DCE-diff: Diffusion Model for Synthesis of Early and Late Dynamic Contrast-Enhanced MR Images from Non-Contrast Multimodal Inputs

Kishore Kumar M[1]† Sriprabha Ramanarayanan[1]∗ Sadhana S[1] Arunima Sarkar[1]
Matcha Naga Gayathri[1] Keerthi Ram[2] Mohanasankar Sivaprakasam[1,2]

[1]Indian Institute of Technology Madras (IITM), India
[2]Healthcare Technology Innovation Center (HTIC), India

†kishore.m@htic.iitm.ac.in
∗ee19D013@smail.iitm.ac.in

Paper ID 69

Abstract

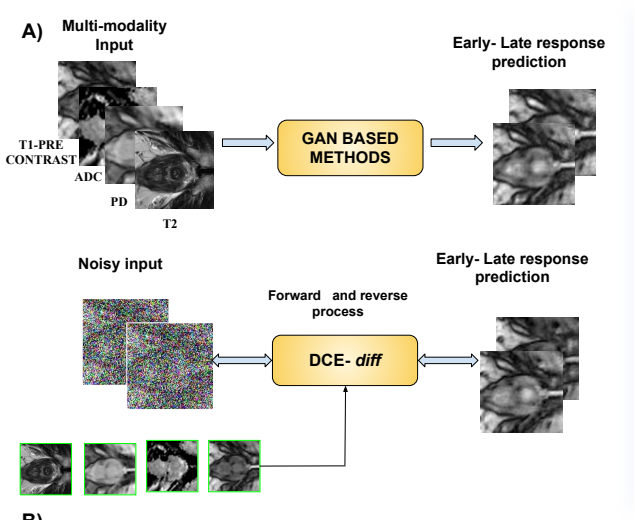
Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) is pivotal in delineating abnormal lesions and cancerous regions in the anatomy of interest. However, DCE-MRI requires the injection of gadolinium (Gad)-based contrast agents during acquisition which is known to have potential toxic effects, posing radiological concerns. Previous deep learning models employed for synthesizing DCE-MRI images consider unimodal structural MRI inputs lacking information about perfusion or perform early to late response predictions requiring Gad-based MRI sequences as input to drive the synthesis. In this work, we consider the heterogeneity in (i) the multimodal MRI structural inputs offering diverse and complementary anatomical features, (ii) the scanner settings and acquisition parameters, and (iii) the importance of incorporating the perfusion information in Apparent Diffusion Coefficient (ADC) data, which is essential to learn the hyperintense features for DCE-MRI synthesis. We propose DCE-diff, a deep generative diffusion model for multimodal image-to-image mapping from non-contrast structural MRI sequences and ADC maps to synthesize early and late response DCE-MRI images to circumvent Gad contrast injection to patients. Comparative studies using ProstateX and Prostate-MRI datasets against previous methods show that our model demonstrates (i) better synthesis quality with improvement margins of +0.85 dB in PSNR, +0.04 in SSIM, -22.8 in FID, and -0.02 in MAE (ii) better adaptability to different scanner data with deviated settings, showcasing a +8.7 dB improvement in PSNR, +0.22 in SSIM, -40.4 in FID, and -0.1 in MAE, and (iii) the importance of ADC maps in the DCE-MRI synthesis.

Keywords *- Dynamic Contrast-Enhanced MRI, Diffusion models, Prostate, Multimodal Image-to-Image translation*

1. Introduction

Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) is a medical image scanning system that measures perfusion, blood flow, and tissue characteristics, and highlights the cancerous lesions with hyperintensity. It captures the increased tissue perfusion and permeability by use of a Gadolinium-based (Gad) contrast agent. DCE-MRI consists of early-phase and late-phase contrast-enhanced images that accentuate the contrast uptake over time during acquisition. However, Gad retention [1] is still one of the biggest radiological concerns due to contraindications like Nephrogenic Systemic Fibrosis (NSF) and hypersensitivity reactions [2]. Various deep learning methods [3, 4] have been proposed to synthesize DCE-MRI images to overcome the toxic effects caused by Gad-based imaging.

Typically, DCE-MRI protocol is characterized by the acquisition of multiple heterogeneous MRI sequences which provide diverse and complementary perspectives to aid radiological decision-making. Among these sequences, the Apparent Diffusion coefficient (ADC) image is of prime importance in DCE-MRI as it contains essential information about the perfusion of contrast in the organ of interest. Moreover, DCE-MRI demonstrates many variations arising from cross-scanner patient demographics, devices, and acquisition parameters [5]. For example, the b-values of Diffusion Weighted Imaging (DWI) sequences from the Philips Achieva scanner might be vastly different from those of the Siemens Magnetron scanner [6, 7]. These differences



B)

Methods \ Aspects	No Dependency on DCE	Multimodal Inputs	No Task specific loss	No Hyper parameter tuning
Early-to-Late [8]	×	×	×	×
TSGAN [21]	✔	×	×	×
Ours(DCE-diff)	✔	✔	✔	✔

Figure 1. A) Generalized concept diagram of GANs and DCE-diff. ADC is incorporated into the input to effectively capture perfusion information and generate DCE-MRI images. All the models were trained on the ProstateX dataset and tested on ProstateX and, also on Prostate-MRI to analyze their performance in the presence of data diversity. B) Comparison of various aspects between DCE-diff and other models.

create heterogeneity in the ADC data distribution across scanners. Existing methods primarily focus on DCE-MRI late-response image synthesis from early-response DCE images, relying on Gad contrast injection [8], and use unimodal structural MRI inputs for synthesis [9]. In this work, we consider the data heterogeneity by effectively utilizing the multimodal non-contrast images to harness the anatomical information from structural MRI sequences and perfusion information from ADC images to synthesize early- and late-phase DCE-MRI.

