Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Parameter-Efficient Mask-Conditioned Magnetic Resonance Image Synthesis for Segmentation Augmentation

Xin Tong

tongxin@tjcu.edu.cn

Tianjin University of Commerce

Jianye An

Tianjin University of Commerce

Chunhui Liang

Tianjin University of Commerce

Dongfang Chen

Tianjin University of Commerce

Research Article

Keywords: magnetic resonance image synthesis, medical image segmentation, synthetic data augmentation, diffusion models, mask-conditioned generation, low-rank adaptation

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11479149/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Parameter-Efficient Mask-Conditioned Magnetic Resonance Image Synthesis for Segmentation Augmentation

Xin Tong*, Jianye An, Chunhui Liang, and Dongfang Chen

School of Science, Tianjin University of Commerce, 409 Guangrong Road, Tianjin 300134, China

Emails: tongxin@tjcu.edu.cn; lxyajy@tjcu.edu.cn; lxylch@tjcu.edu.cn; chendongfang@tjcu.edu.cn

*Corresponding author: Xin Tong (tongxin@tjcu.edu.cn)

Abstract

When real training images for magnetic resonance imaging (MRI) segmentation are limited, existing masks can be reused to synthesize additional image appearances. However, paired image-translation methods may favor pixel fidelity over distributional similarity, while full diffusion-model fine-tuning is costly and generic conditioning does not explicitly encode mask geometry. To address these limitations, we propose ControlNet-guided Low-Rank Adaptation for MRI synthesis (CLoRA-MRI), a mask-conditioned latent diffusion framework that adapts both the denoising and spatial-conditioning pathways. Only 2.96% of the framework parameters are trainable in the BraTS configuration. We evaluate image quality and segmentation utility on BraTS2020 brain MRI and ACDC cardiac MRI across four fractions of real training slices. On BraTS T1 contrast-enhanced MRI, CLoRA-MRI achieves a lower FID than Pix2Pix (55.65 versus 162.42), although Pix2Pix has higher paired pixel fidelity. With 10% real training slices, adding synthetic pairs raises mean Dice from 0.7208 to 0.7401 and reduces HD95 from 48.43 to 45.34 pixels in a within-cohort, slice-level BraTS evaluation. On 50 held-out ACDC patients, a separately trained adapter raises mean Dice from 0.7661 to 0.8263 and reduces HD95 from 15.40 to 13.63 pixels at the same fraction. Gains diminish as real-image availability increases. These results support targeted use of parameter-efficient, mask-conditioned MRI synthesis when real-image coverage is low.

Keywords: magnetic resonance image synthesis; medical image segmentation; synthetic data augmentation; diffusion models; mask-conditioned generation; low-rank adaptation.

1 Introduction

Magnetic resonance imaging (MRI) supports brain-tumor diagnosis and treatment planning; T1 contrast-enhanced (T1ce) scans reveal enhancing tissue. Segmentation requires annotated datasets that are costly to assemble [1, 2]. When valid masks already exist, synthesis can diversify image appearance without replacing expert annotations.

Conventional augmentation perturbs existing images but cannot create new appearances conditional on a specified mask. Conditional generators can synthesize an image from that mask. Pix2Pix demonstrated paired mask-to-image translation [3], although adversarial training can be unstable and deterministic objectives may favor averaged outputs [4].

Diffusion models have substantially improved image generation quality [5, 6]. Nevertheless, adapting a large text-to-image diffusion model to a small medical dataset raises two practical difficulties. First, full fine-tuning is expensive and can overfit. Second, an image-level or textual condition does not provide the spatial precision required for lesion-controlled synthesis. ControlNet introduces a dedicated spatial-conditioning pathway [7], whereas LoRA provides parameter-efficient adaptation through trainable low-rank updates [8].

We ask whether controllable synthetic image–mask pairs improve segmentation under a limited adaptation budget. Perceptual quality alone need not translate into utility, so we evaluate

generation alongside downstream segmentation across real-image data regimes. We call this framework ControlNet-guided Low-Rank Adaptation for MRI synthesis (CLoRA-MRI).

The contributions of this work are:

- We formulate controllable MRI synthesis as a budget-constrained augmentation problem: segmentation masks specify spatial anatomy, while dual-branch LoRA adapts a pretrained latent diffusion model using 37.2M trainable parameters (2.96% of the complete model) in the rank-128 BraTS configuration.
- We evaluate complementary aspects of synthesis quality, including paired pixel fidelity, perceptual similarity, distributional distance, tumor-region structural similarity, and qualitative mask alignment. We treat tumor-region SSIM as an an auxiliary measure of local paired similarity.
- We assess downstream augmentation across multiple real-data regimes rather than treating FID as a proxy for utility. The results identify a practical utility boundary: synthetic data are most valuable when real-image coverage is low and provide little or no gain once real-data coverage is high.
- We report patient-independent downstream evaluation on ACDC cardiac MRI, showing that this adaptation-and-augmentation protocol extends beyond brain tumors when a target-domain adapter is trained independently.

2 Related Work

2.1 Medical Image Synthesis

GANs have been extensively studied for medical image synthesis. The adversarial learning framework was introduced by Goodfellow et al. [9] and subsequently adapted to many image-generation settings. Dar et al. [10] used conditional adversarial learning for multi-contrast MRI synthesis, whereas Frid-Adar et al. [11] examined synthetic-image augmentation for lesion classification. BrainGAN further illustrates the use of generative adversarial models for brain MRI generation [12]. Pix2Pix learns paired image-to-image mappings with adversarial and reconstruction losses [3], whereas CycleGAN supports unpaired translation through cycle consistency [13]. Despite their utility, GANs can suffer from training instability, mode collapse, and limited output diversity [4].

2.2 Diffusion Models for Medical Imaging

Denoising diffusion probabilistic models learn to reverse a gradual noise process [5]. Latent diffusion reduces the computational burden by performing denoising in a compressed latent space [6]. Wolleb et al. [14] demonstrated diffusion-based medical anomaly synthesis, and Pinaya et al. [15] adapted latent diffusion to brain imaging. Khader et al. [16] studied latent-space diffusion for 3D medical image generation. CoLa-Diff uses conditional latent diffusion for multi-modal MRI synthesis [17]. SegGuidedDiff explicitly follows multi-class segmentation masks during sampling [18]. More recent MRI-specific methods include MU-Diff for brain-lesion synthesis [19] and a slice-based latent-diffusion pipeline for multi-modal MRI and mask generation [20]. These studies motivate controllable synthesis, but differ in conditioning, modalities, and training protocol from the low-budget setting addressed here.

2.3 Spatial Conditioning and Parameter-Efficient Adaptation

ControlNet augments a pretrained diffusion backbone with a trainable copy of the encoder that receives an additional spatial condition [7]. Zero-initialized convolutions allow conditioning

residuals to be introduced without immediately destroying the pretrained representation. LoRA constrains a weight update to a low-rank product, reducing the number of trainable parameters [8]. Dutt et al. [21] identified parameter-efficient fine-tuning as an important opportunity in medical image analysis, and Chambon et al. [22] examined the adaptation of pretrained foundation models to medical domains. SeLoRA extends low-rank adaptation specifically for medical image synthesis [23]. The present work uses both ControlNet and LoRA: ControlNet communicates tumor-mask geometry, while LoRA adapts the UNet and ControlNet attention layers to MRI appearance.

