

PCaPaint: Prostate Cancer Inpainting by Mitigating Shortcut Learning

Levente Lippenszky¹[0009-0000-7835-2071], Hongxu Yang²[0000-0003-2447-094X],
Marcell Dömötör³, Krisztian Koos³[0000-0003-3136-8978], and László
Ruskó¹[0009-0003-5073-8978]

¹ Science & Technology Org. AI & ML, GE HealthCare, Budapest, Hungary

² Science & Technology Org. AI & ML, GE HealthCare, Eindhoven, Netherlands

³ Science & Technology Org. AI & ML, GE HealthCare, Szeged, Hungary

levente.lippenszky@gehealthcare.com

Abstract. The development of AI systems for tumor-specific applications is limited by the scarcity of labeled data. Synthetic tumor inpainting offers a promising approach but faces challenges for prostate cancer MRI which contains high-resolution multi-sequence data. Although methods leveraging latent diffusion models (LDMs) enable large-volume synthesis, they are prone to shortcut learning, simply reproducing the condition image created by masking the lesion region. In this work, we introduce PCaPaint, a prostate cancer inpainting method based on LDMs that explicitly addresses this failure mode. To overcome shortcut learning that compromises synthetic tumor texture, we propose a simple yet efficient conditioning strategy in which the condition image is filled with Gaussian noise, and we provide theoretical justification. In addition, we propose a novel training objective for LDM that emphasizes the error within the lesion region. Furthermore, we introduce a multi-sequence latent design, in which T2w scans and DWI&ADC scans are compressed using two separate autoencoders to preserve their distinct frequency characteristics. Extensive experiments demonstrate that the generated synthetic data improves downstream performance in prostate lesion segmentation, patient-level classification and lesion-level detection. Furthermore, our method significantly outperforms a recent state-of-the-art LDM-based tumor inpainting method both in downstream performance and in synthetic image quality.

Keywords: Prostate cancer · Inpainting · Shortcut learning.

1 Introduction

The development of AI systems in healthcare is often constrained by the scarcity of labeled data [7, 11]. This challenge is even more pronounced in tumor-specific applications, where pathological cases are far less common than normal patient data [21]. Furthermore, developing robust deep learning models for these applications requires substantial amount of data with sufficient diversity [12, 19, 5].

However, collecting these datasets is expensive, and the difficulty of accurately annotating tumors intensifies this challenge [3, 20].

Synthetic tumor inpainting offers a promising approach for generating diverse lesion data while eliminating the need for manual annotation. However, prostate biparametric MRI (bpMRI) synthesis poses significant challenges as it involves high-resolution 3D data and three MRI sequences with substantially different frequency characteristics. In recent years, tumor inpainting methods based on diffusion models [8] have emerged covering a large variety of lesion types. LeFusion [21] incorporates forward-diffused background context into the reverse diffusion process, but its image-space design restricts its use for prostate bpMRI. Other state-of-the-art (SOTA) lesion inpainting methods, such as DiffTumor [3] and SynBT [20], leverage latent diffusion models (LDMs) [14]. These approaches use a healthy image as condition where the lesion region is zeroed out, and the LDM is trained to reconstruct the lesion. While latent representations enable prostate bpMRI synthesis, this conditioning strategy can lead to shortcut learning, causing the model to reconstruct the healthy image rather than inpaint lesion texture. Fig. 1 illustrates shortcut learning in DiffTumor [3]. Fig. 1 (a) shows a ground truth T2-weighted (T2w) training example with a real lesion, and (b) the corresponding healthy condition image with the lesion region filled with zeros (gray). After 200,000 training steps, the generated output in (c) exhibits the shortcut failure mode, reproducing the healthy image instead of synthesizing lesion texture. Fig. 1 (d) and (e) show the same behavior when the conditioning image is filled with -1 values (black).

In this work, we propose PCaPaint, a novel prostate cancer inpainting method based on LDMs that generates realistic synthetic lesions in healthy prostate bpMRI cases, while explicitly addressing shortcut learning. Our main **contributions** are as follows.