One of the approaches to generate early- and late-response DCE-MRI images is to use Generative Adversarial Networks (GAN) to learn their conditional distribution given the multimodal non-contrast inputs - T2-Weighted (T2W) MRI, Proton Density (PD), T1 pre-contrast images and ADC maps. GANs have the potential to generate high-fidelity outputs and support efficient sampling. However, training GANs can be challenging as they might lead to mode collapse if the hyper-parameters and regularizers aren't carefully selected [10].

Diffusion models, on the other hand, overcome the above disadvantages and demonstrate the ability to generate high-quality images. They possess favorable characteristics such as comprehensive distribution coverage, consistent training objective, and scalability. Employing a method of gradually reducing noise from images, their training objective can be represented as a re-weighted variational lower bound [10].

Inspired from these merits and the benefits offered in various image-to-image mapping tasks [11], we propose a diffusion model for synthesizing early and late dynamic contrast MRI images from multimodal MRI inputs. Leveraging the benefits of a common architecture and avoiding the use of task-specific losses (shown in Figure 1), our approach differs from these methods [11, 12] by incorporating multi-modal inputs (T2, PD, ADC, T1 pre-contrast) to synthesize early and late DCE images, investigating the applicability of diffusion models in many-to-many image translation problems in MRI. Our contributions are summarized as follows:

- We propose DCE-diff, an image-to-image diffusion model for generating early- and late- DCE-MRI images from multimodal non-contrast MRI images, namely T2-W, PD, ADC, and T1-pre-contrast.
- Our approach demonstrates the importance of using ADC images in the DCE MRI image synthesis process, by utilizing the perfusion information provided by the computed ADC maps.
- Extensive experiments comparing against three GAN-based approaches, a sequence-based convLSTM model, and a transformer-based image translation benchmark show that our proposed model generates early- and late-response images with notable improvement margins of (i) +0.64 dB and +0.85 dB in PSNR, +0.03 and +0.04 in SSIM, -0.01 and -0.02 in MAE, -21.87 and -22.8 in FID for ProstateX dataset and (ii)+6.8 dB and +8.7 dB in PSNR, +0.17 and +0.22 in SSIM, -0.1 and -0.11 in MAE, -52.37 and -40.4 in FID when evaluated on Prostate-MRI dataset without retraining, highlighting the robustness of our model across diverse scanner settings and imaging domains.

2. Related Work

2.1. Medical image-to-image translation

Generative adversarial networks, a class of neural network architecture, have shown great potential in performing image-to-image translation, super-resolution, and in-painting tasks since their inception. The pioneering technique Pix2Pix [13], utilizing conditional GANs, is designed to tackle image-to-image translation across diverse tasks. Similarly, CycleGAN [14] demonstrates unpaired image translation between two domains through cycle consistency loss. However, both methods have been

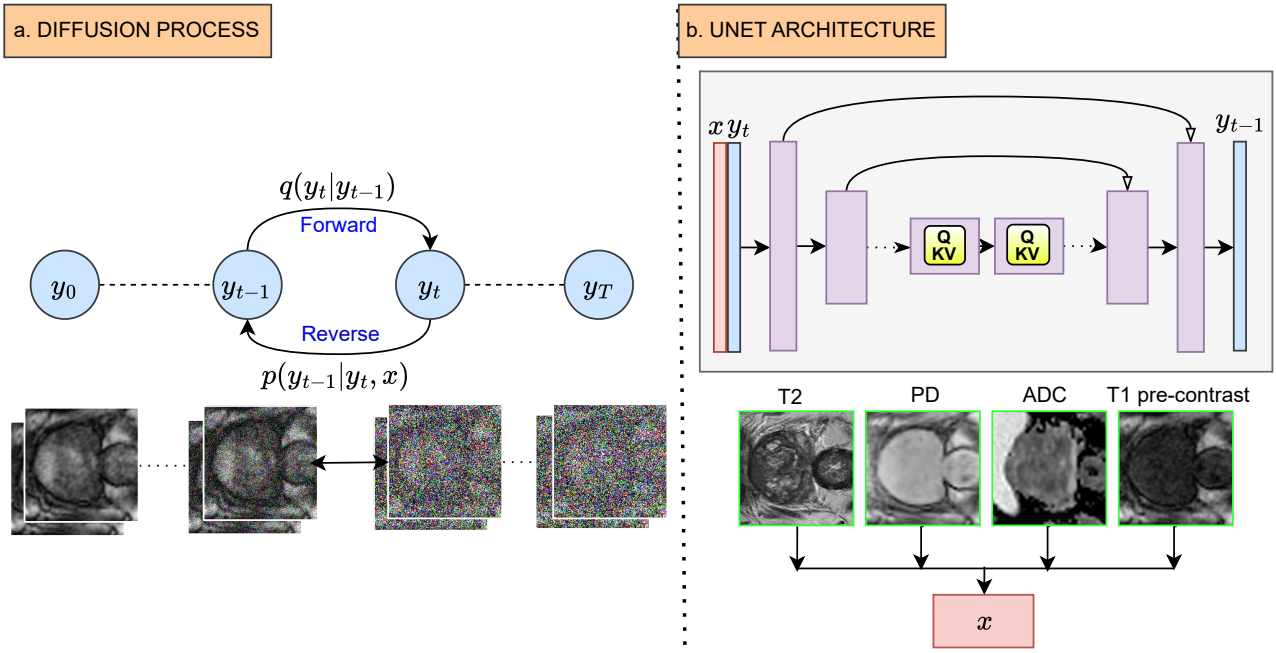


Figure 2. Diffusion Model for DCE-MRI. a) Diffusion process: The forward diffusion process q (left to right) gradually adds Gaussian noise to the target image. The reverse inference process p (right to left) iteratively denoises the target image conditioned on a source image x. $q(y_t|y_{t-1})$ is forward process whereas $p(y_{t-1}|y_t, x)$ is the denoising reverse process. y_t; $t = 0, ..., T$ are the noisy early and late response DCE images in the diffusion process. b) The denoising U-Net has skip connections and self-attention layers and is conditioned with multi-model input x (T2W, PD, ADC, T1 pre-contrast).