2.4 Synthetic Data for Segmentation

The practical value of generated medical images should ultimately be assessed through downstream tasks. Akbar et al. [24] compared GANs and diffusion models for brain-tumor segmentation with synthetic MRI. Al Khalil et al. [25] evaluated synthetic data for robust cardiac-MRI segmentation, whereas a later study used conditional synthesis and Fourier-domain adaptation for multi-modal brain-tumor segmentation [26]. Qiu et al. [27] proposed a noise-consistent diffusion method that jointly addresses synthesis and segmentation. Kebaili et al. [20] evaluated downstream tumor segmentation using synthetic multi-modal MRI. Goceri [28] reviewed the broader role of medical-image augmentation and the need to evaluate it against task-specific objectives. Recent work in *Pattern Recognition* has examined related low-annotation settings from complementary directions. Chen et al. [29] use multi-scale diffusion features and image-level consistency for semi-supervised medical segmentation; unlike that feature-guidance approach, our method synthesizes paired images for existing masks. Wan et al. [30] generate model-guided hard samples for downstream medical segmentation, whereas our generator is conditioned directly on anatomical masks and adapted through a restricted parameter budget. The effect of synthetic data is not uniform: it can help when real samples are scarce, but gains may disappear or reverse when generated images add little information beyond a sufficiently large real dataset. Our experiments explicitly study this behavior across several data fractions.

From a pattern-recognition perspective, mask-conditioned MRI synthesis couples conditional distribution learning, anatomical constraint preservation, and label-preserving data augmentation. The central question is therefore not image realism alone, but whether a resource-efficient conditional generator preserves discriminative structures well enough to improve recognition under limited labeled data. This framing motivates our joint evaluation of global image statistics, condition alignment, and downstream segmentation utility.

3 Method

3.1 Overview

Given a multi-region segmentation mask $\mathbf{c}$, CLoRA-MRI generates a corresponding MRI slice $\hat{\mathbf{x}}$. The primary BraTS instantiation produces T1ce brain MRI, while the ACDC instantiation produces cardiac cine MRI using a separately trained adapter. The framework consists of a frozen variational autoencoder (VAE), a pretrained Stable Diffusion v1.5 UNet, a ControlNet branch, and LoRA modules inserted into the attention projections of both the UNet and ControlNet. Generated image–mask pairs can subsequently be concatenated with real labeled data to train a downstream segmentation network.

3.2 Stable Diffusion Components

Stable Diffusion v1.5 performs denoising in a compressed latent space rather than directly in pixel space. For a 128×128 input, the diffusion state has dimensions $16 \times 16 \times 4$, substantially reducing memory use. The pretrained backbone contains a VAE, a conditional UNet, and a

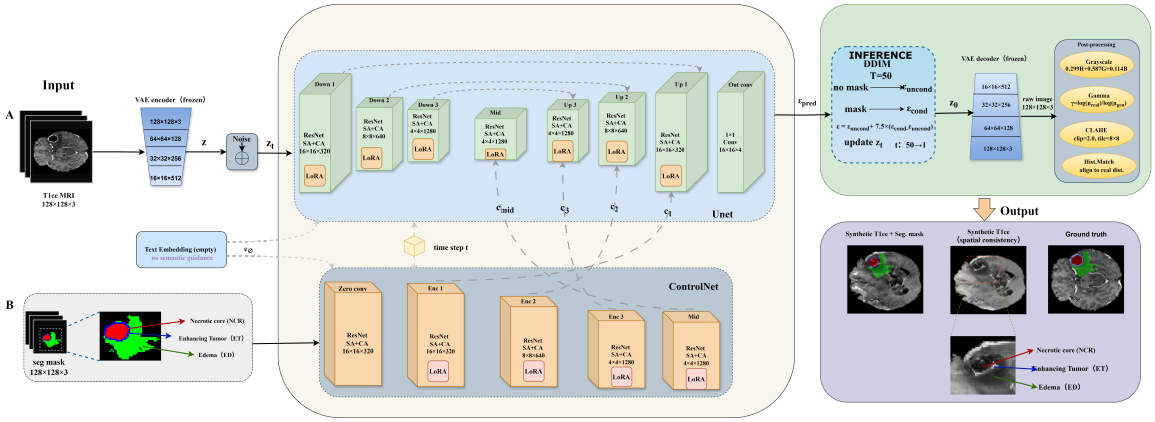


Figure 1: Overview of CLoRA-MRI. A BraTS-style multi-region mask is supplied to ControlNet as a spatial condition. LoRA modules adapt the attention layers of both the diffusion UNet and ControlNet. The resulting image–mask pairs can be used as additional training samples for a downstream segmentation model.

CLIP text encoder. The VAE, text encoder, and base UNet weights remain frozen throughout training.

The natural-image pretraining supplies generic image priors, while LoRA redirects the attention layers toward MRI appearance. Because no semantic text description is required, the text encoder receives an empty prompt and produces a fixed embedding $\mathbf{e}_\varnothing \in \mathbb{R}^{77\times768}$. Consequently, the segmentation mask is the only spatially informative condition.

3.3 Latent Diffusion Backbone

The VAE encoder $\mathcal{E}$ maps a real image $\mathbf{x}$ to a latent representation

$$\mathbf{z}_0 = \kappa\,\mathcal{E}(\mathbf{x}), \tag{1}$$

where κ denotes the fixed VAE latent scaling factor. At diffusion step t, Gaussian noise $\boldsymbol{\epsilon}$ is added:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}. \tag{2}$$

The denoising network predicts the sampled noise from $\mathbf{z}_t$, t, and the spatial condition.

For Stable Diffusion v1.5, $\kappa = 0.18215$. After reverse diffusion, the VAE decoder reconstructs

$$\hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}}_0/\kappa)\,. \tag{3}$$

The UNet alternates residual blocks and spatial transformers. Its self-attention layers model dependencies within latent feature maps, whereas cross-attention receives the fixed empty-prompt embedding. Mask information is supplied separately through ControlNet.

3.4 ControlNet Spatial Conditioning

The segmentation condition is represented as a three-channel color-coded mask. For BraTS, the channels encode edema, necrotic/non-enhancing tumor core, and enhancing tumor; for ACDC, the same interface encodes RV, Myo, and LV. ControlNet processes this mask and produces residual features at multiple resolutions. If $\mathbf{h}_l$ is the frozen UNet feature at level l and $\mathbf{r}_l(\mathbf{c})$ is the corresponding ControlNet residual, the conditioned feature is

$$\tilde{\mathbf{h}}_l = \mathbf{h}_l + \mathcal{Z}_l(\mathbf{r}_l(\mathbf{c}))\,, \tag{4}$$

where $\mathcal{Z}_l$ is a zero-initialized 1×1 convolution.

The ControlNet encoder mirrors the frozen UNet encoder and provides residuals at 16×16, 8×8, and 4×4, together with a middle-block residual. Their channel dimensions are 320, 640, 1280, and 1280. Fine-resolution features communicate lesion position and local shape, while coarse features affect global context. Because the residual projections are initialized to zero, training begins from the behavior of the pretrained model rather than an abruptly perturbed denoiser.

3.5 Low-Rank Adaptation

For a frozen weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, LoRA parameterizes its update as

$$\mathbf{W} = \mathbf{W}_0 + \frac{\alpha_{\mathrm{L}}}{r}\mathbf{B}\mathbf{A}, \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$ are trainable, r is the LoRA rank, and α_{L} is the LoRA scaling coefficient. LoRA is applied to the query, key, value, and output projections of all attention blocks in both branches. At $r = 128$, the UNet contains 25.5M trainable LoRA parameters and ControlNet contains 11.7M, totaling 37.2M.