1. We introduce a simple yet efficient conditioning strategy where the condition image is filled with Gaussian noise and provide theoretical justification that it mitigates shortcut learning. In addition, we propose a training objective for LDM that emphasizes the error within the lesion region to further mitigate this failure mode.
2. We introduce a novel multi-sequence latent design, in which T2w scans and DWI&ADC scans are compressed using two separate autoencoders to preserve their distinct frequency characteristics.
3. Extensive experiments show that our synthetic data improves performance in three downstream tasks. Moreover, our method significantly outperforms a SOTA LDM-based tumor inpainting method both in downstream performance and in synthetic image quality.

2 Methods

2.1 Conditioning with Noise Fill

We provide theoretical justification that filling the lesion region in the condition image with noise mitigates shortcut learning. Proposition 1 assumes a diffusion

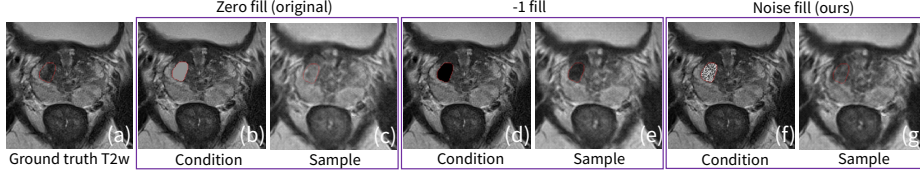


Fig. 1. DiffTumor [3] samples under different conditioning strategies. Panels show the condition and sample for zero fill (gray), -1 fill (black) and the proposed noise fill.

model [8], but the reasoning extends to LDM [14]. This is illustrated qualitatively in Fig. 1 (f) and (g) where DiffTumor [3] generates improved lesion texture, and quantitatively in Section 3.2.

Let $\mathbf{x}_0$ denote the lesion image and $\mathbf{h}$ the healthy condition image. The healthy image is defined as $\mathbf{h} = (1 - \mathbf{m}) \odot \mathbf{x}_0 + \mathbf{m} \odot \boldsymbol{\eta}$, where $\mathbf{m}$ is the lesion mask and $\boldsymbol{\eta}$ is the tensor used to fill the lesion region. In LDM-based tumor inpainting methods, $\boldsymbol{\eta}$ is the all-zero tensor [20, 3]. We define the lesion difference as $\mathbf{d} := \mathbf{x}_0 - \mathbf{h} = \mathbf{m} \odot (\mathbf{x}_0 - \boldsymbol{\eta})$. The diffusion forward process defines the noisy lesion image at timestep t as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ and $\alpha_t = 1 - \beta_t$, with $\{\beta_t\}_{t=1}^T$ denoting the noise schedule and T the number of time steps. The diffusion model $\boldsymbol{\epsilon}_\theta$ predicts the noise as $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{h}, \mathbf{m})$. Fig. 1 (b)-(e) shows that DiffTumor learns a shortcut and denoises $\mathbf{x}_t$ by simply reconstructing the healthy image $\mathbf{h}$ when the fill tensor is constant. We assume the diffusion model exploits a shortcut by approximating the analytically computable residual $\mathbf{r}_t$ as its noise prediction

$$\mathbf{r}_t = \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{h}}{\sqrt{1 - \bar{\alpha}_t}}. \quad (1)$$

This solution depends only on the inputs $(\mathbf{x}_t, \mathbf{h}, \mathbf{m}, t)$ available to the network and the fixed noise schedule.

Proposition 1. Assume the network exploits the shortcut $\hat{\boldsymbol{\epsilon}} = \mathbf{r}_t$. Consider (i) constant fill $\boldsymbol{\eta} = \mathbf{c}$ and (ii) Gaussian noise fill $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{c}, \sigma^2 \mathbf{I})$, where $\mathbf{c}$ is a constant tensor and $\sigma > 0$ denotes the standard deviation. Then, the difference of the expected training losses satisfies

$$\mathcal{L}_{\text{noise}} - \mathcal{L}_{\text{const}} = \kappa \sigma^2 \mathbb{E}_{\mathbf{x}_0} [n_m(\mathbf{x}_0)] > 0, \quad (2)$$

where $t \sim \text{Uniform}(\{1, \dots, T\})$, $\kappa_t := \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}}$ and $\kappa = \mathbb{E}_t [\kappa_t^2] > 0$, $n_m(\mathbf{x}_0) > 0$ is the number of masked voxels for $\mathbf{x}_0$. Noise fill increases the loss for the shortcut solution, encouraging the network to learn the intended denoising.