noted to exhibit limitations in the diversity of translated outputs, and their performance may not reach optimal levels, as shown in [15]. Specifically, generative models are gaining traction in medical image-to-image translation due to their capacity to tackle challenging medical image analysis problems such as medical image de-noising [16], reconstruction [17], segmentation [18], detection [19], and classification. MedGAN [20] demonstrates PET to CT translation, denoising, and motion correction using a combination of non-adversarial, style transfer, and perceptual loss. However, it suffers from a reliance on pixel-paired training. Based on loss-correction theory, Reg-GAN [15] employs a registration framework along with the generator to translat multi-modal images, driven by a deformable registration loss. Another recent method ResViT [9] uses a vision transformer as the backbone to generate the missing modality in structural MRI. A conditional GAN-based method in [8] synthesizes late-response from early-response in breast DCE-MRI optimized through a contrast enhancement loss. TSGAN [21] synthesizes contrast-enhanced images from pre-contrast in breast DCE-MRI with the help of a local discriminator and segmentation mask as guidance. All of the above methodologies rely on GANs, each carrying its inherent limitations stemming from the dual-network training procedure. Additionally, GANs frequently encounter training challenges such as mode collapse and struggle with generalization across diverse datasets.

2.2. Diffusion models

Recently, Diffusion models [22] have received a surge of interest and emerged with impressive results in image generation [23], super-resolution [10], and image editing [24] applications. Conditional diffusion models [25] are extended upon these by conditioning on inputs such as images, text, and audio. Diffusion models are appreciated for their mode coverage, quality of sample generation, and robustness to out-of-distribution data [26]. DDPM [27] demonstrates the effectiveness of diffusion models in high-quality image generation tasks. Conversely, UNIT-DDPM [12] proposes unpaired image translation to learn a joint distribution of domains. However, it leads to undesired artifacts and sub-optimal translation quality compared to paired approaches. The DDIM-based diffusion model in [28] performs MRI-to-CT translation volumetrically, encompassing the full three-dimensional structure of the imaging data. SynDiff [29] proposes a novel adversarial conditional diffusion model for unpaired medical datasets. However, its performance is limited due to its dependency on Cyclegan results for training the diffusion model. Besides, Palette[11] a conditional diffusion model operating in image space utilizes the

advantages of self-attention mechanisms and shows extensive results on various image restoration tasks. Nevertheless, previous approaches have primarily explored single-modality conditions to constrain the training process. In this work, we present a conditional diffusion model that leverages multimodal inputs to generate highly convincing contrast translations in DCE-MRI images.

3. Methodology

Diffusion models involve forward and reverse processes. The forward process gradually adds noise to the input data until it is transformed into pure Gaussian noise. In the reverse process, a denoising network predicts the noise at each time step. Given a noisy image $\widetilde{\boldsymbol{y}}$, the denoising model learns a reverse process that inverts the forward process. The goal is to recover the target image $\boldsymbol{y}_0$ from $\widetilde{\boldsymbol{y}}$.

$$\widetilde{\boldsymbol{y}} = \sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{1}$$

We parameterize our neural network model $f_\theta(x, \widetilde{\boldsymbol{y}}, \gamma)$ to condition on the input x, a noisy image $\widetilde{\boldsymbol{y}}$, and the current noise level γ. So, given a training output image $\boldsymbol{y}$, we generate a noisy version $\widetilde{\boldsymbol{y}}$ and train the neural network f_θ to denoise $\widetilde{\boldsymbol{y}}$ given $\boldsymbol{x}$ and a noise level indicator γ.The loss function (L_{simple})[27] is defined as follows:

$$L_{\text{simple}}(\theta) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\gamma}}\Big\| f_\theta(\boldsymbol{x}, \underbrace{\sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}}_{\widetilde{\boldsymbol{y}}}, \gamma) - \boldsymbol{\epsilon}\Big\|_p^p, \tag{2}$$

Learning involves optimizing this objective to predict the noise vector $\boldsymbol{\epsilon}$. Here $p = 2$, i.e L_2 Norm.

Image-to-image diffusion models are conditional diffusion models of the form $P(y|x)$, where both x and y are images; in our case, x is a multi-modal input images and y is the target DCE contrast image. In our approach, both early response and late response DCE MRI images of the target modality are combined and fed into the forward process of the diffusion model, as illustrated in Figure 2. During this forward process, noise is added to the target modality images. Subsequently, in the reverse process, the denoising model receives four input images (T2, PD, ADC, T1 pre-contrast) concatenated with two randomly generated noisy images. The model then predicts the noise present in the target modalities (early response and late response). During inference, sampling occurs over 1000 steps, generating the target images from the Gaussian noise given the input images and the random noise images. U-Net architecture is used for the denoising model. The U-Net design incorporates a series of residual layers and downsampling convolutions, succeeded by another series of residual layers featuring upsampling convolutions. Skip connections are employed to link layers with the same spatial dimensions. Furthermore, a global attention layer at the 16×16 resolution with a single head is integrated, along with a timestep embedding projection in each residual block. The training and inference algorithm is given in Section 8 (Appendix). Thereby, our model utilizes a multi-modal input image (x) and generates the target DCE contrast image (y) (both early and late response DCE MRI images).

