# Enhancing Low Back Pain Assessment with Diffusion Models for Lumbar Spine MRI Segmentation

**Maria Monzon**[*1,2] MMONZON@ETHZ.CH
[1] *Biomedical Data Science Lab, ETH Zürich, Zürich, Switzerland*
[2] *Swiss Institute of Bioinformatics (SIB), Lausanne, 1015, Switzerland*

**Thomas Iff**[*1] THOMAS.IFF@INF.ETHZ.CH

**Ender Konukoglu**[3] ENDER.KONUKOGLU@VISION.EE.ETHZ.CH
[3] *Computer Vision Lab, ETH Zürich, Zürich, Switzerland*

**Catherine R. Jutzeler**[1,2] CATHERINE.JUTZELER@HEST.ETHZ.CH

**Editors:** Accepted for publication at MIDL 2025

## Abstract

This study introduces a diffusion-based framework for robust and accurate semantic segmentation of lumbar spine MRI scans from patients with low back pain (LBP), regardless of whether the scans are T1- or T2-weighted. We compared with advanced models for segmenting vertebrae, intervertebral discs (IVDs), and spinal canal using the SPIDER dataset. The results showed that SpineSegDiff achieved a segmentation performance comparable to that of the state-of-the-art non-diffusion nnUnet, particularly in improving the identification of degenerated IVDs. In addition, the uncertainty maps generated by our model provide valuable insights for clinical review, enhancing the robustness and reliability of the segmentation results. The potential of diffusion models to enhance the diagnosis and management of LBP through more precise analysis of pathological spine MRI is underscored by our findings

**Keywords:** Diffusion Models, Lumbar Spine MRI, presegmentation

## 1. Introduction and Diffusion models for Medical Image Segmentation

Low Back Pain (LBP) is a leading cause of global disability (Dionne et al., 2006), expected to affect 800 million people by 2050 (Ferreira et al., 2023), imposing a significant economic burden on individuals and society (Kent and Keating, 2005; Marto et al., 2023). Diagnosis of LBP is particularly challenging due to the various pathophysiological mechanisms involved (Fourney et al., 2011), including social, genetic, biophysical and psychological factors. The multifaceted complexity nature of LBP requires a comprehensive assessment, where lumbar spine Magnetic Resonance Imaging (MRI) is a crucial diagnostic tool. However, manual MRI interpretation is time-consuming and subject to inter-rater variability, potentially compromising diagnostic precision and consistency.

Convolutional neural networks (CNNs) have shown promise in overcoming these challenges (Maier et al., 2019) and thus enhancing the diagnostic value of lumbar spine MRI for a more quantitative interpretation (Galbusera et al., 2019). Recent advances include methods for the automatic location of intervertebral discs (IVD) or vertebrae (Lootus et al., 2014;

[*] Contributed equally

Windsor et al., 2020; He et al., 2021; Lessmann et al., 2019) to detect vertebral fractures (Yeh et al., 2022), to create synthetic lumbar MRI data (Han et al., 2018), and segment MRI of the lumbar spine in different anatomical structures (Zhou et al., 2022; Lu et al., 2018; Li et al., 2021; Mushtaq et al., 2022; Zheng et al., 2022; van der Graaf et al., 2024).

However, automatic spine segmentation is challenging due to the high intraclass similarity between vertebrae (Wang et al., 2022; Sekuboyina et al., 2021) and the large variability in the morphology of the intervertebral disc at all levels. Additionally, degenerative pathologies such as disc herniation, spinal stenosis, and vertebral fractures can significantly distort the normal anatomical structure (Pang et al., 2021; van der Graaf et al., 2024).

Such anatomical distortions present significant challenges to conventional segmentation methodologies and highlight the need for new techniques to effectively handle this variability. While medical image segmentation is traditionally a pixel-wise classification problem  (Yao et al., 2023), it can be conceptualized as an image generation task, with a generative model learning the conditional distribution to output the segmentation map. Denoising diffusion probabilistic models (DDPM) (Ho et al., 2020), traditionally used for image generation, can be adapted for image segmentation (Wolleb et al., 2021) by a conditional problem $p(\mathbf{x}|\mathbf{y})$, with the mask as a generated sample $\mathbf{x}$ conditioned on the input image $\mathbf{y}$. Recently, diffusion models (Ho et al.) showed promising results in medical image analysis Kazerouni et al. (2023); Chung et al. (2022) and also in medical image segmentation Liu et al. (2024); Xing et al. (2023); Wolleb et al. (2021); Kim et al. (2023); Wu et al. (2022) due to their ability to effectively capture the underlying data distributions (Dhariwal and Nichol) and handle noise and variability in medical images  (Li et al., 2024). The inherent ability of diffusion models to model complex and noisy data distributions  (Li et al., 2024) may prove advantageous in capturing the variability in signal intensity, anatomy, and pathological features present in MRI scans of LBP patients.

Motivated by the potential of diffusion models to handle variability in LBP MRI scans, this study presents the following contributions: (i) explore diffusion models for unified semantic segmentation of lumbar spine MRI, focusing on their effectiveness with T1 and T2-weighted scans; (ii) develop a 2D diffusion-based segmentation model for lumbar spine segmentation to handle of pathological cases; and (iii) the adaptation of presegmentation strategy that combines initial segmentation and diffusion models for efficient segmentation model training.

## 2. Methods: Diffusion models for Medical Image Segmentation

This study presents a 2D diffusion-based framework to segment the central slice of lumbar spine MRI scans, aligned with the clinical evaluation of LBP. It leverages DDPMs, generative models that reconstruct data by reversing gradual noise addition. The forward process iteratively, over $T$ timesteps, adds Gaussian noise to mask sample $\mathbf{x}_0$, $\mathbf{x}_1, ..., \mathbf{x}_T$:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{1}$$

where $\bar{\alpha}_t$ is an increasing variance scheduler and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma)$ identically distributed Gaussian noise with standard deviation $\sigma$. As time step $t$ increases ($T \to \infty$), the mask loses its distinctive features, approaching an isotropic Gaussian distribution $\mathbf{x}_T$.

The reverse diffusion process aims to progressively denoise Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to recover the segmentation mask $\mathbf{x}_0$, conditioned on the MRI scan $\mathbf{y}$. By parameterizing the transition probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ as a Gaussian distribution (Sohl-Dickstein et al.), we can train a diffusion model by minimizing a loss function that compares the estimated noise $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{y})$ and actual noise $\boldsymbol{\epsilon}$ at each timestep $t$ (Öttl et al., 2024):

$$L_t = \mathbf{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t, \mathbf{y}))\|^2\right] \tag{2}$$

There are two primary approaches to diffusion-based segmentation in medical imaging: an iterative approach that predicts and removes noise $\boldsymbol{\epsilon}_t$ sequentially (Wolleb et al., 2021), and a direct inference method that generates the final segmentation mask $\hat{\mathbf{x}}_0$ from a partially noised input $\mathbf{x}_t$ (Wu et al., 2022; Xing et al., 2023). Although the iterative denoising process is computationally intensive, sampling efficiency can be optimized using Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020). DDIM enhances sampling by enabling generation at set timesteps, substantially reducing iterations and computational resources.