The low-rank matrices are initialized asymmetrically: $\mathbf{A}$ is randomly initialized and $\mathbf{B}$ is initialized to zero, so the initial update is exactly zero. LoRA is inserted into the `to_q`, `to_k`, `to_v`, and `to_out.0` projections of each spatial transformer. Applying LoRA to both branches permits domain adaptation in the primary denoising pathway and the mask-conditioning pathway.

3.6 Training Objective and Inference

The model minimizes the standard diffusion noise-prediction loss:

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{\mathbf{z}_0,t,\boldsymbol{\epsilon},\mathbf{c}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right]. \tag{6}$$

During BraTS inference, DDIM sampling [31] is used with 50 denoising steps; ACDC synthesis uses 20 steps. Classifier-free guidance uses guidance scale $g = 7.5$; this symbol is intentionally distinguished from the VAE scaling factor κ.

At each reverse step, the conditional and unconditional predictions are combined as

$$\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\epsilon}}_{\mathrm{uncond}} + g\,(\hat{\boldsymbol{\epsilon}}_{\mathrm{cond}} - \hat{\boldsymbol{\epsilon}}_{\mathrm{uncond}}). \tag{7}$$

The unconditional branch receives no mask. A larger g strengthens the condition but may reduce diversity or amplify artifacts; both domain-specific synthesis pipelines use $g = 7.5$.

3.7 Reference-Assisted Intensity Analysis

We additionally examine grayscale conversion, automatic gamma correction, CLAHE, and histogram matching as a reference-assisted analysis of intensity bias. Gamma estimation and histogram matching use the paired real image; all downstream augmentation experiments use raw generated images.

For completeness, the analysis consists of four stages. The decoder output is first converted to grayscale:

$$I_{\mathrm{gray}} = 0.299R + 0.587G + 0.114B. \tag{8}$$

A gamma value is then estimated from foreground mean intensities:

$$\gamma = \mathrm{clip}\left(\frac{\log(\mu_{\mathrm{real}} + \epsilon)}{\log(\mu_{\mathrm{gen}} + \epsilon)},\, 0.5,\, 3.0\right). \tag{9}$$

The corrected image $I_{\mathrm{gray}}^{\gamma}$ is processed using CLAHE with clip limit 2.0 and an 8×8 tile grid, followed by histogram matching to the paired reference. This analysis therefore remains separate from the reference-free synthesis pipeline.

3.8 Downstream Segmentation

For downstream experiments, real and synthetic image–mask pairs are combined by natural concatenation:

$$\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{syn}}. \tag{10}$$

A standard 2D U-Net [32] with base width 32 is trained using the sum of cross-entropy and Dice losses. Synthetic images are used only in the training partition; evaluation is performed on real images.

The BraTS segmentation implementation contains three mutually exclusive foreground labels—edema, necrotic/non-enhancing core, and enhancing tumor—plus background. Its models are optimized for 30 epochs. The ACDC implementation instead uses RV, Myo, and LV as the three foreground classes and is optimized for 50 epochs with a batch size of 8. Both implementations use AdamW, an initial learning rate of 10^{-4}, cosine decay to 10^{-6}, weight decay 10^{-4}, and gradient clipping at 1.0. Within each dataset, all compared data conditions use the same network and optimization schedule.

4 Experimental Setup

4.1 BraTS2020 Synthesis

BraTS2020 contains 369 multi-institutional glioma cases with four MRI modalities [2]. We use T1ce images and their tumor annotations. For generator development, eight informative axial slices per case are selected, producing 2,952 slices at 128×128 resolution. The selection emphasizes slices with substantial tumor content and reduces redundancy, but under-represents tumor-free and low-burden slices. This design therefore targets lesion-conditioned synthesis rather than whole-volume generation.

The diffusion model is trained for 30 epochs using AdamW ($\text{lr} = 10^{-4}$, cosine decay to 10^{-7}), mixed precision, an effective batch size of 8, and gradient clipping at 1.0. All experiments are run on an NVIDIA GeForce RTX 3070 Ti Laptop GPU.

4.2 Synthesis Metrics and Baselines

We report PSNR, SSIM [33], LPIPS [34], and FID [35]. Tumor-region SSIM (T-SSIM) is SSIM restricted to the tumor mask. T-SSIM measures local paired similarity but does not directly establish segmentation-mask faithfulness.

Baselines include the unadapted Stable Diffusion backbone, a LoRA-adapted model without ControlNet, and Pix2Pix. These comparisons isolate spatial conditioning and contrast diffusion synthesis with a paired GAN under the same experimental pipeline.

4.3 BraTS Downstream Protocol

The available BraTS downstream results were obtained by collecting 23,352 tumor-containing slices and randomly assigning 80% to training and 20% to validation. Synthetic images were generated from the maximum-tumor slice of all 369 cases, with ten random seeds per mask, giving 3,690 samples. Experiments were repeated with three initialization seeds.

This is a slice-level, in-cohort protocol: slices from one patient may occur in both partitions, and synthetic masks are drawn from all 369 cases. Accordingly, the BraTS results describe in-cohort segmentation; patient-independent testing is provided separately on ACDC.

4.4 ACDC Patient-Independent Protocol

ACDC contains cardiac cine MRI from 150 patients with annotations of the right ventricle (RV), myocardium (Myo), and left ventricle (LV) [36]. Patients 001–100 are used for model

development. Their foreground-containing ED and ES slices are pooled and assigned to a fixed 80/20 internal training/validation partition using a slice-level random split. The 10%, 25%, 50%, and 100% conditions retain the corresponding prefixes of the fixed, shuffled real training-slice index list; the percentages therefore denote *real-training-slice fractions*, not patient fractions.

An ACDC-specific rank-64 LoRA adapter is trained for 50 epochs from the same pretrained Stable Diffusion and ControlNet backbones. The implementation does not initialize this adapter from the BraTS LoRA checkpoint; accordingly, the ACDC study evaluates target-domain adaptation of the framework rather than direct BraTS-to-ACDC weight transfer. For synthesis, the foreground-maximal slice is selected from each ED and ES frame of every development patient, yielding 200 conditioning masks. Five stochastic samples are generated per mask, forming a fixed set of 1,000 synthetic image–mask pairs. This complete synthetic set is concatenated with each real-slice subset.

Patients 101–150 form a held-out external test set and are not used for generator adaptation, segmentation training, or model selection. The fixed synthetic set uses masks from all 100 development patients, so the real-data fractions change image availability while retaining the full development-mask pool. The internal split is slice-level and may include different slices of one patient on both sides; patients 101–150 remain fully independent for external testing.

4.5 Evaluation Measures

Synthesis is evaluated from complementary perspectives. PSNR measures paired pixel error, while SSIM evaluates luminance, contrast, and structural agreement [33]. LPIPS compares deep perceptual features [34]. FID compares the mean and covariance of Inception features from real and generated image sets [35]. Because Inception was trained on natural images and the sample set is relatively small, FID is treated as a comparative distributional indicator rather than a clinical-quality score.

