

Conformal Lesion Segmentation for 3D Medical Images

Binyu Tan^{1*}, Zhiyuan Wang^{1*}, Jinhao Duan², Kaidi Xu³,
Heng Tao Shen^{1,4}, Xiaoshuang Shi^{1†}, Fumin Shen^{1†}

¹University of Electronic Science and Technology of China

²University of North Carolina at Chapel Hill

³City University of Hong Kong

⁴Tongji University

{binyutan2024, yhzywang, xsshi2013}@gmail.com jinhao@cs.unc.edu
kaidixu@cityu.edu.hk shenhengtao@hotmail.com fshen@uestc.edu.cn

Abstract

Medical image segmentation serves as a critical component of precision medicine, enabling accurate localization and delineation of pathological regions, such as lesions. However, existing models empirically apply fixed thresholds (e.g., 0.5) to differentiate lesions from the background, offering no statistical guarantees on key metrics such as the false negative rate (FNR). This lack of principled risk control undermines their reliable deployment in high-stakes clinical applications, especially in challenging scenarios like 3D lesion segmentation (3D-LS). To address this issue, we propose a risk-constrained framework, termed Conformal Lesion Segmentation (CLS), that calibrates data-driven thresholds via conformalization to ensure the test-time FNR remains below a target tolerance ε under desired risk levels. CLS begins by holding out a calibration set to analyze the threshold setting for each sample under the FNR tolerance, drawing on the idea of conformal prediction. We define an FNR-specific loss function and identify the critical threshold at which each calibration data point just satisfies the target tolerance. Given a user-specified risk level α , we then determine the approximate $1 - \alpha$ quantile of all the critical thresholds in the calibration set as the test-time confidence threshold. By conformalizing such critical thresholds, CLS generalizes the statistical regularities observed in the calibration set to new test data, providing rigorous FNR constraint while yielding more precise and reliable segmentations. We validate the statistical soundness and predictive performance of CLS on six 3D-LS datasets across five backbone models, and conclude with actionable insights for deploying risk-aware segmentation in clinical practice.

Introduction

Recent progress in deep learning and computer vision (Chen et al. 2025) has facilitated the development of numerous automated segmentation models for medical imaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI) (Moglia et al. 2025; Sun et al. 2025), achieving expert-level performance across various clinical scenarios (Wu et al. 2025). Despite these gains, current models typically depend on a fixed, heuristic threshold (i.e., 0.5)

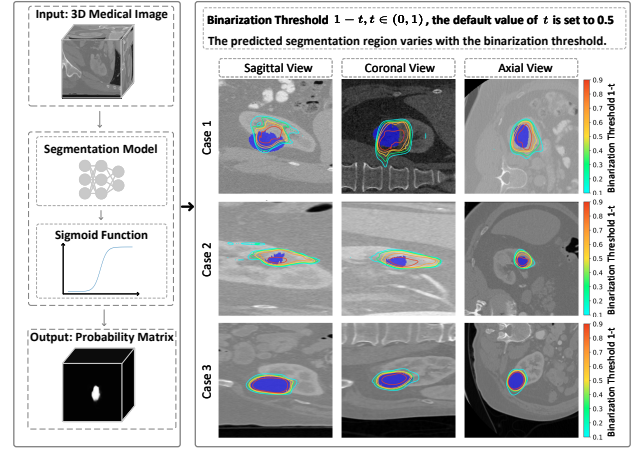


Figure 1: Illustration of 3D-LS tasks.

to delineate target structures from background, lacking statistically valid guarantees for critical safety metrics such as the false negative rate (FNR) (He et al. 2025). This limitation compromises their trustworthiness in risk-sensitive clinical environments, where robust risk control is essential (Moglia et al. 2025). In particular, under-segmentation—failing to fully capture lesion boundaries—can result in pathological regions being missed and thus left untreated (Jalalifar et al. 2022; Zhao et al. 2025). For instance, in early-stage tumor screening, missing lesions smaller than 5 mm can have serious consequences for patient outcomes, as false negatives directly compromise diagnostic and therapeutic decisions (Korhonen et al. 2021; Luo et al. 2023). These concerns highlight the need for statistically grounded segmentation frameworks, especially in challenging scenarios such as 3D lesion segmentation (3D-LS) (Ni et al. 2025; de Grauw et al. 2025), as illustrated in Figure 1, where even small variations in the decision threshold can lead to substantial differences in segmentation outcomes across all three spatial dimensions.

Split (inductive) conformal prediction (SCP) (Papadopoulos et al. 2002; Bates et al. 2021) has recently emerged as a principled solution to address these shortcomings. SCP offers distribution-free, model-agnostic guarantees on ground-

*These authors contributed equally.

†Corresponding Authors.

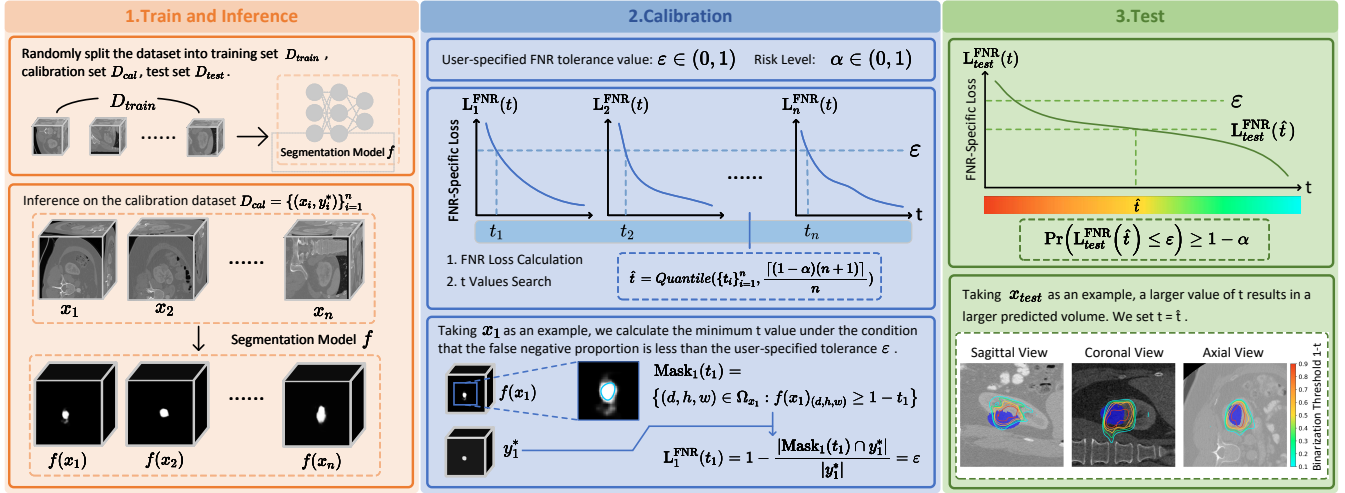


Figure 2: Overview of the CLS framework.

truth coverage, assuming data exchangeability. This framework reserves a held-out calibration set to compute nonconformity scores that quantify the discrepancy between model predictions and ground-truth labels. By computing the quantile of these scores at a user-specified risk level α as the test-time selection threshold, SCP then constructs a prediction set for each test data ensuring that the true label is included with probability at least $1 - \alpha$. While SCP has demonstrated strong performance in classification settings (Angelopoulos et al. 2020), its adaptation to binary image segmentation for FNR control remains a previously under-explored topic.

In this paper, we introduce Conformal Lesion Segmentation (CLS), a novel extension of SCP to image segmentation, which constrains the test-time false negative rate (FNR) to lie below a user-specified tolerance ε , with high probability $1 - \alpha$. As shown in Figure 2, CLS builds upon a pretrained segmentation model and focuses on designing a task-specific nonconformity score that reflects lesion-level false negative risk under the given FNR constraint. Unlike classification tasks, where the nonconformity score is typically defined as one minus the predicted probability of the true class, segmentation requires careful consideration of how thresholding affects lesion detectability. Concretely, for each calibration sample, CLS identifies the maximum decision threshold such that the proportion of false negatives remains below the target tolerance ε . This threshold reflects a critical point: increasing it further would begin to exclude true lesion regions, thereby violating the constraint; lowering it might include more lesions but at the cost of increased false positives. The collection of these per-sample critical thresholds over the calibration set forms a distribution of nonconformity scores, from which CLS computes the approximate $1 - \alpha$ quantile to determine a statistically calibrated test-time threshold. This calibrated threshold ensures, with probability at least $1 - \alpha$, that the FNR on unseen test examples remains within the user-defined tolerance ε .