Proof. Substituting $\mathbf{x}_t$ into $\mathbf{r}_t$ gives

$$\mathbf{r}_t = \frac{\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon} - \sqrt{\bar{\alpha}_t} \mathbf{h}}{\sqrt{1 - \bar{\alpha}_t}} = \boldsymbol{\epsilon} + \frac{\sqrt{\bar{\alpha}_t} (\mathbf{x}_0 - \mathbf{h})}{\sqrt{1 - \bar{\alpha}_t}} = \boldsymbol{\epsilon} + \kappa_t \mathbf{d}. \quad (3)$$

The residual is the sum of the true noise and the scaled lesion difference. The shortcut behavior dominates as the lesion mask $\mathbf{m}$ approaches the empty mask, since $\mathbf{r}_t$ then converges to ϵ . The training loss $\mathcal{L}$ takes the following form

$$\mathbb{E} [\|\hat{\epsilon} - \epsilon\|^2] = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{m}, \epsilon, \boldsymbol{\eta}} [\|\kappa_t \mathbf{d}\|^2] = \mathbb{E}_t [\kappa_t^2] \mathbb{E}_{\mathbf{x}_0, \mathbf{m}, \boldsymbol{\eta}} [\|\mathbf{m} \odot (\mathbf{x}_0 - \boldsymbol{\eta})\|^2] . \quad (4)$$

We define the diagonal matrix $\mathbf{M} := \text{diag}(\text{vec}(\mathbf{m}))$ that performs the same masking operation as $\mathbf{m}$ when the images are treated as vectors. We assume the lesion mask is provided for each $\mathbf{x}_0$ via manual annotation, and we treat $\mathbf{M} = \mathbf{M}(\mathbf{x}_0)$ as deterministic given $\mathbf{x}_0$. Using the law of total expectation and $\boldsymbol{\eta} \perp \mathbf{x}_0$,

$$\mathcal{L} = \kappa \mathbb{E}_{\mathbf{x}_0} [\mathbb{E}_{\boldsymbol{\eta}} [\|\mathbf{M}(\mathbf{x}_0)(\mathbf{x}_0 - \boldsymbol{\eta})\|^2 \mid \mathbf{x}_0]] . \quad (5)$$

By the bias-variance decomposition [1], the expected squared L_2 norm of a random vector $\mathbf{X}$ can be decomposed as $\mathbb{E} [\|\mathbf{X}\|^2] = \text{tr}(\text{Cov}(\mathbf{X})) + \|\mathbb{E}[\mathbf{X}]\|^2$. To decompose the inner expectation in Eq. (5), we define $\mathbf{Z} := \mathbf{M}(\mathbf{x}_0)(\mathbf{x}_0 - \boldsymbol{\eta})$. Then $\mathbb{E}[\mathbf{Z} \mid \mathbf{x}_0] = \mathbf{M}(\mathbf{x}_0)(\mathbf{x}_0 - \boldsymbol{\mu}_{\boldsymbol{\eta}})$ and $\text{Cov}(\mathbf{Z} \mid \mathbf{x}_0) = \mathbf{M}(\mathbf{x}_0)\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\mathbf{M}(\mathbf{x}_0)^T$ where $\boldsymbol{\mu}_{\boldsymbol{\eta}} := \mathbb{E}[\boldsymbol{\eta}]$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} := \text{Cov}(\boldsymbol{\eta})$. Applying the decomposition yields

$$\mathbb{E} [\|\mathbf{Z}\|^2 \mid \mathbf{x}_0] = \text{tr}(\text{Cov}(\mathbf{Z} \mid \mathbf{x}_0)) + \|\mathbb{E}[\mathbf{Z} \mid \mathbf{x}_0]\|^2 . \quad (6)$$

