

SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY —
ELECTRICAL AND COMPUTER
ENGINEERING

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Electrical Engineering and Information
Technology

**Counterfactual Diffusion-based Image
Editing on Brain MRI**

Malek Ben Alaya

SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY —
ELECTRICAL AND COMPUTER
ENGINEERING

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Electrical Engineering and Information
Technology

**Counterfactual Diffusion-based Image
Editing on Brain MRI**

Author: Malek Ben Alaya
Examiner: Reinhard Heckel
Supervisor: Cosmin I. bercea, Daniel M. Lang, Julia Schnabel
Submission Date: 30.09.2024

I confirm that this master's thesis is my own work and I have documented all sources and material used.

Munich, Submission date

Author

Abstract

Denoising diffusion probabilistic models have demonstrated significant success in generating high-fidelity images in synthesis and editing tasks. In the field of biomedicine, these models enable counterfactual image editing, generating pairs of images where one simulates hypothetical conditions. For instance, they can be used to model the progression of diseases like stroke lesions. However, current image editing methods often struggle to produce realistic biomedical counterfactuals. They either fail to adequately represent indirect pathological effects, such as brain atrophy, or excessively modify the scan, compromising its resemblance to the original. To address these issues, we introduce *MedEdit*, a conditional diffusion model designed for medical image editing. *MedEdit* induces pathology in specific regions while maintaining a balance between modeling disease progression and preserving the integrity of the original scan. We validate *MedEdit* using the Atlas v2.0 stroke and BraTS 2023 tumor datasets, evaluating performance with Frechet Inception Distance and Dice scores. In the case of strokes, *MedEdit* surpasses state-of-the-art diffusion models like Palette (by 45%) and SDEdit (by 61%). Additionally, clinical assessments by a board-certified neuroradiologist confirm that *MedEdit* produces stroke scans that are indistinguishable from real ones. In the case of tumors, however, results fall short. Additionally, we extend the conditioning process of *MedEdit* to pathology masks in the form of bounding boxes instead of strictly delineated masks to increase its flexibility and generalisability. In this context, results show promise. Furthermore, we identify key areas of improvement, which we suggest for future work.

Contents

Abstract	iii
1 Introduction	1
2 Background	4
2.1 Brain Anatomy	4
2.2 Generative distribution learning	5
2.3 Evaluation approaches and metrics	9
2.4 Applications	13
2.4.1 Brain Strokes	13
2.4.2 Brain Tumors	15
3 Method	18
4 Experiments	20
4.1 Brain Strokes	20
4.2 Brain Tumors	22
5 Results	23
5.1 Brain Strokes	23
5.1.1 Ablation study	24
5.1.2 Extending the conditioning process of MedEdit	26
5.2 Brain Tumors	27
5.2.1 Ablation study	29
5.2.2 Extending the conditioning process of MedEdit	32
6 Discussion	35
7 Conclusion	37
List of Figures	39
List of Tables	40

Contents

Bibliography	41
---------------------	-----------

1 Introduction

Counterfactuals refer to hypothetical scenarios that depict what could have happened under different conditions, often used to explore "what if" questions by altering certain aspects of reality. In general, counterfactual reasoning involves generating data that represents alternative outcomes. For instance, in image generation tasks, counterfactuals might involve transforming an existing image to show how it would look if certain conditions were changed — such as visualizing a car in a different color or what a room would look like with different furniture. In the medical imaging context, counterfactuals typically involve generating hypothetical medical images or scenarios to explore how a patient's condition might have evolved under different circumstances. These hypothetical constructs enable researchers and clinicians to simulate disease progression, test hypotheses, and potentially develop better treatment strategies. By generating counterfactual images, generative models can provide insights into how diseases evolve or how a patient's condition could have changed if different medical interventions had been taken.

Generated counterfactuals can serve multiple purposes in predictive modeling. They can be used for tasks such as data augmentation to improve the performance of machine learning models on underrepresented populations [Shi+18], stress-testing models to uncover biases caused by population or acquisition shifts [Pér+24], and offering counterfactual explanations to interpret the decision-making processes of classification models [Ata+22].

Recent efforts in biomedical counterfactual image editing have mainly focused on chest X-ray datasets [Gu+23; Pér+24], which are plentiful, feature paired scans showing disease progression, and provide multimodal data, including both images and text. Conversely, brain imaging research has largely concentrated on generating healthy scans from pathological ones, primarily to tackle lesion localization problems [Ber+23; San+22; Wya+22].

In the context of image-to-image translation and counterfactual generation in biomedical imaging, paired and unpaired methods refer to the nature of the training data used to learn mappings between two image domains. Paired methods rely on datasets where corresponding image pairs from two different domains are available. For example, in medical imaging, this could mean having a set of images showing the progression of a disease over time (e.g., a baseline scan of a healthy brain paired with a follow-up scan

showing disease onset). The availability of these paired samples allows models to learn a direct mapping between the two image domains in a supervised manner. In contrast, unpaired methods do not require corresponding image pairs between domains. This is particularly useful when paired data is unavailable, as is often the case in medical imaging, where it may be difficult to obtain scans of the same patient before and after disease onset. Some studies in the biomedical domain tackle counterfactual generation for Alzheimer’s disease progression using paired images [NJ+24; PAR24]. However, counterfactual disease editing for brain imaging from unpaired images remains largely unexplored.

In recent years, diffusion-based methods for unpaired image-to-image translation have been widely adopted in applications that focus on the natural image domain, particularly ones involving datasets like ImageNet [Den+09] and Places2 [Zho+17]. For example, SDEdit [Men+22] was developed to convert rough sketches into photorealistic images by first applying Gaussian noise to the image and then gradually denoising it using a pre-trained diffusion model. Similarly, Palette [Sah+22] focuses on inpainting, where missing regions of an image are predicted and filled in by conditioning the diffusion process on the visible parts of the image. To the best of our knowledge, such diffusion-based methods have not yet been adopted in the medical domain.

In this work, we introduce a conditional diffusion-based image editing method designed to generate realistic counterfactual pathological brain Magnetic Resonance Imaging (MRI) scans, leveraging unpaired data during model training. Unlike existing approaches, our method is capable of modeling indirect pathological changes induced by certain conditions (e.g., brain atrophy resulting from a stroke) while maintaining high fidelity to the original factual scan. We anticipate that this approach will pave the way for further exploration in counterfactual biomedical image editing, particularly in the context of novel imaging modalities and pathology types. The key contributions of this work are outlined below:

- We evaluate state-of-the-art image editing and inpainting techniques for generating realistic counterfactual brain MRI scans featuring stroke and tumor lesions.
- We introduce MedEdit, a novel biomedical counterfactual image editing approach capable of simulating brain diseases and modeling their realistic consequences, while preserving high fidelity to the original scan.
- We validate our findings through an anonymized clinical reader study performed by a board-certified neuroradiologist, evaluating the realism, fidelity to the original scan, and the accurate representation of pathological changes, including induced diseases and their indirect effects.

- We extend *MedEdit* by conditioning its editing process on pathology masks represented as bounding boxes rather than strictly defined masks, laying the ground for increasing the flexibility and generalisability of the method.
- We draw attention to the urgent need of computational metrics that align better with clinical metrics.

2 Background

2.1 Brain Anatomy

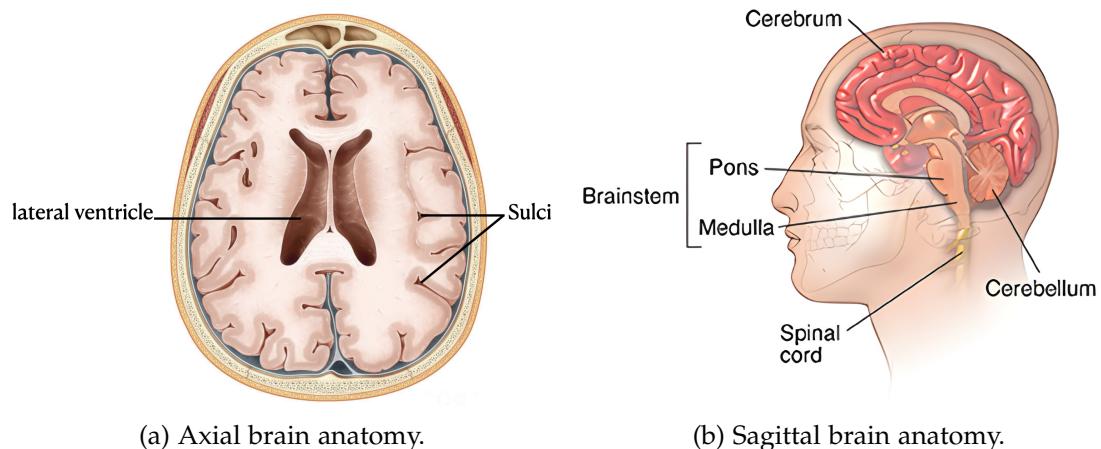


Figure 2.1: Anatomy of the brain, from axial (a) and sagittal (b) perspectives, taken from [Pix24] and [Hea24], respectively.