We evaluate CLS on six 3D-LS benchmarks utilizing five popular 3D medical image segmentation models. Empiri-

cal results demonstrate that CLS consistently enforces the FNR constraint within the predefined tolerance ε at the user-specified risk level α . By rigorously calibrating the test-time threshold, CLS achieves significantly lower FNRs compared to those obtained employing a fixed, heuristic threshold of 0.5 across all datasets. Unlike heuristic uncertainty notions, α serves as a statistically rigorous parameter that provides statistically rigorous control over the allowable constraint violation rate. Beyond FNR control, we further analyze how different models vary in their predicted region sizes across varying risk levels, offering a practical and interpretable tool for benchmarking uncertainty-aware segmentation models.

Our main contributions are summarized as follows:

- We propose Conformal Lesion Segmentation (CLS) that effectively applies SCP to binary segmentation settings.
- We derive novel nonconformity measures from false negative risk-constrained critical thresholds, facilitating statistically rigorous segmentation with FNR control.
- We establish a novel metric for benchmarking model performance specific to uncertainty-aware segmentation.

Related Work

3D Lesion Segmentation. Recently, specialized 3D segmentation models have been developed to tackle challenges specific to 3D-LS tasks. Notable examples include multi-pathway CNNs with CRFs for multiple sclerosis (Saeed et al. 2025), two-pathway 3D CNNs that incorporate contextual MRI information for stroke lesions (Bal et al. 2024), and lightweight CNN-Transformer hybrids like LW-CTrans for small lesion segmentation (Kuang et al. 2025). Transformer-based 3D architectures, such as BrainSegFounder (Cox et al. 2024) and MedSAM2 (Ma et al. 2025), integrate multi-modal inputs and large-scale pretraining to enhance anatomical representation, while ProLesA-Net (Zaridis et al. 2024) enhances prostate lesion segmentation via multi-channel 3D convolutions. Nonetheless, 3D models still inherit the conventional 2D paradigm, applying fixed, heuristic thresholds

(e.g., 0.5) to produce binary masks. These thresholds are typically uncalibrated and lack statistical guarantees on clinically critical metrics such as the false negative rate (FNR), limiting their reliability in high-stakes medical scenarios.

Split Conformal Prediction. SCP is applicable to any pre-trained model to construct sets that are guaranteed to contain the ground truth with a user-specified probability (Angelopoulos and Bates 2021; Wang et al. 2024b, 2025a), under the exchangeability condition (Wang et al. 2025b). Prior studies have effectively applied SCP to image classification scenarios (Liu et al. 2025). Under semantic segmentation settings, recent work views segmentation models as a grid of pixel-level classifiers and constructs prediction sets using a pixel-wise SCP approach (Zhi et al. 2025). We provide the full conformal procedure under classification settings in the appendix. Yet, in binary segmentation tasks like 3D-LS (binary classification), calibrating statistically rigorous prediction sets that validly distinguish foreground lesion regions (class 1) from background (class 0) remains challenging.

Method

Notations and Problem Formulation

Formally, we begin by partitioning the dataset into three disjoint subsets: a training set $\mathcal{D}_{train}$, a calibration set $\mathcal{D}_{cal}$, and a test set $\mathcal{D}_{test}$. Since SCP is compatible with any pre-trained model, we first train a segmentation model $f(\cdot)$ using the training set $\mathcal{D}_{train}$. Let $x_i \in \mathbb{R}^{D \times H \times W}$ denote a 3D image. This model takes x_i as input and produces a confidence map $\hat{y}_i = f(x_i) \in [0, 1]^{D \times H \times W}$, where each element represents the predicted probability of the corresponding voxel belonging to a lesion. The corresponding ground-truth annotation is given by a binary mask $y_i^* \in \{0, 1\}^{D \times H \times W}$, where a value of 1 indicates the presence of pathological tissue at the corresponding location in x_i . Given a decision threshold $t \in [0, 1]$, the predicted lesion region is defined as:

$$\text{Mask}_i(t) = \{(d, h, w) \in \Omega_{x_i} : f(x_i)_{(d,h,w)} \geq 1 - t\}, \quad (1)$$

where $\Omega_{x_i} \subset \mathbb{Z}^3$ is the spatial index domain of the 3D image x_i , (d, h, w) indexes the voxel coordinates, and $f(x_i)_{(d,h,w)}$ is the predicted confidence/probability score. Locations with predicted confidence above $1 - t$ are classified as foreground (lesion), while the rest are considered background.

As previously discussed, heuristically setting a fixed decision threshold $1 - t$ (e.g., 0.5) provides no formal guarantee on the false negative risk. To further illustrate this limitation, we evaluate the pretrained Med3D model (Chen, Ma, and Zheng 2019) on the KiTS21 dataset (Heller et al. 2023), and present the average value distribution of elements in the output probability matrices in Figure 3a. Notably, a substantial portion of predicted probabilities corresponding to ground-truth lesion voxels (i.e., label = 1) fall below 0.5, and these probabilities are distributed relatively uniformly within the lesion regions. Such behavior underscores the inadequacy of using a fixed threshold across samples. **Our objective** is to derive a statistically valid threshold $1 - \hat{t}$ on a calibration set $\mathcal{D}_{cal} = \{(x_i, y_i^*)\}_{i=1}^n$ such that the test-time false negative rate risk remains below a user-specified tolerance level

ε with high probability, formally formulated as:

$$\Pr(R(\hat{t}) \leq \varepsilon) \geq 1 - \alpha, \quad (2)$$

where $R(\hat{t})$ represents the FNR on fresh test instances (the expectation of the false negative proportion) with the decision threshold of $1 - \hat{t}$, and α is a predefined risk level that reflects the maximum acceptable violation/error rate.

Conformal Lesion Segmentation

This section starts with a commonly adopted mild assumption in the SCP framework. We then define the FNR-specific loss and introduce a novel nonconformity score tailored for FNR control on the calibration set. On this basis, we derive a rigorously calibrated test-time decision threshold and establish its statistical validity. Finally, we present the complete workflow of the proposed CLS framework.

Exchangeable data distribution. As a foundational yet non-restrictive assumption, we posit that the n calibration samples $\{(x_i, y_i^*)\}_{i=1}^n$ and each test point (x_{test}, y_{test}^*) in $\mathcal{D}_{test}$ are exchangeable, which underlies the theoretical validity of SCP-based approaches (Angelopoulos and Bates 2021). Notably, exchangeability is a weaker assumption than independent and identically distributed (i.i.d.) data points. We provide a detailed discussion of our assumption in the appendix.

Given the exchangeability between the given test instance and the calibration data points, the calibration set $\mathcal{D}_{cal}$ can be leveraged as a collection of observed data. This enables us to calibrate the segmentation threshold by enforcing an FNR constraint on the calibration set, and to transfer the established statistical guarantees to fresh, unseen data points. To align the nonconformity score with the underlying false negative risk, we define a *threshold-dependent FNR-specific loss function* for each calibration data point:

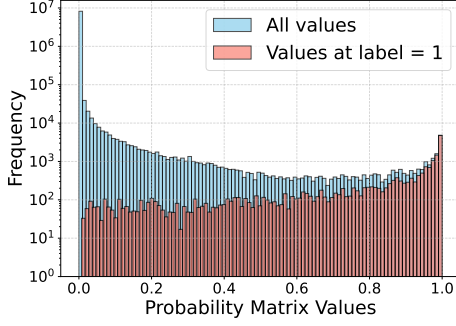
$$\begin{aligned} L_i^{\text{FNR}}(t) &= 1 - \frac{|\text{Mask}_i(t) \cap y_i^*|}{|y_i^*|} \\ &= 1 - \frac{\left| \left\{ \begin{array}{c} (d,h,w) \in \Omega_{x_i} : \\ f(x_i)_{(d,h,w)} \geq 1-t, y_i^*_{(d,h,w)} = 1 \end{array} \right\} \right|}{\left| \left\{ (d,h,w) \in \Omega_{x_i} : y_i^*_{(d,h,w)} = 1 \right\} \right|}, \end{aligned} \quad (3)$$

where $y_i^*_{(d,h,w)}$ is the ground-truth label at voxel (d, h, w) and the loss reflects the proportion of false negatives among all ground-truth lesion voxels, evaluated at threshold $1 - t$. A lower loss indicates that a larger fraction of the ground-truth lesion voxels are successfully identified by the model.