## 2.1. SpineSegDiff

The SpineSegDiff model presents a novel two-dimensional dual-encoder architecture specifically designed for the semantic segmentation of lumbar spine MRI scans, functioning on 320x320 images without the need for sliding-window inference. The model architecture (Fig.1) combines a U-shaped backbone with a dedicated image encoder for multiscale feature extraction. These dual-encoder features enhance the Denoising UNet embedding, enriching the model's representation capacity during training (Xing et al., 2023).

The SpineSegDiff directly generates the segmentation mask rather than iteratively denoising patterns. To enhance segmentation accuracy, SpineSegDiff uses a composite loss, integrating MSE denoising for reconstruction, Dice Loss for boundary alignment, and Binary Cross-Entropy for calibrating probabilities between the predicted mask $\hat{\mathbf{x}}_0$ and the ground truth $\mathbf{x}$. The sampling process leverages the stochastic nature of DDIM, generating intermediate samples $S = 15$ and computing the arithmetic mean between multiple samples $\bar{\mathbf{x}}_t$ in each time step $t$. The final prediction $\hat{\mathbf{x}}_f$ is calculated as a weighted sum of these samples in the last $T_S = 10$ timesteps, with weights exponentially scaled by time:

$$\hat{\mathbf{x}}_f = \sum_{t=1}^{T_s} e^{-\alpha\left(\frac{T_s-t}{T_s}\right)} \cdot \bar{\mathbf{x}}_t \qquad \text{where} \qquad \bar{\mathbf{x}}_t = \frac{1}{S}\sum_{s=1}^{S}\mathbf{x}_s \tag{3}$$

where $\alpha = T_s/2$ sets the decay rate, assigning more weight to later timestep predictions.

### 2.1.1. Uncertainty Based Heatmaps

Diffusion models offer a key advantage through their probabilistic nature, enabling uncertainty estimation in predictions (Wolleb et al., 2021). This study introduces a novel approach for visualizing uncertainty in models that directly infer segmentation masks $\hat{\mathbf{x}}_\mathbf{0}$. These uncertainty-based heatmaps may be useful for clinical assessment of LBP, as they highlight regions where the model's predictions may be less reliable in identifying degenerated spinal structures. We generate uncertainty-based heatmaps, by computing entropy $\hat{\mathbf{h}}_t$

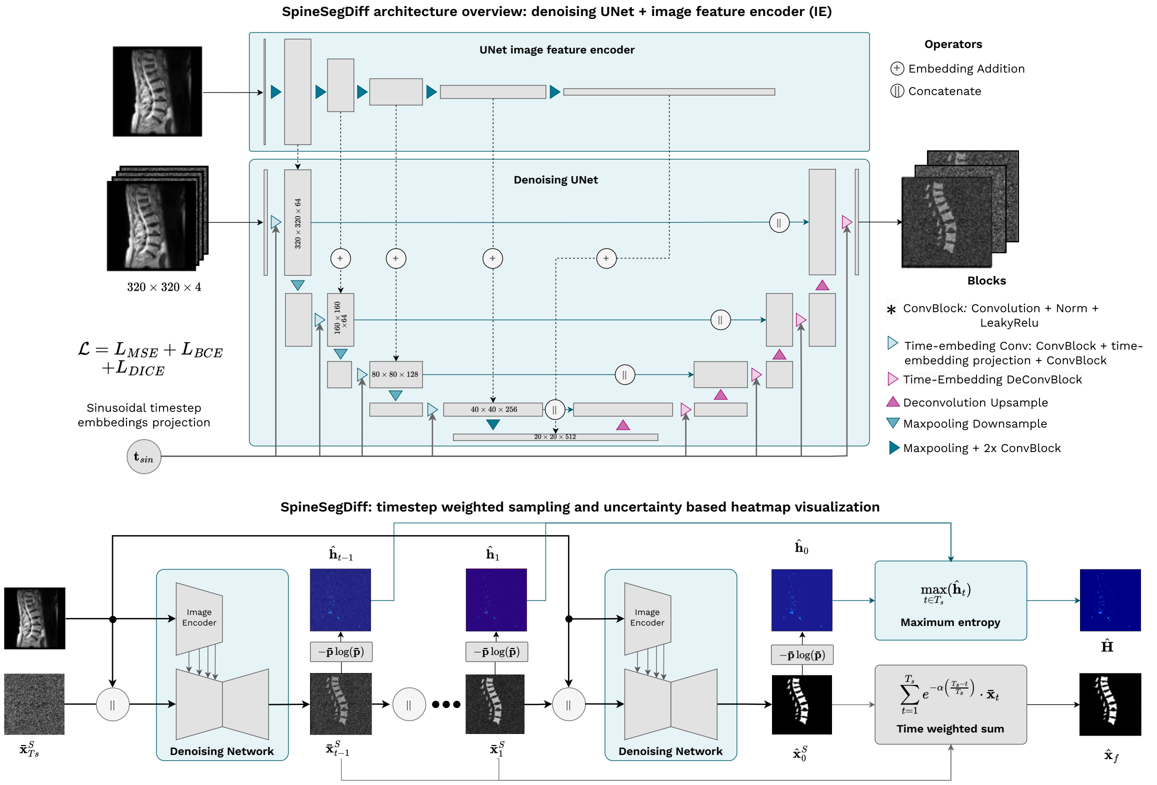
SpineSegDiff architecture overview: denoising UNet + image feature encoder (IE)
UNet image feature encoder
Denoising UNet
320 × 320 × 4
320 × 320 × 64
160 × 160 × 64
80 × 80 × 128
40 × 40 × 256
20 × 20 × 512
$\mathcal{L} = L_{MSE} + L_{BCE} + L_{DICE}$
Sinusoidal timestep embbedings projection
$\mathbf{t}_{sin}$
Operators
Embedding Addition
Concatenate
Blocks
ConvBlock: Convolution + Norm + LeakyRelu
Time-embeding Conv: ConvBlock + time-embedding projection + ConvBlock
Time-Embedding DeConvBlock
Deconvolution Upsample
Maxpooling Downsample
Maxpooling + 2x ConvBlock
SpineSegDiff: timestep weighted sampling and uncertainty based heatmap visualization
$\hat{\mathbf{h}}_{t-1}$
$\hat{\mathbf{h}}_1$
$\hat{\mathbf{h}}_0$
$-\bar{\mathbf{p}}\log(\bar{\mathbf{p}})$
Image Encoder
Denoising Network
$\bar{\mathbf{x}}^S_{T_s}$
$\bar{\mathbf{x}}^S_{t-1}$
$\bar{\mathbf{x}}^S_1$
$\hat{\mathbf{x}}^S_0$
$\max_{t \in T_s}(\hat{\mathbf{h}}_t)$
Maximum entropy
$\sum_{t=1}^{T_s} e^{-\alpha\left(\frac{T_s - t}{T_s}\right)} \cdot \bar{\mathbf{x}}_t$
Time weighted sum
$\hat{\mathbf{H}}$
$\hat{\mathbf{x}}_f$