The brain, enclosed within the protective cranial vault, is divided into several key regions responsible for various functions. Among others, we cite the cerebellum, which coordinates balance and fine motor control. The brainstem, containing the pons and medulla oblongata, regulates vital autonomic functions like breathing and heart rate [Hea24]. Additionally, the ventricles circulate cerebrospinal fluid, and grooves called sulci increase the brain's surface area, allowing for more efficient neural activity [Ana24]. These key regions will be mentioned in later stages of the work. We further list more details about them below:

- The cranial vault is the space in the skull that encases and protects the brain. It is formed by the bones of the skull and serves as the protective dome surrounding the brain tissue.
- The brainstem is the lower portion of the brain, which connects the brain to the spinal cord and controls vital functions such as breathing, heart rate, and

blood pressure. It contains the pons and medulla oblongata, which work together to regulate autonomic processes and relay signals between the brain and body [Hea24].

- The cerebellum, located behind the brainstem, is involved in the coordination of voluntary movements, balance, and fine motor control. It consists of two hemispheres connected by a narrow midline structure called the vermis. The cerebellum also contributes to motor learning and cognitive functions like attention and language [Hea24].
- Sulci are the grooves or fissures on the brain's surface that increase its surface area to permit a more efficient neural activity and allow for more gray matter to fit within the skull [Ana24].
- The ventricles are a network of interconnected, fluid-filled cavities within the brain. They are responsible for the production, transport, and removal of cerebrospinal fluid (CSF), which cushions the brain and spinal cord [Ana24].
- The midline of the brain is an imaginary line that divides the brain into symmetrical left and right halves [hea24]. It is a reference point for neuroanatomy and helps to describe structures in relation to their position within the brain.

2.2 Generative distribution learning

Generative models are machine learning models that focus on learning the underlying distribution of the data they are trained on, allowing them to model the data itself. These models are typically built on deep neural network architectures. In contrast, discriminative models are designed to predict specific quantities based on input data. One of the key capabilities of generative models is their ability to synthesize novel data that, while distinct from real-world examples, appears convincingly realistic. Deep generative learning has lately emerged as a significant area of research, allowing for a broad set of applications. Generative models are extensively utilized in image synthesis and various image-processing tasks, including editing [Kaw+23], inpainting [Sah+22] and super-resolution [Tia+22]. Furthermore, the data representations learned by the neural networks in generative models, along with the synthetic data they produce, can be leveraged to train and fine-tune other machine learning models for a variety of downstream tasks. This capability is particularly valuable in situations where labeled data is limited. For broader adoption in real-world applications, it is desirable for generative models to meet the following key criteria [XKV22]:

- High-quality sampling: Many applications benefit from high-quality generation. For example, in speech generation, clear and understandable audio is important. Likewise, in image generation, outputs that closely resemble natural images are typically preferred.
- Mode coverage and sample diversity: When training data includes a wide range of diversity or complexity, it is helpful for a generative model to capture that variation effectively while still maintaining the quality of its outputs.
- Fast sampling: For interactive applications, such as real-time image editing, quick and resource-efficient generation is often desirable, enabling smoother and more responsive user experiences.

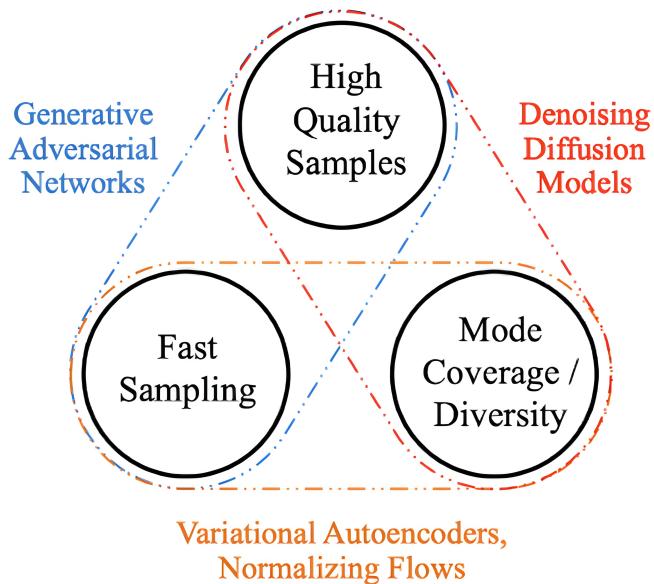


Figure 2.2: Key requirements Generative models are desired to fulfill [XKV22]. No single approach from GANs, VAEs or diffusion models excels at fulfilling all these requirements.

Generative Adversarial Networks (GANs) [Goo+14], Variational Autoencoders (VAEs) [KW22], and Diffusion Models [HJA20] each address the key requirements of generative models—high-quality sampling, mode coverage, and fast sampling—in different ways, with varying levels of success. An illustration of how each approach fulfills these requirements is provided in Figure. 2.2. GANs have been widely recognized

for their ability to produce high-quality samples, especially in tasks like image generation [Goo+14]. Their adversarial training process enables them to generate visually convincing outputs. However, GANs often struggle with mode collapse, where the model generates a limited variety of samples, failing to capture the full diversity of the training data [Met+17]. Furthermore, GANs can be challenging to train and may require significant computational resources to reach stability and quality [Bar18]. VAEs offer strong theoretical foundations and are efficient in capturing the diversity of data. They aim to model the entire data distribution, which helps prevent mode collapse and ensures broader coverage [XKV22]. However, they generally do not match GANs in terms of the sharpness or realism of the generated samples. Diffusion models have recently emerged as one of the most promising approaches in generative modeling. These models work by gradually adding noise to data and then learning to reverse this process, generating samples from random noise. Diffusion models have demonstrated exceptional performance in generating high-quality samples, often surpassing both GANs and VAEs in terms of fidelity and realism [DN21]. Moreover, they excel at covering diverse modes within the data, avoiding the issues of mode collapse often seen in GANs [XKV22]. While diffusion models were initially slower compared to GANs, recent advancements have significantly improved their sampling efficiency, making them more suitable for real-world applications that require fast generation [SH22]. Given these strengths, diffusion models have lately emerged as the preferred approach in generative modeling, combining high-quality output with wide mode coverage and growing computational efficiency.

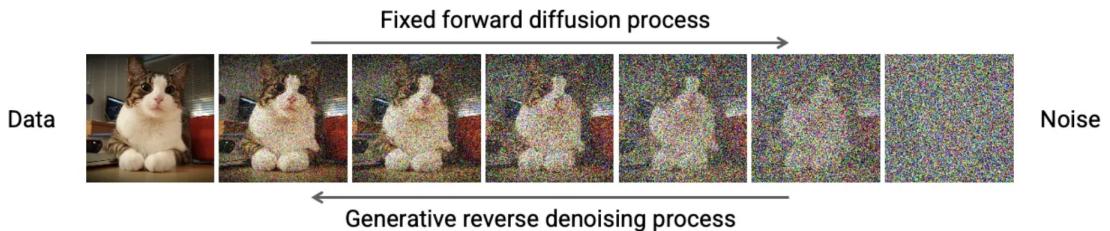


Figure 2.3: General scheme of DDPMS: in the fixed forward process, data is diffused through gradual addition of infinitesimal amounts of noise. By learning to reverse this process, diffusion models can generate samples starting from random noise.