Substituting into Eq. (5) gives

$$\mathcal{L} = \kappa \mathbb{E}_{\mathbf{x}_0} \left[\text{tr}(\mathbf{M}(\mathbf{x}_0)\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\mathbf{M}(\mathbf{x}_0)^T) + \|\mathbf{M}(\mathbf{x}_0)(\mathbf{x}_0 - \boldsymbol{\mu}_{\boldsymbol{\eta}})\|^2 \right] . \quad (7)$$

In (i) constant fill, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \mathbf{0}$ and $\boldsymbol{\mu}_{\boldsymbol{\eta}} = \mathbf{c}$, which implies that

$$\mathcal{L}_{\text{const}} = \kappa \mathbb{E}_{\mathbf{x}_0} [\|\mathbf{M}(\mathbf{x}_0)(\mathbf{x}_0 - \mathbf{c})\|^2] . \quad (8)$$

In (ii) noise fill, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \sigma^2 \mathbf{I}$, $\boldsymbol{\mu}_{\boldsymbol{\eta}} = \mathbf{c}$ and $\text{tr}(\mathbf{M}(\mathbf{x}_0)\sigma^2 \mathbf{I}\mathbf{M}(\mathbf{x}_0)^T) = \sigma^2 n_m(\mathbf{x}_0)$, which yields

$$\mathcal{L}_{\text{noise}} = \kappa \mathbb{E}_{\mathbf{x}_0} [\sigma^2 n_m(\mathbf{x}_0) + \|\mathbf{M}(\mathbf{x}_0)(\mathbf{x}_0 - \mathbf{c})\|^2] . \quad (9)$$

Subtracting Eq. (8) from Eq. (9) completes the proof.

2.2 PCaPaint

Autoencoders In the first stage, T2w scans and DWI&ADC scans are compressed separately to reflect their distinct characteristics: T2w contains rich high-frequency anatomical detail, while ADC and DWI exhibit lower-frequency intensity patterns. We employ the vector-quantized variational autoencoder with adversarial training (VQGAN) [6] for both autoencoders, $i \in \{1, 2\}$. We denote the T2w image as $\mathbf{x}^{(1)} \in \mathbb{R}^{H \times W \times D}$, where H, W and D represent the height, width, and depth, respectively. We further define $\mathbf{x}^{(2)} \in \mathbb{R}^{H \times W \times D \times 2}$, representing the DWI and ADC volumes concatenated along the channel dimension. The input MRI tensor $\mathbf{x}^{(i)}$ is mapped to a continuous latent representation

$\mathbf{z}_e^{(i)} \in \mathbb{R}^{h \times w \times d \times c}$ via the encoder $f^{(i)}$, where h, w, d denote the latent space dimensions and c is the number of channels. In the vector quantization step, each c -dimensional latent vector in $\mathbf{z}_e^{(i)}$ is replaced by its nearest codebook entry from the learned codebook $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^K$, resulting in the quantized latent $\mathbf{z}_q^{(i)} = q(\mathbf{z}_e^{(i)}) \in \mathbb{R}^{h \times w \times d \times c}$, where K is the codebook size. Lastly, the decoder $g^{(i)}$ reconstructs the quantized latent such that $\hat{\mathbf{x}}^{(i)} = g^{(i)}(\mathbf{z}_q^{(i)})$. The loss is defined as $\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{commit}} + \lambda_{\text{percep}} \mathcal{L}_{\text{percep}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}$, where $\mathcal{L}_{\text{recon}} = \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|_1$ and $\mathcal{L}_{\text{commit}} = \beta \|\text{sg}[\mathbf{z}_q^{(i)} - \mathbf{z}_e^{(i)}]\|_2^2$, with $\text{sg}[\cdot]$ denoting the stop-gradient operator and β the commitment cost. The perceptual loss $\mathcal{L}_{\text{percep}}$ measures feature similarity between $\mathbf{x}^{(i)}$ and $\hat{\mathbf{x}}^{(i)}$ in a pretrained network, and $\mathcal{L}_{\text{adv}}$ computes the least-squares patch adversarial loss using a patch discriminator [10].