Figure 1: SpineSegDiff architecture overview (top): the 2D MRI scan ($\mathbf{y}$) is concatenated with the partially noised mask to generate the segmentation image $\mathbf{x}_T$. The framework combines a multi-scale image encoder (IE) with a UNet-based denoising model. Segmentation inference (bottom) employs a multi-sample step-weighted sum, simultaneously generating uncertainty-based heatmaps

at the time step $t$ during the DDIM sampling:

$$\hat{\mathbf{h}}_t = -\bar{\mathbf{p}}_t \cdot \log\left(\bar{\mathbf{p}}_t\right) \qquad \text{where} \qquad \bar{\mathbf{p}}_t = \frac{e^{\bar{\mathbf{x}}_t}}{\sum_{c=1}^{K} e^{\bar{\mathbf{x}}_{t,c}}} \tag{4}$$

and $\mathbf{p}_t$ represents the softmax probability map normalized for the number of classes ($K = 4$) for each diffusion timestep $t$. The final uncertainty-based heatmap, is then computed as the maximum of each time-entropy heatmap for each spinal structure:

$$\hat{\mathbf{H}} = \max_{t \in T_s}(\hat{\mathbf{h}}_t) \tag{5}$$

### 2.1.2. Presegmentation training with nnUnet

SpineSegDiff training is significantly accelerated through the implementation of a presegmentation strategy (Guo et al., 2022). Unlike traditional original presegmentation approach where the diffusion models learn to denoise patterns, our SpineSegDiff with presegmentation directly estimates the final segmentation mask $\mathbf{x}_0$. The complete system, SpineSegDiff with presegmentation, is composed of nnU-Net followed by a SpineSegDiff architecture (see Appendix Figure 4). The workflow consists of two main stages: The initial segmentation $\hat{\mathbf{x}}_{\text{pre}}$ is predicted with the pre-trained baseline nnU-net model (Isensee et al., 2020) and SpineSegDiff takes this partially noised presegmentation as input and learns to recover the original segmentation mask ($x_0$) through a shortened diffusion process. This presegmentation strategy significantly reduces the number of diffusion steps needed, as SpineSegDiff only needs to refine an already reasonable segmentation rather than starting from random noise.

## 3. Experimental Results

### 3.1. Dataset and Implementation Details

The analysis used sagittal MRI of the lumbar spine from a multicenter cohort of 218 patients (63% female) from SPIDER (van der Graaf et al., 2024) dataset (Appendix A.1). Scans were then realigned to the RAS+ coordinate system for consistent orientation. MRI volumes were normalized to intensity (98th percentile, scaled to 255), followed by resampling at 1 mm resolution and resizing to 320×320 pixels. Ground truth labels for semantic segmentation were created by combining vertebrae annotations (starting from L5) and onehot encoded into three structures: spinal canal (SC), vertebral bodies (VB), and IVD.

The models were trained in a high-performance cluster using one RTX 4090 GPU for the 2D case and a single v100 GPU for 3D models. The models were implemented with Pytorch and MONAI (Jorge Cardoso et al., 2022) frameworks. The 2D models were trained and evaluated only on the central slice of the data, whereas the 3D models were trained and evaluated on the entire volume. The optimal epochs for diffusion models were determined by the segmentation precision (Bertels et al., 2019) in the first-fold validation set, where 2500 epochs were used for SpineSegDiff training. The diffusion models training time steps were set to $T = 1000$ with a linear variance noise schedule from $\beta_1 = 10^{-4}$ to 0.02 The rest of training hyperparameters for all the compared modes are summarized in Appendix A.1.

### 3.2. Evaluating Diffusion Models for MRI Contrast-Independent Segmentation

The performance of the model was evaluated using the Dice score with 5-fold cross-validation. The cross-validation split ensured that scans from the same patients were consistently assigned to the same split. 18 series oblique MRI scans were excluded from the evaluation set but retained for training. Diffusion models' capability to segment both T1- and T2-weighted MRI scans without contrast-specific training was evaluated. The models were trained on individual T1w and T2w contrasts, as well as a combined dataset (T1w + T2w).

For baseline comparison, we trained nnU-Net (Isensee et al., 2020), which also served as our presegmentation model, to assess its performance on multi-contrast segmentation without specific optimization. We also compared our approach (”SpineSegDiff”) with several diffusion models: a 2D adaptation of the diffusion U-Net (Xing et al., 2023) architecture (”Diff-UNet 2D”), and the Implicit Image Segmentation Diffusion Model (”IISDM”) (Wolleb et al., 2021) and SpineSegDiff Architecture model without the additional image

Table 1: Quantitative comparison of segmentation performance using mean DICE score for spinal structures (spinal canal, vertebrae, and IVDs) trained on distinct contrast configurations: T1-weighted only (T1w), T2-weighted only (T2w), and combined T1w+T2w dataset.

| **Model** | **Data** | **Spinal Canal** | **Vertebrae** | **IVD** | **mDICE** |
|---|---|---|---|---|---|
| SpineSegDiff | T1w + T2w | 0.92 ± 0.04 | 0.92 ± 0.02 | 0.90 ± 0.05 | 0.913 |
| SpineSegDiff w/o IE | T1w + T2w | 0.92 ± 0.04 | 0.91 ± 0.03 | 0.89 ± 0.05 | 0.909 |
| Diff-UNet 2D | T1w + T2w | 0.92 ± 0.05 | 0.91 ± 0.03 | 0.89 ± 0.05 | 0.906 |
| IISDM | T1w + T2w | 0.90 ± 0.03 | 0.92 ± 0.05 | 0.89 ± 0.04 | 0.903 |
| nnU-Net | T1w + T2w | 0.91 ± 0.03 | 0.92 ± 0.03 | 0.84 ± 0.05 | 0.890 |
| SpineSegDiff | T1w | 0.93 ± 0.04 | 0.91 ± 0.03 | 0.89 ± 0.05 | 0.908 |
| SpineSegDiff w/o IE | T1w | 0.92 ± 0.03 | 0.90 ± 0.04 | 0.88 ± 0.06 | 0.905 |
| Diff-UNet 2D | T1w | 0.9 ± 0.02 | 0.92 ± 0.02 | 0.89 ± 0.04 | 0.908 |
| IISDM | T1w | 0.87 ± 0.10 | 0.91 ± 0.04 | 0.89 ± 0.05 | 0.890 |
| nnU-Net | T1w | 0.91 ± 0.02 | 0.91 ± 0.03 | 0.84 ± 0.06 | 0.887 |
| SpineSegDiff | T2w | **0.93 ± 0.04** | 0.92 ± 0.04 | **0.90 ± 0.04** | **0.917** |
| SpineSegDiff w/o IE | T2w | 0.92 ± 0.04 | 0.92 ± 0.03 | 0.90 ± 0.05 | 0.913 |
| Diff-UNet 2D | T2w | 0.92 ± 0.02 | **0.93 ± 0.02** | 0.89 ± 0.03 | 0.917 |
| IISDM | T2w | 0.86 ± 0.12 | 0.91 ± 0.04 | 0.89 ± 0.05 | 0.887 |
| nnU-Net | T2w | 0.91 ± 0.03 | 0.92 ± 0.03 | 0.85 ± 0.06 | 0.893 |

encoder (”SpineSegDiff w/o IE”). The experiment was expanded to 3D lumbar spine segmentation to evaluate if 2D diffusion models can match nnU-Net in 3D (Appendix Tab.5).