Denoising Diffusion Probabilistic Models (DDPMs) [HJA20] are a class of generative models designed to sample from complex data distributions by learning to denoise progressively corrupted samples. These models follow a forward diffusion process, in

which Gaussian noise is incrementally added to the original sample x_0 over time steps $t = 1, \dots, T$, governed by the following equation:

$$q(x_t | x_{t-1}) \sim \mathcal{N}\left(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}\right). \quad (2.1)$$

Here, the noise schedule $\beta_{1:T}$ increases with each time step t , ensuring that by step T , the sample x_T approaches pure Gaussian noise. Exploiting the independence of noise addition at each step, the forward process can be simplified to:

$$q(x_t | x_0) \sim \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, \sqrt{1 - \bar{\alpha}_t}\mathbf{I}\right), \quad (2.2)$$

which can also be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\bar{\epsilon}_t, \quad (2.3)$$

where $\bar{\epsilon}_{1:T} \sim \mathcal{N}(0, I)$ represents the noise, and $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ captures the cumulative effect of the noise schedule. To generate new images, the forward diffusion process is reversed, transforming Gaussian noise into a sample from the learned distribution. Although reversing this process exactly is intractable, a variational approximation is applied by minimizing the denoising objective during training [HJA20]:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, t, \epsilon} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2. \quad (2.4)$$

The variational approximation is captured by the following equations:

$$x_{t-1} = \hat{\mu}_t(x_t, t) + \sigma_t z_t, \quad (2.5)$$

$$\hat{\mu}_t(x_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right), \quad (2.6)$$

where $\epsilon_\theta(x_t, t)$ is a learned estimate of the noise $\bar{\epsilon}_t$ that corrupted the original image x_0 to produce x_t , typically parameterized by a U-Net architecture [HJA20; RFB15b]. The term $z_{1:T} \sim \mathcal{N}(0, I)$ introduces Gaussian noise, while $\sigma_{1:T}$ controls the noise level at each step. For $\sigma_t = 0$, the process becomes deterministic, referred to as a Denoising Diffusion Implicit Model (DDIM) [SME22]. For probabilistic models with $\sigma_t = \sqrt{\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t}$, the process is known as a DDPM [27]. Throughout this work, we adopt this value for σ_t . A general scheme of how DDPMS operate is provided in Figure 2.3.

Learning conditional distributions with DDPMs requires adjusting the denoiser network ϵ_θ to incorporate a conditioning signal, such as a bounding box, text prompt, or

semantic map. This conditioning can be implemented using cross-attention or by directly appending the signal to the input channels of the denoiser network ϵ_θ [Rom+22]. Consequently, the equations above are updated by introducing the conditional signal c , modifying $\epsilon_\theta(\mathbf{x}_t, t)$ to $\epsilon_\theta(\mathbf{x}_t, c, t)$.

Generic image editing with RePaint. Re-paint [Lug+22] modifies the reverse diffusion process to enable targeted inpainting of specific regions by sampling from the joint distribution of a learned image set. The method operates in two key stages. First, it conditions the inpainting of unknown areas on the known regions, leading to the following formulation for a single reverse step:

$$x_{t-1}^{\text{known}} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (2.7)$$

$$x_{t-1}^{\text{unknown}} \sim \mathcal{N}(\mu_\theta(x_t, t), \sigma_t^2\mathbf{I}) \quad (2.8)$$

$$x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1 - m) \odot x_{t-1}^{\text{unknown}}, \quad (2.9)$$

where m is the mask specifying the region to be inpainted. This step ensures semantic coherence between the inpainted and unmasked regions. However, discrepancies may still occur between the two areas. To resolve this, the second stage, called resampling, reintroduces harmony by repeating the conditioning process. More specifically, this is done by diffusing x_{t-1} back to x_t through:

$$x_t \sim \mathcal{N}\left(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}\right), \quad (2.10)$$

and reapplying the conditioning process mentioned in the first component. This “resampling” process can be iterated multiple times to further align the inpainted and known regions.

2.3 Evaluation approaches and metrics

nnU-Net as pathology segmentation model. When generating counterfactual brain images conditioned on pathology masks, one can use a lesion segmentation model to assess whether a pathology was indeed generated in the desired area. One model that can be used for such purpose is the nnU-Net [Ise+21]. It represents an advanced, self-configuring variant of the U-Net architecture [RFB15a], designed to automatically adapt its processing pipeline to novel semantic segmentation tasks. Unlike traditional approaches, which often require the design of specialized networks, loss functions, or training protocols, nnU-Net autonomously configures preprocessing, training, and post-processing steps for any given task. nnU-Net configures its segmentation pipelines

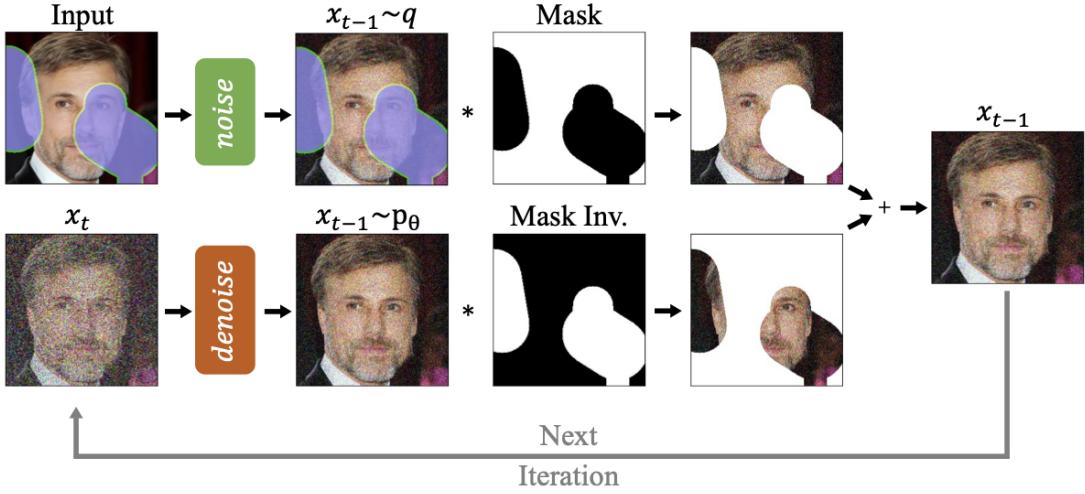


Figure 2.4: Overview of RePaint, taken from [Lug+22]: it modifies the standard denoising process in order to condition on the given image content. The known regions are guaranteed to stay the same as they get sampled from the forward process (top). The unknown regions to be inpainted are sampled from the DPPM (bottom).

based on a three-step recipe defined by three distinct parameter groups: fixed parameters, rule-based parameters, and empirical parameters. Fixed parameters remain constant across datasets and experimental runs, requiring no task-specific modification. As an example, we cite the loss function, which is a combination of the Dice coefficient and cross-entropy. In contrast, rule-based parameters are adaptive, adjusting to the specific characteristics of each dataset. For instance, intensity normalization strategies vary depending on the imaging modality. For MRI images, z-score normalization is applied based on the mean and standard deviation of individual images while for CT images, global percentile clipping is applied followed by z-score normalization using global mean and standard deviation. Empirical parameters, on the other hand, are derived entirely from the data itself. An example of this is the selection of the most effective model from an ensemble of 2D, 3D, and 3D-cascaded models, a decision that is based on the performance metrics obtained during cross-validation on the dataset. As with the original U-Net, robust data augmentation plays a critical role in nnU-Net. The augmentation strategies employed include random rotations, scaling, Gaussian noise and mirroring. nnU-Net has established itself as the state-of-the-art model for medical image segmentation, demonstrating superior performance over specialized segmentation pipelines across a wide range of tasks, which is demonstrated through a

dominance of nn-Unet based models in the leaderboards of segmentation challenges of internationally renowned scientific conferences such as the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) conference in recent years.

Dice score. The Dice score is a widely utilized metric in segmentation tasks, commonly used to evaluate the similarity between a predicted segmentation and the corresponding ground truth. Mathematically, it is defined as:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN}, \quad (2.11)$$

where TP , FP , and FN represent the true positive, false positive, and false negative predictions, respectively. Alternatively, the Dice score can be interpreted as twice the area of overlap (i.e., the intersection of true positive predictions) divided by the sum of the prediction and ground truth (i.e., their union), as illustrated in Figure 2.5. The Dice score ranges between 0 and 1, where a value of 0 indicates no overlap between the prediction and ground truth, while a value of 1 represents a perfect match.