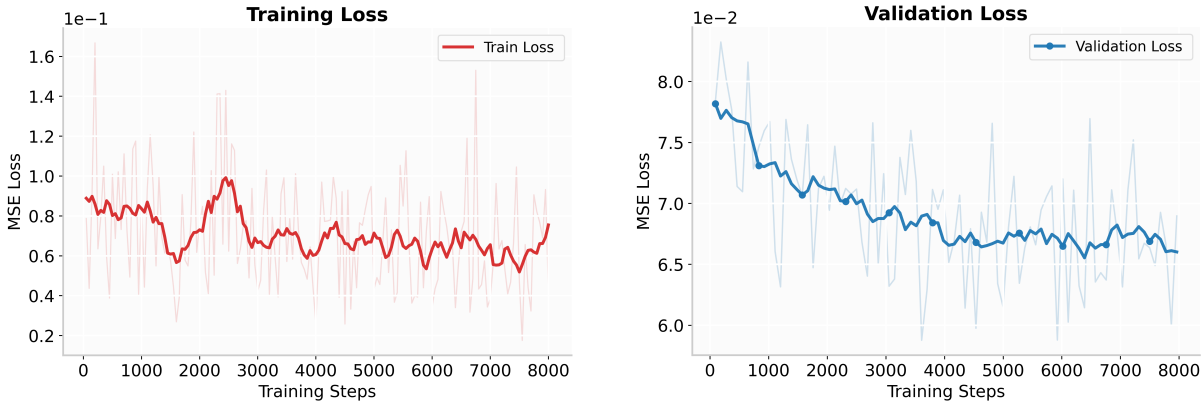
T-SSIM is computed within the tumor mask. It answers whether the generated and paired real images have similar local structure in the specified region. Spatial condition adherence is additionally examined through the ControlNet ablation and qualitative mask overlays.

Downstream segmentation is evaluated using foreground Dice and the 95th-percentile Hausdorff distance (HD95). Dice measures region overlap; HD95 is more sensitive to boundary outliers. In the ACDC implementation, class-wise Dice is computed after pooling pixels and slices within each mini-batch, followed by averaging over the three foreground classes and then over mini-batches. HD95 is computed in 2D for each slice and foreground class; when either the prediction or reference region is empty, a value of 128 pixels is assigned. These slice/class values are subsequently averaged over classes, mini-batches, and the test loader. Thus, the reported ACDC values are slice-wise implementation metrics on patient-independent test data, not patient-level volumetric Dice or 3D HD95. All reported standard deviations are computed across three training runs. Given the small number of runs and the non-identical 10% BraTS baseline seed set, we emphasize effect sizes and consistency rather than formal significance claims.

5 Results

5.1 Generator Training Convergence

Figure 2 shows generator optimization over 30 epochs. The training objective fluctuates because every iteration samples a random diffusion timestep and noise realization, but its smoothed trend decreases. Validation loss falls from approximately 7.8×10^{-2} to 6.6×10^{-2} and does not diverge from the training curve. This behavior indicates stable optimization under the LoRA-constrained parameterization and motivates evaluation of the generated images and their downstream utility.



(a) Training loss.

(b) Validation loss.

Figure 2: Generator convergence for the rank-128 configuration. High-frequency variation is expected from random diffusion timesteps and noise samples.

5.2 Synthesis Quality

Table 1: Quantitative synthesis comparison on BraTS2020. The upper block reports deployable raw outputs. The lower block is a reference-assisted intensity analysis using paired real images.

Setting	Method	PSNR↑	SSIM↑	T-SSIM↑	LPIPS↓	FID↓
Raw output	SD baseline	5.68	0.063	0.945	0.691	350.75
	Without ControlNet	13.99	0.506	0.956	0.240	55.91
	CLoRA-MRI	15.43	0.511	0.966	0.202	**55.65**
	Pix2Pix	**19.05**	**0.713**	**0.980**	**0.168**	162.42
Reference-assisted	SD baseline	9.72	0.275	0.930	0.556	321.74
	Without ControlNet	15.01	0.470	0.957	0.268	77.86
	CLoRA-MRI	16.22	0.527	0.963	0.220	75.39
	Pix2Pix	**18.67**	**0.627**	**0.975**	**0.179**	135.30

Table 1 shows that Pix2Pix obtains the best paired pixel metrics, whereas CLoRA-MRI obtains a substantially lower FID. This disagreement is important: deterministic paired regression favors pixel alignment, while diffusion sampling better covers the image distribution. We therefore avoid treating any single metric as a complete measure of medical realism.

Removing ControlNet reduces PSNR from 15.43 to 13.99 and T-SSIM from 0.966 to 0.956, consistent with stronger local conditioning when ControlNet is included.

Reference-assisted processing improves PSNR and SSIM for CLoRA-MRI but worsens LPIPS and FID. Histogram matching makes individual images more similar to their paired references while altering the feature distribution. This mixed result is why post-processing is excluded from the practical pipeline.

5.3 LoRA Rank Ablation

Table 2: LoRA-rank ablation for raw generated images. Rank 128 achieves the strongest aggregate performance across the evaluated settings.

r	PSNR↑	SSIM↑	T-SSIM↑	LPIPS↓	FID↓
4	12.61	0.274	0.959	0.288	99.59
8	12.95	0.291	0.960	0.275	89.98
16	13.41	0.332	0.961	0.258	81.29
32	14.57	0.461	0.963	0.229	63.34
64	14.75	0.484	0.964	0.218	60.88
128	**15.43**	**0.511**	**0.966**	**0.202**	**55.65**

Performance improves across the evaluated range, with the largest practical trade-off occurring between ranks 64 and 128. Rank 64 may be preferable under tighter memory constraints, whereas rank 128 is used for the remaining BraTS synthesis and augmentation experiments. The separately trained ACDC instantiation uses rank 64.

Table 3: LoRA-rank ablation after reference-assisted intensity alignment using paired real images.

r	PSNR↑	SSIM↑	T-SSIM↑	LPIPS↓	FID↓
4	14.94	0.517	0.954	0.281	115.95
8	15.27	0.526	0.956	0.270	111.29
16	15.46	**0.537**	0.958	0.256	100.86
32	15.73	0.519	0.960	0.245	86.81
64	15.92	0.528	0.961	0.232	83.62
128	**16.22**	0.527	**0.963**	**0.220**	**75.39**

The post-processed results do not improve monotonically for every metric. Rank 16 has the highest post-processed SSIM, whereas rank 128 has the best PSNR, LPIPS, and FID, demonstrating that rank selection depends on the desired quality criterion.

5.4 ControlNet Ablation

The comparison with the model without ControlNet isolates the value of explicit mask conditioning. For raw outputs, adding ControlNet improves PSNR by 1.44 dB, LPIPS by 0.038, and T-SSIM by 0.010, while FID changes only from 55.91 to 55.65. The small FID difference is unsurprising because FID does not use the conditioning mask and can therefore overlook spatial mismatch. Paired and local metrics are more sensitive to whether the generated abnormality occurs near the requested region.

Because T-SSIM compares against a paired image, future work could complement this analysis with independent mask-recovery Dice and boundary-distance measurements on generated images.

