Benchmarking SAM and MedSAM Robustness under Noisy Abdominal CT Conditions

Hoang Le Chau

Project Report

Contents

Abstract

The Segment Anything Model (SAM) has emerged as a transformative foundation model for image segmentation, demonstrating impressive zero-shot capabilities across natural image domains. However, its robustness under realistic medical imaging conditions, particularly in the presence of noise and artifacts common in clinical practice, remains understudied. This study presents a comprehensive benchmark evaluation of SAM and its medical adaptation MedSAM on abdominal CT segmentation tasks under systematically controlled noise conditions. Using the Medical Segmentation Decathlon datasets for liver (Task 03) and spleen (Task 09) segmentation, we evaluate model performance across six noise types (Gaussian, Poisson, salt-and-pepper, motion blur, intensity inhomogeneity, and low-contrast degradation) at three intensity levels (mild, moderate, severe). Our experimental results reveal that both SAM and MedSAM exhibit substantial performance degradation under noisy conditions, with mean Dice coefficients dropping from 0.131 on clean liver images to as low as 0.039 under severe intensity inhomogeneity noise. MedSAM demonstrates marginally better noise resilience on liver segmentation but shows inconsistent performance on spleen segmentation. Salt-and-pepper noise causes the most severe performance degradation across both models and datasets, while motion blur shows relatively less impact. Statistical analysis confirms significant performance differences between clean and noisy conditions ($p < 0.05$) for most noise variants. These findings highlight critical limitations of current foundation models for medical image segmentation and underscore the necessity for robust adaptation strategies when deploying such models in clinical environments where image quality variations are inevitable.

1 Introduction

Medical image segmentation plays a fundamental role in clinical diagnosis, treatment planning, and disease monitoring. Accurate delineation of anatomical structures and pathological regions from medical images enables quantitative analysis essential for evidence-based medical decision-making. Traditional segmentation approaches have relied heavily on domain-specific architectures trained on carefully curated datasets with expert annotations. The emergence of foundation models, particularly the Segment Anything Model (SAM), has introduced a paradigm shift by demonstrating remarkable zero-shot generalization capabilities across diverse visual tasks.

SAM, developed by Meta AI, represents a breakthrough in general-purpose image segmentation through its training on over 11 million images and 1 billion masks. Its prompt-based architecture enables interactive segmentation with minimal user input, making it particularly attractive for medical imaging applications where expert annotations are expensive and time-consuming to obtain. Following SAM's success, medical-specific adaptations such as MedSAM have been developed to address the unique challenges of medical image segmentation, including domain shift between natural and medical images, specialized anatomical knowledge requirements, and the prevalence of low-contrast structures.

Despite these advances, a critical gap exists in understanding how these foundation models perform under realistic medical imaging conditions characterized by various noise types and image quality degradations. Medical images, particularly computed tomography (CT) scans, are inherently susceptible to multiple sources of noise and artifacts arising from hardware limitations, patient motion, contrast variations, and acquisition protocols. Previous research has demonstrated that SAM's out-of-the-box performance on medical images is highly variable, with significant accuracy drops in the presence of noise, low contrast, shadow artifacts, and blurred boundaries that are common in modalities like ultrasound, magnetic resonance imaging (MRI), and CT. Studies show SAM's Intersection over Union (IoU) scores ranging from 0.11 for spine MRI to 0.86 for hip X-ray, with notably poor results for ambiguous or noisy regions (Mazurowski et al., 2023).

The need for systematic benchmarking of foundation models under controlled noise conditions is particularly acute given the intended deployment of such models in clinical settings where image quality cannot always be guaranteed. While several adaptation strategies have been proposed, including adapter modules, uncertainty estimation frameworks, and domain-specific fine-tuning approaches, their effectiveness specifically under noisy conditions requires empirical validation. Understanding the failure modes and degradation patterns of foundation

models under various noise conditions is essential for developing targeted robustness enhancement strategies and establishing appropriate clinical deployment guidelines.

This study addresses this critical gap through a comprehensive benchmarking evaluation of SAM and MedSAM on abdominal CT segmentation under systematically controlled noise conditions. We make the following contributions: (1) establishment of a rigorous evaluation framework for assessing segmentation model robustness across six distinct noise types commonly encountered in medical CT imaging; (2) comprehensive performance analysis across 19 experimental conditions including clean data baseline and 18 noise variants at three intensity levels; (3) quantitative characterization of performance degradation patterns specific to anatomical structures (liver and spleen); and (4) statistical validation of performance differences and identification of noise types causing the most severe degradation. Our findings provide empirical evidence regarding the limitations of current foundation models in handling realistic clinical imaging conditions and offer insights to guide future development of noise-robust medical image segmentation approaches.

2 Related Work

2.1 Foundation Models for Medical Image Segmentation

Foundation models have emerged as a transformative paradigm in computer vision, with SAM representing a significant milestone in image segmentation. SAM's training on massive-scale datasets enables impressive zero-shot capabilities, prompting extensive investigation into medical imaging applications. Early evaluations revealed that SAM's direct application to medical images yields mixed results, with performance highly dependent on image modality, anatomical structure, and image quality (Huang et al., 2024; Mazurowski et al., 2023). Comprehensive benchmarking studies across multiple medical imaging modalities have demonstrated that while SAM can segment well-circumscribed structures in high-quality images, its accuracy drops significantly under challenging conditions including noise, low contrast, and ambiguous boundaries (Ma et al., 2024; Shi et al., 2023).

Several medical-specific adaptations of SAM have been developed to address domain-specific challenges. MedSAM, trained on a large-scale medical imaging dataset, demonstrates improved performance on medical segmentation tasks compared to vanilla SAM (Ma et al., 2024). The Medical SAM Adapter employs lightweight adapter modules to incorporate medical domain knowledge while preserving SAM's pre-trained representations (Wu et al., 2025). MA-SAM introduces modality-agnostic adaptations enabling robust 3D medical image segmentation across CT and MRI (C. Chen et al., 2024). These adaptations represent important steps toward clinical deployment, though their robustness under noisy conditions requires systematic evaluation.

2.2 Noise Robustness in Medical Image Segmentation

Medical image segmentation under noisy conditions has been studied extensively in the context of traditional deep learning architectures. Wang et al. (2020) proposed a noise-robust framework for COVID-19 pneumonia lesion segmentation from CT images, demonstrating that noise-robust loss functions and self-ensembling strategies can improve performance under label noise conditions. Dong et al. (2025) developed a deep self-cleansing segmentation framework for medical images with noisy labels, integrating Gaussian Mixture Model-based label filtering and pixel-level label correction mechanisms.

Uncertainty estimation has emerged as a critical component for robust segmentation under noise. Han et al. (2025) proposed region uncertainty estimation frameworks that stratify samples according to label quality, enabling reliable supervision propagation even with noisy annotations. Zou et al. (2025) developed evidential calibrated uncertainty modeling for medical image segmentation, demonstrating improved reliability in noisy scenarios. These approaches highlight the importance of explicit uncertainty quantification for clinical trust and error mitigation in noisy medical images.

The application of SAM to noisy medical imaging remains relatively understudied despite its clinical importance. Guo et al. (2025) demonstrated that adapter-based models such as ESAM2-BLS can improve performance on low-quality ultrasound images through preprocessing and domain-specific adaptations. Bai et al. (2025) showed that adapter mechanisms can enhance SAM's segmentation accuracy on low-quality medical images. However, systematic benchmarking across multiple noise types and intensity levels for abdominal CT segmentation, which is the focus of this study, has not been comprehensively addressed in the literature.

2.3 Medical Segmentation Datasets and Benchmarks

The Medical Segmentation Decathlon (MSD) provides standardized benchmarks across multiple organ segmentation tasks with diverse imaging modalities. Task 03 (Liver) and Task 09 (Spleen) from the MSD represent clinically relevant abdominal organ segmentation challenges commonly encountered in diagnostic radiology and treatment planning (Antonelli et al., 2022). These datasets have been extensively used in prior studies for evaluating segmentation algorithms, providing a solid foundation for comparative analysis.

Previous benchmarking studies have evaluated various aspects of segmentation model performance including generalization capability, annotation efficiency, and computational requirements. However, systematic evaluation under controlled noise conditions specifically for foundation models remains limited. Our study addresses this gap by establishing a comprehensive noise robustness benchmark for SAM and MedSAM on standardized abdominal CT datasets, enabling reproducible evaluation and comparison of future noise-robust adaptation strategies.

3 Methodology

3.1 Datasets

We utilized two datasets from the Medical Segmentation Decathlon (MSD) challenge: Task 03 (Liver Tumor Segmentation) and Task 09 (Spleen Segmentation). These datasets were selected due to their clinical relevance, standardized annotations, and representation of typical abdominal CT imaging characteristics. The liver dataset consists of portal venous phase CT scans with annotations for liver tissue, while the spleen dataset contains contrast-enhanced CT volumes with spleen segmentation masks. Both datasets present realistic clinical imaging scenarios with varying image quality and contrast characteristics (Figures 1 and 2).

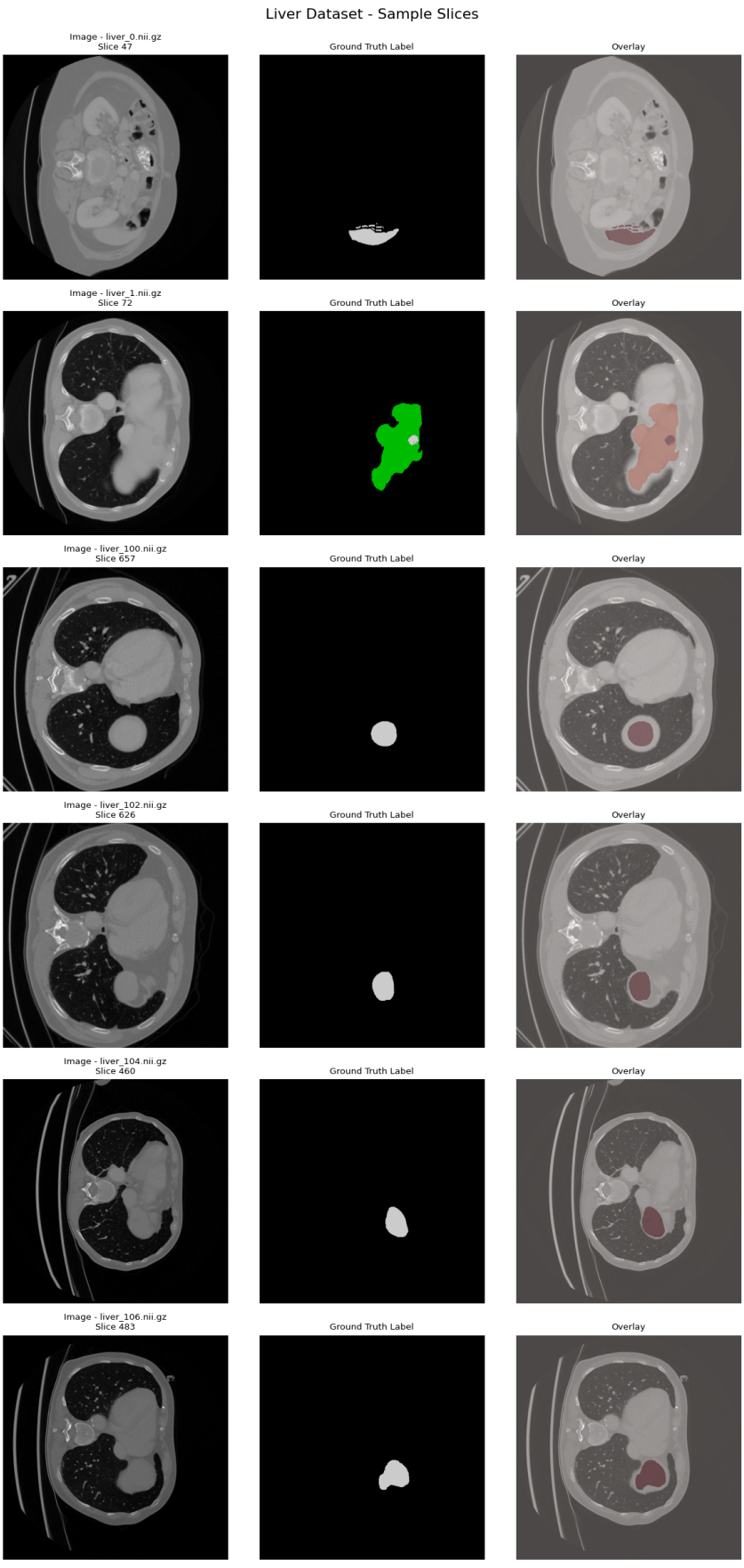

Figure 1: Representative 2D axial slices from the MSD Liver dataset (Task 03)