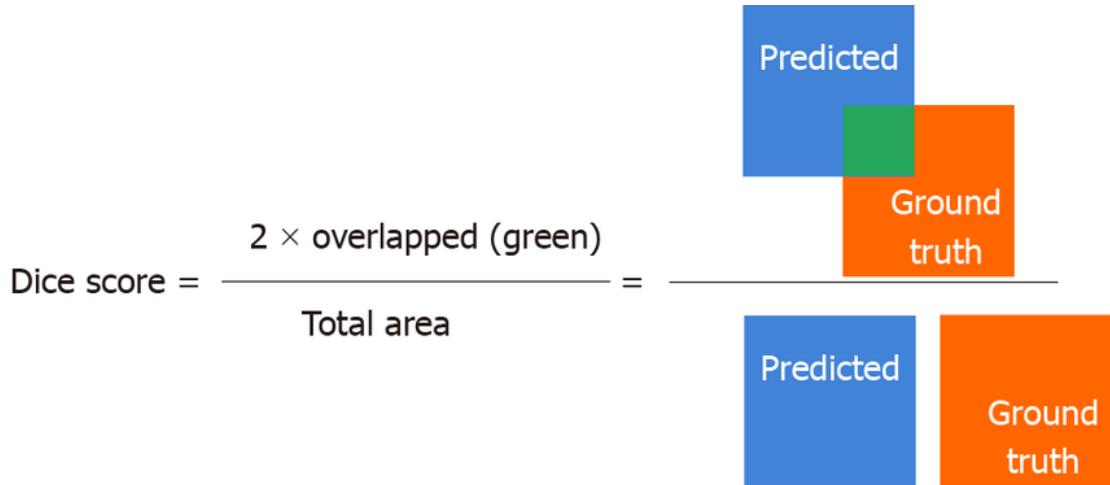


Figure 2.5: Visual interpretation of the Dice score, taken from [Kwo+23].

Intersection over Union. The Intersection over Union (IoU) is a fundamental metric used in image segmentation tasks to quantify the degree of overlap between a predicted segmentation and the corresponding ground truth. Mathematically, IoU is defined as:

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \quad (2.12)$$

The IoU represents the ratio of the intersection between the predicted segmentation and the ground truth to their union. This metric provides a direct measure of the quality of segmentation by evaluating how much the predicted regions overlap with the actual regions of interest. The IoU ranges between 0 and 1, where a value of 0 indicates no overlap between the prediction and ground truth, and a value of 1 signifies a perfect match, meaning the prediction completely aligns with the ground truth. As a stricter metric compared to others like the Dice score, IoU penalizes false positives and false negatives more severely, making it particularly useful in applications where minimizing both types of errors is crucial. Figure 2.6 provides a visual representation of IoU, demonstrating the intersection and union of the predicted and actual regions.

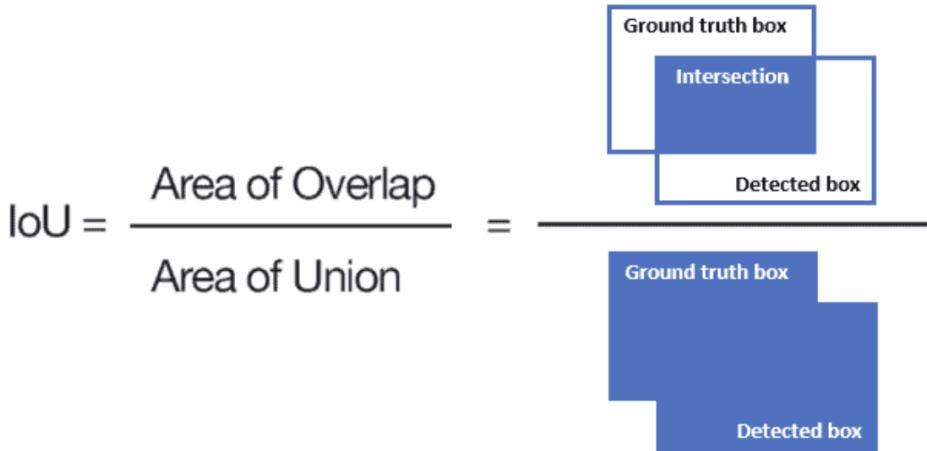


Figure 2.6: Visual interpretation of the IoU score, taken from [Bae24].

FID score. The Fréchet Inception Distance (FID) [Heu+17] is a widely adopted metric for evaluating the quality of image generation models. It has emerged as a standard measure for assessing various generative models, including GANs, VAEs and diffusion models. In contrast to its predecessor, the Inception Score (IS) [Sal+16], FID offers a more comprehensive evaluation by assessing both the fidelity and diversity of generated images through a comparative analysis of their distribution against that of real images. The FID metric operates by quantifying the similarity between two distributions, e.g. the distribution of real images and the distribution of generated images. Both distributions are modeled as multivariate Gaussians in a high-dimensional feature space. These features are extracted from an intermediate layer (typically the pool3 layer) of the Inception-v3 network [Sze+15], a convolutional neural network pre-trained on ImageNet. This approach allows for a nuanced comparison that captures both statistical

and perceptual similarities between the real and generated image sets. Let $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_g, \Sigma_g)$ represent respectively the distributions of real and generated images in the extracted feature space mentioned earlier, where μ denotes the mean vector and Σ the covariance matrix. The FID is then defined as [Heu+17]:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr} \left(\Sigma_r + \Sigma_g - 2 (\Sigma_r \Sigma_g)^{1/2} \right). \quad (2.13)$$

This metric quantifies both the disparity in mean values, which is indicative of image fidelity, and the divergence in covariance matrices, which reflects the diversity of the distributions.

2.4 Applications

Here, we introduce two different practical applications of our proposed method, namely the simulation of pathological conditions in brain strokes and tumors. Each part addresses the clinical significance of the respective condition, explores its potential indirect pathological effects, and details the dataset we aim to use for training and evaluation.

2.4.1 Brain Strokes

Why they are important. Brain strokes rank among the most devastating neurological conditions. Ischemic and hemorrhagic strokes are particularly severe, being among the most common and impactful cerebrovascular events in adults [WHO24]. They exhibit significant variability in location, size, and outcomes, with potentially life-altering consequences [Wik24b]. Strokes, in general, pose significant challenges. They require rapid diagnosis and treatment, and they can lead to long-term disability or death if not addressed promptly. The complexity of stroke management stems from the time-sensitive nature of interventions and the brain's limited capacity for regeneration. Despite these obstacles, years of extensive research to improve prevention, acute treatment, and rehabilitation have yielded notable progress [Wik24b]. Over the years, advanced treatments and therapies have revolutionized stroke care. However, strokes remain a leading cause of disability and are among the top three causes of death worldwide [WHO24].

Indirect pathological effects of strokes. After a stroke, the brain is unable to repair some of the lost neurons. This cell death triggers brain atrophy in the region affected by the stroke. The atrophy can be localized to the area where the stroke occurred or, in some cases, extend to other parts of the brain over time [Wik24a]. As brain atrophy

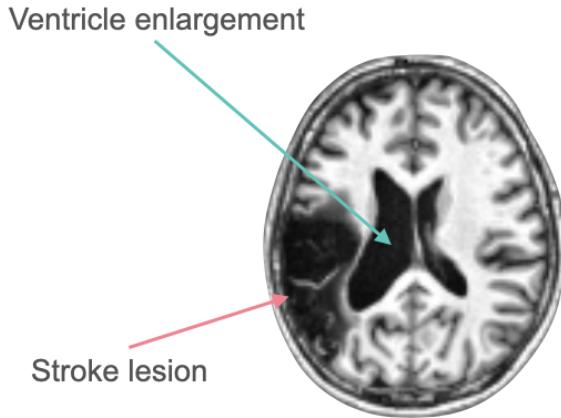


Figure 2.7: Illustration of possible indirect pathological effects of strokes exemplified by the enlargement of the left ventricle. The scan is taken from Atlas v2.0 [Lie+22].

progresses, the loss of brain tissue can cause the ventricles to enlarge, a phenomenon known as ventricular enlargement [Con11]. This occurs because the shrinking of brain tissue creates more space within the skull, which the ventricles expand into. This is shown in Figure 2.7.

Dataset description. The Atlas v2.0 [Lie+22] dataset, comprising T1-weighted brain MRI scans exhibiting stroke lesions, serves as the foundation for training our diffusion models. This dataset has undergone a comprehensive lesion tracing and preprocessing pipeline, including co-registration to a standardized anatomical template. The scans were acquired using diverse clinical protocols and various scanners from 44 distinct research cohorts. Data acquisition was performed on both 1.5-Tesla and 3-Tesla MR scanners, with the majority of scans maintaining high resolution (1 mm^3). Notably, four cohorts present at least one dimension with a resolution between $1\text{-}2 \text{ mm}^3$. With the exception of two cohorts, each dataset was collected using consistent parameters on a single scanner. In the training dataset ($N=655$), 61.9% of subjects exhibit a solitary lesion, while 38.1% exhibit multiple lesions. Among the latter group, 7.2% demonstrate unilateral involvement (confined to either the left or right hemisphere), 18.5% show bilateral distribution, and 12.4% exhibit at least one lesion in the cerebellum or brainstem. The lesion distribution demonstrates approximate hemispheric symmetry: 57.1% of subjects present with at least one left hemisphere lesion, 58.8% with at least one right hemisphere lesion, and 22.9% with involvement of the cerebellum or brainstem.