5.5 Qualitative Synthesis

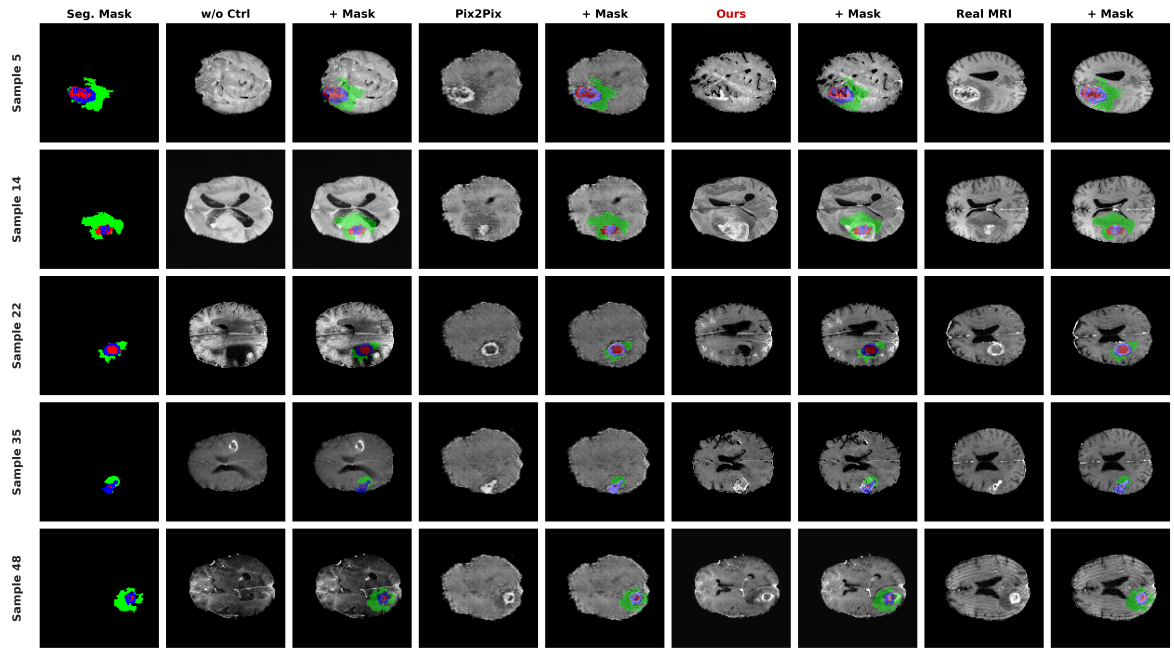


Figure 3: Representative BraTS synthesis results showing image texture, lesion appearance, and alignment with the conditioning masks.

Qualitatively, the unadapted Stable Diffusion model fails to reproduce MRI appearance. LoRA adaptation improves texture, but omitting ControlNet can weaken correspondence between the requested lesion location and generated abnormality. Pix2Pix outputs are more directly aligned with paired targets but appear smoother. CLoRA-MRI produces sharper stochastic samples while preserving the broad location of the mask condition.

The displayed examples also reveal occasional intensity overexposure, weak lesion contrast, and local mismatch between mask boundaries and generated texture. These observations motivate further improvements to intensity control and spatial consistency.

5.6 Effect of Reference-Assisted Intensity Alignment

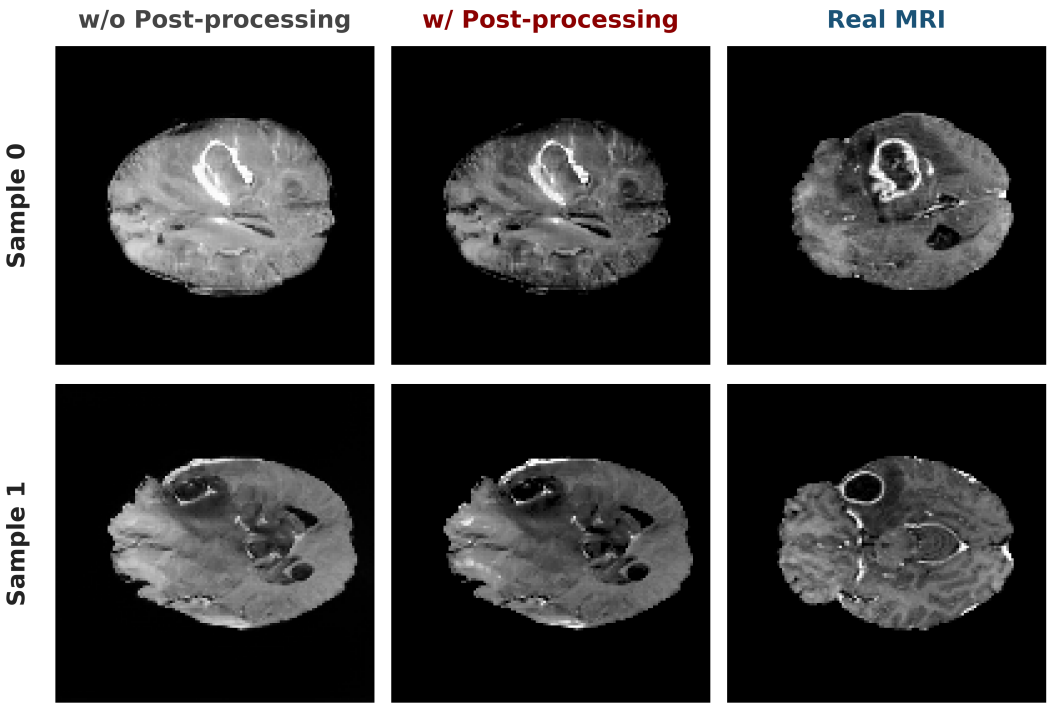


Figure 4: Raw output, reference-assisted post-processed output, and paired real MRI for two examples. The first case benefits from brightness correction, whereas the second shows mild over-correction.

Figure 4 illustrates why the intensity pipeline is not treated as a universally beneficial component. In one case it reduces a clear brightness bias and improves tissue contrast. In another, the raw image is already close to the reference intensity range and gamma correction makes the result too dark. Quantitatively, CLoRA-MRI improves from 15.43 to 16.22 dB PSNR and from 0.511 to 0.527 SSIM, while LPIPS changes from 0.202 to 0.220, FID from 55.65 to 75.39, and T-SSIM from 0.966 to 0.963. The mixed metric changes support using the raw outputs in subsequent augmentation experiments.

5.7 BraTS In-cohort Segmentation Analysis

Table 4: BraTS in-cohort slice-level segmentation results (mean ± standard deviation over three runs). HD95 is measured in pixels at 128×128.

(a) Overall segmentation performance

Ratio	Method	Mean Dice	HD95↓
10%	Real only	0.7208±0.0065	48.43±0.83
	Real + CLoRA-MRI	**0.7401±0.0013**	**45.34±0.52**
	Real + Pix2Pix	0.7384±0.0066	46.52±1.08
25%	Real only	0.8198±0.0019	40.80±0.26
	Real + CLoRA-MRI	**0.8219±0.0009**	**40.48±0.26**
	Real + Pix2Pix	0.8201±0.0016	40.86±0.14
50%	Real only	**0.8627±0.0014**	37.70±0.16
	Real + CLoRA-MRI	0.8608±0.0013	**37.60±0.12**
	Real + Pix2Pix	0.8620±0.0021	37.87±0.14
100%	Real only	**0.8941±0.0013**	35.51±0.16
	Real + CLoRA-MRI	0.8913±0.0014	**35.45±0.06**
	Real + Pix2Pix	0.8933±0.0015	35.52±0.09

(b) Region-wise Dice

Ratio	Method	Edema	N./N.E. core	Enhancing
10%	Real only	0.6660	0.7154	0.7811
	Real + CLoRA-MRI	**0.6808**	**0.7408**	0.7986
	Real + Pix2Pix	0.6761	0.7372	**0.8021**
25%	Real only	0.7715	0.8304	0.8575
	Real + CLoRA-MRI	**0.7722**	**0.8344**	**0.8589**
	Real + Pix2Pix	0.7721	0.8299	0.8582
50%	Real only	**0.8327**	**0.8728**	**0.8827**
	Real + CLoRA-MRI	0.8284	0.8726	0.8813
	Real + Pix2Pix	0.8316	0.8720	0.8824
100%	Real only	**0.8785**	**0.9017**	**0.9020**
	Real + CLoRA-MRI	0.8735	0.8996	0.9007
	Real + Pix2Pix	0.8777	0.9010	0.9012

At 10% real data, CLoRA-MRI augmentation increases mean Dice by 0.0193 and reduces HD95 by 3.09 pixels. At 25%, the Dice gain is 0.0021. At 50% and 100%, augmentation does not improve mean Dice. The pattern supports the hypothesis that synthetic data are most useful when real-data coverage is limited.