Latent Diffusion Model In the second stage, we train an LDM [14] conditioned on the healthy image and the lesion mask. Lesion bpMRI $\mathbf{x}_0$ is encoded to the latent space using the VQGAN encoders, $\mathbf{z}_0 = \text{concat}(\mathbf{z}_0^{(1)}, \mathbf{z}_0^{(2)})$, where $\mathbf{z}_0^{(i)} = f^{(i)}(\mathbf{x}_0^{(i)})$ and $i \in \{1, 2\}$. In the forward process, $\mathbf{z}_0$ is gradually transformed into standard Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ following a Markov process. The reverse process is parametrized by a denoising U-Net ϵ_θ that predicts the noise added at each time step. The healthy image is constructed by filling the lesion region with Gaussian noise. Formally, we generate standard Gaussian noise $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, scale it by $\sigma > 0$ and map it to the $(-1, 1)$ range using the tanh function, $\boldsymbol{\eta} = \tanh(\sigma \boldsymbol{\xi})$. Then, $\mathbf{h}$ is defined as $\mathbf{h} = (1 - \mathbf{m}) \odot \mathbf{x}_0 + \mathbf{m} \odot \boldsymbol{\eta}$ and encoded into latent $\mathbf{z}_h$ where $\mathbf{z}_h = \text{concat}(\mathbf{z}_h^{(1)}, \mathbf{z}_h^{(2)})$, $\mathbf{z}_h^{(i)} = f^{(i)}(\mathbf{h}^{(i)})$ and $i \in \{1, 2\}$. Furthermore, we downsample the lesion mask to match the latent space dimensions, $\tilde{\mathbf{m}} = \text{down}(\mathbf{m})$. The denoising U-Net receives $\mathbf{z}_h$ and $\tilde{\mathbf{m}}$ as conditions concatenated to $\mathbf{z}_t$ along the channel dimension. To further mitigate shortcut learning, we employ a lesion-weighted training objective that amplifies the error inside the lesion region:

$$\mathbb{E}_{\mathbf{x}_0, \mathbf{m}, \epsilon, \boldsymbol{\xi}, t} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{z}_h, \tilde{\mathbf{m}})\|^2 + \lambda \|\tilde{\mathbf{m}} \odot \epsilon - \tilde{\mathbf{m}} \odot \epsilon_\theta(\mathbf{z}_t, t, \mathbf{z}_h, \tilde{\mathbf{m}})\|^2], \quad (10)$$

where λ controls the contribution of the lesion-focused term.

Sampling In the third stage, we generate prostate lesion masks incorporating medical knowledge and inpaint synthetic lesions onto non-lesion bpMRIs. Lesion synthesis is performed by sampling the LDM using DDIM [17], conditioned on the healthy image and the lesion mask. Given peripheral zone (PZ) and transition zone (TZ) segmentation masks, we sample lesion center with probabilities of 0.75 from PZ and 0.25 from TZ [13]. We then sample a target lesion volume in voxels constrained to lie between minimum and maximum fractions of the whole gland volume. An ellipsoidal lesion is generated by drawing random axis ratios and scaling the ellipsoid to match the target volume while accounting for voxel spacing. To create realistic, irregular boundaries, we perturb the ellipsoid surface using Gaussian-filtered noise field scaled by the mean radius. Finally, we clip the mask to the gland by removing voxels outside the gland segmentation.