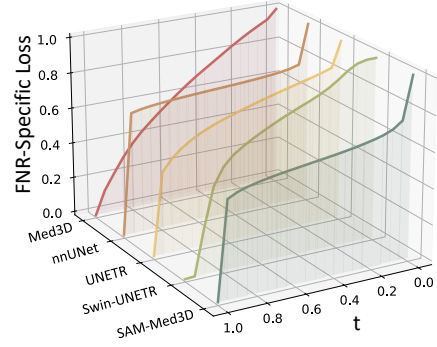
The monotonicity of the FNR-specific loss (i.e., the false negative proportion of each sample under a given decision threshold $1 - t$) with respect to t is immediate from its definition, as also illustrated in Figure 3b. Leveraging this property, we then develop the *nonconformity score* as

$$t_i = \inf \{t : \forall t' \geq t, L_i^{\text{FNR}}(t') \leq \varepsilon\}, \quad (4)$$

which represents the lowest feasible decision threshold for each calibration sample under which the false negative proportion remains within the specified tolerance ε . This critical point t_i minimizes the predicted lesion area $\text{Mask}_i(t_i)$ to enhance precision subject to the risk constraint. Subsequently, we sort all nonconformity scores $\{t_i\}_{i=1}^n$ on the calibration



(a) Value Distribution in Probability Matrix.



(b) FNR-Specific Loss vs. t .

Figure 3: **(a)** The output probability matrices contain a substantial number of voxels with predicted probabilities below 0.5. Lesion-region probabilities are relatively uniformly distributed, causing nearly half of ground-truth lesion voxels to be missed under a fixed threshold of 0.5. **(b)** The FNR-specific loss is monotonically non-increasing with respect to t .

set in ascending order such that $t_1 \leq \dots \leq t_n$, and compute their $\frac{\lceil (1-\alpha)(1+n) \rceil}{n}$ quantile:

$$\hat{t} = \inf \left\{ t : \frac{|\{i : t_i \leq t\}|}{n} \geq \frac{\lceil (1-\alpha)(1+n) \rceil}{n} \right\}. \quad (5)$$

Theorem 1 (Statistically rigorous FNR constraint). *Suppose the given test instance (x_{test}, y_{test}^*) and $(x_i, y_i^*)_{i=1, \dots, n}$ are exchangeable, we employ $\hat{t}$ as the test-time decision threshold and the resulting predicted lesion region is $\text{Mask}_{test}(\hat{t})$. Then the false negative proportion $L_{test}^{FNR}(\hat{t})$ satisfies*

$$\Pr(L_{test}^{FNR}(\hat{t}) \leq \varepsilon) \geq 1 - \alpha. \quad (6)$$

This is the same property of risk control as Eq. (2). Below, we establish its statistical rigor.

Proof of Theorem 1. Under the condition that $s_{i=1, \dots, n}$ are in ascending order, $\hat{t}$ can be reformulated as

$$\hat{t} = t_{n, \frac{\lceil (1-\alpha)(1+n) \rceil}{n}} = t_{\lceil (1-\alpha)(1+n) \rceil}. \quad (7)$$

By the definition of t_i , if $L_{test}^{FNR}(\hat{t}) \leq \varepsilon$, it can obtain

$$\hat{t} \geq t_{test}. \quad (8)$$

Since $(x_1, y_1^*), \dots, (x_n, y_n^*), (x_{test}, y_{test}^*)$ are supposed to be exchangeable, it has

$$\Pr(t_{test} \leq t_i) = \frac{i}{n+1}. \quad (9)$$

Finally, it can obtain

$$\begin{aligned} \Pr(L_{test}^{FNR}(\hat{t}) \leq \varepsilon) &= \Pr(\hat{t} \geq t_{test}) \\ &= \Pr(t_{\lceil (1-\alpha)(1+n) \rceil} \geq t_{test}) \\ &= \frac{\lceil (1-\alpha)(1+n) \rceil}{n+1} \\ &\geq 1 - \alpha \end{aligned} \quad (10)$$

This completes the proof of Theorem 1 and establishes the statistical validity of the test-time decision threshold.

Workflow of CLS. Given a target FNR tolerance ε , we compute nonconformity scores $\{t_i\}_{i=1}^n$ on the calibration set to determine the threshold settings under the risk constraint. We proceed to compute the approximate $1 - \alpha$ quantile $\hat{t}$ of these critical scores, corresponding to the user-specified risk level α . At test time, $\hat{t}$ is employed as the decision threshold for unseen instances. With probability at least $1 - \alpha$, the FNR on the test set is guaranteed to remain below ε . We provide the corresponding pseudocode in the appendix.

Experiments

Experimental Settings

Datasets. We consider six fully annotated 3D medical segmentation datasets from the ULS23 Segmentation Challenge (de Grauw et al. 2025), each covering a different anatomical region or organ system: KiTS21 (Heller et al. 2023), LiTS (Bilic et al. 2023), NIH-LN ABD (Roth et al. 2014), LIDC-IDRI (Armato III et al. 2011), MDSC-Colon (Antonelli et al. 2022), and MDSC-Pancreas (Antonelli et al. 2022). More details of the utilized datasets are provided in the appendix.

Backbone Models. We adopt five popular 3D medical image segmentation models with diverse architectural designs, each fine-tuned to achieve a comparable number of parameters: Med3D(Chen, Ma, and Zheng 2019), nnUNet(Isensee et al. 2021), UNETR(Hatamizadeh et al. 2022), Swin-UNETR(Hatamizadeh et al. 2021), and SAM-Med3D(Wang et al. 2023). More details are provided in the appendix.

Evaluation Metrics. We check whether the *empirical compliance rate* (ECR) (Angelopoulos and Bates 2021), defined as the proportion of test samples whose FNR-specific loss is controlled below ε , exceeds $1 - \alpha$. Beyond FNR control, we also emphasize spatial precision by encouraging compact lesion predictions, as smaller predicted regions are generally more accurate and clinically preferable. We introduce *prediction compactness* (PC), defined as the ratio of the number of predicted lesion voxels to the total number of voxels in the input image. We evaluate PC under the constraint that the FNR remains within the specified tolerance ε , and examine how PC varies with different risk levels α and across models.

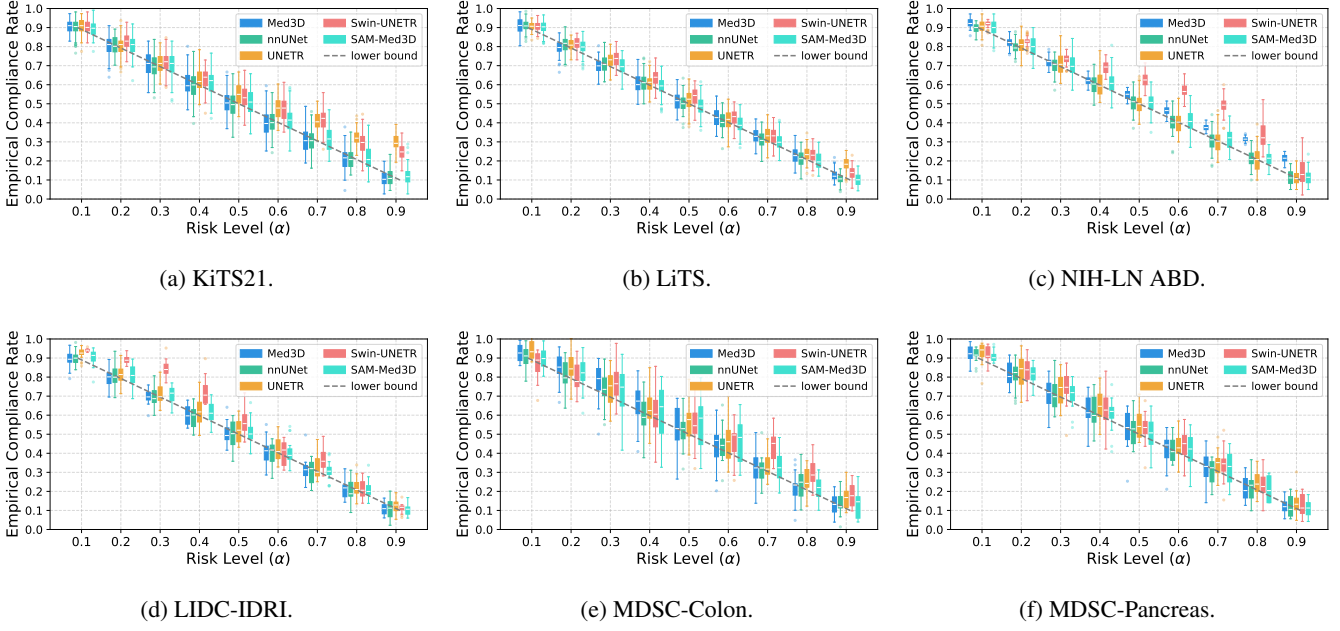


Figure 4: Test-time ECR results on six datasets utilizing five pretrained segmentation models.