The reported regions are the mutually exclusive implementation labels: edema, necrotic/non-enhancing core, and enhancing tumor. They should not be confused with the standard BraTS composite WT, TC, and ET metrics, which would require recomputation.

5.8 Full-Data and Synthetic-Only Analysis

Table 5: BraTS full-data in-cohort analysis. Region columns follow the mutually exclusive implementation labels rather than standard BraTS composite regions.

Method	Dice	Edema	N.E. core	Enhancing	HD95↓
Real only	**0.8933**	**0.8799**	**0.9011**	**0.8989**	**35.97**
Real + CLoRA-MRI	0.8908	0.8740	0.8998	0.8986	36.00
Real + Pix2Pix	0.8925	0.8782	0.9006	0.8987	36.09
CLoRA-MRI only	0.4971	0.4359	0.4452	0.6101	63.24
Pix2Pix only	0.1847	0.1146	0.2012	0.2383	66.54

When all real slices are available, neither synthetic source improves the real-only baseline. The difference between real only and real+CLoRA-MRI is -0.0025 Dice, consistent with diminishing returns at full real-data coverage. In the synthetic-only comparison, CLoRA-MRI achieves 0.4971 Dice versus 0.1847 for Pix2Pix, suggesting that its samples retain more usable variation under this protocol.

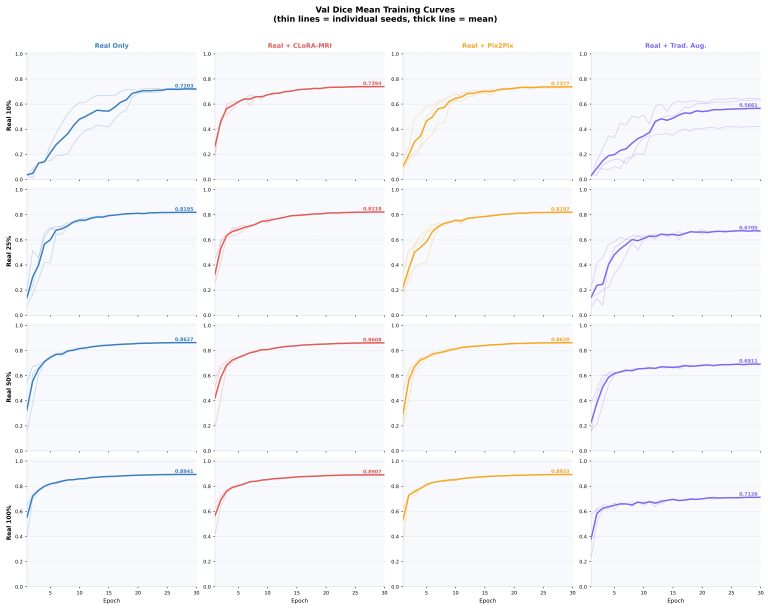
5.9 Sensitivity to Traditional Augmentation Settings

Table 6: Sensitivity analysis of the implemented flip, rotation, and elastic-deformation augmentation pipeline.

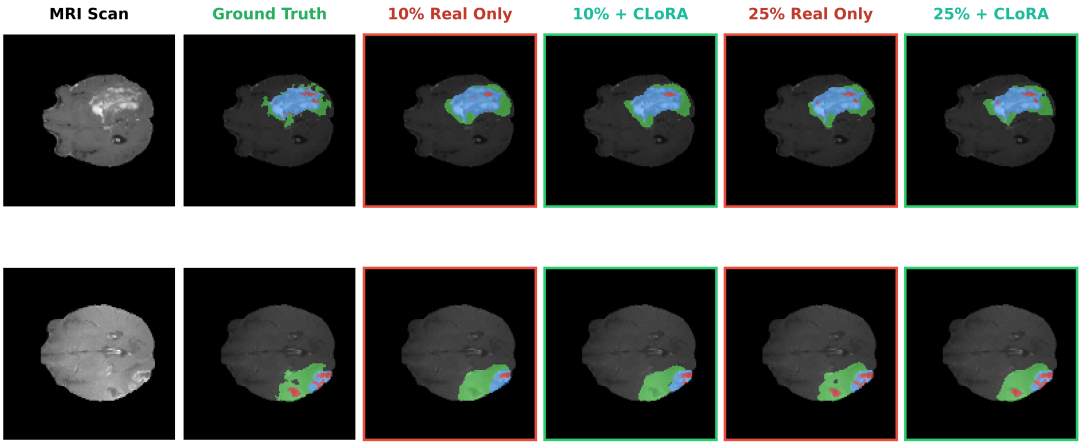
Ratio	Method	Dice	Region 2	HD95↓
10%	Real only	0.7208±0.0065	0.7154	48.43
	Traditional aug.	0.5673±0.1265	0.4160	58.23
	Real + CLoRA-MRI	**0.7401±0.0013**	**0.7408**	**45.34**
25%	Real only	0.8198±0.0019	0.8304	40.80
	Traditional aug.	0.6760±0.0052	0.6795	43.69
	Real + CLoRA-MRI	**0.8219±0.0009**	**0.8344**	**40.48**
50%	Real only	**0.8627±0.0014**	**0.8728**	37.70
	Traditional aug.	0.6925±0.0014	0.6988	40.87
	Real + CLoRA-MRI	0.8608±0.0013	0.8726	**37.60**
100%	Real only	**0.8941±0.0013**	**0.9017**	35.51
	Traditional aug.	0.7150±0.0037	0.7243	40.14
	Real + CLoRA-MRI	0.8913±0.0014	0.8996	**35.45**

The implemented flip, rotation, and elastic-deformation pipeline underperforms markedly, including a failed necrotic/non-enhancing-core run at 10%. Conventional augmentation is sensitive to interpolation, deformation strength, label handling, and sampling probability. The comparison highlights the importance of augmentation configuration when evaluating synthetic data against transformation-based baselines.

5.10 BraTS Training Dynamics and Visual Predictions



(a) Validation Dice trajectories across data ratios.



(b) Representative real-image segmentation predictions.

Figure 5: BraTS in-cohort training dynamics and representative segmentation predictions across real-data fractions.

At 10% and 25% data, the synthetic-augmentation curves rise faster during early epochs and end slightly above the real-only curves. At 50% and 100%, the trajectories converge to similar values. The qualitative examples show tighter predictions in selected low-data cases, consistent with the numerical improvement in mean Dice.

5.11 Patient-Independent ACDC Segmentation

Table 7: ACDC segmentation under different real-training-slice fractions, evaluated on held-out patients 101–150 (mean ± standard deviation over three runs). Each augmented condition uses the same 1,000 synthetic image–mask pairs. HD95 is the implemented 2D slice-wise measure in pixels at 128×128.