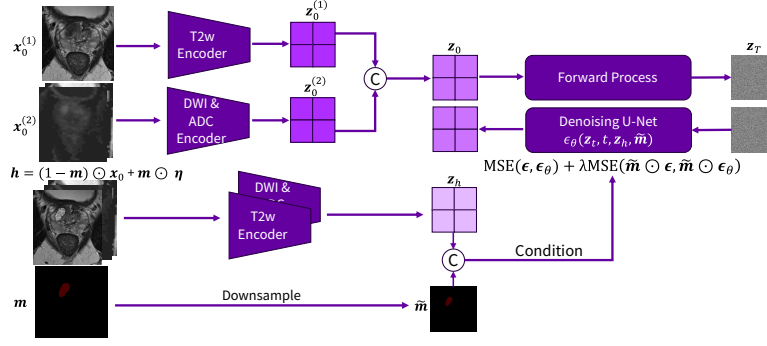


Fig. 2. Training the latent diffusion model (LDM). The lesion bpMRI $\mathbf{x}_0$ is encoded by the VQGAN encoders to form the latent $\mathbf{z}_0$. LDM is conditioned on the latent healthy image $\mathbf{z}_h$ and the downsampled lesion mask $\tilde{\mathbf{m}}$. The healthy image $\mathbf{h}$ is constructed by replacing the lesion region with Gaussian noise. LDM is trained with an objective that amplifies the error in the lesion region.

3 Experimental Results

3.1 Experimental Setup

Data Training set consisted of 900 bpMRI cases from the publicly available PI-CAI [16] dataset, including 252 cases with clinically significant prostate cancer (csPCa). The test set comprised of 599 bpMRI cases from PI-CAI, of which 172 were csPCa cases. Preprocessing was performed using the `picaiprep` repository [15]. Each scan was resampled to a resolution of $0.5 \times 0.5 \times 3.0 \text{ mm}^3$ and cropped or padded to a spatial size of $256 \times 256 \times 32$ voxels. Image intensities were rescaled to the $[-1, 1]$ range.

Implementation Details For PCaPaint, we trained the two VQGANs using the VQ-VAE [18] implementation of MONAI 1.5.0 [2] for 100,000 steps using the Adam optimizer. The learning rate was set to 5×10^{-5} for the T2w VQGAN and to 10^{-4} for the DWI&ADC VQGAN. We set the embedding dimensions to 64 and the codebook size to 8192 for both autoencoders. The denoising U-Net was trained using the DiffTumor [3] implementation for 200,000 steps with Adam and a learning rate of 5×10^{-5} . The standard deviation of the Gaussian noise fill was set to $\sigma = 0.5$ such that most sampled values lie within the $[-1, 1]$ range. The weight of the second loss term in Eq. (10) was empirically set to $\lambda = 150$. For DiffTumor, we trained a single VQGAN to compress the bpMRIs with a learning rate of 5×10^{-5} . To provide larger capacity for the autoencoder, we increased the embedding dimensions to 128. We applied the same training setup as for PCaPaint to ensure a fair comparison. Trainings were performed on an NVIDIA A100 Tensor Core GPU with 40 GB memory.

Table 1. Synthetic image quality evaluation for DiffTumor [3], DiffTumor with noise fill (NF) and PCaPaint. SSIM and PSNR were computed between ground truth and synthetic images for test lesion bpMRIs for the entire image, foreground (fg) and background (bg). 95% bootstrap CI widths are in brackets. Best and second-best are in **bold** and underline, respectively. * indicates statistically significant differences between DiffTumor and PCaPaint at 5% level using two-sided Wilcoxon signed-rank tests with Holm correction.

MRI	Metric $\uparrow$	DiffTumor	DiffTumor NF	PCaPaint (Ours)
T2w	SSIM	0.77 (0.01)	<u>0.79</u> (0.01)	0.83* (0.02)
	PSNR	20.75 (0.56)	<u>21.58</u> (0.60)	23.11* (0.62)
	PSNR fg	12.38 (0.64)	<u>16.26</u> (0.80)	20.42* (0.97)
	PSNR bg	20.80 (0.55)	<u>21.60</u> (0.60)	23.11* (0.62)
DWI	SSIM	0.51 (0.06)	0.51 (0.06)	0.64* (0.05)
	PSNR	<u>21.39</u> (0.66)	21.16 (0.67)	25.54* (1.26)
	PSNR fg	15.58 (0.98)	<u>16.21</u> (0.89)	17.98* (1.12)
	PSNR bg	<u>21.43</u> (0.66)	21.18 (0.67)	25.59* (1.27)
ADC	SSIM	0.71 (0.04)	0.71 (0.04)	0.79* (0.03)
	PSNR	22.37 (0.52)	<u>22.47</u> (0.57)	25.42* (0.60)
	PSNR fg	15.82 (0.71)	<u>18.31</u> (0.82)	22.27* (0.88)
	PSNR bg	22.42 (0.52)	<u>22.49</u> (0.57)	25.44* (0.61)