Table 1: Test-time FNR (mean $\pm$ std) on the KITS21 dataset under fixed versus CLS-calibrated decision thresholds at varying FNR tolerance levels (ε).

ε / Models	Med3D	nnUNet	UNETR	Swin-UNETR	SAM-Med3D
$t = 0.5$ (Fixed by default)					
	0.5338 $\pm$ 0.2018	0.5122 $\pm$ 0.1971	0.5804 $\pm$ 0.2110	0.5642 $\pm$ 0.1571	0.4935 $\pm$ 0.1478
$t = \hat{t}$ calibrated at the risk level of 0.2					
0.1	0.0533 $\pm$ 0.0169	0.0649 $\pm$ 0.0201	0.0576 $\pm$ 0.0196	0.0466 $\pm$ 0.0136	0.0469 $\pm$ 0.0165
0.2	0.1032 $\pm$ 0.0212	0.0933 $\pm$ 0.0238	0.1145 $\pm$ 0.0263	0.1070 $\pm$ 0.0290	0.0956 $\pm$ 0.0286
0.3	0.1654 $\pm$ 0.0413	0.1764 $\pm$ 0.0346	0.1693 $\pm$ 0.0267	0.1501 $\pm$ 0.0377	0.1779 $\pm$ 0.0328
0.4	0.2253 $\pm$ 0.0836	0.2378 $\pm$ 0.0753	0.2201 $\pm$ 0.0523	0.2540 $\pm$ 0.0688	0.2473 $\pm$ 0.0623
0.5	0.4211 $\pm$ 0.0501	0.4345 $\pm$ 0.0679	0.3928 $\pm$ 0.0540	0.4022 $\pm$ 0.0617	0.3895 $\pm$ 0.0669

While the denominator in PC can alternatively be formulated using the ground-truth lesion volume, such a choice only introduces a constant scaling factor per sample and does not affect the relative ranking of models. Thus, PC provides a robust and interpretable measure for comparing the spatial efficiency of lesion predictions under risk constraints.

Empirical Evaluations

We set the split ratio between the calibration set and the test set to 0.5 by default. Each experimental group is evaluated over 100 random splits of the calibration and test samples.

Statistical Validity of CLS. We begin by demonstrating the statistical rigor of Theorem 1. As illustrated in Figure 4, utilizing each decision threshold calibrated by Eq. (5) effectively constrains the test-time ECR metric under various user-specified risk levels on all six 3D-LS datasets across five pretrained segmentation models. Notably, we expect the ECR results on the test set to approach but remain above the

$1 - \alpha$ lower bound (Angelopoulos and Bates 2021; Wang et al. 2025b) under the strict assumption of data exchangeability. In 100 test runs, we occasionally observe that the ECR falls below the theoretical lower bound of $1 - \alpha$. While the guarantee provided by the SCP framework is statistically rigorous (Ye et al. 2024; Wang et al. 2024a), minor violations can occur in practice due to finite-sample variability.

Comparison with Heuristic Thresholding. We conduct a comprehensive empirical study to compare the performance of CLS with conventional heuristic thresholding in test-time risk-sensitive segmentation. As presented in Table 1, under a fixed threshold of 0.5, all five pretrained segmentation models exhibit substantially high test-time FNRs on the KiTS21 dataset, ranging from 0.4935 (SAM-Med3D) to 0.5804 (UNETR). These results underscore the limitations of heuristic decision thresholds, which fail to adapt to distributional uncertainty and lead to frequent false negatives—an issue particularly critical in clinical settings. By

Table 2: Comparison of test-time ECR (mean $\pm$ std) on six 3D-LS benchmarks using heuristic (fixed $t = 0.5$) and CLS-calibrated thresholds under a risk level of $\alpha = 0.2$.

Datasets / Models	Med3D	nnUNet	UNETR	Swin-UNETR	SAM-Med3D
$t = 0.5$ (Fixed by default)					
KITS21	0.4117 $\pm$ 0.0454	0.4072 $\pm$ 0.0368	0.4549 $\pm$ 0.0378	0.4081 $\pm$ 0.0351	0.5623 $\pm$ 0.0432
LITS	0.3635 $\pm$ 0.0206	0.3384 $\pm$ 0.0283	0.3877 $\pm$ 0.0187	0.3134 $\pm$ 0.0289	0.6002 $\pm$ 0.0210
NIH-LN ABD	0.4275 $\pm$ 0.0322	0.4156 $\pm$ 0.0294	0.4311 $\pm$ 0.0243	0.4078 $\pm$ 0.0305	0.5975 $\pm$ 0.0372
LIDC-IDRI	0.4783 $\pm$ 0.0466	0.4802 $\pm$ 0.0458	0.4512 $\pm$ 0.03964	0.4233 $\pm$ 0.0478	0.5788 $\pm$ 0.0401
MDSC-Colon	0.3029 $\pm$ 0.0261	0.3489 $\pm$ 0.0297	0.3677 $\pm$ 0.0313	0.3167 $\pm$ 0.0212	0.5499 $\pm$ 0.0397
MDSC-Pancreas	0.4588 $\pm$ 0.0453	0.4725 $\pm$ 0.0376	0.3935 $\pm$ 0.0428	0.4360 $\pm$ 0.0451	0.6277 $\pm$ 0.0380
$t = \hat{t}$ calibrated at the risk level of 0.2					
KITS21	0.8080 $\pm$ 0.0513	0.8027 $\pm$ 0.0454	0.8041 $\pm$ 0.0506	0.8289 $\pm$ 0.0429	0.8031 $\pm$ 0.0523
LITS	0.7988 $\pm$ 0.0387	0.8103 $\pm$ 0.0393	0.8086 $\pm$ 0.0278	0.8147 $\pm$ 0.0322	0.8104 $\pm$ 0.0404
NIH-LN ABD	0.8156 $\pm$ 0.0227	0.8198 $\pm$ 0.0275	0.8024 $\pm$ 0.0301	0.8078 $\pm$ 0.0337	0.8254 $\pm$ 0.0375
LIDC-IDRI	0.8137 $\pm$ 0.0523	0.8006 $\pm$ 0.0502	0.7969 $\pm$ 0.0496	0.8125 $\pm$ 0.0433	0.8094 $\pm$ 0.0511
MDSC-Colon	0.8034 $\pm$ 0.0378	0.8023 $\pm$ 0.0343	0.8165 $\pm$ 0.0376	0.8166 $\pm$ 0.0424	0.8154 $\pm$ 0.0344
MDSC-Pancreas	0.8012 $\pm$ 0.0243	0.7943 $\pm$ 0.0298	0.8034 $\pm$ 0.0220	0.8104 $\pm$ 0.0322	0.8060 $\pm$ 0.0348

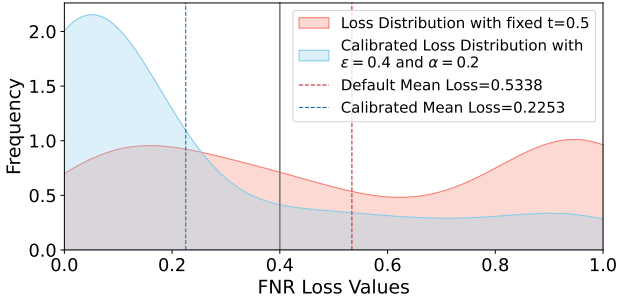


Figure 5: Distribution of FNR-specific loss before (fixed threshold at 0.5) and after CLS calibration ($\varepsilon = 0.4$, $\alpha = 0.2$), illustrating the reduction in average loss and variance.