4. Dataset Description and Implementation Details

We have used the ProstateX [6] dataset obtained from Siemens 3T scanner, comprising studies of 346 patients acquired without an endo-rectal coil, totaling 5520 images. Out of these, 4416 and 1104 images were used for training and testing respectively. Each patient data consists of T2W, ADC, T1 pre-contrast, PD, and DCE-MRI images. To ensure alignment across modalities, we have registered the images using the SimpleITK rigid registration framework. The prostate organ is cropped to a dimension of $160 \times 160 \times 16$.

For the evaluation of unseen data, we have utilized the Prostate-MRI [7] dataset sourced from scans acquired with a Philips 3T Achieva scanner employing an endorectal coil. This dataset comprises 26 patients, each registered and cropped to the same dimension as ProstateX. We have fixed the middle time point and the last time point sequence as the early- and late-response images in the series of DCE-MRI acquisitions.

For GAN-based methods, all models are trained for 200 epochs with a batch size of 4 and a learning rate of 0.0001. For DCE-diff, we employ the standard Adam optimizer with a consistent learning rate of 1e-4 and incorporate a linear learning rate warmup schedule spanning 10k steps. We train the diffusion model for 260k iterations. During training, we employ a linear noise scheduler ranging from 1e-6 to 0.01 over 2000 time-steps. Additionally, we utilize 1000 refinement steps with a linear scheduler ranging from 1e-4 to 0.09 during inference. All Models are implemented in PyTorch v1.12 on a 24GB RTX 3090 GPU.

Evaluation Metrics: For benchmarking and comparison with existing methods, we present several automated metrics. Specifically, we provide Fréchet Inception Distance (FID), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) for quantitative comparison.

5. Results and Discussion

5.1. Comparison analysis with other methods for DCE-MRI synthesis

To ensure a comprehensive evaluation of our model, we conducted extensive experiments with various

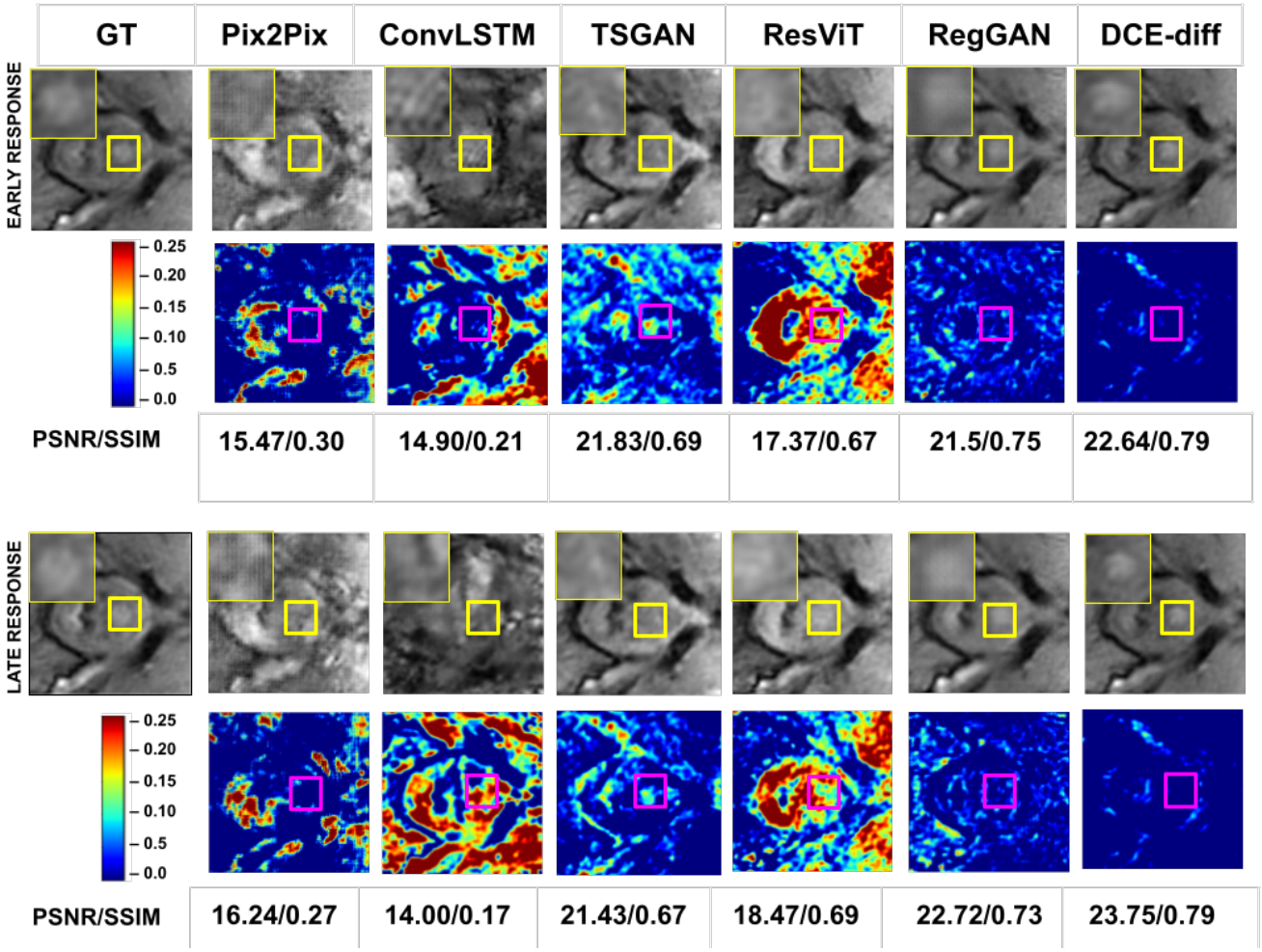


Figure 3. Visualization of the synthesized early & late DCE-MRI timepoints between DCE-diff & other models for ProstateX dataset. Row 2 and row 4 correspond to residue images between ground truth (GT) and predicted output. Note that our model able to reconstruct finer structural details and contrast uptake better than other baselines. Yellow and pink boxes represent the region of interest.