(a) Overall segmentation performance

Real fraction	Method	Mean Dice	HD95↓
10%	Real only	0.7661±0.0057	15.40±0.70
	Real + CLoRA-MRI	**0.8263±0.0026**	**13.63±0.70**
25%	Real only	0.8217±0.0036	12.89±0.90
	Real + CLoRA-MRI	**0.8486±0.0027**	**12.09±0.28**
50%	Real only	0.8590±0.0005	**10.56±0.10**
	Real + CLoRA-MRI	**0.8661±0.0024**	10.61±0.19
100%	Real only	0.8787±0.0039	**9.79±0.18**
	Real + CLoRA-MRI	**0.8807±0.0010**	9.85±0.28

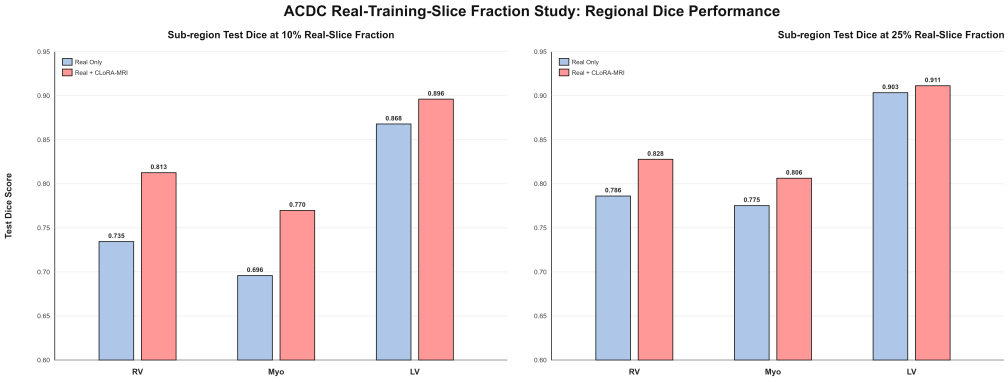
(b) Structure-wise Dice

Real fraction	Method	RV	Myo	LV
10%	Real only	0.7345	0.6959	0.8680
	Real + CLoRA-MRI	**0.8127**	**0.7698**	**0.8962**
25%	Real only	0.7862	0.7753	0.9035
	Real + CLoRA-MRI	**0.8279**	**0.8064**	**0.9114**
50%	Real only	0.8336	0.8192	**0.9243**
	Real + CLoRA-MRI	**0.8493**	**0.8250**	0.9238
100%	Real only	0.8617	0.8419	**0.9324**
	Real + CLoRA-MRI	**0.8662**	**0.8440**	0.9319

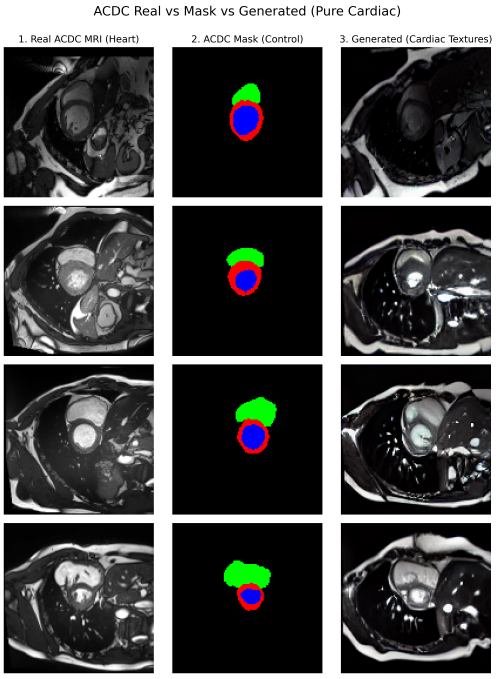
On ACDC, the benefit is largest at the 10% real-slice fraction: mean Dice improves by 0.0602 and HD95 decreases by 1.77 pixels. Improvements remain visible at 25%, while Dice gains narrow at 50% and 100%. Thin and boundary-sensitive structures, particularly the myocardium, benefit substantially in the lowest-data setting.

The per-structure changes at 10% are +0.0782 for RV, +0.0739 for myocardium, and +0.0282 for LV. The larger gain for myocardium is plausible because a thin wall is difficult to learn from a small sample and is sensitive to missing shape variation. At 25%, the mean Dice gain remains +0.0269. At 50% and 100%, Dice still increases slightly, but HD95 worsens by approximately 0.05–0.06 pixels. This overlap–boundary trade-off suggests that synthetic images may improve average region recognition while introducing small local edge artifacts.

The across-run Dice standard deviation also decreases at 10%, from 0.0057 for real-only training to 0.0026 with synthetic augmentation. Unlike the BraTS analysis, this result is measured on patient-disjoint held-out cases. Together with the mean improvement, this supports the value of augmentation when real-image coverage is limited.



(a) Per-structure Dice at 10% and 25% real-slice fractions.



(b) Examples of mask-conditioned cardiac synthesis.

Figure 6: ACDC downstream and qualitative results. The held-out test patients are not used for generator adaptation or segmentation training.

6 Discussion

6.1 What the Combined Evaluation Shows

FID, LPIPS, PSNR, and SSIM characterize generated images, whereas downstream Dice and HD95 assess their utility as supervision. CLoRA-MRI achieves a lower FID than Pix2Pix and improves segmentation most in the low-data regimes, while Pix2Pix has stronger paired pixel metrics. These complementary results motivate joint evaluation of image distribution, condition alignment, and downstream task utility.

6.2 Position Relative to Recent Studies

Recent studies address related tasks. SegGuidedDiff [18] uses segmentation guidance for diffusion, whereas our evaluation couples a pretrained latent diffusion backbone to a spatial condition pathway and a limited-parameter LoRA adaptation. Kebaili et al. [20] study multimodal MRI and mask synthesis; their work also uses a slice-based design. SeLoRA [23] adapts low-rank capacity for text-conditioned medical image generation, whereas our fixed-rank adaptation is spatially mask-conditioned. Fourier domain adaptation [26] instead targets distribution shift in cardiac MRI. Our contribution focuses on the combination of spatial mask conditioning, parameter-efficient adaptation, and downstream augmentation across real-image data fractions.

6.3 Why Augmentation Helps Primarily at Low Real-Image Fractions

At 10% BraTS data, the segmentation model sees approximately 1,868 real slices, whereas the synthetic set contributes 3,690 samples obtained by sampling ten realizations from each selected mask. These samples reuse existing anatomical labels but vary texture and intensity. The likely benefit is improved coverage of appearance conditional on a valid shape, rather than the creation of new lesion geometries.

As the real fraction increases, the fixed synthetic collection may become redundant. The slice-level protocol does not control patient coverage, so additional patient-specific appearances may also enter the real set. The diminishing-return pattern in both domains does not support automatic use of synthesis at full data scale.

6.4 Relationship Between BraTS and ACDC Evidence

The two downstream studies provide complementary evidence. BraTS supplies detailed in-cohort comparisons with Pix2Pix across data fractions; ACDC evaluates on patients held out at the patient level. The ACDC generator uses an independently trained target-domain adapter, demonstrating application of the same framework to a second anatomy.

6.5 Role of ControlNet and LoRA

ControlNet supplies spatial mask information; LoRA adapts image appearance with few trainable parameters. Ablations show weaker mask-conditioned synthesis without ControlNet and limited adaptation at very low ranks. Their combination supports controlled MRI synthesis under a restricted adaptation budget.

The rank study further suggests that adapting a natural-image prior to MRI is not a trivial low-dimensional shift. Performance improves substantially between ranks 16 and 32 and continues to improve through 128 for most raw-output metrics. Rank 64 offers a practical compromise for restricted hardware, while rank 128 is the best overall evaluated configuration.

6.6 Deployment Interpretation

The practical pipeline accepts a mask, samples one or more images, and uses the same mask as a training label. Raw generation requires no paired reference image. If intensity normalization is desired, it can be estimated from training-set statistics, an unpaired normalizer, or a learned target-domain mapping. The present results most directly support data augmentation for model development when valid masks are available but real-image coverage is limited.