contrast, CLS develops nonconformity scores based on a tailored FNR-specific loss, enabling the derivation of statistically rigorous thresholds under a user-specified risk level (e.g., $\alpha = 0.2$). This calibration leads to substantial and consistent FNR reductions across all models and tolerance levels (ε). For instance, at a moderate tolerance of $\varepsilon = 0.2$, CLS reduces the FNR of Med3D from 0.5338 to 0.1032, achieving an 80.7% relative reduction. Similarly, significant improvements are observed for nnUNet (from 0.5122 to 0.0933), UNETR (from 0.5804 to 0.1145), Swin-UNETR (from 0.5642 to 0.1070), and SAM-Med3D (from 0.4935 to 0.0956). Furthermore, CLS consistently satisfies the specified FNR tolerance across a wide range of risk levels ($\varepsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$), demonstrating both statistical rigor and adaptability. Notably, even under the strictest constraint ($\varepsilon = 0.1$), FNRs remain below target across all models (e.g., 0.0466 for Swin-UNETR), while relaxed tolerances are met with less conservative thresholds, offering a flexible trade-off between coverage and safety.

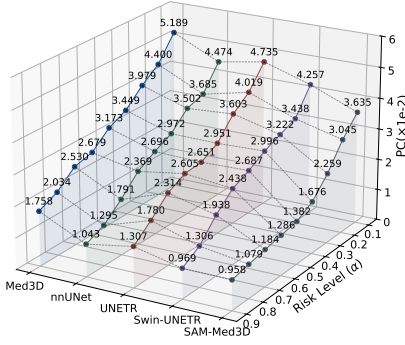
To understand the underlying calibration behavior from

a probabilistic lens, we visualize the distribution of FNR-specific loss values before and after applying CLS calibration. As illustrated in Figure 5, the loss distribution under a fixed threshold (red) is skewed rightward with a high mean of 0.5338, indicating systematic lesion under-segmentation. After CLS calibration with $\varepsilon = 0.4$ and $\alpha = 0.2$ (blue), the distribution shifts sharply leftward, with the mean reduced to 0.2253. This reflects CLS’s ability to tightly control false negatives by adjusting thresholds based on sample-specific risk, resulting in globally improved test-time performance.

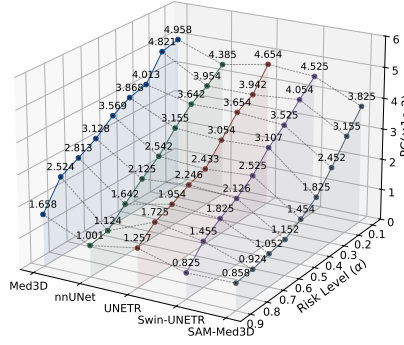
We further examine the generalization capability of CLS across multiple datasets. As shown in Table 2, when using a fixed threshold, ECR values remain low across six 3D lesion segmentation (3D-LS) datasets, with several models performing below 0.40 (e.g., 0.3029 for Med3D on MDSC-Colon, 0.3134 for Swin-UNETR on LiTS). After applying CLS calibration at a risk level of $\alpha = 0.2$, ECR improves substantially and consistently, achieving gains of over +0.45 absolute improvement on average. For example, the ECR of Med3D on LIDC-IDRI increases from 0.4783 to 0.8137, while Swin-UNETR on KiTS21 jumps from 0.4081 to 0.8289, indicating a near doubling in lesion coverage. These results confirm that CLS not only enforces failure rate constraints but also enhances practical segmentation quality.

Importantly, gains are model-agnostic and task-invariant, demonstrating the robustness of CLS across both CNN-based and Transformer-based backbones, as well as across diverse organ and modality domains. By jointly controlling FNR and maximizing coverage, CLS offers a principled and practically effective framework for deploying segmentation models under clinically meaningful risk constraints.

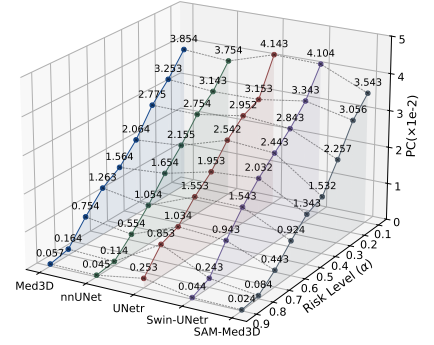
Prediction Compactness as a Risk-aware Benchmark. To further characterize model performance beyond FNR control, we examine PC across six datasets, comparing five segmentation models under varying risk levels. As illustrated in Figure 6, the variation of PC with respect to the risk level α reveals how each model handles uncertainty. For most mod-



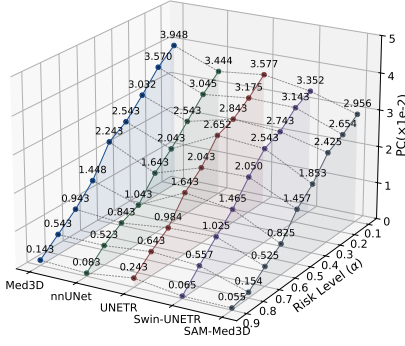
(a) KiTS21.



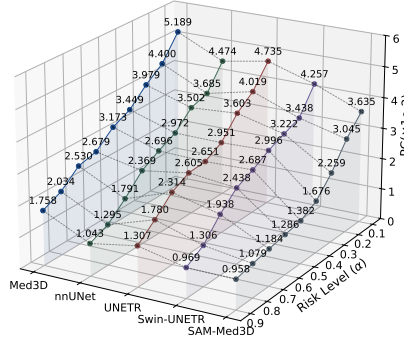
(b) LiTS.



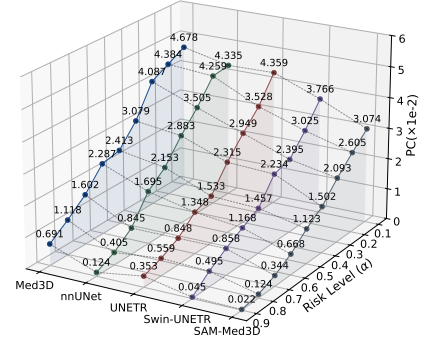
(c) NIH-LN ABD.



(d) LIDC-IDRI.



(e) MDSC-Colon.



(f) MDSC-Pancreas.

Figure 6: PC results on six 3D-LS datasets.

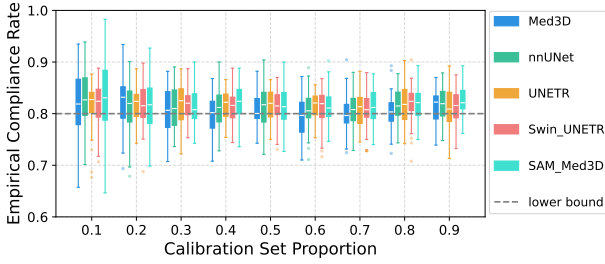


Figure 7: ECR results using Med3D on KiTS21 under various calibration-to-test splits, with $\varepsilon = 0.4$ and $\alpha = 0.2$.

els, PC decreases monotonically with increasing α , consistent with the intuition that lower risk levels allow for more aggressive (i.e., spatially expansive) lesion delineation. This compactness improvement is not achieved at the expense of coverage: by construction, all results satisfy the predefined FNR tolerance ε , confirming that the reduction in PC reflects improved spatial precision under reliable conditions. In clinical segmentation tasks, such compact predictions are crucial for reducing false alarms and annotation overhead. Importantly, PC curves provide a model-agnostic benchmark for assessing the uncertainty structure and calibration efficiency. Models exhibiting a flatter PC curve (i.e., less sensitive to increases in α) tend to allocate predictions more conservatively, indicating higher confidence concentration.

By contrast, models with steep PC growth may reflect less calibrated uncertainty handling. These findings collectively demonstrate that CLS enables not only statistically guaranteed FNR control, but also introduces a robust, interpretable spatial metric—prediction compactness—under risk-aware evaluation. This highlights the utility of CLS in benchmarking and improving model uncertainty behavior in medical segmentation.