Table 1. Quantitative Comparison of the generated early- and late-response DCE-MRI images between DCE-diff and other models, for ProstateX dataset

Model	**EARLY RESPONSE**				**LATE RESPONSE**			
	PSNR↑	**SSIM↑**	**MAE↓**	**FID↓**	**PSNR↑**	**SSIM↑**	**MAE↓**	**FID↓**
ConvLSTM	14.92 ± 1.50	0.23 ± 0.04	0.13	118.70	15.27 ± 2.56	0.23 ± 0.06	0.13	115.48
Pix2Pix	15.21 ± 5.49	0.28 ± 0.18	0.11	65.86	15.29 ± 1.36	0.25 ± 0.07	0.12	32.85
RegGAN	20.56 ± 0.02	0.59 ± 0.02	0.05	23.79	20.09 ± 0.02	0.58 ± 0.02	0.06	22.61
TSGAN	21.16 ± 3.50	0.62 ± 0.10	0.06	23.75	20.46 ± 2.64	0.59 ± 0.09	0.07	24.66
ResViT	21.46 ± 0.04	0.63 ± 0.04	0.06	32.46	20.88 ± 0.04	0.62 ± 0.05	0.06	30.06
DCE-diff(ours)	**22.10 ± 1.79**	**0.67 ± 0.05**	**0.04**	**10.59**	**21.73 ± 1.95**	**0.65 ± 0.06**	**0.05**	**7.26**

baseline methods, including GAN-based approaches such as Pix2Pix [13], RegGAN [15], and TSGAN [21], as well as transformer-based GAN (ResViT) [9] and traditional ConvLSTM [30] models. Our experimental setup includes (i) a comparative analysis against baseline GAN methods for contrast translation, (ii) a comparative assessment of the proposed model and other baseline methods on a different dataset, and (iii) an analysis of the significance of ADC images in generating the early- and late-DCE-MRI images. The quantitative comparison of our proposed approach with other baseline methods is presented in Table 1. Our findings reveal the following observations: Our model