Figure 2.8 shows a visual lesion distribution of the dataset.

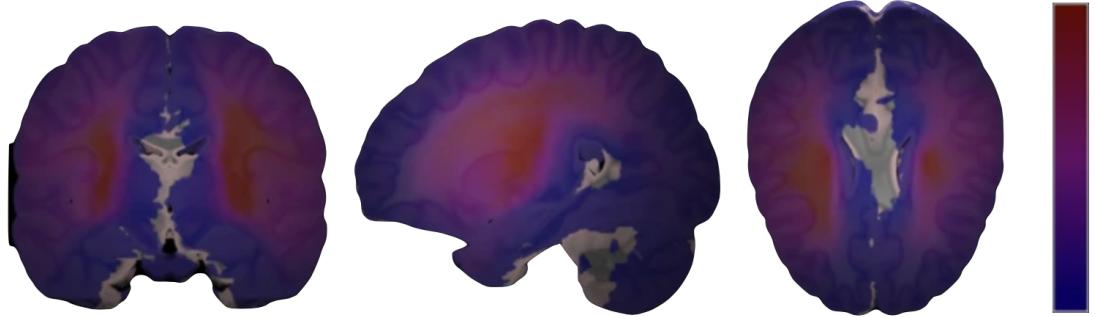


Figure 2.8: Probabilistic Lesion Overlap Map on the Montreal Neurological Institute (MNI) Template, taken from [Lie+22]. Visualization of the lesion overlap across all subjects overlaid on the MNI template, with hotter colors representing more subjects with lesions at that position.

2.4.2 Brain Tumors

Why they are important. Brain tumors represent one of the most serious forms of cancer. Among these, glioblastoma and related high-grade gliomas are particularly malignant. These tumors are the most prevalent and aggressive primary malignancies affecting the central nervous system in adult populations. They are characterized by significant variability in their presentation, structure, and cellular composition, with patients typically surviving only about 15 months after diagnosis. Brain tumors present numerous challenges in clinical practice. Their diagnosis and treatment are complex, and they often resist standard therapeutic approaches. This resistance is partly due to difficulties in delivering medications effectively to the brain and the tumors' diverse characteristics across imaging, structural, and genetic analyses. Despite these significant obstacles, ongoing research efforts to enhance diagnostic techniques, tumor characterization, and treatment strategies have led to some improvements. For instance, in the United States, mortality rates associated with brain tumors have seen a modest decline of 7% over the past three decades.¹

Indirect pathological effects of tumors. The mass effect, a consequence of intracranial tumors, is responsible for various indirect non-pathological effects. This phenomenon is characterized by the displacement of the soft tissue of the brain due to the presence

¹Facts mentioned in this paragraph are all taken from [Syn24].

of space-occupying lesions [cla24]. The brain, encased within the rigid cranial vault, has limited capacity to accommodate additional volume. Consequently, any intracranial lesion that occupies space may lead to increased intracranial pressure and subsequent displacement of cerebral tissues. This displacement, known as the mass effect, can manifest in several ways. Three primary consequences of the mass effect are particularly noteworthy [cla24]:

- Sulcal effacement: The obliteration or narrowing of the brain's sulci
- Ventricular effacement: The compression or distortion of the ventricular system
- Midline shift: The lateral displacement of midline brain structures

These effects are of significant clinical importance and can serve as valuable indicators of disease progression or treatment efficacy. In the context of neuroimaging research and development, it is imperative that counterfactual-generating models accurately simulate these mass effect phenomena to ensure the production of realistic and clinically relevant synthetic images.

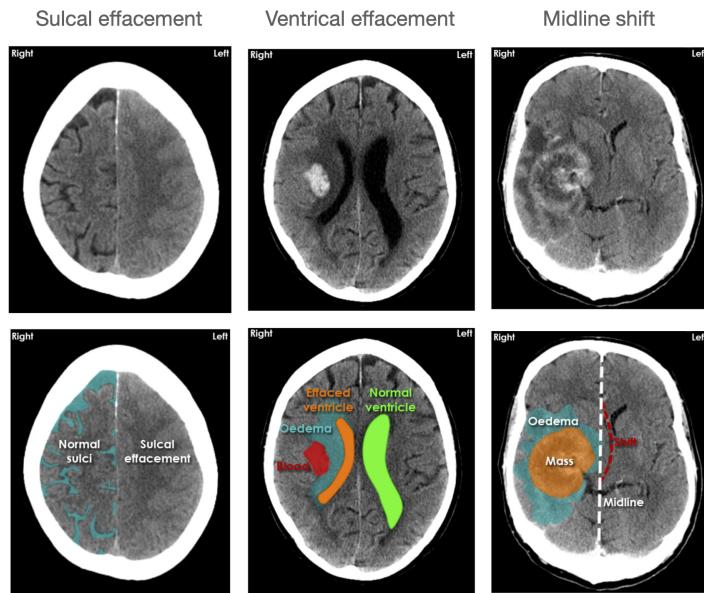


Figure 2.9: Visual illustration of possible indirect pathological effects that can be caused by tumors, taken from [cla24].

Dataset description. We use the BraTS 2023 dataset [Bai+21], which contains multi-parametric brain MRI scans of 4 modalities, namely native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2) and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). The scans were acquired with different clinical protocols and various scanners from multiple data contributing institutions. All scans have been annotated manually, by one to four raters, following the same annotation protocol, and their annotations were approved by experienced neuro-radiologists. Annotations comprise the Gadolinium-enhancing tumor (ET), the peritumoral edematous tissue (ED), and the necrotic tumor core (NCR). The dataset has undergone a preprocessing pipeline, which includes co-registration to the same anatomical template, interpolation to the same resolution (1 mm³) and skull-stripping.

3 Method

We introduce *MedEdit*, a conditional diffusion-based counterfactual image editing algorithm that is designed to strike a balance between accurately modeling indirect pathological changes and maintaining high fidelity to the original scan when simulating a pathology. Building on the RePaint algorithm (refer to Section 2.2), MedEdit shifts from a class-conditional to a mask-conditioned inpainting process, enabling precise pathology simulation. Additionally, it implements a mask selection technique for identifying the target inpainting region, effectively capturing potential indirect pathological changes (see Section 2.4).

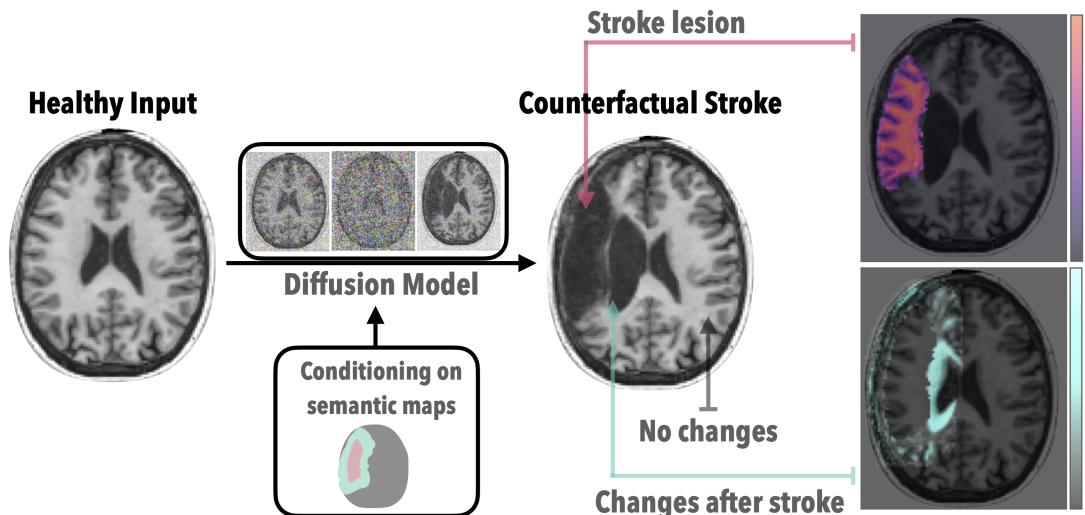


Figure 3.1: Overview of *MedEdit*, taken from [Ala+24]. It performs conditional edits on prior scans to generate counterfactual stroke images, simulating both direct and indirect pathological effects.