Robustness to Calibration-Test Split Ratios. To assess the robustness of CLS under varying calibration data availability, we evaluate its performance using different calibration-test split ratios. As shown in Figure 7, CLS maintains strict control of the test-time ECR across all configurations, even when only 10% of the data is reserved for calibration. These results highlight a critical advantage of our method: statistical guarantees remain valid even in highly imbalanced calibration settings—a desirable trait for scalable and trustworthy deployment of risk-aware medical AI systems.

Conclusion

In this paper, we present CLS, a risk-controlled lesion segmentation framework, which constructs a novel nonconformity score based on a tailored FNR-specific loss and establishes statistically rigorous decision thresholds via conformal calibration. Unlike heuristic thresholding, CLS consistently enforces test-time FNR control across diverse 3D-LS benchmarks, thereby substantially reducing the under-segmentation risk, which is an essential requirement for

practical, safety-critical clinical applications. Beyond reliability, we introduce the *prediction compactness* metric, serving as a novel, interpretable benchmark to quantify spatial precision and model uncertainty under formal risk constraints. We further demonstrate that CLS remains robust and effective even with limited calibration samples, supporting its applicability in resource-constrained settings. Overall, CLS offers a principled, flexible, and practical foundation for deploying uncertainty-aware segmentation models with statistical guarantees—bridging the gap between theoretical soundness and real-world clinical reliability.

References

- Angelopoulos, A.; Bates, S.; Malik, J.; and Jordan, M. I. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.
- Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
- Antonelli, M.; Reinke, A.; Bakas, S.; et al. 2022. The medical segmentation decathlon. *Nature Communications*, 13(1): 4128.
- Armato III, S. G.; McLennan, G.; Bidaut, L.; et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2): 915–931.
- Bal, A.; Banerjee, M.; Chaki, R.; and Sharma, P. 2024. A robust ischemic stroke lesion segmentation technique using two-pathway 3D deep neural network in MR images. *Multimedia Tools and Applications*, 83(14): 41485–41524.
- Bates, S.; Angelopoulos, A.; Lei, L.; et al. 2021. Distribution-free, Risk-controlling Prediction Sets. *Journal of the ACM (JACM)*, 68(6): 1–34.
- Bilic, P.; Christ, P.; Li, H. B.; et al. 2023. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84: 102680.
- Chen, J.; Liu, Y.; Wei, S.; et al. 2025. A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. *Medical Image Analysis*, 100: 103385.
- Chen, S.; Ma, K.; and Zheng, Y. 2019. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*.
- Cox, J.; Liu, P.; Stolte, S. E.; et al. 2024. BrainSegFounder: Towards 3D foundation models for neuroimage segmentation. *Medical Image Analysis*, 97: 103301.
- de Grauw, M.; Scholten, E. T.; Smit, E. J.; et al. 2025. The ULS23 challenge: A baseline model and benchmark dataset for 3D universal lesion segmentation in computed tomography. *Medical Image Analysis*, 102: 103525.
- Farinhas, A.; Zerva, C.; Ulmer, D.; and Martins, A. F. 2023. Non-exchangeable conformal risk control. *arXiv preprint arXiv:2310.01262*.
- Hatamizadeh, A.; Nath, V.; Tang, Y.; et al. 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, 272–284.
- Hatamizadeh, A.; Tang, Y.; Nath, V.; et al. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584.
- He, Y.; Guo, P.; Tang, Y.; et al. 2025. VISTA3D: A unified segmentation foundation model for 3D medical imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20863–20873.
- Heller, N.; Isensee, F.; Trofimova, D.; et al. 2023. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*.
- Isensee, F.; Jaeger, P. F.; Kohl, S. A.; et al. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2): 203–211.
- Jalalifar, S. A.; Soliman, H.; Sahgal, A.; and Sadeghi-Naini, A. 2022. Impact of tumour segmentation accuracy on efficacy of quantitative MRI biomarkers of radiotherapy outcome in brain metastasis. *Cancers*, 14(20): 5133.
- Korhonen, K. E.; Zuckerman, S. P.; Weinstein, S. P.; et al. 2021. Breast MRI: false-negative results and missed opportunities. *Radiographics*, 41(3): 645–664.
- Kuang, H.; Wang, Y.; Tan, X.; et al. 2025. LW-CTrans: A lightweight hybrid network of CNN and Transformer for 3D medical image segmentation. *Medical Image Analysis*, 102: 103545.
- Liu, K.; Sun, T.; Zeng, H.; Zhang, Y.; et al. 2025. Spatial-aware conformal prediction for trustworthy hyperspectral image classification. *arXiv preprint arXiv:2409.01236*.
- Luo, X.; Yang, Y.; Yin, S.; Li, H.; et al. 2023. False-negative and false-positive outcomes of computer-aided detection on brain metastasis: Secondary analysis of a multicenter, multi-reader study. *Neuro-Oncology*, 25(3): 544–556.
- Ma, J.; Yang, Z.; Kim, S.; Chen, B.; et al. 2025. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*.
- Moglia, A.; Cavicchioli, M.; Mainardi, L.; and Cerveri, P. 2025. Deep learning for pancreas segmentation on computed tomography: a systematic review. *Artificial Intelligence Review*, 58(8): 220.
- Ni, G.; Cao, K.; Qin, X.; Zeng, X.; et al. 2025. Advanced 3D retinal lesion segmentation using channel-spatial attention-guided multi-scale feature aggregation. *Biomedical Optics Express*, 16(5): 2093–2110.
- Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gamerman, A. 2002. Inductive Confidence Machines for Regression. In *Machine Learning: ECML 2002*, 345–356.
- Roth, H. R.; Lu, L.; Seff, A.; Cherry, K. M.; et al. 2014. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 520–527.
- Saeed, R.; Ansari, S. U.; Hanif, M.; et al. 2025. Multi-Pathway 3D CNN With Conditional Random Field for Automated Segmentation of Multiple Sclerosis Lesions in MRI. *IEEE Access*, 13: 62154–62164.

Sun, Y.; Wang, L.; Li, G.; Lin, W.; and Wang, L. 2025. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nature Biomedical Engineering*, 9(4): 521–538.

Wang, H.; Guo, S.; Ye, J.; Deng, Z.; et al. 2023. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. *arXiv preprint arXiv:2310.15161*.

Wang, Q.; Geng, T.; Wang, Z.; Wang, T.; Fu, B.; and Zheng, F. 2024a. Sample then identify: A general framework for risk control and assessment in multimodal large language models. *arXiv preprint arXiv:2410.08174*.

Wang, Z.; Duan, J.; Cheng, L.; Zhang, Y.; et al. 2024b. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. *arXiv preprint arXiv:2407.00499*.

Wang, Z.; Duan, J.; Wang, Q.; Zhu, X.; et al. 2025a. COIN: Uncertainty-Guarding Selective Question Answering for Foundation Models with Provable Risk Guarantees. *arXiv preprint arXiv:2506.20178*.

Wang, Z.; Wang, Q.; Zhang, Y.; Chen, T.; et al. 2025b. Sconu: Selective conformal uncertainty in large language models. *arXiv preprint arXiv:2504.14154*.

Wu, J.; Wang, Z.; Hong, M.; Ji, W.; Fu, H.; et al. 2025. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis*, 102: 103547.

Xiao, H.; Li, L.; Liu, Q.; Zhu, X.; and Zhang, Q. 2023. Transformers in medical image segmentation: A review. *Biomedical Signal Processing and Control*, 84: 104791.

Ye, F.; Yang, M.; Pang, J.; Wang, L.; et al. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37: 15356–15385.

Zaridis, D. I.; Mylona, E.; Tsiknakis, N.; et al. 2024. ProLesA-Net: A multi-channel 3D architecture for prostate MRI lesion segmentation with multi-scale channel and spatial attentions. *Patterns*, 5(7): 100992.

Zhao, L.; Wang, T.; Chen, Y.; Zhang, X.; et al. 2025. A novel framework for segmentation of small targets in medical images. *Scientific Reports*, 15(1): 9924.

Zhi, Z.; Feng, C.; Daneshmend, A.; Orlu, M.; et al. 2025. Seeing and Reasoning with Confidence: Supercharging Multimodal LLMs with an Uncertainty-Aware Agentic Framework. *arXiv preprint arXiv:2503.08308*.