3.2 Image Quality Evaluation

Synthetic image quality was evaluated for DiffTumor, DiffTumor with noise fill (NF) and PCaPaint by inpainting the lesion region of the 172 real bpMRIs in the test set. We computed structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR) between the ground truth and synthetic images. In addition, we calculated PSNR separately for the foreground and background regions. Table 1 demonstrates that PCaPaint generates statistically significantly higher quality images compared to DiffTumor across MRI sequences and image regions. Foreground (fg) results show that DiffTumor with NF outperforms DiffTumor, while PCaPaint achieves the best image quality, highlighting the benefits of our noise fill conditioning and lesion-focused loss. PCaPaint also achieves the best performance in background (bg) metrics, underscoring the advantage of compressing bpMRIs using two separate autoencoders.

3.3 Prostate Lesion Downstream Tasks

We evaluated the utility of synthetic prostate lesion bpMRIs for three downstream tasks on the test set: csPCa segmentation, patient-level csPCa classification and lesion-level csPCa detection. Evaluation metrics included Dice score for segmentation, area under the receiver operating characteristic curve (AUC) for classification and average precision (AP) for detection. The latter two tasks follow the PI-CAI challenge protocol and were evaluated using the `picai_eval` repository [15]. First, we trained nnU-Net [9] with default configuration on the training set. We then trained separate nnU-Net models on training sets augmented with synthetic lesions generated by DiffTumor and PCaPaint at a ratio

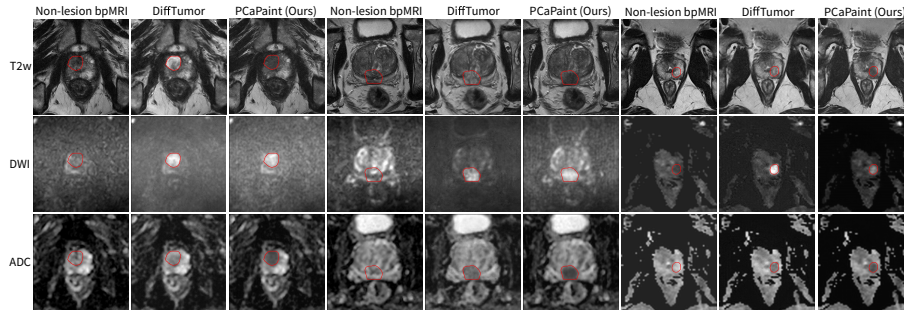


Fig. 3. Three non-lesion bpMRI examples inpainted with synthetic lesions by DiffTumor [3] and PCaPaint using generated lesion masks.

of three synthetic lesions per real lesion [4]. Fig. 3 illustrates three non-lesion bpMRI cases, each inpainted with synthetic lesion by both methods, using lesion masks described in Section 2.2. Results demonstrate that our method correctly inpaints lesion textures with modality-specific intensity—darker in T2w, brighter in DWI, and darker in ADC—and outperforms DiffTumor. Statistical significance for segmentation was assessed using two-sided Wilcoxon signed-rank test. For classification and detection, significance was determined by bootstrapping AUC and AP differences, CIs excluding zero indicated significance. Table 2 shows that synthetic examples from PCaPaint yield the best performance across all downstream tasks, significantly outperforming DiffTumor. Although the Dice improvement is small, paired differences in the Wilcoxon test consistently favored PCaPaint over DiffTumor: 87 positives, 60 negatives and 25 ties across the 172 test lesion cases.

Table 2. nnU-Net performance on csPCa segmentation, patient-level csPCa classification and lesion-level csPCa detection using real and synthetic data. 95% bootstrap CI widths are in brackets. Best performance is in **bold**. * indicates statistically significant differences between Real+DiffTumor and Real+PCaPaint at 5% level.