Conditional inpainting. We incorporate the conditions into the diffusion model by integrating the masks as supplementary input channels fed into the denoiser network

Algorithm 1 Biomedical counterfactual image editing with MedEdit.

```

1:  $x_T \sim \mathcal{N}(0, I)$ ,  $x_0$  : prior scan,  $b$  : brain mask,  $p$  : pathology mask
2:  $k$  : kernel size
3:  $m = dilute(p, k)$ 
4: for  $t = T, \dots, 1$  do
5:   for  $u = 1, \dots, U$  do
6:      $\epsilon \sim \mathcal{N}(0, I)$  if  $t > 1$  else  $\epsilon = \mathbf{0}$ 
7:      $x_{t-1}^{\text{known}} = \sqrt{\bar{\alpha}_t}x_0 + (1 - \bar{\alpha}_t)\epsilon$ 
8:      $z \sim \mathcal{N}(0, I)$  if  $t > 1$  else  $z = \mathbf{0}$ 
9:      $x_{t-1}^{\text{unknown}} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, b, p, t) \right) + \sigma_t z$ 
10:     $x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1 - m) \odot x_{t-1}^{\text{unknown}}$ 
11:    if  $u < U$  and  $t > 1$  then
12:       $x_t \sim \mathcal{N}(\sqrt{1 - \beta_{t-1}}x_{t-1}, \beta_{t-1}I)$ 
13:    end if
14:  end for
15: end for
16: return edited version of  $x_0$ 

```

of the diffusion model. This is shown in line 9 of Algorithm 1:

$$x_{t-1}^{\text{unknown}} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, b, p, t) \right) + \sigma_t z, \quad (3.1)$$

where b and p are the brain and the pathology masks, respectively.

Mask selection. We select the inpainting mask in such a way that it enables the modeling of indirect pathological changes during the editing process. Simply setting $m = p$ would confine inpainting to the pathology's specified region, neglecting the broader effects it might induce in other areas of the brain. This straightforward approach, which we term naïve RePaint, is used as a baseline in our experiments. To capture the necessary indirect pathological changes, we define m as a diluted version of the original pathology mask p , where the dilution kernel size k determines the spatial extent of the modeled indirect effects. A comprehensive description of *MedEdit* is provided in Algorithm 1.

4 Experiments

We carry out a thorough set of experiments to evaluate the performance of our proposed method against state-of-the-art image editing and inpainting approaches for simulating stroke and tumor effects on brain scans. Our evaluation focuses on key aspects such as realism, consistency with the desired pathological changes, fidelity to the original scan, and the accurate representation of indirect pathological effects like brain atrophy. We use clinical and computational metrics, including FID and Dice scores, to assess the outcomes. The proposed method is compared to Palette [Sah+22], SDEdit [Men+22], and naïve RePaint. Notably, we adapt SDEdit to perform conditional denoising based on brain and pathology masks, similar to the adaptation by Couairon et al. [Cou+22] for text-prompt-based denoising. Additionally, in a separate experiment, we extend our method by conditioning the editing process on pathology masks represented as bounding boxes, rather than strictly defined masks. We apply the same modification to the baseline methods for comparison. This adjustment is made with the aim of enhancing the flexibility of our approach, allowing it to function without reliance on precisely delineated pathology masks during the editing phase, but instead on more generalized forms of masks.

4.1 Brain Strokes

Dataset Preparation. We utilize the Atlas v2.0 training dataset, which comprises 655 T1-weighted brain MRI scans. As preprocessing, we normalize the mid-axial slices (95th) to the 98th percentile by clamping pixel intensities above this threshold, followed by max normalization to map intensities within the [0,1] range. Scans are then padded and resized to a resolution of 128×128 pixels. We do the same for the pathology masks except for the intensity normalization. Of the 655 images, only 443 contain visible pathologies. We stratify this pathological subset by lesion size into three distinct groups: small, medium, and large. The small group ($N=111$) consists of lesions within the lowest 25th percentile, with areas smaller than 18.5 pixels. The large group ($N=111$) corresponds to the top 25th percentile, with lesion areas exceeding 371 pixels. The medium group ($N=221$) comprises scans with lesions of intermediate size. The pathological subset is further divided into training ($N=389$) and testing ($N=54$) sets for

evaluation.

Implementation details. We train a diffusion model conditioned on brain and pathology masks to generate pathological brain scans, which is subsequently used to produce counterfactuals for SDEdit, naïve RePaint, and MedEdit. The U-Net architecture from [HJA20] is employed, with $T = 1000$ diffusion steps and a linear noise schedule β_t ranging from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ as specified in [HJA20]. The model is trained for 1500 epochs. In MedEdit, a dilution kernel size of $k = 25$ is applied, with four resampling steps. For naïve RePaint, we use three resampling steps. For SDEdit, we use an encoding ratio of 0.2, i.e. diffuse the prior scan for 200 timesteps. Further, we do the same process while extending the conditioning of the diffusion model to pathology masks in form of a bounding box instead of strictly defined pathology masks. The bounding box masks are computed by fitting them to the original pathology masks. Here, we use a dilution kernel size of $k = 20$, as fitting the bounding box to the original pathology mask already yields an increase in the mask size.

Evaluation. At inference time, for both types of conditioning experiments, we generate counterfactuals by randomly pairing each pathology mask from the test set with a non-pathological scan, forming triplets of (prior scan, brain mask of the prior, pathology mask).

Computational metrics. We assess the realism of the generated counterfactuals by calculating the Fréchet Inception Distance (FID) between the generated and real pathological test distributions across 10 bootstrapped runs. In each run, the counterfactuals remain fixed, while samples from the test set are resampled via bootstrapping. FID scores are computed using features extracted by a ResNet50 backbone [He+15], pre-trained on RadImageNet [Mei+22]. To quantify the adherence to the intended pathological changes, we employ nnUNet to segment pathology lesions in the generated counterfactuals and calculate the Dice score by comparing the predicted lesions to the ground truth pathology masks. For the bounding-box conditioning experiments, the Dice score is calculated by comparing bounding boxes fitted to the predicted lesions against the ground truth pathology masks. Additionally, we compute the IoU in these experiments to reinforce the robustness of our results, as the metric is widely used in bounding box prediction tasks such as object detection [Red+16]. We note that nnUNet is trained on the same training split as the diffusion model, using default hyperparameters over 200 epochs.

Clinical metrics: For clinical evaluation, a board-certified neuroradiologist assesses the validity of the computational metrics and further evaluates the fidelity of the counterfactuals relative to the original scans and their ability to model indirect pathological changes. We randomly select 20 counterfactuals from each benchmarked method,

stratified by pathology size, and mix them with 20 real test samples, resulting in a set of 100 scans. In the first stage of the clinical evaluation, the neuroradiologist rates the realism of the counterfactuals on a scale from 0 to 4. In the second stage, the same counterfactuals, along with their corresponding prior scans, are assessed based on three criteria: adherence to the intended pathological change (Path.), fidelity to the original scan (Fidel.), and accurate modeling of indirect pathological effects (Ind-Path.), each rated on a scale from 0 to 4. These ratings from the clinical assessment are referred to as clinical metrics throughout the evaluation.

4.2 Brain Tumors

Dataset Preparation. We make use of the T1-Gd scans ($N=1260$) of BraTS 2023, from which we normalize the mid-axial (90th) slices to the 99.5th percentile as done for strokes, central crop them to a size of 200×200 pixels and resize them to a size of 128×128 pixels. We do the same for the pathology masks except for the intensity normalization. Additionally, we fuse all subregions of a pathology, namely ED, ET and NCR, to a single region. Of the total 1260 images, only 1063 contain a pathology. We stratify the pathological subset with respect to the pathology size into three pathology groups, namely small, medium and large. The small group ($N=265$) comprises the first 25th percentile, consisting of lesions smaller than 135.75 pixels. The large group ($N=265$) encompasses the top 25th percentile, including lesions larger than 397 pixels. The medium group ($N=527$) includes the remaining scans with lesions of intermediate sizes. We further split the pathological subset into a train ($N=937$) and test set ($N=126$).

Implementation details. For SDEdit, we use an encoding ratio of 0.6, as it yields optimal computational metrics. All other parameters are kept same as for brain strokes.