Appendix A: Additional Related Work

Background of Split Conformal Prediction. We provide a detailed introduction to the standard split conformal prediction (SCP) procedure in classification tasks (Angelopoulos and Bates 2021; Wang et al. 2025b). SCP provides a principled framework for transforming any heuristic or model-dependent notion of uncertainty into a statistically rigorous one. Given a held-out calibration set of size N , we compute the nonconformity score (NS) for each data point as one minus the softmax probability assigned to its ground-truth class. These scores are then sorted in ascending order, and we select the $\lceil (N+1)(1-\alpha) \rceil / N$ quantile as the threshold.

For a new test instance, we evaluate the softmax outputs across all classes. Any class whose softmax probability exceeds the derived threshold is included in the prediction set. Under the assumption of exchangeability, this construction guarantees that the true label will be included in the prediction set with approximate probability $1 - \alpha$. This procedure yields valid marginal coverage on finite-sample test data, offering statistically meaningful uncertainty estimates. The complete framework is presented as follows:

1. Given the calibration dataset $\{(X_i, Y_i^*)\}_{i=1}^n$ (i.i.d.) and a pretrained model $\hat{f}(\cdot)$ that produces probabilistic predictions ($\hat{f}(X_i) \in [0, 1]^K$, representing a probability distribution over K classes for input X_i). The predicted probability assigned to the ground-truth class Y_i^* is denoted by $\hat{f}(X_i)_{Y_i^*}$.
2. Define the nonconformity score for each calibration sample as a measure of uncertainty associated with its true class: $s_i = s(X_i, Y_i^*) = 1 - \hat{f}(X_i)_{Y_i^*}$, where $\hat{f}(X_i)_{Y_i^*}$ denotes the predicted probability for the ground-truth label ($\{s_1 \leq s_2 \leq \dots \leq s_n\}$).
3. Compute the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of $\{s_i\}_{i=1}^n$: $\hat{q} = \inf \left\{ q : \frac{|\{i: s_i \leq q\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\} = s_{\lceil (n+1)(1-\alpha) \rceil}$.
4. Construct the prediction set for X_{test} : $\mathcal{C}(X_{test}) = \{y \in [K] : s(X_{test}, y) \leq \hat{q}\}$.
5. The event $Y_{test}^* \in \mathcal{C}(X_{test})$ is equivalent to the condition $s(X_{test}, Y_{test}^*) \leq \hat{q}$. As long as this inequality holds, the true label is guaranteed to be included in the prediction set $\mathcal{C}(X_{test})$. Consequently, it can obtain a prediction set that successfully captures the ground-truth class.
6. Owing to the exchangeability of the $n+1$ data points (the n calibration samples and one test instance), it has: $\mathbb{P}(s_{test} \leq s_{(i)}) = \frac{i}{n+1}$.
7. Based on the exchangeability assumption, it can obtain that the probability of the prediction set covering the true label satisfies: $\mathbb{P}(Y_{test}^* \in \mathcal{C}(X_{test})) = \mathbb{P}(s_{test} \leq \hat{q}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \geq 1 - \alpha$. This guarantees a marginal coverage level of at least $1 - \alpha$ on the test distribution.

Appendix B: Assumptions

Exchangeable data distribution. We assume that the inputs to the segmentation model are independently and identically drawn from a fixed underlying distribution. This assumption

is reasonable for many standard scenarios, such as the 3D medical imaging tasks for lesion segmentation explored in our experiments. However, it is important to note that this assumption does not hold in the presence of distribution shifts between the calibration and testing phases.

Throughout the paper, we utilize the following formal definition of exchangeable data distribution (Farinhas et al. 2023; Wang et al. 2025b), which is a weaker assumption than independent and identically distributed (i.i.d.) data. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, respectively. A data distribution in $\mathcal{X} \times \mathcal{Y}$ is said to be exchangeable if and only if $P((X_1, Y_1), \dots, (X_n, Y_n)) = P((X_{\pi(1)}, Y_{\pi(1)}), \dots, (X_{\pi(n)}, Y_{\pi(n)}))$ holds for any finite collection $(X_i, Y_i)_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ and for any permutation π of $1, \dots, n$. It is important to note that every i.i.d. (independent and identically distributed) sequence is also exchangeable, since $P((X_1, Y_1), \dots, (X_n, Y_n)) = \prod_{i=1}^n P(X_i, Y_i)$.

The proposed method exhibits potential for extension to settings involving certain types of distribution shift, which are common in real-world applications due to variations in data acquisition protocols, patient populations, or imaging modalities. One possible direction involves adapting the conformal calibration procedure by incorporating instance-wise or domain-adaptive weighting schemes when computing the nonconformity scores. These weighting schemes can be carefully designed to ensure that the resulting t-values retain the super-uniformity property relative to the target distribution, thereby preserving the validity guarantees under distributional changes and enhancing the robustness of the method in practical deployment.

Appendix C: CLS Algorithm Description

The algorithm begins with a calibration dataset $\mathcal{D}_{cal} = (x_i, y_i^*)_{i=1}^n$, where each x_i denotes an input image and y_i^* is the corresponding ground-truth segmentation mask. This calibration set is obtained as one realization from 100 random data splits to enhance the robustness and reliability of the evaluation. For each calibration sample, we employ a pretrained segmentation model $f(\cdot)$ to compute the corresponding confidence map $\hat{y}_i = f(x_i) \in [0, 1]^{D \times H \times W}$.

For each calibration sample, we perform a binary search to identify the smallest threshold t_i such that the corresponding FNR-specific loss satisfies $L_i^{\text{FNR}}(t_i) \leq \varepsilon$, within a specified numerical tolerance δ (e.g., 10^{-4}). This process yields a set of thresholds $\{t_i\}_{i=1}^n$, where each t_i is individually calibrated to ensure that the user-defined false negative tolerance ε is satisfied for its respective sample.

Subsequently, we compute the $\frac{\lceil (1-\alpha)(1+n) \rceil}{n}$ quantile of the sorted threshold set as the global decision threshold $\hat{t}$, which is then applied to segment unseen test instances. This guarantees, under the conformal prediction framework, that the resulting segmentation satisfies the false negative risk constraint $\Pr(L_{test}^{\text{FNR}}(\hat{t}) \leq \varepsilon) \geq 1 - \alpha$.

Algorithm 1: CONFORMAL LESION SEGMENTATION

Require: Calibration set $\mathcal{D}_{cal} = \{(x_i, y_i^*)\}_{i=1}^n$
 Pretrained segmentation model $f(\cdot)$
 FNR-specific loss $L_i^{\text{FNR}}(t) = 1 - \frac{|\text{Mask}_i(t) \cap y_i^*|}{|y_i^*|}$
 User-specified FNR tolerance $\varepsilon \in (0, 1)$
 Risk level $\alpha \in (0, 1)$
 Fresh test instances x_{test}

Ensure: Segmentation mask $\text{Mask}_{test}(\hat{t})$
 such that $\Pr(L_{test}^{\text{FNR}}(\hat{t}) \leq \varepsilon) \geq 1 - \alpha$

- 1: **for** $i = 1$ to n **do**
- 2: $\hat{y}_i \leftarrow f(x_i)$
- 3: Initialize $t_{\min} \leftarrow 0, t_{\max} \leftarrow 1$
- 4: Set numerical tolerance δ (e.g., $1e-4$)
- 5: **while** $|L_i^{\text{FNR}}(t) - \varepsilon| > \delta$ **do**
- 6: $t \leftarrow \frac{t_{\min} + t_{\max}}{2}$
- 7: $\text{Mask}_i(t) \leftarrow \{(d, h, w) \in \Omega_{x_i} : \hat{y}_{i(d, h, w)} \geq 1 - t\}$
- 8: $L_i^{\text{FNR}}(t) \leftarrow 1 - \frac{|\text{Mask}_i(t) \cap y_i^*|}{|y_i^*|}$
- 9: **if** $L_i^{\text{FNR}}(t) > \varepsilon$ **then**
- 10: $t_{\min} \leftarrow t$
- 11: **else**
- 12: $t_{\max} \leftarrow t$
- 13: **end if**
- 14: $t_i \leftarrow t$
- 15: **end while**
- 16: **end for**
- 17: Sort thresholds $\{t_i\}_{i=1}^n$ in ascending order
- 18: $\hat{t} \leftarrow \inf \left\{ t : \frac{|\{i: t_i \leq t\}|}{n} \geq \frac{\lceil (1-\alpha)(1+n) \rceil}{n} \right\}$
- 19: $\text{Mask}_{test}(\hat{t}) \leftarrow \left\{ \begin{array}{l} (d, h, w) \in \Omega_{x_{test}} : \\ f(x_{test})_{(d, h, w)} \geq 1 - \hat{t} \end{array} \right\}$
- 20: **return** $\text{Mask}_{test}(\hat{t})$

Table 3: Summary statistics of 3D-LS datasets.