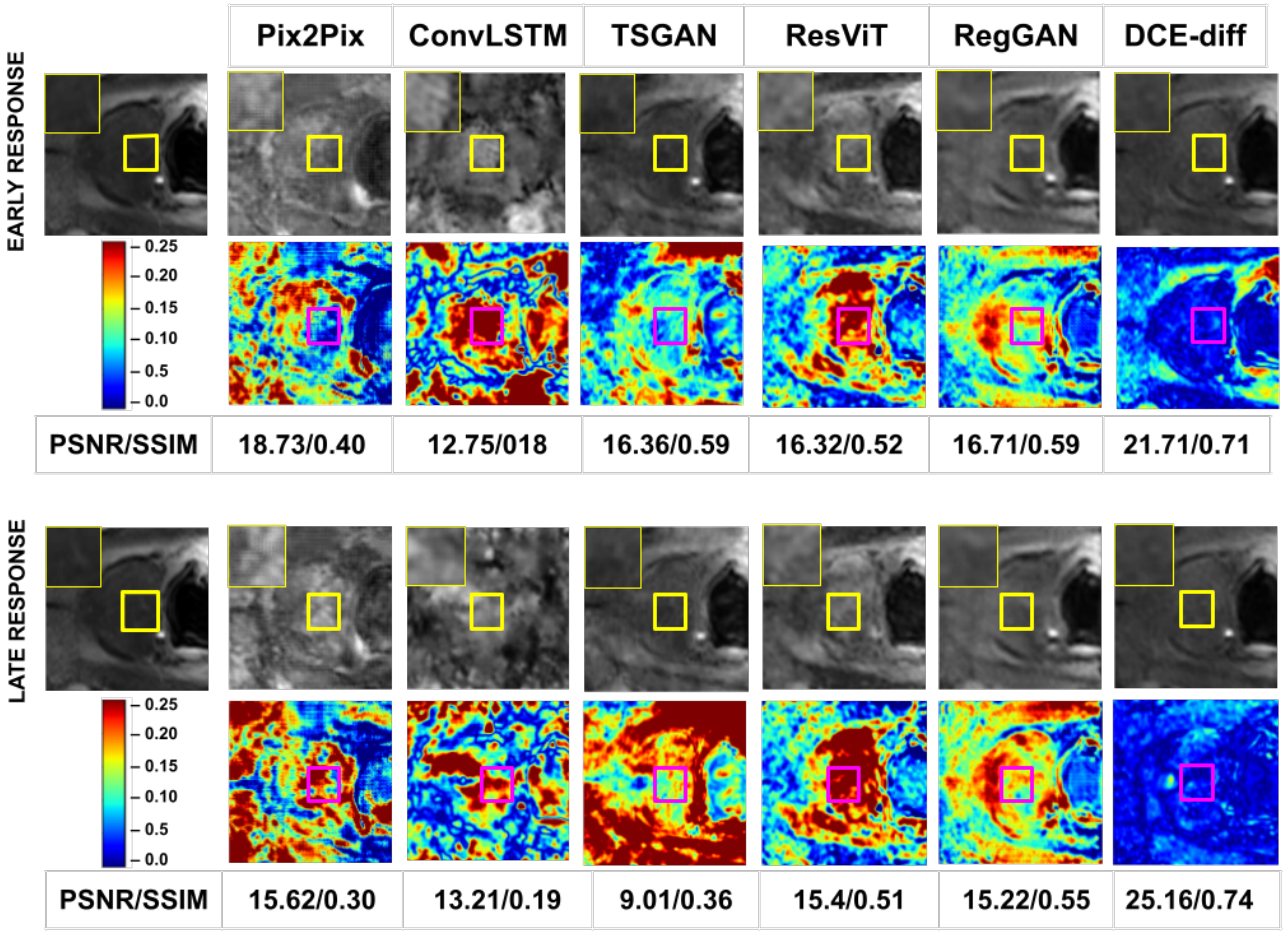


Figure 4. Visualization of the synthesized early & late DCE-MRI timepoints between DCE-diff & other models for Prostate-MRI dataset. Our model excels in capturing temporal contrast enhancement patterns, even when presented with unseen data. Yellow and pink boxes represent the region of interest.

consistently outperforms all other baseline methods across all evaluation metrics for early- and late-response images.

Quantitative results reveal that our model surpasses the second best-performing baseline model (ResViT [9]) by an improvement margin of -21.87 in FID score, +0.64 dB in PSNR, +0.04 in SSIM, and -0.02 in MAE for early-response and -22.8 in FID score, +0.85 dB in PSNR, +0.036 in SSIM and -0.017 in MAE metrics for late-response synthesis. This superiority of performance is further illustrated in the visual results depicted in Figure 3. The residue images, where darker hues indicate lesser error, in rows 2 and 4 represent the difference between the ground truth and the predicted image. Our model DCE-diff, exhibits the least residual error, especially in the region of interest (bounded by a pink box) than other models. The generated image from our model effectively retains structural information and hyper-intensity patterns of the contrast-enhanced image compared to other baseline models, demonstrating superior performance for both early and late responses. We note that conditioning the diffusion process using anatomical MRI sequence together with the perfusion MRI images, namely ADC, controls the generation process by providing additional information with desired attributes or characteristics.

5.2. Evaluation on Deviated Data Domain

Prostate-MRI shows a domain shift from ProstateX in terms of the perfusion information given by the different b-values of the corresponding Diffusion-weighted images (refer Section 4). Table 2 shows the performance of various models evaluated on the Prostate-MRI dataset. From the table, we observe that the proposed diffusion model demonstrates superior adaptation capabilities compared to other models. The results reveal that our model surpasses the second best-performing baseline model (RegGAN [15]) by an improvement margin of -52.37 in FID score, +6.89 dB in PSNR,+0.1 in SSIM, and -0.1 in MAE for early-response and -40.4 in FID score, +8.78 in PSNR, +0.22 in SSIM, and -0.11 in MAE for late-response. The visual results

Table 2. Quantitative Comparison of the generated early- and late-response DCE-MRI images between DCE-diff and other models, for Prostate-MRI dataset. Note that the models are trained on the ProstateX dataset and evaluated on the Prostate-MRI dataset.

Model	EARLY RESPONSE				LATE RESPONSE			
	PSNR↑	SSIM↑	MAE↓	FID↓	PSNR↑	SSIM↑	MAE ↓	FID↓
ConvLSTM	9.31 ± 2.7	0.14 ± 0.05	0.19	90.51	11.71 ± 2.34	0.18 ± 0.05	0.14	94.66
TSGAN	10.74 ± 3.0	0.32 ± 0.11	0.16	104.40	7.62 ± 2.27	0.22 ± 0.07	0.24	93.85
Pix2Pix	11.41 ± 3.31	0.19 ± 0.07	0.15	119.42	10.03 ± 3.58	0.15 ± 0.06	0.17	136.39
ResViT	12.99 ± 1.68	0.34 ± 0.08	0.19	94.99	14.54 ± 2.67	0.42 ± 0.12	0.14	82.21
RegGAN	14.96 ± 1.73	0.43 ± 0.08	0.14	84.79	14.54 ± 1.68	0.42 ± 0.08	0.14	66.29
DCE-diff(ours)	**21.79 ± 2.47**	**0.60 ± 0.08**	**0.04**	**32.425**	**23.32 ± 2.58**	**0.64 ± 0.08**	**0.03**	**25.83**

Table 3. Ablative study on the importance of ADC

DCE Response	ADC comparison	PSNR	SSIM	MAE
Early Response	w/o ADC	21.64	0.65	0.05
	with ADC	**22.09**	**0.67**	**0.04**
Late Response	w/o ADC	21.15	0.64	0.06
	with ADC	**21.71**	**0.65**	**0.05**

illustrated in Figure 4, show that the model can synthesize DCE-MRI images under drifts in the perfusion information.

A key factor contributing to this robustness is the utilization of conditional inputs in the diffusion model architecture indicating its ability to effectively learn complementary and reusable features from multimodal conditioning information. The conditional inputs provide additional contextual information related to perfusion using the ADC images. Furthermore, the diffusion model's unique training objective, which gradually reduces noise from images, enhances its capability to capture underlying patterns and features. This, coupled with its comprehensive distribution coverage and consistent training objective, enables more efficient adaptation to variations in data distribution. The combination of these attributes emphasizes the diffusion model's superior performance on the Prostate-MRI dataset, showcasing its effectiveness in addressing domain shift and achieving better results than other models.