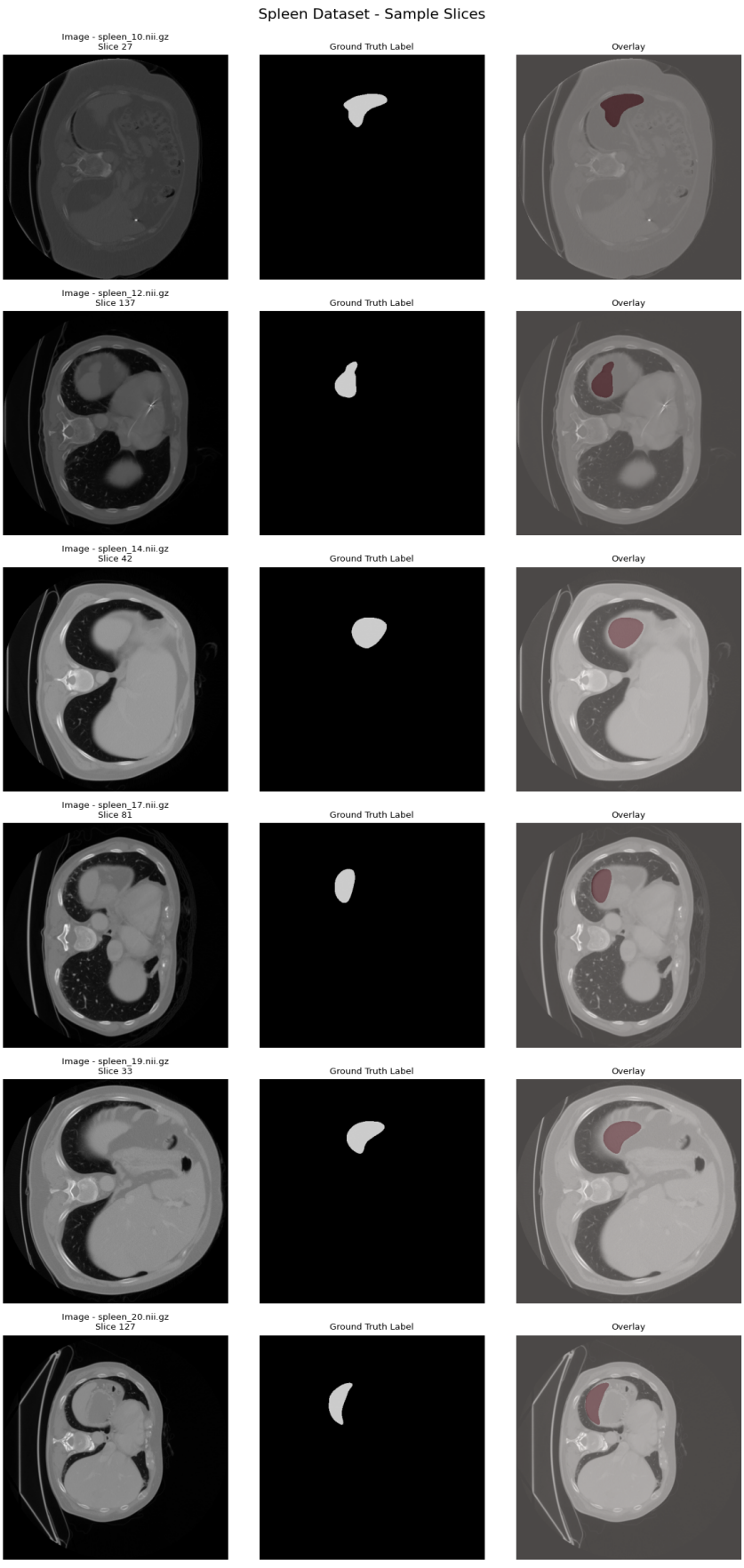

Figure 2: Representative 2D axial slices from the Spleen dataset (Task 09)

For experimental consistency, we extracted 50 representative 2D axial slices from each dataset, ensuring sufficient anatomical coverage while maintaining computational feasibility for comprehensive noise robustness evaluation. Slice selection prioritized images containing substantial foreground tissue to enable meaningful segmentation assessment. The extracted slices were normalized to 512×512 pixel resolution to match SAM's input requirements and facilitate standardized processing across all experimental conditions.

3.2 Noise Injection Framework

To systematically evaluate model robustness, we implemented a comprehensive noise injection pipeline simulating six distinct types of image degradation commonly encountered in medical CT imaging:

Gaussian Noise: Additive white Gaussian noise with zero mean and varying standard deviations was applied to simulate electronic noise from CT detector systems. Intensity levels were calibrated to represent mild, moderate, and severe noise conditions corresponding to different imaging protocols and dose levels.

Poisson Noise: Shot noise following Poisson distribution was introduced to simulate quantum noise inherent to photon-counting CT systems. This represents the fundamental physical noise arising from the discrete nature of X-ray photons.

Salt-and-Pepper Noise: Random intensity spikes (salt) and drops (pepper) were introduced to simulate impulse noise from detector malfunctions or transmission errors. This noise type is particularly challenging as it creates high-frequency artifacts.

Motion Blur: Simulated patient motion during acquisition by applying directional blur kernels. Motion artifacts are common in clinical practice, especially for uncooperative patients or lengthy acquisition protocols.

Intensity Inhomogeneity: Gradual intensity variations across the image were introduced using smooth bias fields to simulate field inhomogeneity effects common in CT imaging, particularly at field edges or in the presence of metallic implants.

Low-Contrast Degradation: Contrast reduction was applied through histogram transformation to simulate scenarios with suboptimal contrast agent administration or low-dose imaging protocols.

Each noise type was systematically applied at three intensity levels (mild, moderate, severe) representing progressively challenging imaging conditions. The intensity parameters were calibrated based on noise characteristics observed in clinical CT images to ensure realistic simulation. This yielded 18 distinct noise conditions per dataset in addition to clean baseline images, resulting in 19 experimental conditions total for comprehensive robustness evaluation (Figures 3 and 4).

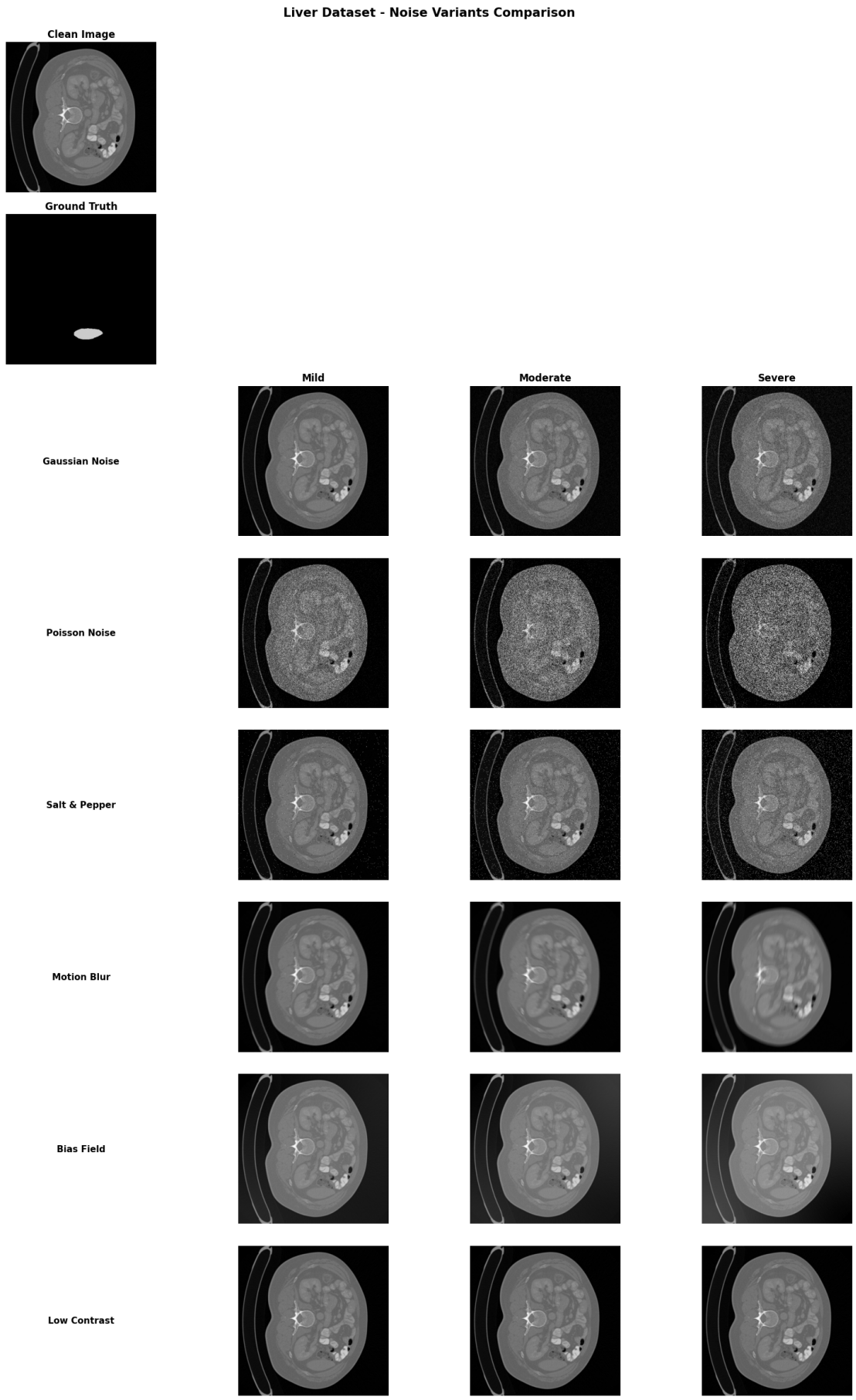

Figure 3: Grid visualization of all 18 noise variants applied to representative samples from the liver dataset.

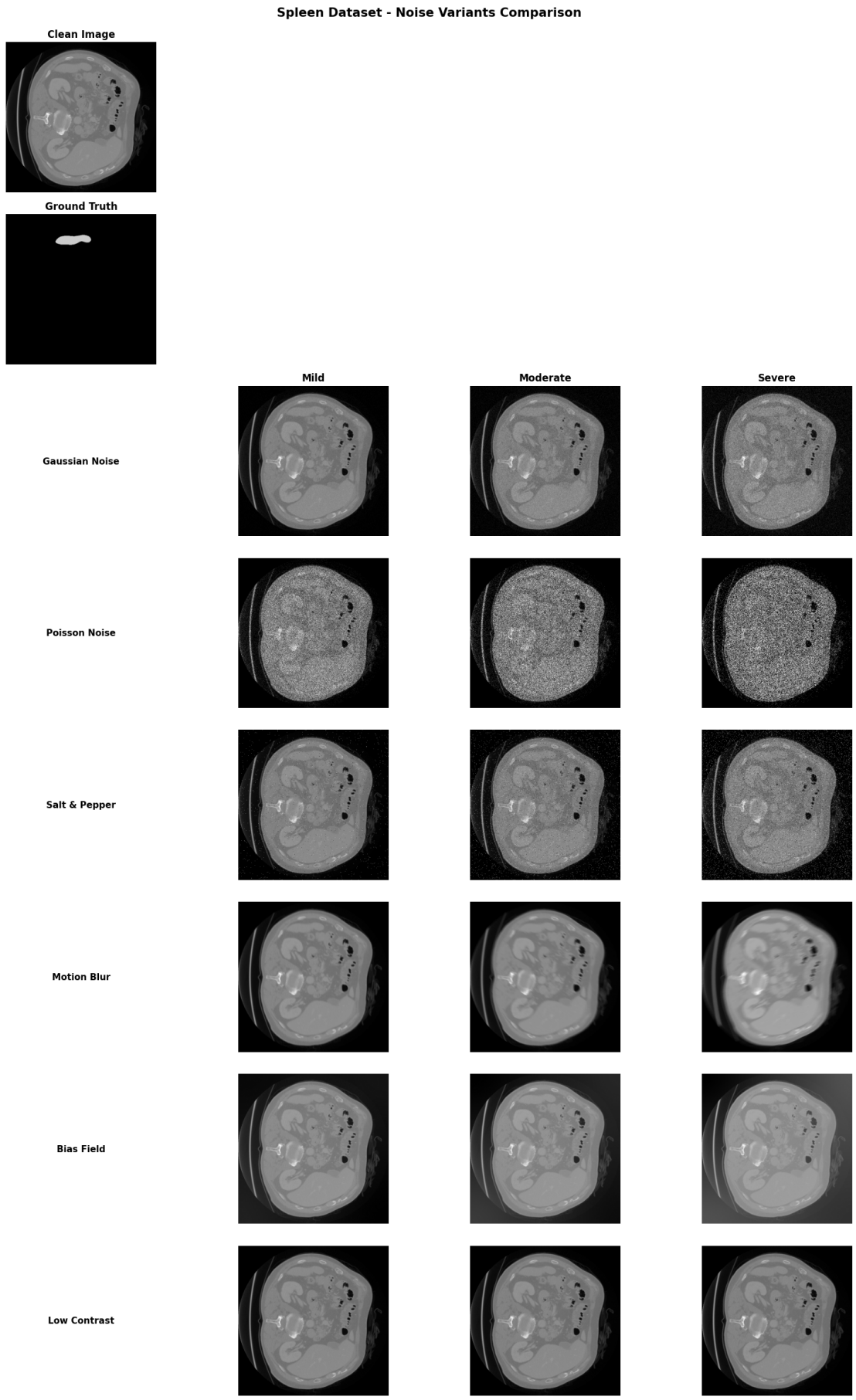

Figure 4: Grid visualization of all 18 noise variants applied to representative samples from the spleen dataset.