The results are summarized in Table 1 indicated that diffusion models that directly infer segmentation masks achieve comparable or slightly better results than non-diffusion approaches across different contrast configurations. The performance improvements are most notable in the segmentation of IVDs, which may be particularly challenging due to their variable morphology in pathological conditions. Figure 3 shows qualitative comparisons between SpineSegDiff and baseline methods trained on both contrast (T1w + T2W). The uncertainty-based heatmaps (right) highlight regions where segmentation predictions exhibit higher entropy.

### 3.2.1. Statistical Evaluation of Performance on Pathologies

We analyzed how different pathologies affect the segmentation performance of the SpineSegDiff model, trained using T1w+T2w data across spine structures. Pathologies such as modic changes (bone marrow alterations), disc herniation (displacement of IVD material) and spondylolisthesis (forward displacement of a vertebra), disc narrowing, and overall disc degeneration evaluated through the Pfirrman grading, which are prevalent in lumbar spine conditions, were considered due to their potential impact on model performance. The pathology distribution of the study cohort is detailed in the appendix A.1.

Figure 2 illustrates the statistical analysis, showing Dice scores between patients with and without these conditions box plots and t-test results that highlight the relationship between these pathologies and model performance. To address the issue of multiple comparisons, we applied the Benjamini-Hochberg p-values correction to control the false discovery rate at $\alpha = 0.05$. The figure indicates that pathologies like spondylolisthesis and disc narrowing significantly impact segmentation. Upper endplate changes affected IVD segmentation ($p = 0.0310$), while lower endplate changes impacted both IVD ($p = 0.0120$) and SC ($p = 0.0337$). Spondylolisthesis had widespread effects on SC ($p = 0.0048$), VB ($p = 0.0039$), and IVD ($p < 0.0001$) segmentation scores. Disc herniation only significantly affected SC segmentation ($p = 0.0263$), and disc degeneration significantly affected IVD segmentation ($p = 0.0003$).

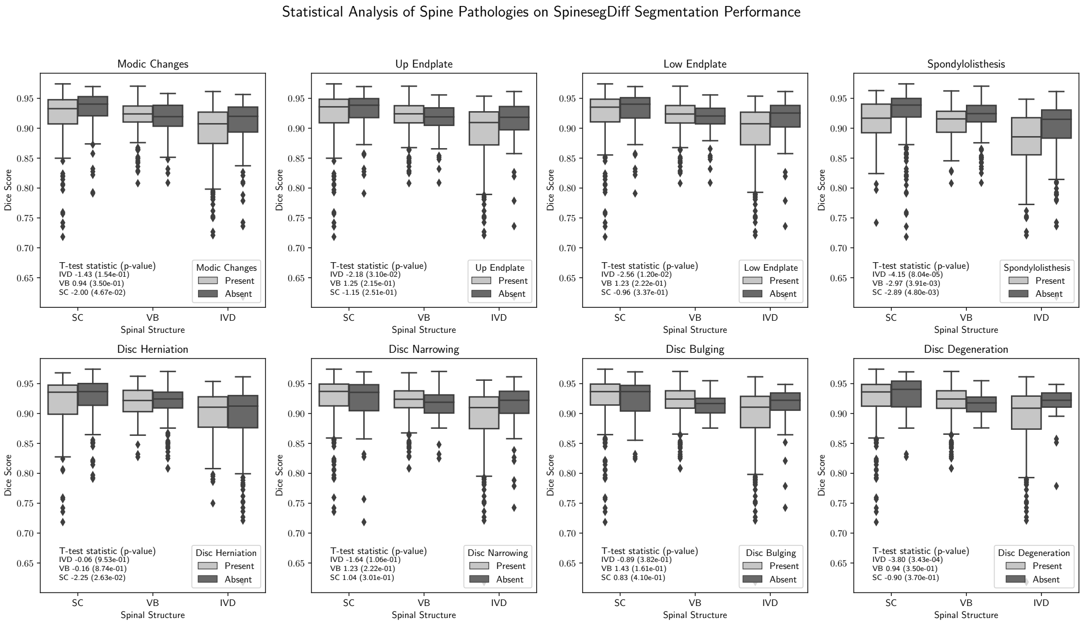

Figure 2: Statistical analysis of segmentation performance in the presence of specific spinal pathologies in each subplot, including modic changes, spondylolisthesis, disc herniation, disc narrowing, disc bulging, and disc degeneration. Significant differences ($p < 0.005$) identified via T-tests with Benjamini-Hochberg correction.

### 3.3. presegmentation Time Diffusion Steps

To evaluate the effectiveness of the presegmentation strategy, we conducted an ablation study to determine the optimal number of timesteps $t$ that balance computational efficiency and segmentation accuracy. Various time-step configurations were tested, and the results were compared to a baseline model using 1000 steps starting from the noised presegmentation, summarized in Table 2. The ablation study revealed that the preconditioning strategy significantly reduced the number of time steps needed while maintaining the 2D segmentation performance.

Table 2: Evaluation of the diffusion timesteps ($T$) on presegmentation, with $T = 0$ representing the baseline non-diffusion segmentation approach.

| | $T = 0$ | $T = 30$ | $T = 100$ | $T = 300$ | $T = 500$ | $T = 1000$ |
|---|---|---|---|---|---|---|
| **SC** | 0.91 ± 0.03 | **0.92 ± 0.05** | 0.92 ± 0.06 | 0.92 ± 0.06 | 0.92 ± 0.06 | 0.92 ± 0.07 |
| **VB** | 0.92 ± 0.03 | **0.92 ± 0.04** | 0.91 ± 0.04 | 0.91 ± 0.04 | 0.91 ± 0.04 | 0.91 ± 0.03 |
| **IVD** | 0.84 ± 0.05 | **0.89 ± 0.05** | 0.89 ± 0.06 | **0.89 ± 0.05** | 0.89 ± 0.06 | **0.89 ± 0.05** |