5.3. Ablative study on ADC

We conduct an ablative study to assess the importance of ADC maps in enhancing DCE MRI predictions so that they contain the tissue perfusion information necessary for clinical studies. The qualitative results, shown in Figure 5, illustrate a noticeable enhancement in the prediction quality with the least residual error. Quantitative results, from Table 3 show improved PSNR and SSIM metrics with ADC as a part of the input. These findings indicate the importance of considering ADC in the learning process, where the model can learn the correlation observed between DCE-MRI and ADC images.

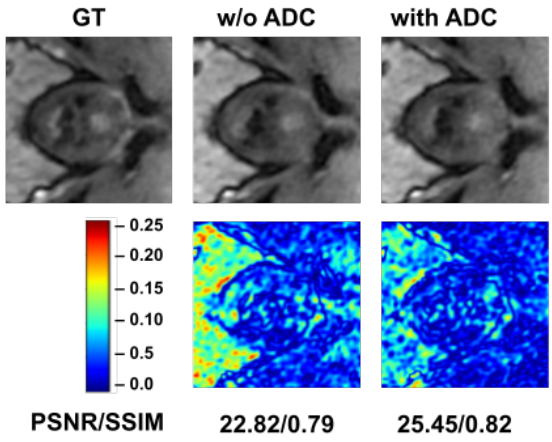


Figure 5. Qualitative results for the significance of ADC. Visual results prove that the generated images are close to the ground truth when ADC is included as a part of the input.

6. Conclusion

We propose, DCE-diff, an image-to-image diffusion model for generating early and late DCE-MRI images from multimodal non-contrast MRI images. Our experiments showcase better improvement margins in the qualitative and quantitative studies against other GAN-based, sequence-based, and transformer-based methods, better reusability of our model to cross-scanner data, and the importance of using ADC perfusion maps in the synthesis process.

The sampling strategy of diffusion models leads to longer inference time and this remains an area of concern. We are investigating these aspects and extending our work to include more clinical scenarios like application-driven synthesis.

7. Acknowledgements

The authors express their gratitude to Dr. Ramesh Venkatesan, Dr. Suresh Joel, and Mr. Harsh from GE Healthcare for their valuable discussions, which significantly enhanced the quality of the paper.

References

[1] B. J. Guo, Z. L. Yang, and L. J. Zhang, "Gadolinium deposition in brain: current scientific evidence and future perspectives," *Frontiers in molecular neuroscience*, vol. 11, p. 335, 2018.

[2] D. M. Schieda N, Krishna S, "Update on gadolinium-based contrast agent-enhanced imaging in the genitourinary system," *AJR Am J Roentgenol*, pp. 1223–1233, 2019.

[3] K. S. Choi, S.-H. You, Y. Han, J. C. Ye, B. Jeong, and S. H. Choi, "Improving the reliability of pharmacokinetic parameters at dynamic contrast-enhanced mri in astrocytomas: a deep learning approach," *Radiology*, vol. 297, no. 1, pp. 178–188, 2020.

[4] R. Osuala, S. Joshi, A. Tsirikoglou, L. Garrucho, W. H. Pinaya, O. Diaz, and K. Lekadir, "Pre-to post-contrast breast mri synthesis for enhanced tumour segmentation," *arXiv preprint arXiv:2311.10879*, 2023.

[5] R. Kushol, C. C. Luk, A. Dey, M. Benatar, H. Briemberg, A. Dionne, N. Dupré, R. Frayne, A. Genge, S. Gibson, S. J. Graham, L. Korngut, P. Seres, R. C. Welsh, A. H. Wilman, L. Zinman, S. Kalra, and Y.-H. Yang, "Sf2former: Amyotrophic lateral sclerosis identification from multi-center mri data using spatial and frequency fusion transformer," *Computerized Medical Imaging and Graphics*, vol. 108, p. 102279, 2023.

[6] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "Spie-aapm prostatex challenge data (version 2) [dataset]." The Cancer Imaging Archive, 2017.

[7] P. Choyke, B. Turkbey, P. Pinto, M. Merino, and B. Wood, "Data from prostate-mri." The Cancer Imaging Archive, 2016.

[8] R. D. Fonnegra, M. L. Hernandez, J. C. Caicedo, and G. M. Díaz, "Early-to-late prediction of dce-mri contrast-enhanced images in using generative adversarial networks," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2023.

[9] O. Dalmaz, M. Yurt, and T. Çukur, "Resvit: residual vision transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[10] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[11] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

[12] H. Sasaki, C. G. Willcocks, and T. P. Breckon, "Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models," *arXiv preprint arXiv:2104.05358*, 2021.

[13] X. Wang, H. Yan, C. Huo, J. Yu, and C. Pant, "Enhancing pix2pix for remote sensing image classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2332–2336, IEEE, 2018.

[14] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 701–710, 2018.

[15] L. Kong, C. Lian, D. Huang, Y. Hu, Q. Zhou, *et al.*, "Breaking the dilemma of medical image-to-image translation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1964–1978, 2021.

[16] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[17] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing mri reconstruction using a generative adversarial network with a cyclic loss," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1488–1497, 2018.

[18] M. Rezaei, K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. Yang, and C. Meinel, "A conditional adversarial network for semantic segmentation of brain tumor," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pp. 241–252, Springer, 2018.

[19] V. Alex, M. S. KP, S. S. Chennamsetty, and G. Krishnamurthi, "Generative adversarial networks for brain lesion detection," in *Medical Imaging 2017: Image Processing*, vol. 10133, pp. 113–121, SPIE, 2017.

[20] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," *Computerized medical imaging and graphics*, vol. 79, p. 101684, 2020.

[21] E. Kim, H.-H. Cho, J. Kwon, Y.-T. Oh, E. S. Ko, and H. Park, "Tumor-attentive segmentation-guided gan for synthesizing breast contrast-enhanced mri without contrast agents," *IEEE journal of translational engineering in health and medicine*, vol. 11, pp. 32–43, 2022.