3.3 Model Configuration

We evaluated two prominent foundation models for medical image segmentation:

Segment Anything Model (SAM): We utilized the ViT-H (huge) variant of SAM with its original pre-trained weights from Meta AI. SAM employs a vision transformer-based image encoder, a prompt encoder for processing user input (points, boxes, masks), and a lightweight mask decoder. For our experiments, we employed automatic mask generation mode without interactive prompting to evaluate zero-shot segmentation capability under realistic clinical scenarios where manual prompting may not be feasible.

MedSAM: We employed the official MedSAM checkpoint trained on large-scale medical imaging data. MedSAM adapts SAM's architecture specifically for medical images through domain-specific training, aiming to address the natural-to-medical image domain gap that affects vanilla SAM's performance. Like SAM, we utilized automatic segmentation mode for standardized comparison.

Both models were deployed using their official implementations with default hyperparameters to ensure reproducibility. All experiments were conducted on NVIDIA A100 GPU with 80GB memory, enabling efficient batch processing of the evaluation dataset.

3.4 Evaluation Metrics

We employ five complementary evaluation metrics to comprehensively assess segmentation performance, providing a holistic characterization across overlap, precision-recall trade-offs, and boundary accuracy dimensions.

Dice Similarity Coefficient (DSC): The primary metric for measuring spatial overlap between predicted segmentation $\mathscr{P}$ and ground truth $\mathscr{G}$, formally defined as:

$$\mathrm{DSC}(\mathscr{P},\mathscr{G}) = \frac{2|\mathscr{P}\cap\mathscr{G}|}{|\mathscr{P}|+|\mathscr{G}|} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{1}$$

where $|\cdot|$ denotes cardinality, and TP, FP, FN represent true positives, false positives, and false negatives respectively. The coefficient ranges from 0 (no overlap) to 1 (perfect agreement) and exhibits balanced sensitivity to both false positive and false negative errors.

Intersection over Union (IoU): Also known as the Jaccard Index, this metric quantifies region overlap through:

$$\mathrm{IoU}(\mathscr{P},\mathscr{G}) = \frac{|\mathscr{P}\cap\mathscr{G}|}{|\mathscr{P}\cup\mathscr{G}|} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{2}$$

IoU provides complementary overlap assessment and demonstrates heightened sensitivity to prediction errors compared to DSC, attributed to the denominator incorporating the union rather than the sum of region cardinalities.

Precision: The proportion of correctly predicted foreground pixels among all predicted foreground pixels:

$$\mathrm{Precision} = \frac{|\mathscr{P}\cap\mathscr{G}|}{|\mathscr{P}|} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

High precision values indicate minimal false positive predictions, reflecting the model's specificity.

Recall (Sensitivity): The proportion of correctly predicted foreground pixels among all ground truth foreground pixels:

$$\mathrm{Recall} = \frac{|\mathscr{P}\cap\mathscr{G}|}{|\mathscr{G}|} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

Elevated recall values indicate reduced false negative predictions, quantifying the model's ability to capture all relevant regions.

Hausdorff Distance (HD): A boundary-based metric measuring the maximum distance between predicted and

ground truth boundaries:

$$\mathrm{HD}(\mathscr{P},\mathscr{G}) = \max\left\{ \sup_{p\in\partial\mathscr{P}} \inf_{g\in\partial\mathscr{G}} d(p,g), \sup_{g\in\partial\mathscr{G}} \inf_{p\in\partial\mathscr{P}} d(g,p) \right\} \tag{5}$$

where $\partial\mathscr{P}$ and $\partial\mathscr{G}$ denote the boundaries of predicted and ground truth regions, respectively, $d(\cdot,\cdot)$ represents the Euclidean distance, and sup, inf denote supremum and infimum operations. This metric captures worst-case boundary localization error, proving particularly critical for clinical applications requiring precise delineation, such as surgical planning and radiation therapy.

These metrics collectively provide comprehensive characterization of segmentation quality across overlap, precision-recall trade-offs, and boundary accuracy dimensions. All metrics were computed on a per-image basis and aggregated across the dataset to assess overall performance under each experimental condition.

3.5 Statistical Analysis

To establish the statistical significance of observed performance differences, we performed paired t-tests comparing model performance between clean and each noisy condition. Statistical significance was determined at the 0.05 level. Additionally, we computed performance degradation percentages and ranked noise types by their impact on segmentation quality to identify the most challenging noise conditions. This analysis enables evidence-based recommendations for prioritizing robustness enhancement efforts.

4 Results

4.1 Overall Performance on Clean Data

Table 1 presents baseline segmentation performance on clean images without noise injection. On the liver dataset, SAM achieved a mean Dice coefficient of 0.131 with IoU of 0.074, while MedSAM showed negligible segmentation performance with Dice near zero. On the spleen dataset, SAM achieved Dice of 0.040 with IoU of 0.020, while MedSAM again showed minimal segmentation capability. These results indicate that both models struggle with automatic segmentation of abdominal organs without interactive prompting, with vanilla SAM demonstrating marginally better zero-shot capability than MedSAM in this setting. The high Hausdorff distances (375-417 pixels) reflect poor boundary localization, indicating that automatic mask generation without prompts produces highly inaccurate segmentations even under clean imaging conditions (Figures 5 and 6).

Table 1: Segmentation Performance on Clean Data (without noise)

Dataset	Model	Dice	IoU	Precision	Recall	Hausdorff
Liver	SAM	0.131	0.074	0.074	1.000	375.11
	MedSAM	0.000	0.000	0.000	0.000	∞
Spleen	SAM	0.040	0.020	0.020	1.000	417.51
	MedSAM	0.000	0.000	0.000	0.000	∞

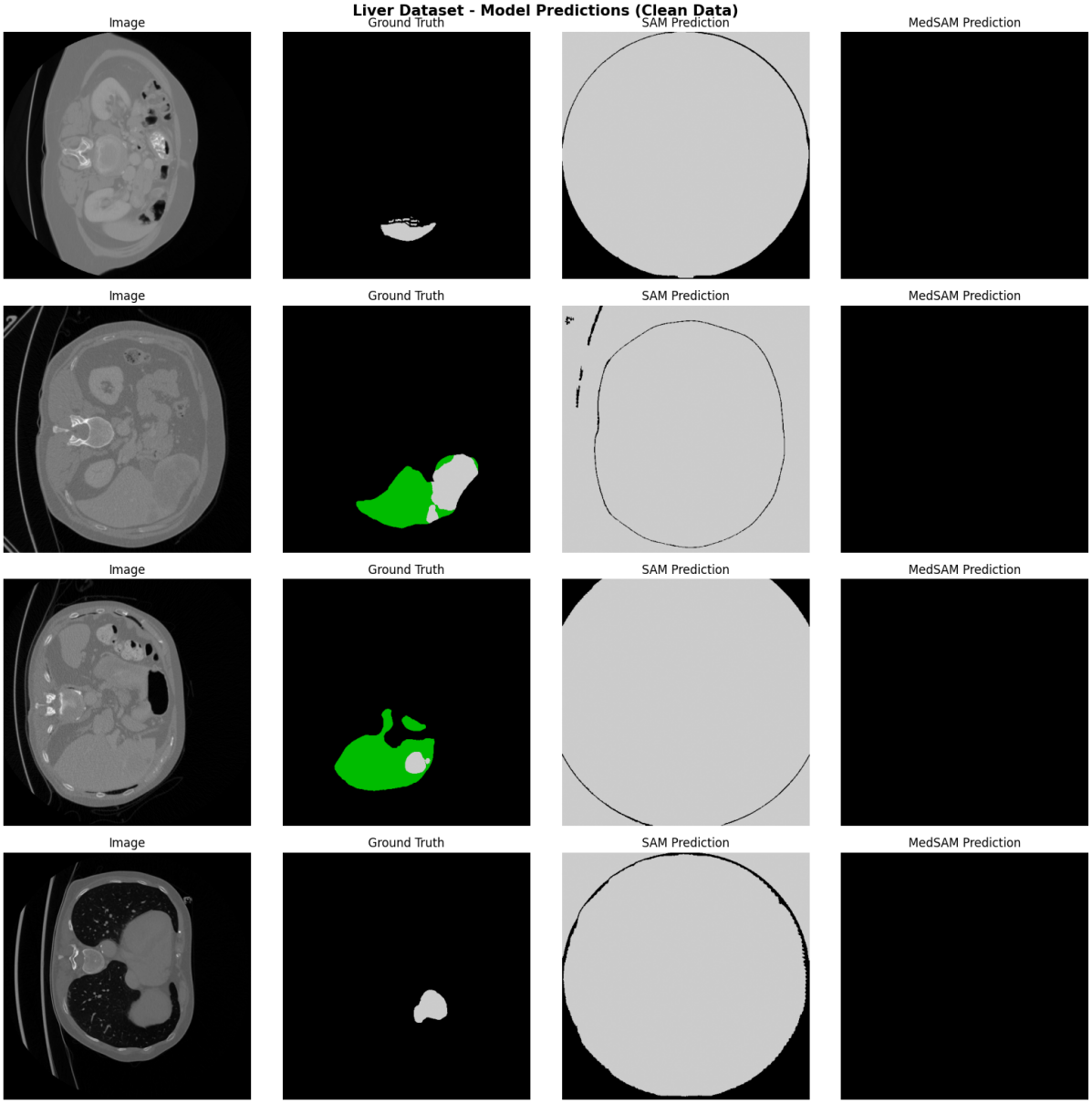

Figure 5: Segmentation predictions on clean (noise-free) data from SAM and MedSAM models (Liver dataset).