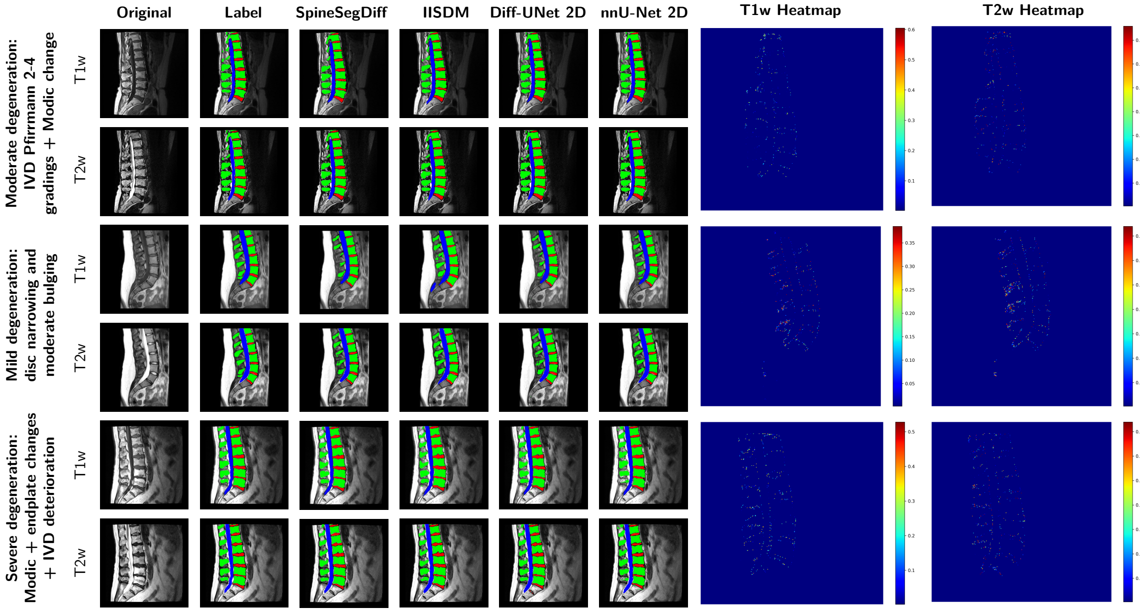
Original
Label
SpineSegDiff
IISDM
Diff-UNet 2D
nnU-Net 2D
T1w Heatmap
T2w Heatmap
Moderate degeneration: IVD Pfirrmann 2-4 gradings + Modic change
T1w
T2w
Mild degeneration: disc narrowing and moderate bulging
T1w
T2w
Severe degeneration: Modic + endplate changes + IVD deterioration
T1w
T2w

Figure 3: The visual comparisons on segmentation results on the central slice produced by the evaluated baseline and diffusion models for the three anatomical structures: spinal canal (blue), vertebrae (green), and intervertebral discs (red), along with the uncertainty maps for SpineSegDiff, where regions of higher uncertainty are denoted by darker red hues. The examples highlight challenging pathological cases with advanced disc degeneration (Pfirrmann grades 4-5), endplate irregularities, and disc narrowing. The uncertainty heatmaps effectively highlight regions of anatomical ambiguity, particularly at the boundaries of spinal structures

## 4. Discussion

Our findings demonstrate the potential of diffusion models, particularly SpineSegDiff, for accurate and efficient segmentation of the lumbar spine in MRI scans. The strong performance of these models, comparable to the state-of-the-art nnU-Net, highlights their ability to capture the complex anatomical structures and variability present in patients with low back pain. The improved segmentation of IVD is particularly noteworthy, as disc degeneration is a common cause of low back pain and accurate delineation of these structures is crucial for diagnosis and treatment planning. Despite the similar numerical performance of nnUNet 3D models, in many clinical settings, only 2D MRI scans of the lumbar spine may be available.

Furthermore, a key advantage of SpineSegDiff is its ability to generate uncertainty-based heatmaps through stochastic sampling, which may provide valuable insights for quality assurance. This approach effectively highlights anatomical regions where the model exhibits variable predictions, particularly at the boundaries of pathological structures. While visually informative, they do not currently provide calibrated statistical confidence intervals. The uncertainty is represented qualitatively rather than as precise probability distributions. Figure 3 shows SpineSegDiff segmentation errors in low-confidence areas.

These uncertainty maps may help clinicians identify regions needing closer examination, minimizing the risk of missing subtle abnormalities.

The statistical analysis reveals that certain degenerative pathologies, particularly spondylolisthesis and disc narrowing, can substantially reduce the accuracy of SpineSegDiff. Spondylolisthesis and disc narrowing exhibit the highest t-statistics and the lowest p-values, which underscores their profound impact on segmentation accuracy relative to other pathological conditions. The presence of these conditions correlates with substantially lower Dice scores.

By leveraging the initial segmentation produced by nnUNet, the study of diffusion time steps ($T$) needed (Table 2) reveals the effectiveness of the presegmentation strategy in maintaining high accuracy while significantly reducing computational requirements, making SpineSegDiff a more practical and efficient solution for lumbar spine segmentation tasks by requiring fewer diffusion steps to achieve accurate segmentation.

Nonetheless, it is important to acknowledge the limitations of our study and the challenges that remain for clinical translation. Despite the multicenter nature of the dataset, with varied sequences and acquisition parameters, further validation is necessary on larger and more diverse populations to establish the generalizability of the models. Additionally, the computational requirements of diffusion models, even with the presegmentation strategy, may still pose barriers to widespread adoption, particularly in resource-limited settings. Future work should focus on further optimizing the models for efficiency and integration into clinical workflows.

## 5. Conclusion

We present diffusion-based models for segmenting lumbar spine MRI scans from patients with LBP. Diffusion models demonstrate promising performance that approaches state-of-the-art results, particularly in the challenging task of identifying degenerated IVD. Uncertainty-based heatmaps offer valuable insights into the segmentation process, thereby improving the reliability of segmentation results. Through the implementation of a presegmentation strategy, SpineSegDiff maintains high accuracy while reducing the number of diffusion time steps, addressing computational limitations.

To fully realize the potential of SpineSegDiff, future research should focus on two key areas. First, efforts should be made to further optimize the model's computational efficiency, making it suitable for clinical implementation. Second, the model should be validated on larger and more diverse datasets to ensure its generalizability between different patient populations and imaging protocols. The present study demonstrates substantial potential;

however, it is acknowledged that the training of diffusion models requires significant computational resources. However, the superior ability to quantify uncertainties intrinsic to diffusion models offers a promising approach for the detection of degenerative changes in IVD among patients suffering from LBP related pathologies.

## Acknowledgments

This research study retrospectively analyzed open access human subject data , exempt from ethical approval according to the open access license of (van der Graaf et al., 2024). This project was supported by grant # 380 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology). The SpineSegDiff model code, along with training and evaluation scripts, and reproducibility instructions, is available at https://gitlab.ethz.ch/BMDSlab/publications/low-back/diffusion-models-for-lumbar-spine-mri-segmentation.

## References

Jeroen Bertels et al. Optimizing the dice score and jaccard index for medical image segmentation: Theory & practice. 2019. URL http://arxiv.org/abs/1911.01685.

Hyungjin Chung et al. Mr image denoising and super-resolution using regularized reverse diffusion. *IEEE Transactions on Medical Imaging*, 42(4):922–934, 2022.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. URL http://arxiv.org/abs/2105.05233v4.

C. E. Dionne, K. M. Dunn, and P. R. Croft. Does back pain prevalence really decrease with increasing age? a systematic review. *Age and Ageing*, 35(3):229–234, 2006.

