drops across all organs.