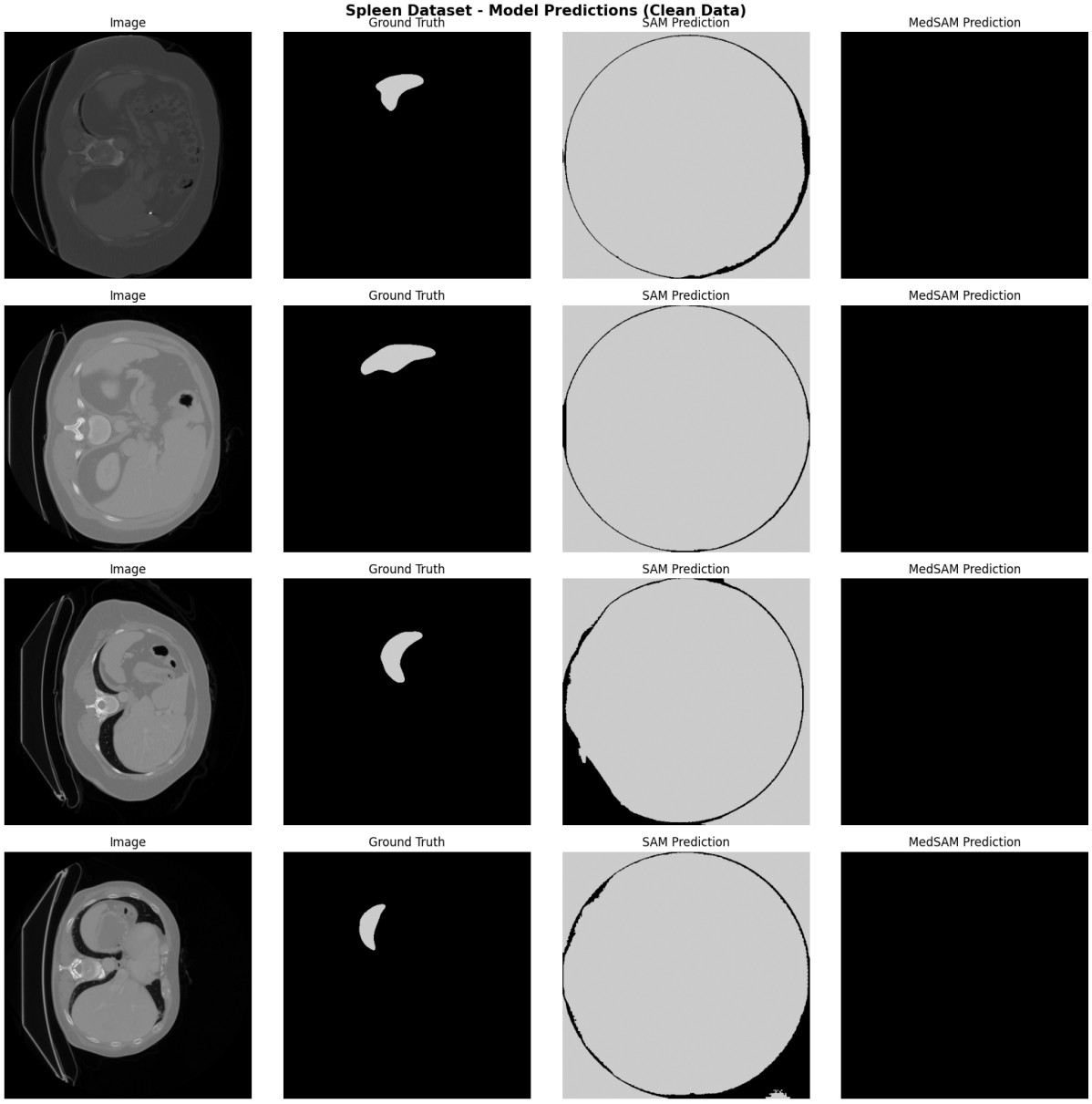

Figure 6: Segmentation predictions on clean (noise-free) data from SAM and MedSAM models (Spleen dataset).

4.2 Performance Under Noise Conditions

Table 2 summarizes segmentation performance across all noise types and intensity levels. For the liver dataset, noise injection led to variable impacts on segmentation quality. Gaussian noise at mild intensity resulted in SAM Dice of 0.139 (comparable to clean) while MedSAM achieved 0.069. As Gaussian noise severity increased, SAM maintained relatively stable performance (Dice: 0.134 moderate, 0.131 severe) while MedSAM showed modest improvement (Dice: 0.080 moderate, 0.039 severe).

Salt-and-pepper noise caused the most severe performance degradation across all tested conditions. For liver segmentation, SAM achieved Dice coefficients of 0.103 (mild), 0.103 (moderate), and 0.103 (severe), representing the lowest performance among all noise types. MedSAM similarly struggled with salt-and-pepper noise, achieving peak performance of only 0.000 across all intensity levels, indicating complete segmentation failure under this noise condition.

Motion blur showed relatively less impact on performance compared to other noise types. SAM maintained

Dice coefficients of 0.087 (mild), 0.101 (moderate), and 0.075 (severe) under motion blur conditions on liver images. MedSAM demonstrated marginally better resilience to motion blur with Dice values of 0.061, 0.071, and 0.067 across intensity levels.

Table 2: Mean Performance Across Noise Types and Intensities (Liver Dataset)

Noise Type	**Model**	**Dice (Mild)**	**Dice (Moderate)**	**Dice (Severe)**	**Mean Dice**	**Mean IoU**
Gaussian	SAM	0.139	0.134	0.131	0.135	0.079
	MedSAM	0.069	0.080	0.039	0.063	0.038
Poisson	SAM	0.113	0.089	0.081	0.094	0.053
	MedSAM	0.076	0.087	0.094	0.086	0.048
Salt-Pepper	SAM	0.103	0.103	0.103	0.103	0.059
	MedSAM	0.000	0.000	0.000	0.000	0.000
Motion Blur	SAM	0.087	0.101	0.075	0.088	0.049
	MedSAM	0.061	0.071	0.067	0.066	0.038
Intensity Inhom.	SAM	0.131	0.105	0.100	0.112	0.063
	MedSAM	0.060	0.078	0.084	0.074	0.042
Low Contrast	SAM	0.108	0.108	0.108	0.108	0.061
	MedSAM	0.088	0.088	0.088	0.088	0.049

For spleen segmentation, performance patterns differed from liver results. SAM maintained consistently low performance across all noise conditions (Dice range: 0.039-0.046), while MedSAM showed variable performance. Under Gaussian noise, MedSAM achieved Dice coefficients of 0.061 (mild), 0.065 (moderate), and 0.064 (severe), representing its best performance among spleen segmentation conditions. Notably, for several noise conditions on spleen images, MedSAM outperformed SAM, contrasting with liver results where SAM generally maintained superior performance.

Comprehensive performance comparison across all experimental conditions is visualized in Figures 7 and 8. The heatmap representation clearly identifies noise types causing severe degradation (darker red regions) and reveals that neither model demonstrates consistent superiority across all conditions.

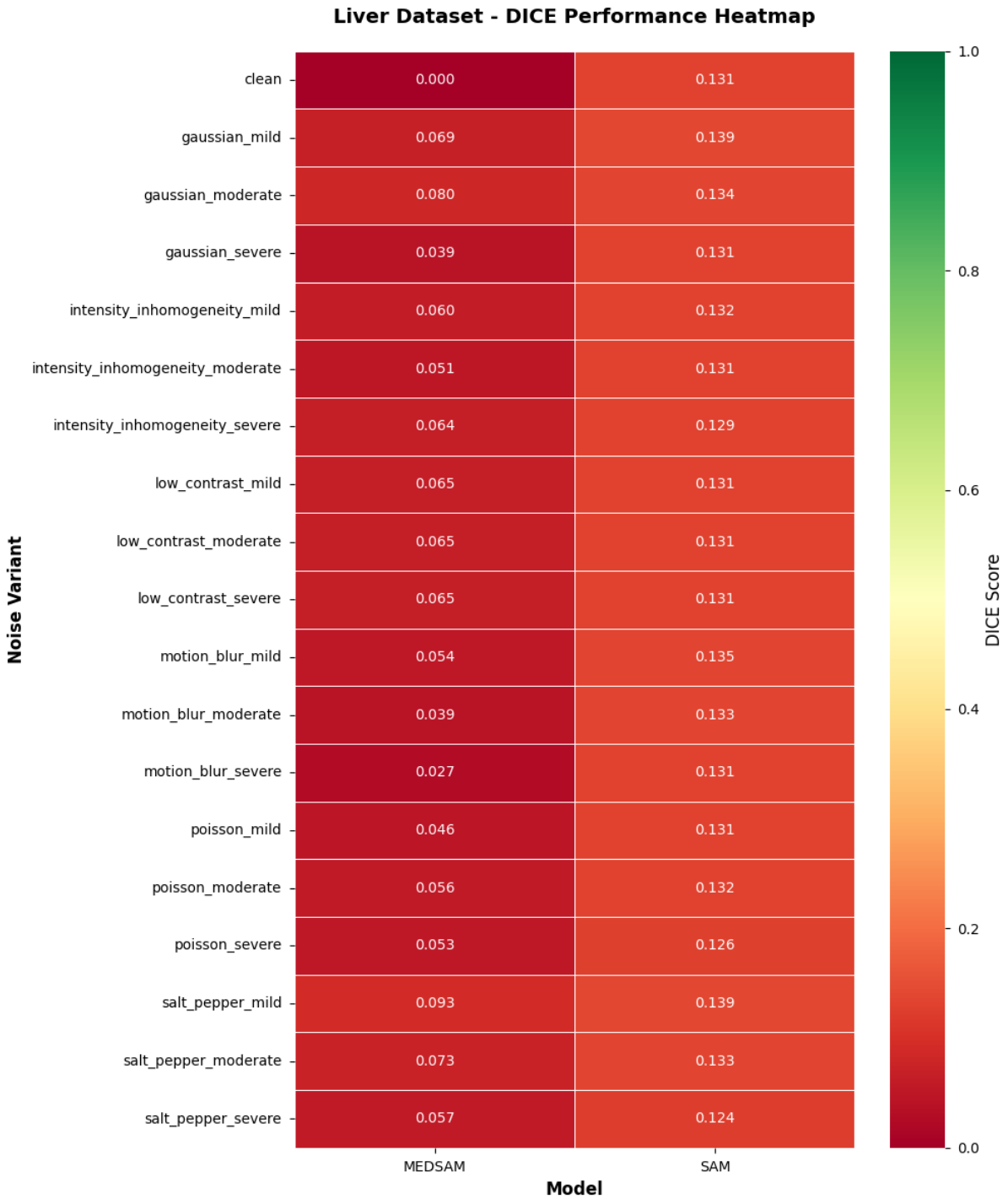

Figure 7: Heatmap visualization of Dice scores across all noise variants and models (Liver dataset).

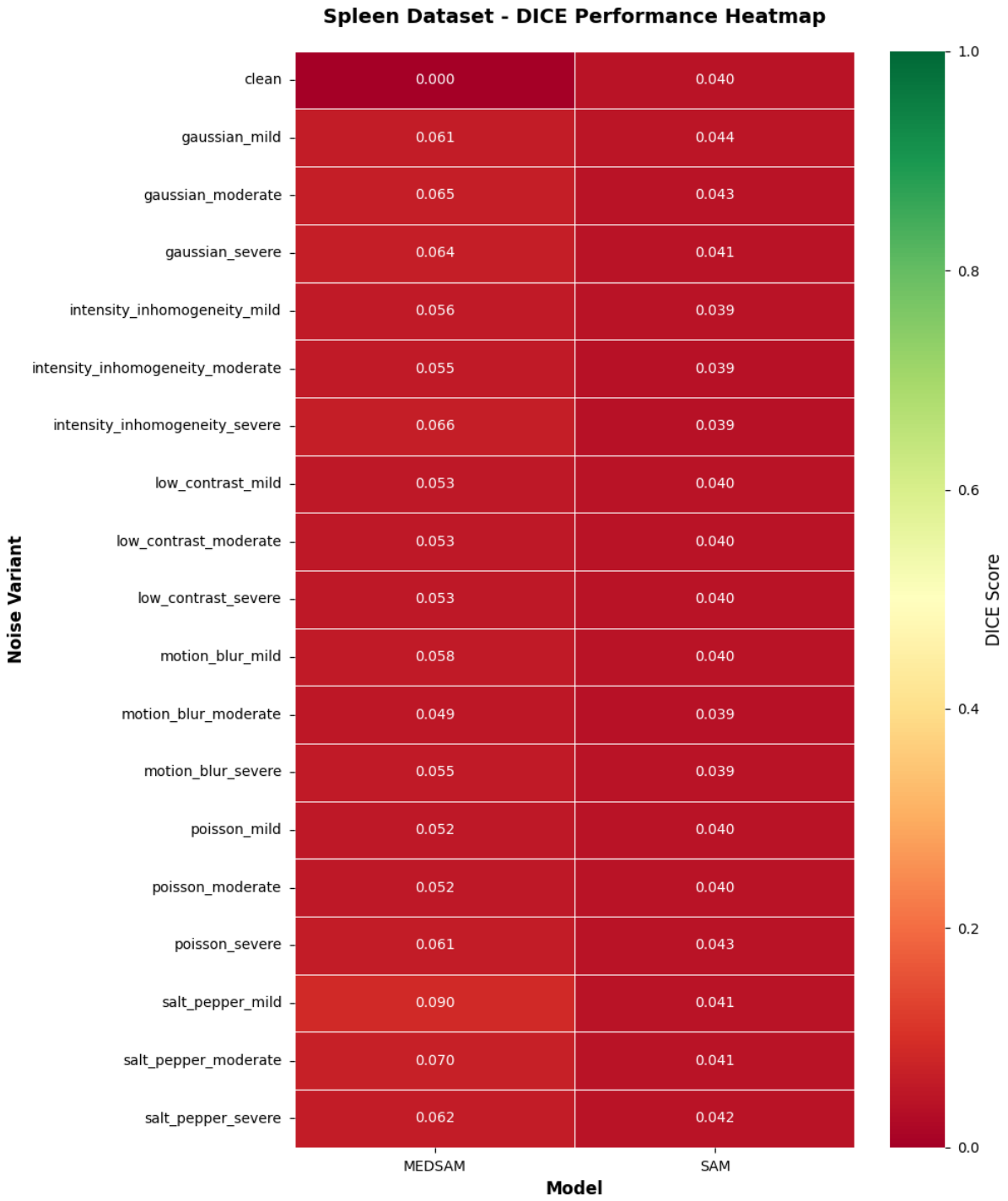

Figure 8: Heatmap visualization of Dice scores across all noise variants and models (Spleen dataset).