Datasets	Data Type	#Series	#Lesions	#Train
KiTS21	Kidney	300	332	249
LiTS	Liver	113	832	624
NIH-LN ABD	Lymph Nodes	85	557	417
LIDC-IDRI	Lung	750	2236	1677
MDSC-Colon	Colon	126	131	98
MDSC-Pancreas	Pancreas	281	283	212

Appendix D: Details of Experimental Settings

D.1 Details of Datasets

As shown in the Table 3, we employ six publicly available 3D medical image segmentation datasets, each targeting a distinct organ type. The column #Series reports the number of annotated 3D imaging volumes (i.e., patient-level scans), #Lesions indicates the total number of labeled lesion instances, and #Train denotes the number of training samples utilized in our experimental setup.

KiTS21 (Heller et al. 2023) is a benchmark dataset released as part of a public challenge designed to advance automatic segmentation of kidneys and renal tumors from

clinical abdominal CT scans. The dataset comprises multi-institutional CT volumes, each annotated independently by three experts, and includes a held-out test set from an external center to rigorously evaluate model generalizability. The dataset offers high-quality manual annotations for both kidney parenchyma and tumors, enabling robust and standardized performance comparisons across methods. Due to the substantial variability in tumor size, shape, and anatomical location, the KiTS21 dataset poses a challenging segmentation task and is widely utilized as a benchmark for evaluating the generalization capability of models in jointly segmenting organs and associated lesions.

LiTS (Bilic et al. 2023) is a widely adopted benchmark dataset for liver and tumor segmentation and detection, comprising 201 contrast-enhanced abdominal CT volumes collected from multiple clinical institutions. Each volume is annotated with pixel-wise labels for the liver and intrahepatic tumors, including both primary and secondary lesions. The dataset poses significant challenges due to the low contrast, ambiguous boundaries, and diverse morphological characteristics of the tumors. Despite these complexities, LiTS remains a standard benchmark in the field and is particularly valuable for assessing model performance on small lesion segmentation and hepatic pathology analysis.

NIH-LN ABD (Roth et al. 2014) is a benchmark dataset developed by the National Institutes of Health (NIH) for evaluating lymph node detection and multi-organ segmentation algorithms. It consists of abdominal CT scans with pixel-level annotations for multiple organs, including the liver, spleen, kidneys, pancreas, and abdominal lymph nodes. The dataset is particularly challenging due to the low contrast, variable size of lymph nodes, and the anatomical complexity of abdominal structures. NIH-LN ABD is widely used in research involving lymph node detection, classification, and multi-organ segmentation in clinically complex contexts.

LIDC-IDRI (Armato III et al. 2011) is a widely used public dataset for the development and evaluation of computer-aided detection and diagnosis systems for pulmonary nodules. It consists of 1018 low-dose thoracic CT scans, each annotated by four experienced radiologists using a two-phase review process (initially blinded, then unblinded). The dataset includes a total of 7371 marked lesions, with 2669 nodules annotated with detailed contours and subjective characteristics. Following prior studies (de Grauw et al. 2025), we selected a subset of 2236 nodules for our analysis, based on the availability of complete annotations and clinical relevance. Rich metadata such as nodule size, location, and boundary definition makes LIDC-IDRI a valuable resource for research on lung nodule segmentation, particularly in the context of inter-observer variability and uncertainty-aware modeling.

MDSC-Colon (Antonelli et al. 2022) is a subtask of the Medical Segmentation Decathlon (MSD), specifically curated to evaluate the effectiveness of segmentation algorithms in colorectal tumor analysis. This dataset consists of contrast-enhanced abdominal CT scans collected from colon cancer patients across multiple clinical institutions, accompanied by high-quality, expert-annotated tumor delineations. The segmentation task is notably challenging due to the sub-

Table 4: Models and parameter sizes.

Models	Parameters (M)
Med3D	43.8365
nnUNet	44.2345
UNETR	44.6862
Swin-UNETR	45.0454
SAM-Med3D	45.5447

stantial heterogeneity in tumor morphology, including irregular shapes, asymmetry, and the presence of diffuse or poorly defined boundaries. MDSC-Colon serves as a rigorous benchmark for assessing a model’s ability to handle anatomically complex and variable lesion presentations.

MDSC-Pancreas (Antonelli et al. 2022) represents another task within the Medical Segmentation Decathlon (MSD), aimed at evaluating segmentation performance on the pancreas and its associated pathological structures, including tumors and cysts. The dataset comprises contrast-enhanced abdominal CT scans acquired from multiple clinical institutions, each annotated with high-precision labels for both the pancreas and relevant lesions. Pancreatic segmentation poses considerable challenges due to the organ’s small size, highly variable shape, and low contrast relative to surrounding anatomical structures.

D.2 Details of Models and Fine-tuning Strategies

We selected five 3D medical image segmentation models, each characterized by a distinct architectural design, and applied customized fine-tuning strategies to align their parameter counts for a fair and consistent comparison. Table 4 provides a summary of the selected models and their respective parameter counts following the fine-tuning strategies.

Med3D (Chen, Ma, and Zheng 2019) is a convolutional neural network (CNN) framework based on an encoder-decoder architecture, specifically adapted for 3D medical image segmentation. Its encoder is derived from the ResNet family and modified to accommodate volumetric data. To ensure a fair comparison with transformer-based models, Med3D increases the number of channels and convolutional layers at each stage, thereby expanding the model’s capacity while maintaining its fully convolutional design. These enhancements enable Med3D to serve as a strong CNN baseline for evaluating downstream 3D segmentation performance.

nnU-Net (Isensee et al. 2021) is a self-configuring, CNN-based segmentation framework built upon the standard U-Net encoder-decoder architecture. Designed to adapt to the characteristics of each target dataset, nnU-Net automatically determines optimal preprocessing steps, network configurations, and training protocols. In this work, we follow prior practice by increasing the number of filters and deepening the convolutional blocks at each resolution level, thereby matching the model’s parameter count with transformer-based counterparts. These adaptations ensure a fair comparison in downstream segmentation performance.

UNETR (Hatamizadeh et al. 2022) integrates Vision Transformers (ViTs) into the classical U-shaped architecture for

3D medical image segmentation, leveraging transformers’ ability to model long-range dependencies while preserving the spatial precision of convolutional decoders. Although convolutional layers excel at capturing local features, they often struggle with global semantic understanding (Xiao et al. 2023). To enable a fair comparison with CNN-based models, UNETR is modified by reducing the hidden size, MLP dimensionality, and number of attention heads in the transformer encoder. These modifications substantially decrease the parameter count while maintaining the core advantages of transformer-based representation learning.

Swin-UNETR (Hatamizadeh et al. 2021) integrates a Swin Transformer-based encoder with a fully convolutional decoder for 3D medical image segmentation. It employs a hierarchical architecture with shifted window self-attention, enabling efficient modeling of both local and global features while significantly reducing computational overhead. To maintain a comparable parameter scale with other models, the network’s depth and feature dimensions are deliberately reduced. These modifications retain the core advantages of window-based attention, while enhancing scalability and reducing the overall computational burden.

SAM-Med3D (Wang et al. 2023) extends the Segment Anything Model (SAM) framework to 3D medical image segmentation by adapting its core components, including the image encoder, prompt encoder, and mask decoder, to volumetric data. The model employs 3D convolutions, learnable 3D absolute positional embeddings, and 3D attention mechanisms to support spatial representation in three dimensions. To reduce model complexity, SAM-Med3D decreases the embedding dimensions and the number of attention heads in the transformer layers. This lightweight design maintains the original ViT-based architecture while improving efficiency and performance in 3D medical segmentation tasks.