Evaluation. We use the same approach to generate and assess the counterfactuals from a computational metrics perspective as for brain strokes. However, we do not assess the generated counterfactuals from a clinical metrics perspective, which is strongly tied to their perceptual quality. More details will be provided on this in later sections.

5 Results

In this section, we present the results of the experiments outlined in the previous section, using both clinical and computational metrics. Additionally, we perform ablation studies to evaluate the impact of individual components of *MedEdit*. Lastly, we report our findings from the bounding box-conditioning experiments.

5.1 Brain Strokes

Here, we investigate the results of the experiments involving strictly defined mask conditioning for strokes. Quantitative and qualitative results are summarized in Table 5.1 and Figure 5.1. From a computational metrics standpoint, *MedEdit* achieves the lowest FID score, demonstrating a significantly stronger alignment with the distribution of real stroke images compared to baseline methods. In that regard, *MedEdit* surpasses SDEdit by approximately 65.6%. In the downstream lesion segmentation evaluation, *MedEdit* markedly outperforms Palette and naïve RePaint, with improvements of approximately 65.8% and 26%, respectively, while maintaining competitive performance against SDEdit, trailing by only a marginal 6%. Notably, *MedEdit* excels in balancing accurate pathology inpainting with generating images closely resembling real stroke scans, as evidenced by its superior score on the combined (1-Dice) * FID metric, surpassing all baseline methods.

From a clinical metrics standpoint, *MedEdit* delivers realism levels nearly indistinguishable from real samples, outperforming SDEdit by 14%, naïve RePaint by 25%, and Palette by 33%. While all methods exhibit comparable adherence to the intended pathological changes, *MedEdit* scores slightly lower in this area. However, our analysis reveals a clear trade-off between modeling indirect pathological effects and preserving healthy brain features. Both naïve RePaint and Palette prioritize the preservation of healthy tissue but fail to adequately capture indirect pathological changes, as illustrated in Figure 5.1. In contrast, *MedEdit* and SDEdit strike a more effective balance, with *MedEdit* surpassing SDEdit by offering better preservation of healthy brain regions while maintaining similar performance in modeling indirect pathological changes. This balance is visually evident in the comparisons shown in Figure 5.1.

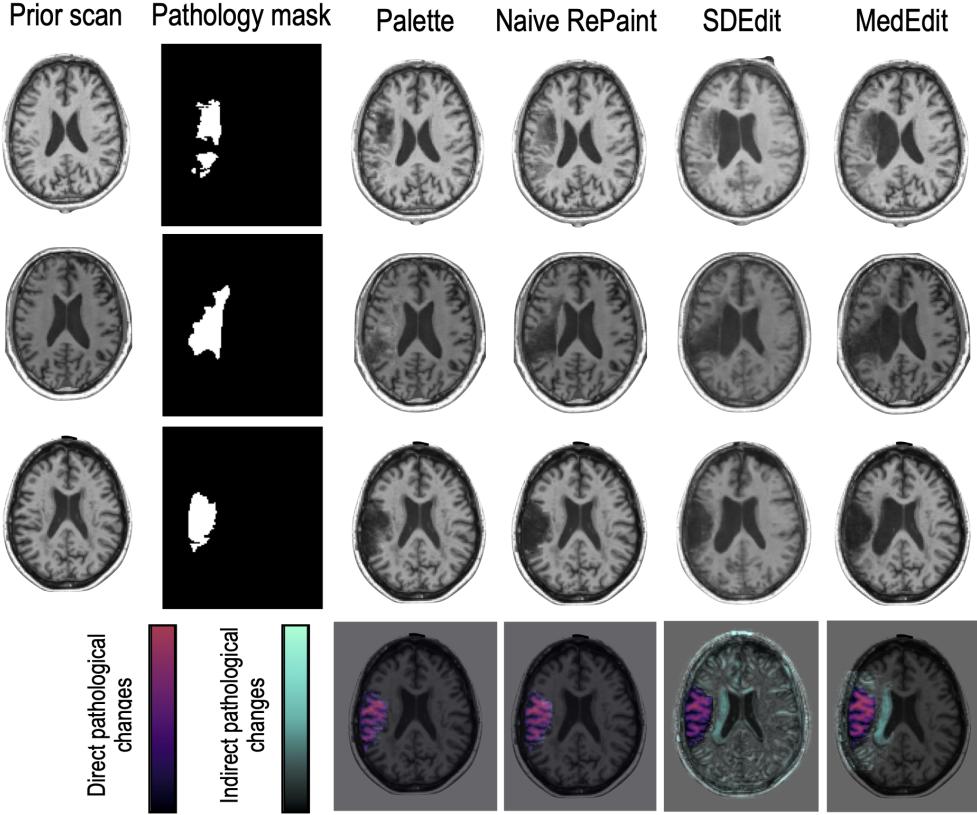


Figure 5.1: Examples of counterfactuals generated by *Palette*, naïve *RePaint*, *SDEdit*, and *MedEdit* are presented. While all methods effectively model the pathology in the final case (bottom row, purple difference maps), *MedEdit* uniquely captures the indirect pathological changes induced by the lesion, highlighted in turquoise, while maintaining a high fidelity to the prior scan. Specifically, *MedEdit* accurately models the stroke-induced enlargement of the ventricle on the affected side, demonstrating its superior ability to account for secondary anatomical effects of the pathology.

5.1.1 Ablation study

Ablating the effect of dilation kernel. For our proposed method, we ablate the effect of the dilation kernel size for a fixed number of resample steps. Results are shown in Table 5.2. The FID scores are practically invariant across all tried kernel sizes, indicating that the dilation kernel size has a negligible impact on the quality of generated images. The Dice coefficient shows a more noticeable improvement as the kernel size increases.

Table 5.1: We assessed perceived realism, fidelity to the original scan (Fidel.), adherence to the intended pathological changes (Path.), and the ability to model indirect pathological effects (Ind-Path.). The top-performing results are highlighted in **bold**, while the second-best results are underlined. Performance variations relative to the highest-performing method and the best baseline method are denoted by \blacktriangle and \blacktriangledown , respectively.

Method	Computational Metrics			Clinical Metrics				
	(1-Dice)	* FID \downarrow	FID \downarrow	Dice \uparrow	Realism \uparrow	Fidel. \uparrow	Path. \uparrow	Ind-Path. \uparrow
Real samples	-	-	-	3.20	-	-	-	-
SDEdit [Men+22]	7.95 \blacktriangle 159%	24.1	0.67	<u>2.80</u>	2.10	3.60	<u>3.00</u>	
Palette [Sah+22]	5.63 \blacktriangle 83%	9.08	0.38	2.40	<u>3.95</u>	<u>3.65</u>	2.00	
Naïve RePaint	<u>4.24</u> \blacktriangle 38%	<u>8.31</u>	0.50	2.55	4.00	<u>3.70</u>	1.85	
<i>MedEdit</i> (ours)	3.07 \blacktriangledown 28%	8.30	<u>0.63</u>	3.20	3.20	3.45	3.15	

It starts at 0.59 for a kernel size of 10 and gradually improves to 0.63 at a kernel size of 25. This suggests that larger dilation kernels result in better adherence to desired pathology.

Table 5.2: Ablating the effect of dilation kernel size in MedEdit for brain strokes. The Dice score increases with the dilation kernel size, which suggests that larger dilation kernels result in better adherence to desired pathology.

k, r = 4	FID	Dice
10	8.16	0.59
15	8.20	0.60
20	8.23	0.61
25	8.17	0.63

Ablating the effect of resample steps. We also ablate the effect of resampling steps for a fixed dilation kernel size. Results are shown in Table 5.3. We observe that the FID score remains relatively stable, with values ranging from 8.17 to 8.61. There is a slight increase in FID as the number of resampling steps increases, but the differences are minor. The Dice score shows a small improvement from 0.62 ($r = 1$) to 0.65 ($r = 3$), suggesting that the adherence to the desired pathology improves with more resampling steps. For $r = 4$, the Dice score slightly drops to 0.63, which may indicate diminishing returns after a certain number of resampling steps. In general, the results suggest that increasing the number of resampling steps can lead to marginal improvements in the Dice score, especially up to 3 resampling steps. However, the FID score remains

almost unchanged, implying that the quality of the generated counterfactuals, from a perceptual standpoint, is not significantly impacted by the number of resampling steps beyond a certain point.