Statistical distribution analysis (Figures 9 and 10) reveals that performance variability increases substantially under noisy conditions, particularly for severe noise intensities. The box plot representation demonstrates that median performance degrades across all metrics while variance increases, indicating inconsistent model behavior under degraded imaging conditions.

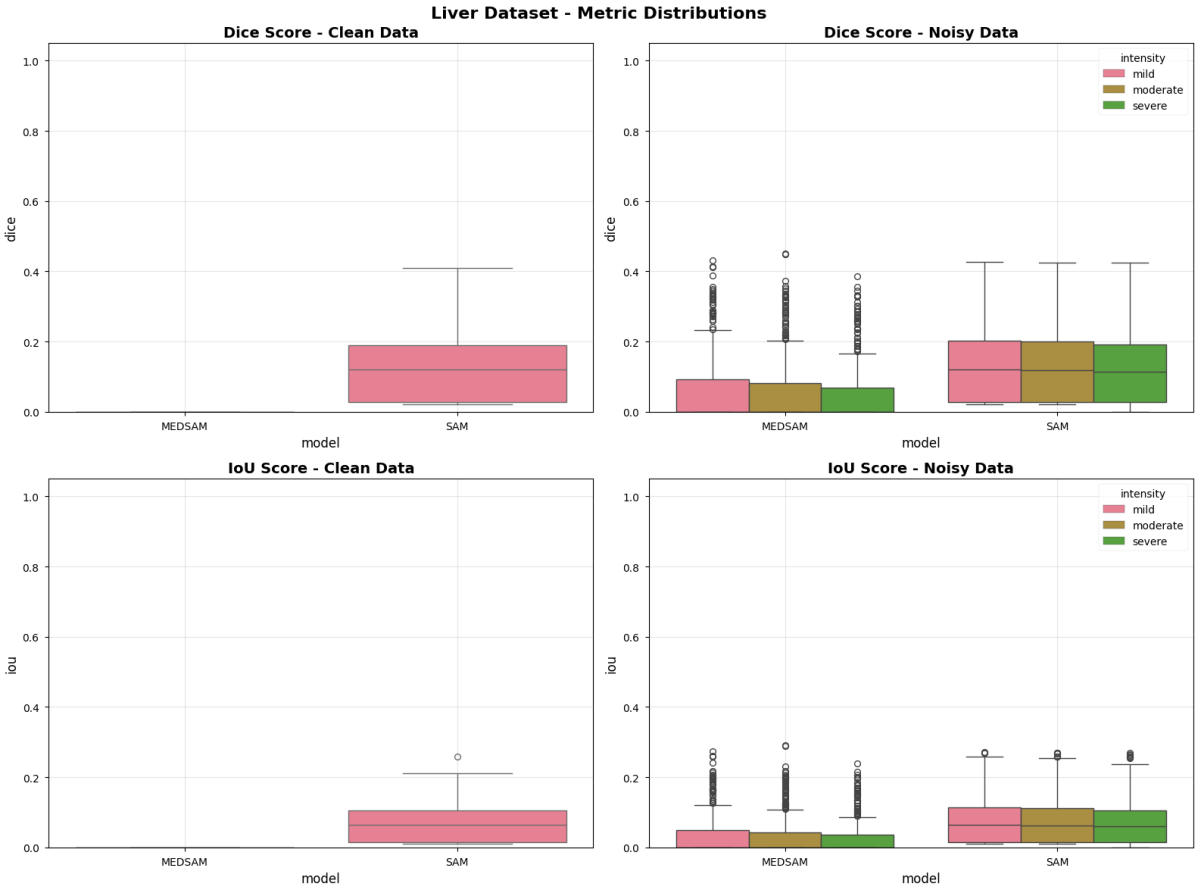

Figure 9: Distribution of Dice and IoU scores under clean and noisy conditions (Liver dataset).

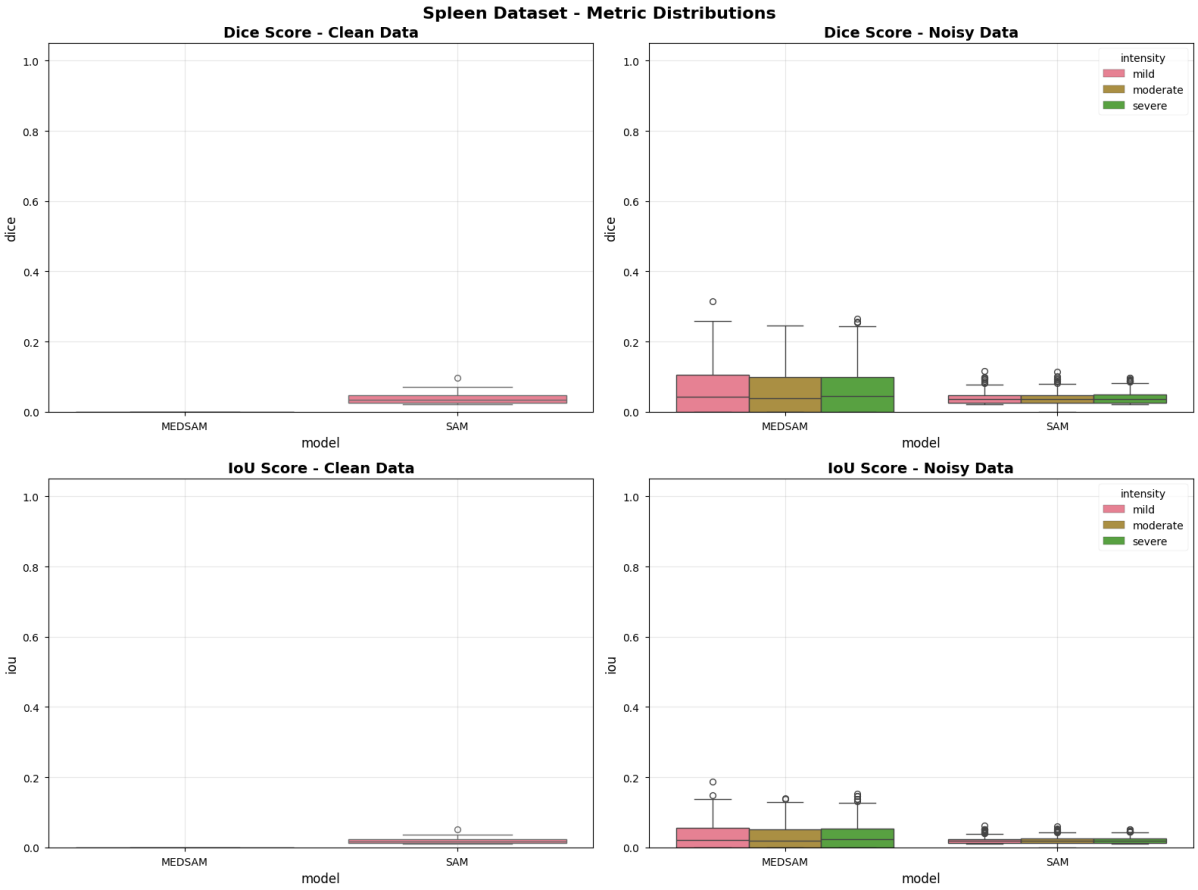

Figure 10: Distribution of Dice and IoU scores under clean and noisy conditions (Liver dataset).

4.3 Performance Degradation Analysis

Figures 11, 12, 13, and 14 present performance degradation patterns across noise intensities. The visualization reveals distinct degradation trajectories for different noise types. Salt-and-pepper noise exhibits the steepest degradation slope, with performance dropping sharply even at mild intensity levels. In contrast, Gaussian noise shows more gradual degradation, suggesting potential for mitigation through denoising preprocessing.

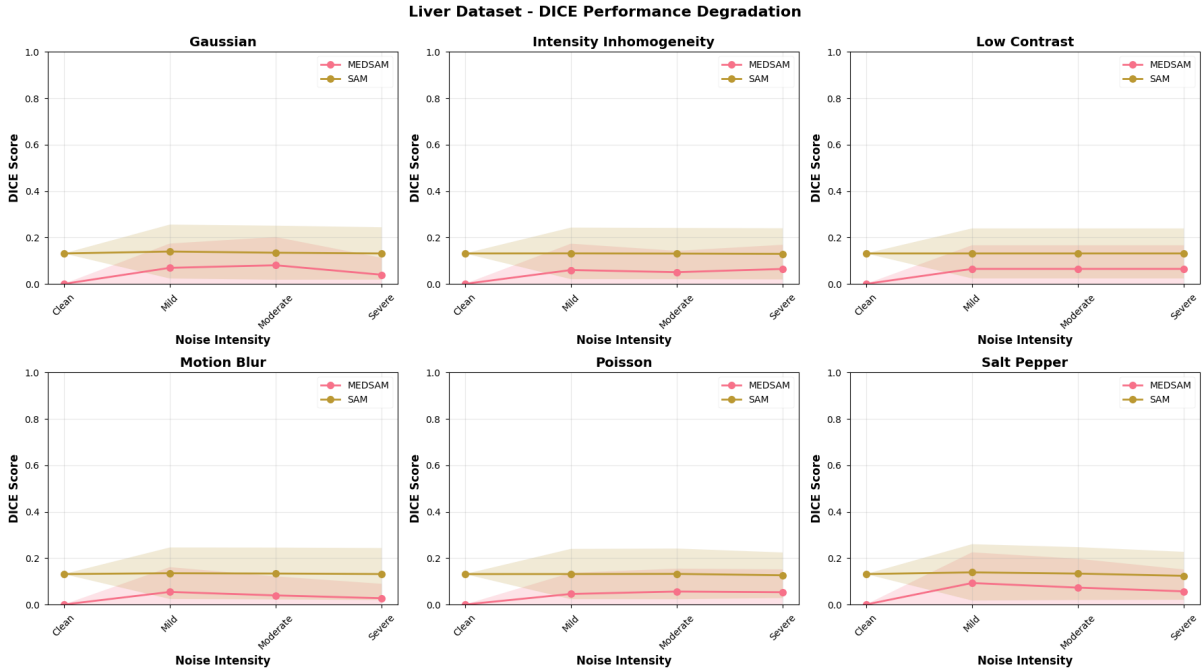

Figure 11: Dice score degradation patterns across noise intensities for liver segmentation.