Downstream Task	Real	Real+DiffTumor	Real+PCaPaint (Ours)
Segmentation, Dice $\uparrow$	0.47 (0.08)	0.47 (0.09)	0.48* (0.08)
Classification, AUC $\uparrow$	0.71 (0.09)	0.71 (0.09)	0.76* (0.08)
Detection, AP $\uparrow$	0.28 (0.13)	0.29 (0.13)	0.34* (0.15)

4 Conclusion

In this work, we propose a prostate cancer inpainting method based on LDMs. We address shortcut learning in existing LDM-based tumor inpainting methods by introducing a conditioning strategy with theoretical justification. As Proposition 1 applies broadly to diffusion-based tumor inpainting, we hypothesize that

our conditioning would benefit other imaging modalities and tumor types. Future work will explore these directions. As our synthetic data improves performance on prostate lesion downstream tasks, it will support the development of more robust AI systems for prostate cancer care.

Acknowledgments. We thank Avinash Gopal for his guidance on applying synthetic data to this domain, and Franciska Rajki for valuable discussions on experimental results. This study was funded under the 2023 Call for “R&D Projects linked to Personalized Medicine and Advanced Therapies within the framework of the ISCHII-CDTI Joint Initiative”, financed through the “Recovery, Transformation and Resilience Plan – Funded by the European Union – NextGenerationEU”.

Disclosure of Interests. This work was conducted by the authors as part of their full-time employment at GE HealthCare.

References

1. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
2. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
3. Chen, Q., Chen, X., Song, H., Xiong, Z., Yuille, A., Wei, C., Zhou, Z.: Towards generalizable tumor synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11147–11158 (2024)
4. Chen, Q., Zhou, X., Liu, C., Chen, H., Li, W., Jiang, Z., Huang, Z., Zhao, Y., Yu, D., He, J., et al.: Scaling tumor segmentation: Best lessons from real and synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24001–24013 (2025)
5. Chou, Y.C., Li, B., Fan, D.P., Yuille, A., Zhou, Z.: Acquiring weak annotations for tumor localization in temporal and volumetric data. Machine Intelligence Research **21**(2), 318–330 (2024)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
7. Grabke, E.P., Taati, B., Haider, M.A.: Mitigating multi-sequence 3d prostate mri data scarcity through domain adaptation using locally-trained latent diffusion models for prostate cancer detection. arXiv preprint arXiv:2507.06384 (2025)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
11. Jin, C., Guo, Z., Lin, Y., Luo, L., Chen, H.: Label-efficient deep learning in medical image analysis: Challenges and future directions. arXiv preprint arXiv:2303.12484 (2023)

12. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21152–21164 (2023)
13. McNeal, J.E., Redwine, E.A., Freiha, F.S., Stamey, T.A.: Zonal distribution of prostatic adenocarcinoma: correlation with histologic pattern and direction of spread. *The American journal of surgical pathology* **12**(12), 897–906 (1988)
14. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
15. Saha, A., Bosma, J., Twilt, J., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., et al.: Artificial intelligence and radiologists at prostate cancer detection in mri—the pi-cai challenge. In: Medical Imaging with Deep Learning, short paper track (2023)
16. Saha, A., Twilt, J.J., Bosma, J.S., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., Huisman, H.: The pi-cai challenge: public training and development dataset. (No Title) (2022), <https://doi.org/10.5281/zenodo.6624726>
17. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
18. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. *Advances in neural information processing systems* **30** (2017)
19. Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al.: Annotation-efficient deep learning for automatic medical image segmentation. *Nature communications* **12**(1), 5915 (2021)
20. Yang, H., Timko, E., Lippenszky, L., Czipczer, V., Ferenczi, L.: Synbt: High-quality tumor synthesis for breast tumor segmentation by 3d diffusion model. In: Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care. pp. 1–10. Springer (2025)
21. Zhang, H., Liu, Y., Yang, J., Wan, S., Wang, X., Peng, W., Fua, P.: Lefusion: Controllable pathology synthesis via lesion-focused diffusion models. arXiv preprint arXiv:2403.14066 (2024)