6.7 Two-Dimensional versus Volumetric Generation

The method synthesizes each axial slice independently, making training feasible on an 8 GB GPU and permitting direct use of a pretrained 2D latent diffusion model. Volumetric use would additionally benefit from consistency across neighboring slices.

Potential extensions include a native 3D latent diffusion model, 2.5D conditioning on neighboring masks, or a second-stage consistency model.

6.8 Limitations

Three aspects define the scope of the evidence. BraTS uses an in-cohort slice-level split and mutually exclusive tissue labels, whereas ACDC provides a patient-disjoint test set but retains masks from all development patients across the real-image fractions; its reported metrics are computed from 2D slices and mini-batches. Generation is also performed independently at 128×128 per slice. Future evaluation would benefit from patient-disjoint BraTS testing, volumetric consistency and metrics, independent expert assessment, and comparisons with recent mask-guided diffusion methods.

7 Conclusion

CLoRA-MRI combines ControlNet mask conditioning with low-rank diffusion adaptation for segmentation augmentation; 2.96% of the BraTS framework's parameters are trainable at rank 128. Image-quality metrics and downstream performance do not move together. On the patient-independent ACDC test, synthetic pairs improve implemented mean Dice most at the 10% real-training-slice fraction, with diminishing benefit as real-image coverage grows. These findings identify a useful low-data operating regime for mask-conditioned augmentation.

Future studies can extend the protocol to patient-disjoint BraTS evaluation, standard volumetric metrics, recent mask-guided diffusion baselines, and independent anatomical assessment. Public code, fixed splits, and testing across scanners would further establish reproducibility and generalizability.

References

[1] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

[2] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci Data*, 4:170117, 2017. doi: 10.1038/sdata.2017.117.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE/CVF CVPR*, pages 1125–1134, 2017.

[4] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *Med Image Anal*, 58:101552, 2019. doi: 10.1016/j.media.2019.101552.

[5] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst*, 33:6840–6851, 2020.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF CVPR*, pages 10684–10695, 2022.

[7] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. IEEE/CVF ICCV*, pages 3836–3847, 2023.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *Proc. ICLR*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Adv Neural Inf Process Syst*, 27:2672–2680, 2014.

[10] S. U. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE Trans Med Imaging*, 38(10):2375–2388, 2019. doi: 10.1109/TMI.2019.2901750.

[11] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. doi: 10.1016/j.neucom.2018.09.013.

[12] H. H. N. Alrashedy, A. F. Almansour, D. M. Ibrahim, and M. A. A. Hammoudeh. Braingan: Brain MRI image generation and classification framework using GAN architectures and CNN models. *Sensors*, 22(11):4297, 2022. doi: 10.3390/s22114297.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE/CVF ICCV*, pages 2223–2232, 2017.

[14] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin. Diffusion models for medical anomaly detection. In *Proc. MICCAI*, volume 13438, pages 35–45. Springer, 2022. doi: 10.1007/978-3-031-16452-1_4.

[15] Walter H. L. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F. Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Brain imaging generation with latent diffusion models. In *Proc. MICCAI Workshop Deep Generative Models*, pages 117–126. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-18576-2_12.

[16] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, Johannes Stegmaier, Christiane Kuhl, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. Denoising diffusion probabilistic models for 3D medical image generation. *Sci Rep*, 13:7303, 2023. doi: 10.1038/s41598-023-34341-2.

[17] L. Jiang, Y. Mao, X. Wang, X. Chen, and C. Li. CoLa-Diff: Conditional latent diffusion model for multi-modal MRI synthesis. In *Proc. MICCAI*, volume 14229, pages 398–408. Springer, 2023. doi: 10.1007/978-3-031-43999-5_38.

[18] N. Konz, Y. Chen, H. Dong, and M. A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *Proc. MICCAI*, volume 15007, pages 88–98, 2024.

[19] Sanuwani Dayarathna, Yicheng Wu, Jianfei Cai, Tien-Tsin Wong, Meng Law, Kh Tohidul Islam, Himashi Peiris, and Zhaolin Chen. MU-Diff: A mutual learning diffusion model for synthetic MRI with application for brain lesions. *npj Artificial Intelligence*, 1(1):11, 2025. doi: 10.1038/s44387-025-00016-8.

[20] A. Kebaili, J. Lapuyade-Lahorgue, P. Vera, and S. Ruan. Multi-modal MRI synthesis with conditional latent diffusion models for data augmentation in tumor segmentation. *Comput Med Imaging Graph*, 123:102532, 2025. doi: 10.1016/j.compmedimag.2025.102532.

[21] R. Dutt, L. Ericsson, P. Sanchez, S. A. Tsaftaris, and T. Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In *Proceedings of the 7th International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pages 406–425. PMLR, 2024. URL `https://proceedings.mlr.press/v250/dutt24a.html`.

[22] Pierre Joseph Marcel Chambon, Christian Bluethgen, Curtis Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. In *NeurIPS 2022 Workshops: Foundation Models for Decision Making*, 2022. URL `https://mlanthology.org/neuripsw/2022/chambon2022neuripsw-adapting/`.

[23] Y. Mao, H. B. Li, W. Pang, G. Papanastasiou, G. Yang, E. Mohammadipasand, T. Harrison-Drummond, and C. Wang. SeLoRA: Self-expanding LoRA for high-quality and efficient medical image synthesis. *Expert Syst Appl*, 314:131569, 2026. doi: 10.1016/j.eswa.2026.131569.

[24] M. U. Akbar, M. Larsson, I. Blystad, and A. Eklund. Brain tumor segmentation using synthetic MR images—a comparison of GANs and diffusion models. *Sci Data*, 11(1):259, 2024. doi: 10.1038/s41597-024-03073-x.

[25] Y. Al Khalil, S. Amirrajab, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer. On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. *Med Image Anal*, 84:102688, 2023. doi: 10.1016/j.media.2022.102688.

[26] Y. Al Khalil, A. Ayaz, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer. Multi-modal brain tumor segmentation via conditional synthesis with Fourier domain adaptation. *Comput Med Imaging Graph*, 112:102332, 2024. doi: 10.1016/j.compmedimag.2024.102332.

[27] K. Qiu, Z. Gao, Z. Zhou, M. Sun, and Y. Guo. Noise-consistent Siamese-diffusion for medical image synthesis and segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15672–15681, 2025. doi: 10.1109/CVPR52734.2025.01461.

[28] E. Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev*, 56:12561–12605, 2023. doi: 10.1007/s10462-023-10453-z.

[29] Y. Chen, Y. Liu, M. Lu, L. Fu, and F. Yang. Multi-consistency for semi-supervised medical image segmentation via diffusion models. *Pattern Recognit*, 161:111216, 2025. doi: 10.1016/j.patcog.2024.111216.

[30] Z. Wan, Z. Gao, M. Sun, Y. Yang, C. Min, H. He, and G. Fu. Rethinking hard training sample generation for medical image segmentation. *Pattern Recognit*, 172:112533, 2026. doi: 10.1016/j.patcog.2025.112533.

[31] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *Proc. ICLR*, 2021. URL `https://openreview.net/forum?id=St1giarCHLP`.

[32] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, volume 9351, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans Image Process*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE/CVF CVPR*, pages 586–595, 2018.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Adv Neural Inf Process Syst*, 30:6626–6637, 2017.

[36] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans Med Imaging*, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502.