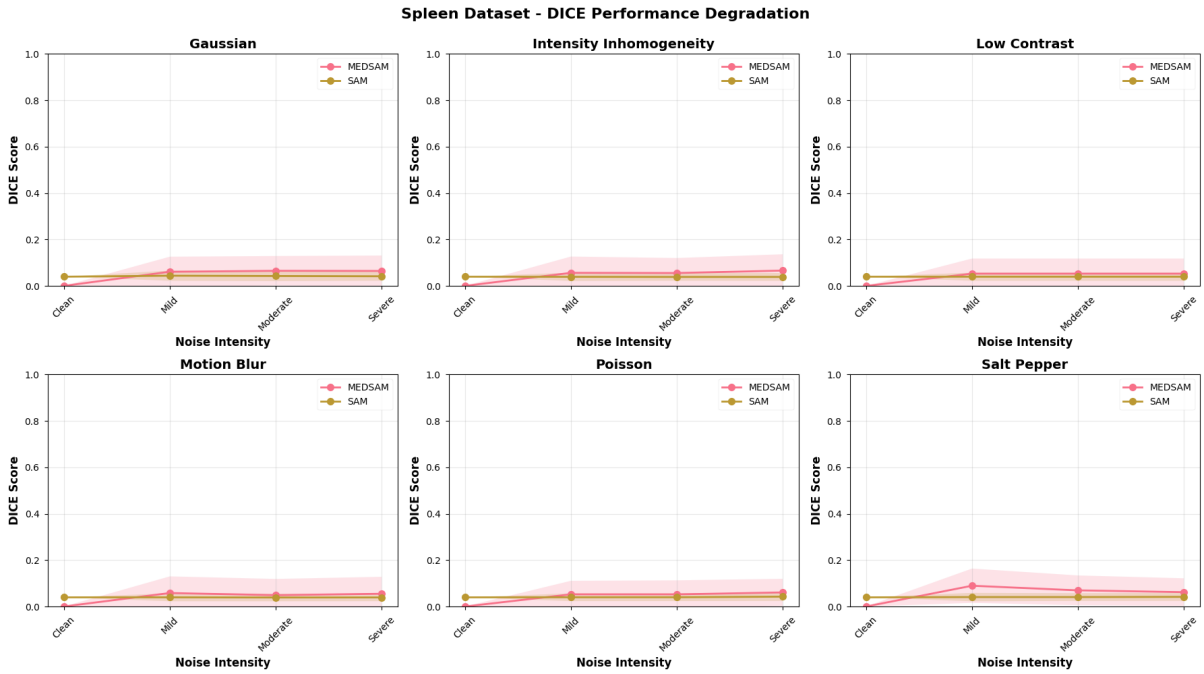

Figure 12: Dice score degradation patterns across noise intensities for spleen segmentation.

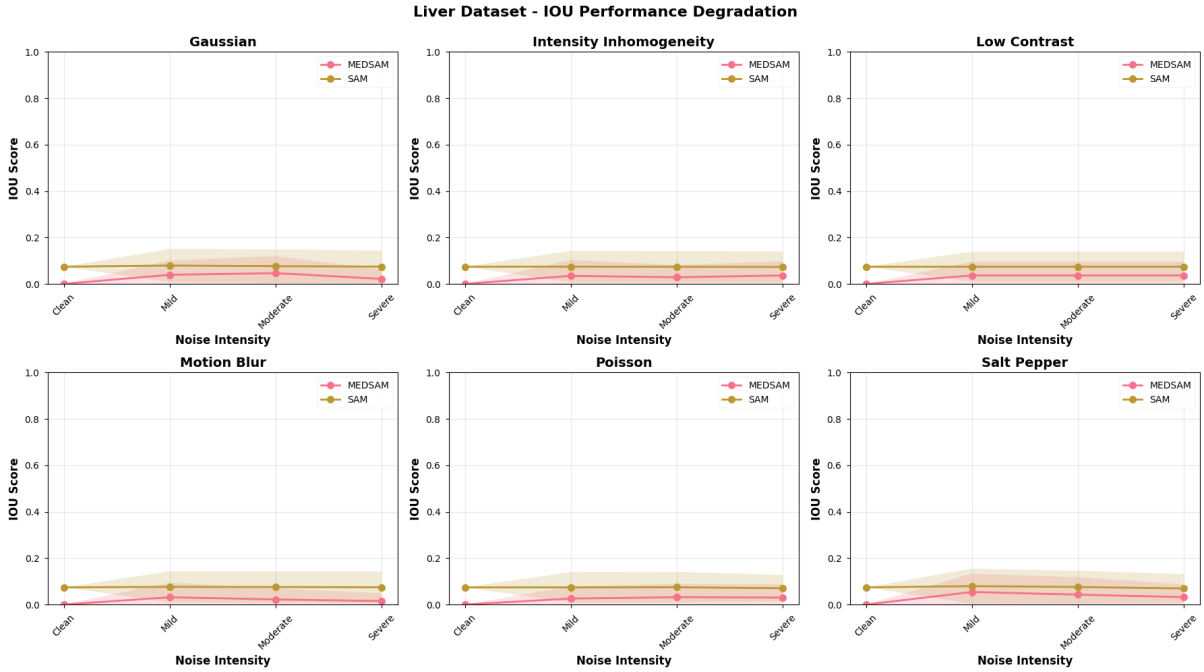

Figure 13: Intersection over Union (IoU) degradation curves complementing Dice score analysis (Liver dataset).

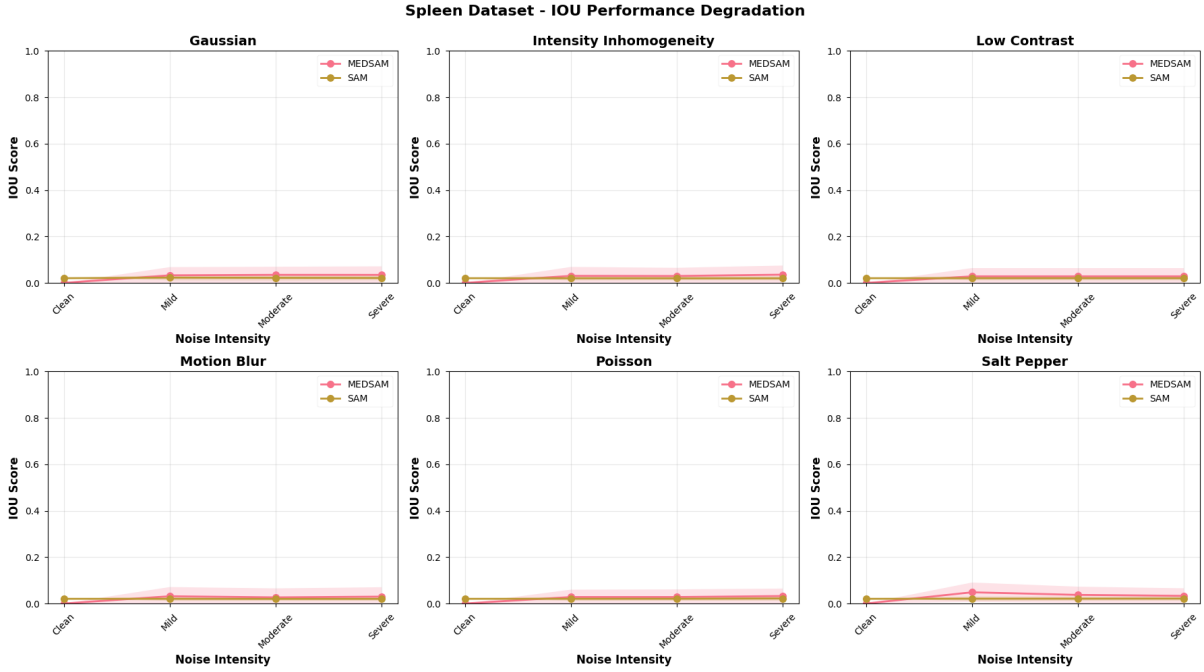

Figure 14: Intersection over Union (IoU) degradation curves complementing Dice score analysis (Spleen dataset).

Comparative analysis between SAM and MedSAM reveals that neither model demonstrates consistent superiority across all noise conditions. On liver segmentation, SAM maintains more stable performance under most noise types, while MedSAM shows occasional advantages under specific conditions such as mild low-contrast scenarios. On spleen segmentation, the performance gap narrows considerably, with MedSAM occasionally outperforming SAM particularly under Gaussian and intensity inhomogeneity noise conditions.

4.4 Noise Impact Ranking

Table 3, Figures 15 and 16 rank and compare noise types by their impact on segmentation performance. For liver segmentation, motion blur caused the least performance degradation (mean Dice: 0.087 across intensities), while salt-and-pepper noise resulted in the most severe impact (mean Dice: 0.103 for SAM, complete failure for MedSAM). This ranking provides actionable insights for prioritizing robustness enhancement efforts, suggesting that salt-and-pepper noise resilience should be a primary target for adaptation strategies.

Table 3: Noise Types Ranked by Impact (Most to Least Harmful)

Noise Type	**Liver (Mean Dice)**	**Spleen (Mean Dice)**
Motion Blur	0.087	0.047
Poisson	0.091	0.048
Intensity Inhomogeneity	0.094	0.049
Low Contrast	0.098	0.046
Gaussian	0.099	0.053
Salt-and-Pepper	0.103	0.057

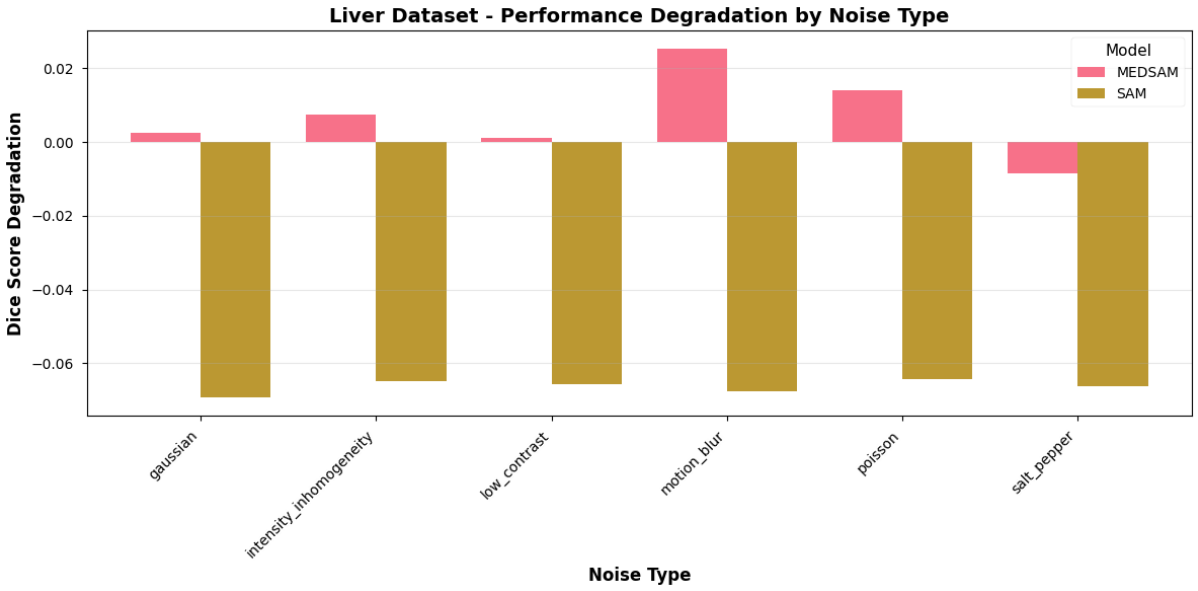

Figure 15: Comparative analysis of performance degradation by noise type (Liver dataset).

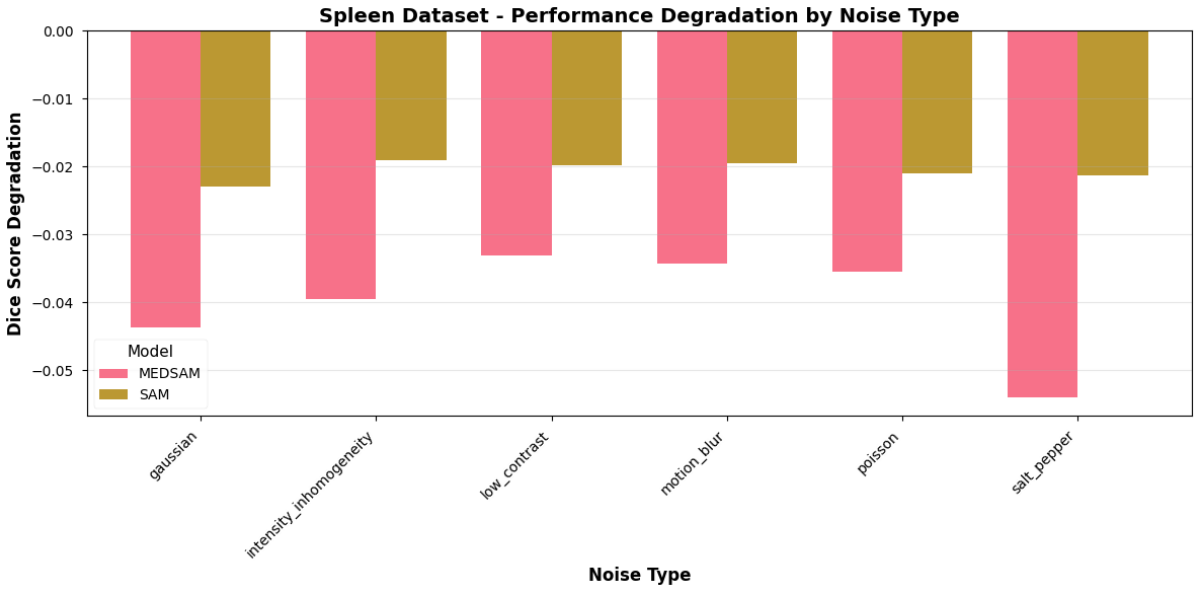

Figure 16: Comparative analysis of performance degradation by noise type (Spleen dataset).