Manuela L. Ferreira et al. Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *The Lancet Rheumatol.*, 5(6):e316–e329, 2023. ISSN 26659913.

Daryl R. Fourney et al. Chronic low back pain: A heterogeneous condition with challenges for an evidence-based approach. *Spine*, 36(21), 10 2011. ISSN 03622436.

F. Galbusera, G. Casaroli, and T. Bassani. Artificial intelligence and machine learning in spine research. *JOR Spine*, 2(1), 2019. doi: 10.1002/jsp2.1044.

Xutao Guo, Yanwu Yang, et al. Accelerating diffusion models via pre-segmentation diffusion sampling for medical image segmentation. 2022. doi: https://doi.org/10.48550/arXiv.2210.17408.

Z. Han, B. Wei, et al. Spine-gan: Semantic segmentation of multiple spinal structures. *Medical Image Analysis*, 50:23–35, 2018.

Jiabo He, Wei Liu, et al. Spineone: A one-stage detection framework for degenerative discs and vertebrae. 2021. doi: 10.48550/arXiv.2110.15082.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. URL https://arxiv.org/abs/2006.11239.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. 6 2020. URL http://arxiv.org/abs/2006.11239.

Fabian Isensee, Paul F. Jaeger, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 12 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z.

M Jorge Cardoso, Wenqi Li, et al. MONAI: An open-source framework for deep learning in healthcare. 2022. doi: 10.48550/arXiv.2211.02701.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, et al. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 8 2023. ISSN 1361-8423. doi: 10.1016/j.media.2023.102846. URL http://www.ncbi.nlm.nih.gov/pubmed/37295311.

Peter M Kent and Jennifer L Keating. The epidemiology of low back pain in primary care. 2005. doi: 10.1186/1746-1340-13-13.

Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. In *The Eleventh International Conference on Learning Representations*, 2023.

N. Lessmann, B. van Ginneken, et al. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Medical Image Analysis*, 53:142–155, 2019. doi: 10.1016/j.media.2019.02.005.

H. Li et al. Automatic lumbar spinal mri image segmentation with a multi-scale attention network. *Neural Computing and Applications*, 33:11589–11602, 2021.

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the Generalization Properties of Diffusion Models. 2024. URL https://arxiv.org/abs/2311.01797.

Zengxin Liu, Caiwen Ma, Wenji She, and Meilin Xie. Biomedical image segmentation using denoising diffusion probabilistic models: A comprehensive review and analysis. *Applied Sciences*, 14(2), 2024. ISSN 2076-3417. doi: 10.3390/app14020632.

M Lootus, T Kadir, and A Zisserman. Vertebrae detection and labelling in lumbar mr images. volume 17, pages 219–230. Springer, 2014.

J.-T. Lu, S. Pedemonte, et al. Deepspine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. *Proceedings of Machine Learning Research*, 85:1–16, 2018.

A. Maier, C. Syben, et al. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 29(2):86–101, 2019.

Carlos Miguel Marto et al. Global and regional estimates of clinical and economic burden of low back pain in high-income countries: a systematic review and meta-analysis. 2023. doi: 10.3389/fpubh.2023.1098100.

Malaika Mushtaq, Muhammad Usman Akram, et al. Localization and Edge-Based Segmentation of Lumbar Spine Vertebrae to Identify the Deformities Using Deep Learning Models. *Sensors 2022, Vol. 22, Page 1547*, 22(4):1547, 2 2022. ISSN 1424-8220.

S. Pang, C. Pang, et al. Spineparsenet: Spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation. *IEEE Transactions on Medical Imaging*, 40(1):262–273, 2021.

Anjany Sekuboyina et al. Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical Image Analysis*, 73:102166, 2021. ISSN 1361-8415.

Jascha Sohl-Dickstein et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. URL https://arxiv.org/abs/1503.03585.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. 10 2020. URL http://arxiv.org/abs/2010.02502.

J. W. van der Graaf, M. L. van Hooff, et al. Lumbar spine segmentation in mr images: a dataset and a public benchmark, 2024.

Dexu Wang et al. Spine segmentation with multi-view gcn and boundary constraint. In *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2136–2139, 2022.

Rhydian Windsor, Amir Jamaludin, et al. A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI. *Lecture Notes in Computer Science*, 12266 LNCS: 712–722, 7 2020. ISSN 16113349. doi: 10.1007/978-3-030-59725-2_69.

Julia Wolleb, Robin Sandkühler, et al. Diffusion Models for Implicit Image Segmentation Ensembles. 12 2021. URL http://arxiv.org/abs/2112.03145.

J. Wu, R. Fu, et al. Medsegdiff: Medical image segmentation with diffusion probabilistic model. 2022. URL https://arxiv.org/abs/2211.00611.

Z. Xing, L. Wan, et al. Diff-unet: A diffusion embedded network for volumetric segmentation. 2023. URL https://arxiv.org/abs/2303.10326v1.

Wenjian Yao et al. From cnn to transformer: A review of medical image segmentation models, 2023. URL https://arxiv.org/abs/2308.05305.

Lee Ren Yeh, Yang Zhang, et al. A deep learning-based method for the diagnosis of vertebral fractures on spine MRI: retrospective training and validation of ResNet. *European Spine Journal*, 31(8):2022–2030, 8 2022. ISSN 14320932.

Hua-Dong Zheng, Yue-Li Sun, et al. Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from mri. *Nat. Comm.*, 13:841, 2022.

N. Zhou, H. Wen, et al. Review of deep learning models for spine segmentation. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 498–507, 2022. doi: 10.1145/3512527.3531356.

Mathias Öttl et al. Analysing diffusion segmentation for medical images, 2024. URL https://arxiv.org/abs/2403.14440.

# Appendix A. Dataset and Implementation Details

## A.1. Degenerative Pathologies

This work uses publicly available SPIDER dataset (van der Graaf et al., 2024) for training and evaluation which includes MRI scans of the lumbar spine from 218 subjects with low back pain. The data includes T1- and T2-weighted images with spatial resolutions from 3.3 x 0.33 x 0.33 mm to 4.8 x 0.90 x 0.90 mm. For the 3D analysis, scans were resampled to a uniform spatial resolution of 1 mm and resized to 64x320x320 voxels. The dataset comprises a multicenter collection of sagittal lumbar MRI obtained from four different hospitals in the Netherlands, with pathological conditions such as spondylolisthesis, disc herniation, and modic changes. In our study, the incidence of present spinal degenerative pathologies was determined if they manifested at any vertebral level and is summarized in the following table.