[22] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, pp. 2256–2265, PMLR, 2015.

[23] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, p. 102846, 2023.

[24] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6007–6017, 2023.

[25] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[26] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, p. 102846, 2023.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[28] R. Graf, J. Schmitt, S. Schlaeger, H. K. Möller, V. Sideri-Lampretsa, A. Sekuboyina, S. M. Krieg, B. Wiestler, B. Menze, D. Rueckert, *et al.*, "Denoising diffusion-based mri to ct image translation enables automated spinal segmentation," *European Radiology Experimental*, vol. 7, no. 1, p. 70, 2023.

[29] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, Ş. Özturk, A. Güngör, and T. Çukur, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Transactions on Medical Imaging*, 2023.

[30] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[31] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

DCE-diff: Diffusion Model for Synthesis of Early and Late Dynamic Contrast-Enhanced MR Images from Non-Contrast Multimodal Inputs

Supplementary Material

8. Appendix

From [31], we adapt the diffusion model training and inference.

The forward diffusion process is a Markovian process that adds noise to the image $\boldsymbol{y}_0 \equiv \boldsymbol{y}$ over T iterations. At a time step t, the addition of noise is given by:

$$q\left(\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_t\right) = \mathcal{N}\left(\boldsymbol{y}_{t-1}; \sqrt{\alpha_t}\boldsymbol{y}_{t-1}, (1-\alpha_t)\, I\right) \quad (3)$$

$$q\left(\boldsymbol{y}_{1:T} \mid \boldsymbol{y}_0\right) = \prod_{t=1}^{T} q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}\right) \quad (4)$$

where α_t are noise schedule hyper-parameters. At $t = T, \boldsymbol{y}_T$ is Gaussian Noise. The forward process can be marginalizable at each step and is given by

$$q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_0\right) = \mathcal{N}\left(\boldsymbol{y}_t; \sqrt{\gamma_t}\boldsymbol{y}_0, (1-\gamma_t)\, I\right) \quad (5)$$

where $\gamma_t = \prod_{t'}^{t} \alpha_t'$.

$$q\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_0, \boldsymbol{y}_t\right) = \mathcal{N}\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}\right) \quad (6)$$

where $\boldsymbol{\mu} = \frac{\sqrt{\gamma_{t-1}}(1-\alpha_t)}{1-\gamma_t}\boldsymbol{y}_0 + \frac{\sqrt{\alpha_t}(1-\gamma_{t-1})}{1-\gamma_t}\boldsymbol{y}_t$ and $\sigma^2 = \frac{(1-\gamma_{t-1})(1-\alpha_t)}{1-\gamma_t}$.

During reverse Process:

$$\widetilde{\boldsymbol{y}} = \sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \quad (7)$$

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{\boldsymbol{\epsilon},\gamma}\|f_\theta(\boldsymbol{x}, \underbrace{\sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}}_{\widetilde{\boldsymbol{y}}}, \gamma) - \boldsymbol{\epsilon}\|_p^p \quad (8)$$

Inference: The model performs inference via the learned reverse process. Since the forward process is constructed so the prior distribution $\mathrm{p}\left(y_T\right)$ approximates a standard normal distribution $\mathcal{N}\left(\boldsymbol{y}_T \mid \boldsymbol{0}, \boldsymbol{I}\right)$, the sampling process can start at pure Gaussian noise, followed by T steps of iterative refinement.

The neural network model f_θ is trained to estimate $\boldsymbol{\epsilon}$, given any noisy image $\widetilde{\boldsymbol{y}}$, and $\boldsymbol{y}_t$. Thus, given $\boldsymbol{y}_t$, we approximate $\boldsymbol{y}_0$ as

$$\hat{\boldsymbol{y}}_0 = \frac{1}{\sqrt{\gamma_t}}\left(\boldsymbol{y}_t - \sqrt{1-\gamma_t} f_\theta\left(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t\right)\right) \quad (9)$$

Substitute the estimate $\hat{\boldsymbol{y}}_0$ into the posterior distribution of $q\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_0, \boldsymbol{y}_t\right)$ to parameterize the mean of $p_\theta\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{x}\right)$ as

Algorithm 1 Training a denoising model f_θ

repeat
$(\boldsymbol{x}, \boldsymbol{y}_0) \sim p(\boldsymbol{x}, \boldsymbol{y})$
$\gamma \sim p(\gamma)$
$\epsilon \sim \mathcal{N}(0, \mathrm{I})$
Take a gradient descent step on
$\nabla_\theta \left\| f_\theta\left(\boldsymbol{x}, \sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \gamma\right) - \boldsymbol{\epsilon}\right\|_p^p$
until converged

Algorithm 2 Inference in T iterative refinement steps

$y_T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ do
$z \sim \mathcal{N}(0, \mathrm{I})$ if $t > 1$, else $z = 0$
$\boldsymbol{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta\left(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t\right)\right) + \sqrt{1-\alpha_t}\boldsymbol{z}$
end for
return y_0

$$\mu_\theta\left(x, \boldsymbol{y}_t, \gamma_t\right) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta\left(x, \boldsymbol{y}_t, \gamma_t\right)\right) \quad (10)$$

The variance $p_\theta\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{x}\right)$ is set to $(1-\alpha_t)$, a default. Now, each iteration of the reverse process can be written as

$$\boldsymbol{y}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta\left(\boldsymbol{x}, \boldsymbol{y}_t, \gamma_t\right)\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \quad (11)$$

where $\epsilon_t \sim \mathcal{N}(0, I)$. This resembles one step of Langevin dynamics for which f_θ provides an estimate of the gradient of the data log density.