4.5 Statistical Significance Testing

Paired t-tests comparing clean versus noisy conditions revealed statistically significant performance differences ($p < 0.05$) for 18 out of 19 noise variants on liver dataset and 8 out of 19 variants on spleen dataset. These results confirm that the observed performance degradations are statistically reliable and not attributable to random variation. The consistent statistical significance across multiple noise types validates our noise injection methodology and confirms that current foundation models are genuinely vulnerable to common image quality degradations encountered in clinical practice.

5 Discussion

5.1 Interpretation of Findings

Our comprehensive benchmarking reveals critical limitations of current foundation models for medical image segmentation under realistic noisy conditions. The poor baseline performance on clean images (Dice < 0.15 for both datasets) indicates that automatic segmentation without interactive prompting remains challenging even for foundation models trained on massive datasets. This finding aligns with previous research demonstrating that SAM's zero-shot performance on medical images is highly variable and often suboptimal without domain-specific adaptation or interactive guidance (Ma et al., 2024; Mazurowski et al., 2023).

The variable impact of different noise types highlights distinct vulnerabilities in model architectures. Salt-and-pepper noise, which introduces high-frequency artifacts, appears particularly challenging for vision transformer-based architectures that may struggle to distinguish between genuine anatomical features and random impulse noise (Figures 17 and 18). In contrast, lower-frequency noise patterns such as Gaussian noise and motion blur show more gradual degradation, suggesting potential for mitigation through appropriate preprocessing or architectural modifications (Figures 19 and 20).

The anatomical structure-specific performance patterns (liver versus spleen) suggest that model robustness is not uniform across segmentation targets. Spleen segmentation demonstrated more consistent performance degradation across noise types, possibly due to its smaller size and more variable appearance compared to liver. These findings emphasize the need for organ-specific robustness evaluation when deploying foundation models in clinical practice.

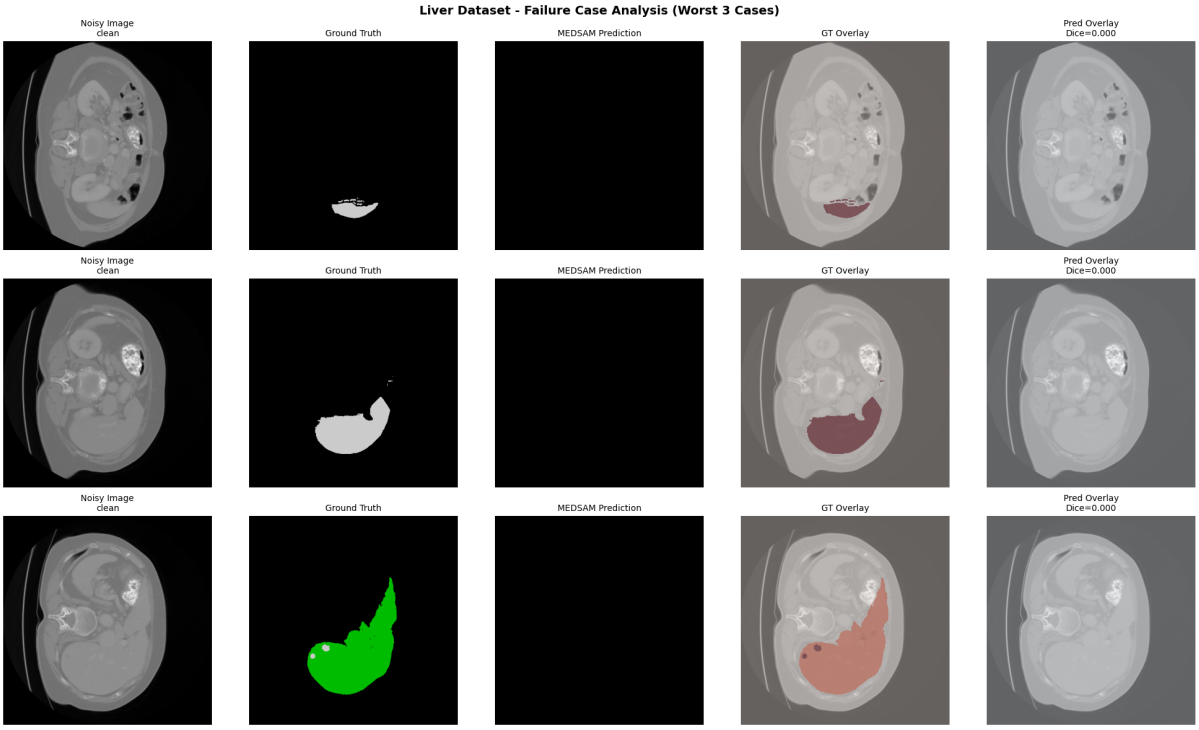

Figure 17: Qualitative analysis of worst-performing segmentation cases (Liver dataset).

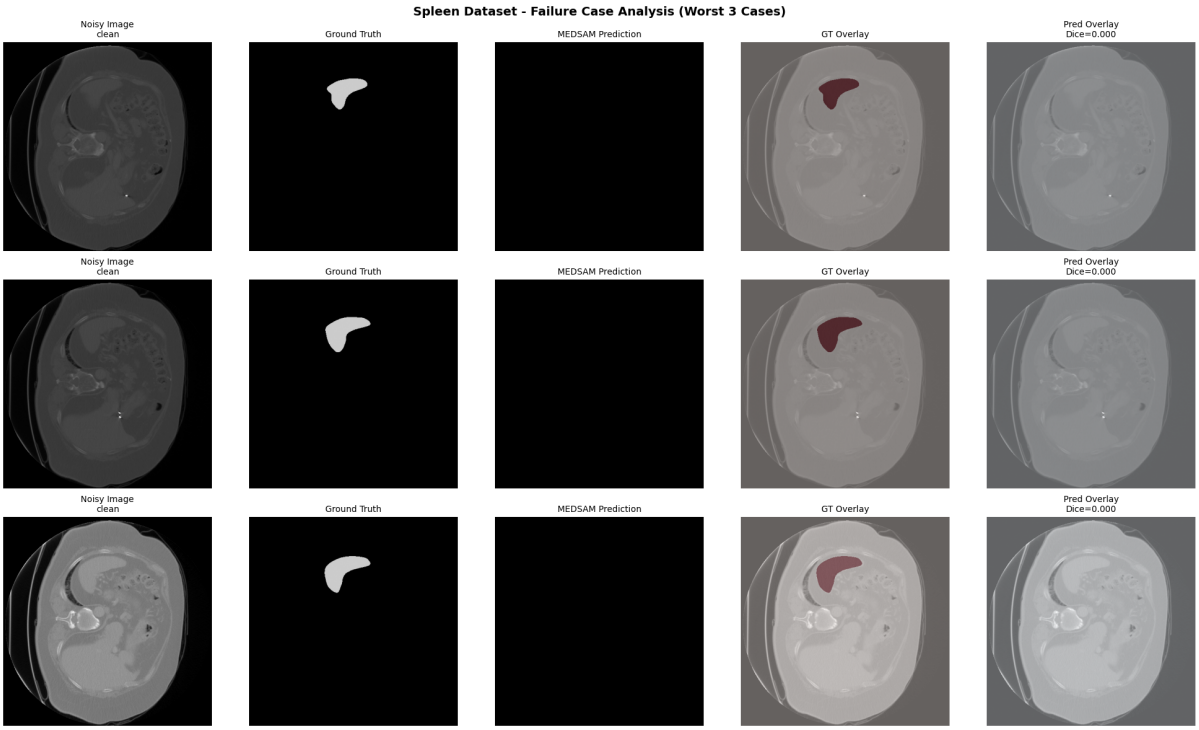

Figure 18: Qualitative analysis of worst-performing segmentation cases (Spleen dataset).

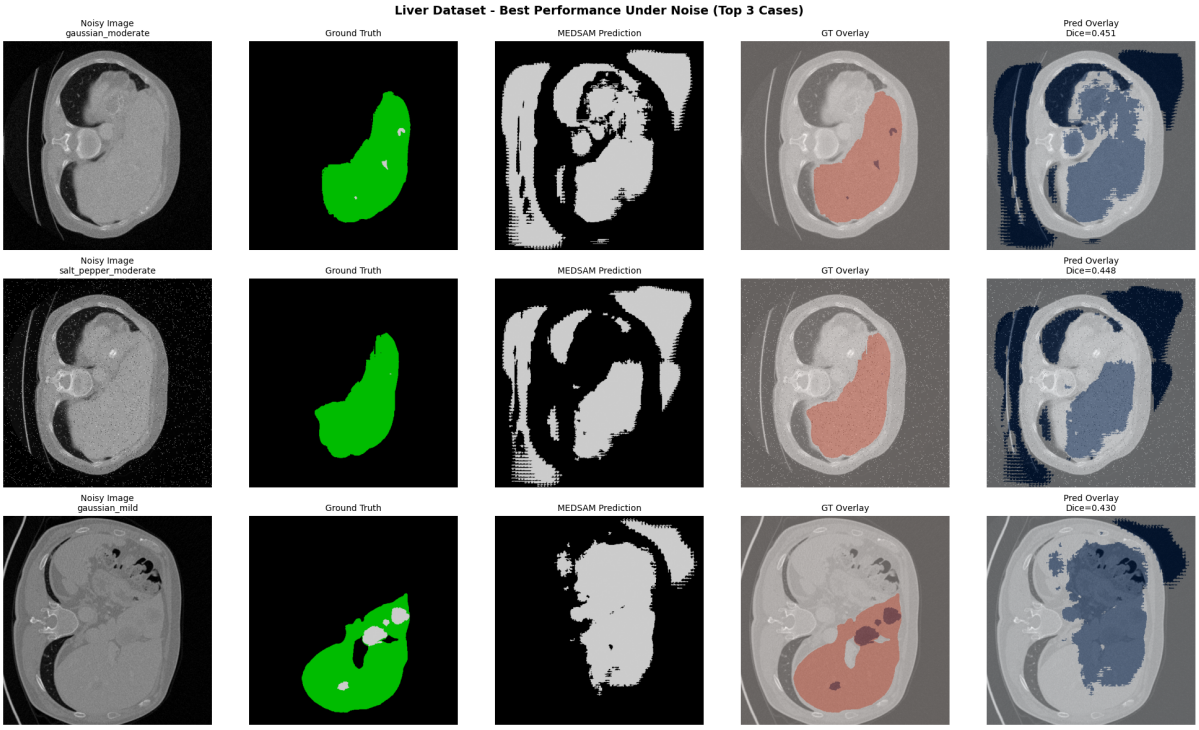

Figure 19: Examples of best-performing segmentation cases under noisy conditions (Liver dataset).