Table 3: Overview of degenerative pathology's presence in the SPIDER dataset

| **Pathology** | **Patients (%)** |
|---|---|
| Spondylolisthesis | 42 (19.27%) |
| Disc Herniation | 72 (33.03%) |
| Modic Changes | 149 (68.34%) |
| Endplate Changes | 177 (81.19%) |
| Disc Narrowing | 193 (88.53%) |
| Disc Bulging | 200 (91.74%) |

## A.2. SpineSegDiff Training

The SpineSegDiff model is trained using a composite loss function that combines Mean Squared Error (MSE), Dice Loss, and Binary Cross-Entropy (BCE) Loss. The total loss is formulated as: $L_{total} = L_{MSE} + L_{Dice} + L_{BCE}$ where each terms are can be decomposed as $L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2$, $L_{Dice} = 1 - \frac{2|\hat{X} \cap X|}{|\hat{X}| + |X|}$, $L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[x_i \log(\hat{x}_i) + (1 - x_i)\log(1 - \hat{x}_i)]$. This loss optimizes the model for pixel accuracy (MSE), segmentation quality (Dice), and probabilistic output (BCE). The training hyperparameters are summarized in the table below:

Table 4: Training hyperparameters for SpineSegDiff

| **Parameter** | **T1w, T2w, T1w+T2w** |
|---|---|
| **Image Size** | 320x320 |
| **Epochs** | 2500 |
| **Batch** | 4 |
| **Optimizer** | AdamW |
| **Learning Rate** | 0.0001 |
| **Training Loss** | MSE + Dice + Cross Entropy |

## A.3. SpinSegDiff with Presegmentation

The presegmentation strategy (Guo et al., 2022) is adapted to augment the efficiency and precision of the diffusion model's sampling process by furnishing an initial segmentation that directs subsequent refinement stages. An initial segmentation $\hat{\mathbf{x}}_{\text{pre}}$ is produced utilizing a pre-trained baseline model. This initial segmentation acts as a prior for the diffusion model, thereby diminishing the number of diffusion steps necessary to attain accurate segmentation. The diffusion segmentation is trained using SpineSegDiff. $\hat{\mathbf{x}}_{\text{pre}}$ undergoes partial noising via a cosine noise scheduler, which introduces noise at a more gradual rate compared to a linear scheduler, thus preserving a greater extent of image features.

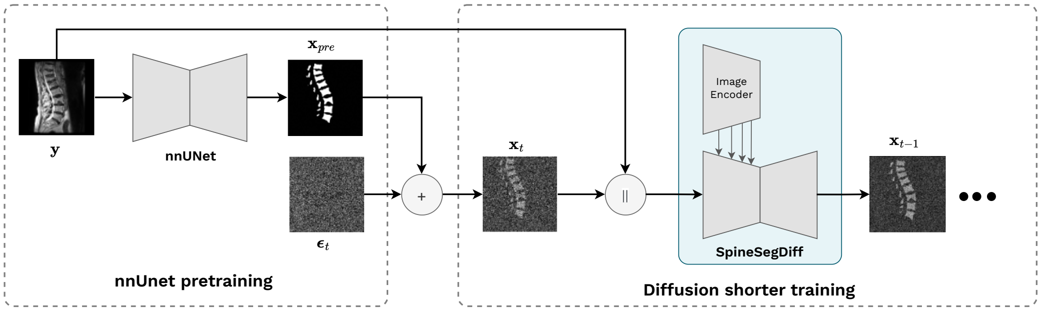

Figure 4: (a) Training pipeline with presegmentation where nnU-Net generates initial mask $\mathbf{x}$pre from MRI input $\mathbf{y}$, followed by partial noising to obtain $\mathbf{x}_T$ for diffusion training.

### A.4. Spinsegdiff Sampling and Uncertainty Maps

The computation of uncertainty maps in SpineSegDiff involves several key steps. Initially, $S$ segmentation masks are generated by repeatedly sampling the diffusion model over the latest $T_S$ timesteps. The detailed pseudo-algorithm is listed:

## Appendix B. Extended Results

### B.1. Impact of Spinal Pathologies on Segmentation Performance: Statistical Analysis

We further detail the analysis of the impact of Spinal Pathologies segmentation performance of the the baseline comparison model of the diffusion models compared to the nnU-Net baseline. The plots presented in this appendix show Dice scores for different spinal structures, such as the spinal canal (SC), vertebral bodies (VB), and intervertebral discs (IVD), in various pathological conditions. Each plot compares the segmentation performance between patients with and without specific pathologies. The t-test statistics and p-values provided in the plots indicate the statistical significance of the differences observe pathologies.

Modic changes, disc narrowing, and spondylolisthesis exhibit substantial influences on segmentation performance, particularly for intervertebral discs (IVDs) and the spinal canal, as evidenced by high t-statistics and low p-values.

```
Algorithm 1: Uncertainty-based Heatmaps
Input: MRI $\mathbf{y}$, Batch $N$, Number of Samples $S$
Output: Final prediction $\hat{\mathbf{x}}_f$
#Extract embeddings from the input MRI
$\mathbf{e_t} \leftarrow$ image_encoder($\mathbf{y}$)
#Generate $S$ number of samples using DDIM sampling
for $i \leftarrow 1$ to $S$ do
    $\mathcal{S}$.append(DDIM_sample(model, $(1, N, P_x, P_y)), \mathbf{y}, \mathbf{e_t})$
end
$\hat{\mathbf{x}}_f \leftarrow$ zeros($(1, N, P_x, P_y)$)
for $t \leftarrow 0$ to $T_s$ do
    $\bar{\mathbf{x}}_t \leftarrow 0$
    for $i \leftarrow 1$ to $S$ do
        $\bar{\mathbf{x}}_t \leftarrow \bar{\mathbf{x}}_t + \mathcal{S}[i][t]$
    end
    $\bar{\mathbf{x}}_t \leftarrow \bar{\mathbf{x}}_t / S$
    # Compute the entropy for each timestep
    $\hat{\mathbf{h}}_t \leftarrow$ compute_entropy($\bar{\mathbf{x}}_t$)
    # Compute timestep scaling weight
    $w_t \leftarrow \exp(-\alpha(T_s - t)/T_s)$
    for $i \leftarrow 1$ to $S$ do
        # Final prediction as the weighted sum
        $\hat{\mathbf{x}}_f \leftarrow \hat{\mathbf{x}}_f + w_t \cdot \bar{\mathbf{x}}_t$
    end
end
return $\hat{\mathbf{x}}_f$
```

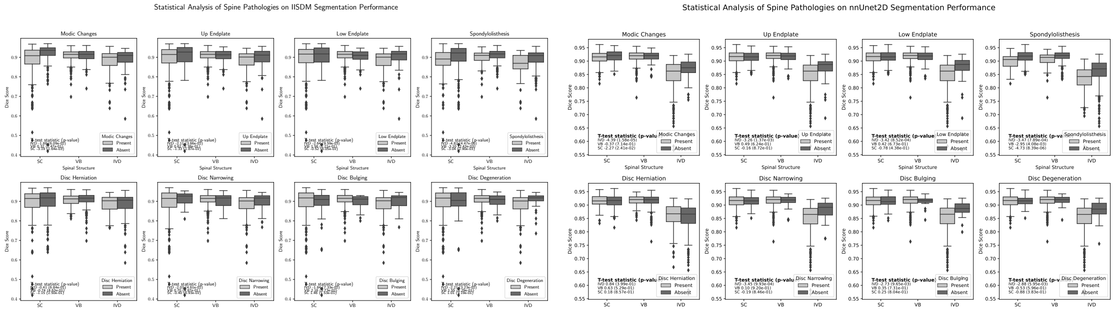

Figure 5: Dice scores boxplot for IISDM (left) and nnUnet (right)

Table 5: A quantitative analysis of Dice scores for 3D spinal volume segmentation of spinal structures (including the spinal canal, vertebrae, and intervertebral disks) using nnU-Net3D and Diff-UNet models across T1-weighted (T1w), T2-weighted (T2w), and combined T1w + T2w imaging modalities.

| **Model** | **Dim** | **Modality** | **Spinal Canal** | **Vertebrae** | **IVD** | **mDICE** |
|---|---|---|---|---|---|---|
| nnU-Net | 3D | T1w | 0.92 ± 0.09 | 0.93 ± 0.02 | 0.84 ± 0.04 | 0.897 |
| nnU-Net | 3D | T2w | 0.93 ± 0.03 | 0.93 ± 0.02 | 0.89 ± 0.04 | 0.917 |
| nnU-Net | 3D | T1w + T2w | **0.93 ± 0.02** | **0.93 ± 0.02** | 0.89 ± 0.04 | 0.917 |
| Diff-UNet | 3D | T1w | 0.92 ± 0.04 | 0.93 ± 0.04 | **0.91 ± 0.03** | **0.920** |
| Diff-UNet | 3D | T2w | 0.92 ± 0.02 | 0.93 ± 0.02 | 0.90 ± 0.03 | 0.917 |
| Diff-UNet | 3D | T1w + T2w | 0.92 ± 0.02 | 0.93 ± 0.02 | 0.89 ± 0.04 | 0.913 |

### B.2. Results of 3D Segmentation

We present a comprehensive analysis of the segmentation performance on full-sized 3D spine volumes. The training was conducted using complete 3D MRI datasets, allowing for a detailed evaluation of model capabilities in capturing complex anatomical structures. The results, as summarized in Table 5, highlight the segmentation accuracy across different spinal components, including the spinal canal, vertebrae, and intervertebral discs (IVD).

Notably, the Diff-UNet model demonstrates superior performance in segmenting IVDs, achieving the highest mean Dice score (mDICE) of 0.920 in the T1-weighted modality. These findings underscore the potential of 3D models to enhance segmentation precision, particularly in the context of detailed volumetric analysis.

## Appendix C. Baseline Comparison Experiments Details

### C.1. nnUnet Baseline

The nnU-Net model (Isensee et al., 2020) is trained using a highly automated and adaptable framework designed for semantic segmentation tasks which informs the configuration of multiple U-Net architectures tailored to the dataset's specific characteristics. The model training involves a multi-step process that includes preprocessing, model configuration, training. nnU-Net employs a five-fold cross-validation strategy to ensure robust performance evaluation. The training utilizes various configurations, such as 2D, 3D full resolution, to optimize segmentation performance across different data modalities. The hyperparameters that were used in the training are summarzed in the following tables:

### C.2. Implicit Image Segmentation Diffusion Model (IISMD)

IISMD (Wolleb et al., 2021) follows DDPM training, adding Gaussian noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the segmentation mask $\mathbf{x}_0$ at each timestep $t \in \{1, \ldots, T\}$ using a linear noise scheduler $\{\alpha_t \in (0,1)\}_{t=1}^{T}$. For denoising, U-Net architecture $f_{\boldsymbol{\theta}}$ estimates noise $\boldsymbol{\epsilon}_t = f_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}, t)$ at each timestep, concatenated with MRI images $\mathbf{y}$, used to guide the generation of the

Table 6: Training Hyperparameters for nnUnet 2D and 3D

| **Parameter** | **T1w** | **T2w** | **T1w+T2w** | **T1w** | **T2w** | **T1w+T2w** |
|---|---|---|---|---|---|---|
| Patch Size | 256x64 | 256x64 | 256x64 | 56x 224x192 | 56x224x192 | 56x224x192 |
| Epochs | 250 | 250 | 250 | 250 | 250 | 250 |
| Batch | 197 | 197 | 197 | 2 | 2 | 2 |
| Optimizer | SGD | SGD | SGD | SGD | SGD | SGD |
| Learning Rate | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| Training Loss | Dice | Dice | Dice | Dice | Dice | Dice |
| | | **nnUnet 2D** | | | **nnUnet 3D** | |

segmentation mask. The parameters $\boldsymbol{\theta}$ are optimized by minimizing the Mean Squared Error (MSE) loss between the estimated noise $\hat{\boldsymbol{\epsilon}}_t$ and the true noise $\boldsymbol{\epsilon}_t$.

In the inference or sampling process, the model takes random noise concatenated with the MRI input image ($\mathbf{x}_y$) and iteratively denoises the segmentation mask by estimating the noise $\hat{\boldsymbol{\epsilon}}_t$ at each timestep. During the sampling procedure, uncertainty maps are synthesized by exploiting the inherent stochasticity present in DDPMs. Through iterative application of IISMD, multiple segmentation masks are produced for a given input image. The uncertainty map is derived by assessing the pixel-wise variance of the masks.

Table 7: Training Hyperparameters for IISDM

| **Hyperparameter** | **T1w, T2w, T1w+T2w** |
|---|---|
| **Image Size** | 320x320 |
| **Epochs** | 2600 |
| **Batch** | 10 |
| **Optimizer** | AdamW |
| **Learning Rate** | 0.0001 |
| **Training Loss** | MSE |

### C.3. DiffUnet

DiffUnet (Xing et al., 2023) is a diffusion-based volumetric segmentation framework for medical volumetric segmentation that directly infers the segmentation mask $\hat{\mathbf{x}}_0$ from a partially noised input $\mathbf{x_t}$. The architecture includes an additional encoder to extract features from MRI scans, which enhances the model during training. The training uses a composite loss function that combines cross-entropy, Dice, and MSE losses to penalize segmentation errors. During the inference phase, Diff-UNet employs the DDIM (Song et al., 2020) sampling algorithm, which accelerates the process while maintaining a balance between speed and accuracy. To further improve robustness, Diff-UNet performs step-uncertainty-based fusion during sampling $\mathbf{u}_i = -\bar{p}_i \log(\bar{p}_i)$, applied to the step-wise predictions to compile the final fused result mask $\hat{\mathbf{x}}$.

Due to the computational load of the diffusion models, the volumetric segmentation for DiffU-Net was performed patch-wise with input size $32 \times 120 \times 120$ and sliding window inference with 0.5 overlap. The training hyperparameters are summarized in: the next table

Table 8: Hyperparameters for DiffU-Net

| | **T1w** | **T2w** | **T1w+T2w** |
|---|---|---|---|
| **Patch Size** | 32x128x128 | 32x128x128 | 32x128x128 |
| **Epochs** | 1350 | 1400 | 700 |
| **Batch** | 4 | 4 | 4 |
| **Optimizer** | AdamW | AdamW | AdamW |
| **Learning Rate** | 0.0001 | 0.0001 | 0.0001 |
| **Training Loss** | MSE + Dice + CE | MSE + Dice + CE | MSE + Dice + CE |