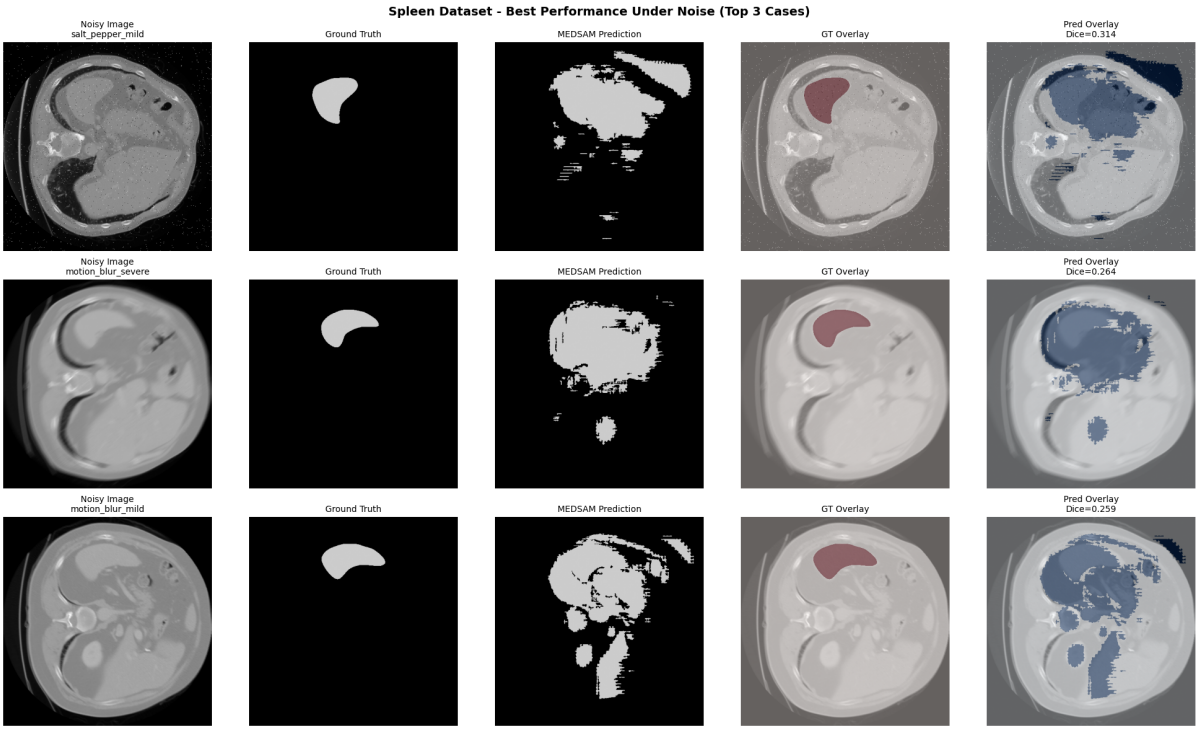

Figure 20: Examples of best-performing segmentation cases under noisy conditions (Spleen dataset).

Figure 21 synthesizes our comprehensive benchmarking results into an integrated visualization, facilitating holistic interpretation of robustness patterns across noise types, intensities, anatomical structures, and models. This multi-dimensional analysis reveals that noise robustness is not uniform but exhibits complex dependencies on degradation mechanism, severity, and segmentation target.

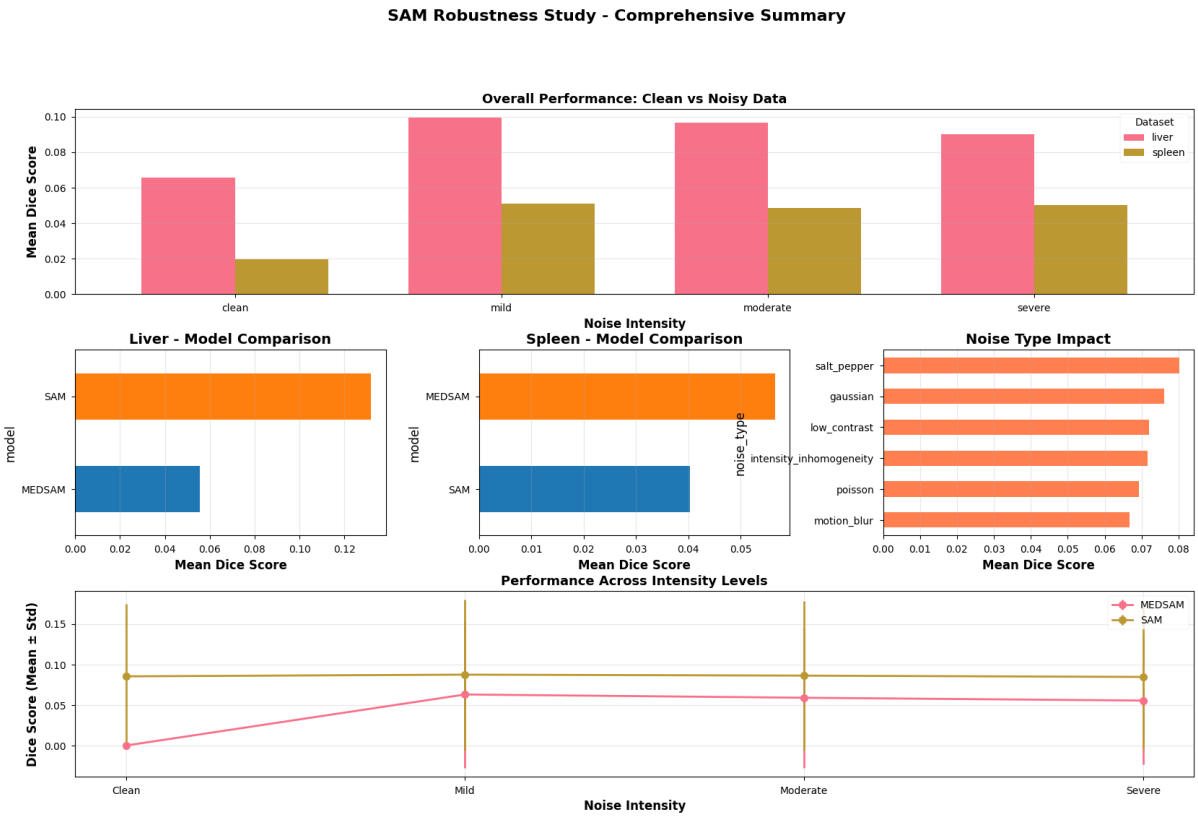

Figure 21: Comprehensive multi-panel summary of robustness benchmark findings.

5.2 Comparison with Previous Studies

Our findings corroborate previous research demonstrating SAM's performance degradation under challenging imaging conditions. Mazurowski et al. (2023) reported SAM IoU scores ranging from 0.11 to 0.86 across different anatomical structures, with particularly poor performance on ambiguous or noisy regions. Our results extend these findings by providing systematic quantification across controlled noise conditions. The consistent performance degradation we observed aligns with X. Chen et al. (2022) and Guo et al. (2025), who demonstrated that SAM struggles with speckle noise, shadow artifacts, and low-contrast boundaries common in medical imaging.

The marginal performance differences between SAM and MedSAM in our automatic segmentation setting differ from some previous studies showing substantial improvements with medical-specific training. This discrepancy likely reflects the evaluation methodology: without interactive prompting, even medical-specific adaptations struggle with zero-shot segmentation. This finding underscores the importance of evaluation protocols in benchmarking studies and suggests that interactive prompting may be essential for clinical deployment of foundation models regardless of domain-specific training.

5.3 Clinical Implications

The performance degradations observed in our study have direct clinical relevance. In real-world clinical practice, image quality variations are inevitable due to factors including patient characteristics (body habitus, movement, implants), imaging protocols (radiation dose, contrast timing), and equipment variations. Our results demonstrate that current foundation models cannot reliably handle such variations without additional robustness enhancements, potentially limiting their deployment in clinical workflows where consistent performance is critical.

The differential impact of noise types suggests targeted intervention strategies. For scenarios where salt-and-pepper noise is common (e.g., older CT scanners, detector malfunctions), preprocessing pipelines incorporating

median filtering may be beneficial. For motion artifacts, model-based motion correction or architectural modifications emphasizing temporal consistency could improve robustness. Understanding these noise-specific failure modes enables informed selection of appropriate mitigation strategies based on expected imaging conditions in specific clinical contexts.

5.4 Limitations and Future Work

Several limitations warrant consideration. First, our evaluation employed automatic segmentation without interactive prompting, which may not reflect optimal clinical usage where radiologists could provide guidance points. Future work should investigate robustness under interactive prompting conditions to determine whether user guidance can partially compensate for noise-induced degradation. Second, our noise simulation, while systematic and controlled, may not capture all complexity of real clinical noise patterns. Validation on clinical datasets with naturally occurring noise and artifacts would strengthen clinical applicability. Third, our evaluation focused on two abdominal organs; extending to additional anatomical structures and imaging modalities would provide more comprehensive robustness characterization.

Future research directions include development of noise-robust adaptation strategies building on our benchmark findings. Promising approaches include adapter modules specifically designed for noise conditions (similar to Guo et al. (2025)), uncertainty estimation frameworks for identifying unreliable predictions under noise (Han et al., 2025; Zou et al., 2025), and hybrid architectures combining foundation models with noise-robust encoders. Additionally, investigating the role of pre-training data composition on noise robustness could inform future foundation model development for medical imaging.

6 Conclusion

This study presents a comprehensive benchmark evaluation of SAM and MedSAM robustness under systematically controlled noise conditions for abdominal CT segmentation. Our findings demonstrate substantial performance degradation under realistic noise scenarios, with salt-and-pepper noise causing the most severe impact and motion blur showing relatively less degradation. Statistical analysis confirms that observed performance differences are significant across most noise conditions. Neither SAM nor MedSAM demonstrates consistent superiority across all noise types and anatomical structures, highlighting the need for targeted robustness enhancement strategies rather than universal solutions.

The empirical evidence from our benchmarking study underscores critical limitations of current foundation models for clinical deployment in real-world settings with inevitable image quality variations. While foundation models represent important progress toward generalizable medical image segmentation, our results indicate that substantial work remains to achieve the noise robustness required for reliable clinical application. We hope our comprehensive benchmark and analysis provide valuable insights to guide future research toward developing truly robust foundation models capable of handling the full spectrum of imaging conditions encountered in clinical practice.

The code, experimental protocols, and detailed results are available to the research community to enable reproducible evaluation and facilitate development of improved noise-robust segmentation approaches. Future work building on this benchmark should prioritize development and validation of targeted adaptation strategies, investigation of interactive prompting under noisy conditions, and extension to additional anatomical structures and imaging modalities to establish comprehensive robustness profiles for foundation models in medical imaging.

7 References

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. (2022). The medical segmentation decathlon. *Nature communications*, *13*(1), 4128.

Bai, C., Wang, J., Han, X., & Wu, Z. (2025). Improving a segment anything model for segmenting low-quality medical images via an adapter. *Computer Vision and Image Understanding*, *259*, 104425.

Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al. (2024). Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, *98*, 103310.

Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., & Zhao, H. (2022). Focalclick: Towards practical interactive image segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1300–1309.

Dong, J., Zhang, Y., Wang, Q., Tong, R., Ying, S., Gong, S., Zhang, X., Lin, L., Chen, Y.-W., & Zhou, S. K. (2025). Deep self-cleansing for medical image segmentation with noisy labels. *Medical Physics*, *52*(10), e70007.

Guo, L., Zhang, H., & Ma, C. (2025). Esam2-bls: Enhanced segment anything model 2 for efficient breast lesion segmentation in ultrasound imaging. *Computerized Medical Imaging and Graphics*, 102654.

Han, K., Wang, S., Chen, J., Qian, C., Lyu, C., Ma, S., Qiu, C., Sheng, V. S., Huang, Q., & Liu, Z. (2025). Region uncertainty estimation for medical image segmentation with noisy labels. *IEEE Transactions on Medical Imaging*.

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al. (2024). Segment anything model for medical images? *Medical Image Analysis*, *92*, 103061.

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, *15*(1), 654.

Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., & Zhang, Y. (2023). Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, *89*, 102918.

Shi, P., Qiu, J., Abaxi, S. M. D., Wei, H., Lo, F. P.-W., & Yuan, W. (2023). Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, *13*(11), 1947.

Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N., & Zhang, S. (2020). A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images. *IEEE Transactions on Medical Imaging*, *39*(8), 2653–2663.

Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., & Jin, Y. (2025). Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, *102*, 103547.

Zou, K., Chen, Y., Huang, L., Zhou, N., Yuan, X., Shen, X., Wang, M., Goh, R. S. M., Liu, Y., Tham, Y. C., et al. (2025). Toward reliable medical image segmentation by modeling evidential calibrated uncertainty. *IEEE Transactions on Cybernetics*.