Table 5.3: Ablating the effect of resample steps in MedEdit for brain strokes: increasing the number of resample steps leads to marginal improvements in the Dice score up to a certain point. However, it has practically no impact on FID score as it remains almost unchanged.

r, k = 25	FID	Dice
1	8.17	0.62
2	8.31	0.64
3	8.61	0.65
4	8.17	0.63

Visually ablating the effect of resample steps. We further show a visual ablation of the effect of resample steps in Figure 5.2. In the provided examples, which are selected to effectively demonstrate the impact of resampling, as the number of resampling steps increases, there is a noticeable refinement in the generated pathology regions in the output images. For $r = 1$, the results show visual artefacts, especially evident in the areas highlighted by the red boxes. The generated region seems to poorly define the affected area. For larger resample steps, the results appear more refined, showcasing more harmony with reduced artefacts.

5.1.2 Extending the conditioning process of MedEdit

Here, we investigate the results of our bounding-box conditioning experiments involving strokes (refer to Section 4). Quantitative and qualitative results are presented in Table 5.4 and Figure 5.3, respectively. From Table 5.4, we observe FID scores in the same range as for the type of mask conditioning mentioned earlier, with *MedEdit* achieving the lowest score, demonstrating a significantly stronger alignment with the distribution of real stroke images compared to baseline methods. In that regard, *MedEdit* surpasses SDEdit by approximately 62%. In the downstream lesion segmentation evaluation, *MedEdit* outperforms the baselines in terms of both Dice and IoU scores. Furthermore, also for this type of conditioning, *MedEdit* outperforms SDEdit in balancing the modeling of indirect pathological changes with maintaining a high fidelity to the prior scan, which is shown in Figure 5.3. Additionally, *MedEdit* excels in balancing accurate pathology inpainting with generating images closely resembling real stroke scans, as

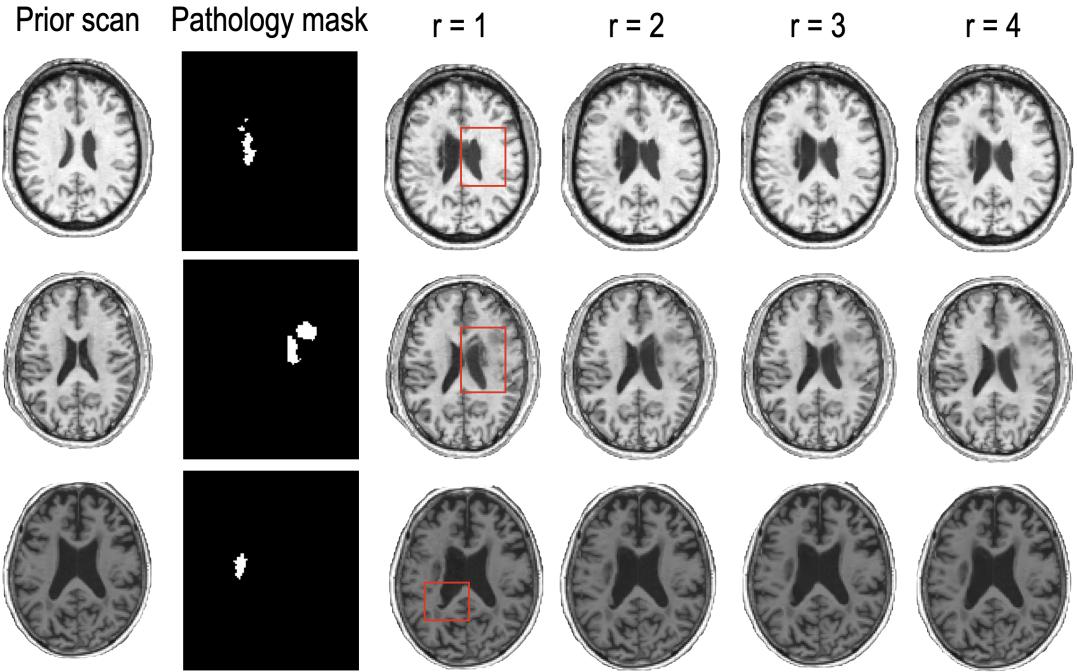


Figure 5.2: Visual Ablation of resample steps in MedEdit: larger resample steps result in more refined counterfactuals, showcasing more harmony and reduced artefacts.

evidenced by its superior score on the combined (1-Dice) * FID metric, surpassing all baseline methods.

5.2 Brain Tumors

Here, we investigate the results of the experiments involving strictly defined mask conditioning for tumors. Quantitative and qualitative results are presented in Table 5.5 and Figure 5.4. From a computational metrics perspective, *MedEdit* ranks highest in simultaneously inpainting pathologies while producing images that are close to the distribution of real stroke scans, which is showcased by its higher score on the combined (1-Dice) * FID metric compared to the baseline methods. Notably, it outperforms the best competing method in that regard, namely naïve RePaint by 20%. In the downstream lesion segmentation evaluation, *MedEdit* substantially outperforms Palette and naïve RePaint, with improvements of approximately 17.18% and 12%, respectively, while having an equal performance compared to SDEdit.

Table 5.4: We assess the computational metrics for the bounding-box conditioned versions of Palette, Naïve RePaint, SDEdit, and MedEdit using the Atlas V2.0 dataset. Best results are shown in **bold** and second-best are underlined. Performance variations relative to the highest-performing method and the best baseline method are denoted by \blacktriangle and \blacktriangledown , respectively.

Method	Computational Metrics			
	(1-Dice) * FID \downarrow	FID \downarrow	Dice \uparrow	IoU \uparrow
SDEdit	9.49 $\blacktriangle 174\%$	22.09	<u>0.57</u>	<u>0.48</u>
Palette	4.08 $\blacktriangle 18\%$	8.5	0.52	0.41
Naïve RePaint	3.73 $\blacktriangle 8\%$	<u>8.48</u>	0.56	0.48
MedEdit (ours)	3.46 $\blacktriangledown 7\%$	8.45	0.59	0.49

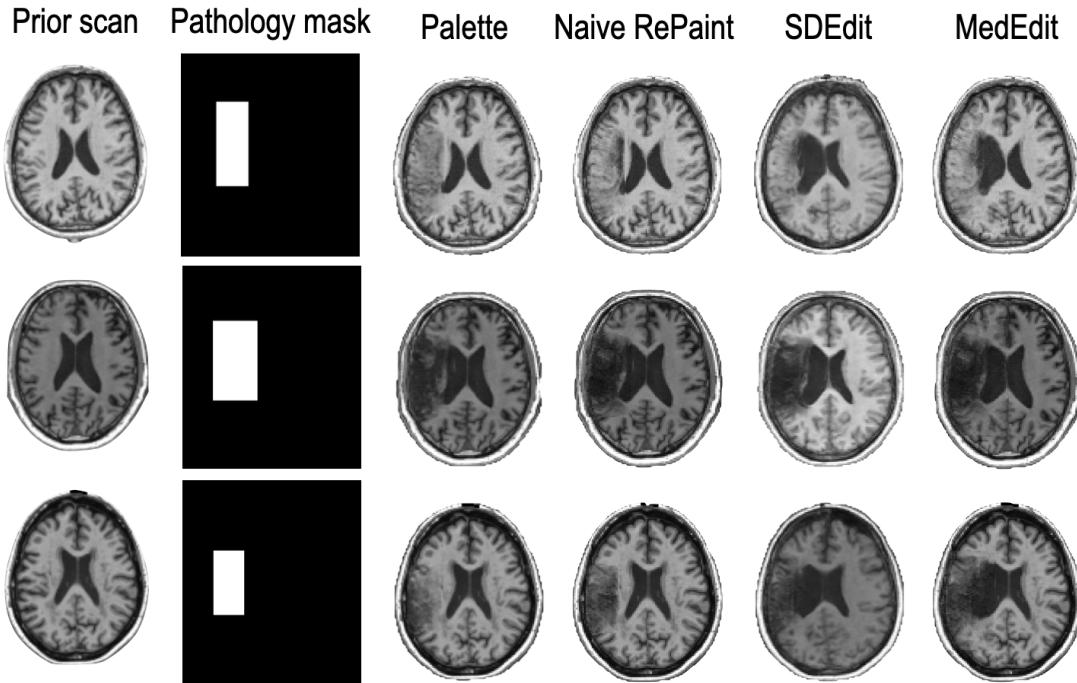


Figure 5.3: Examples of counterfactuals obtained with the bounding-box conditioned versions of Palette, Naïve RePaint, SDEdit and MedEdit for the Atlas V2.0 dataset.

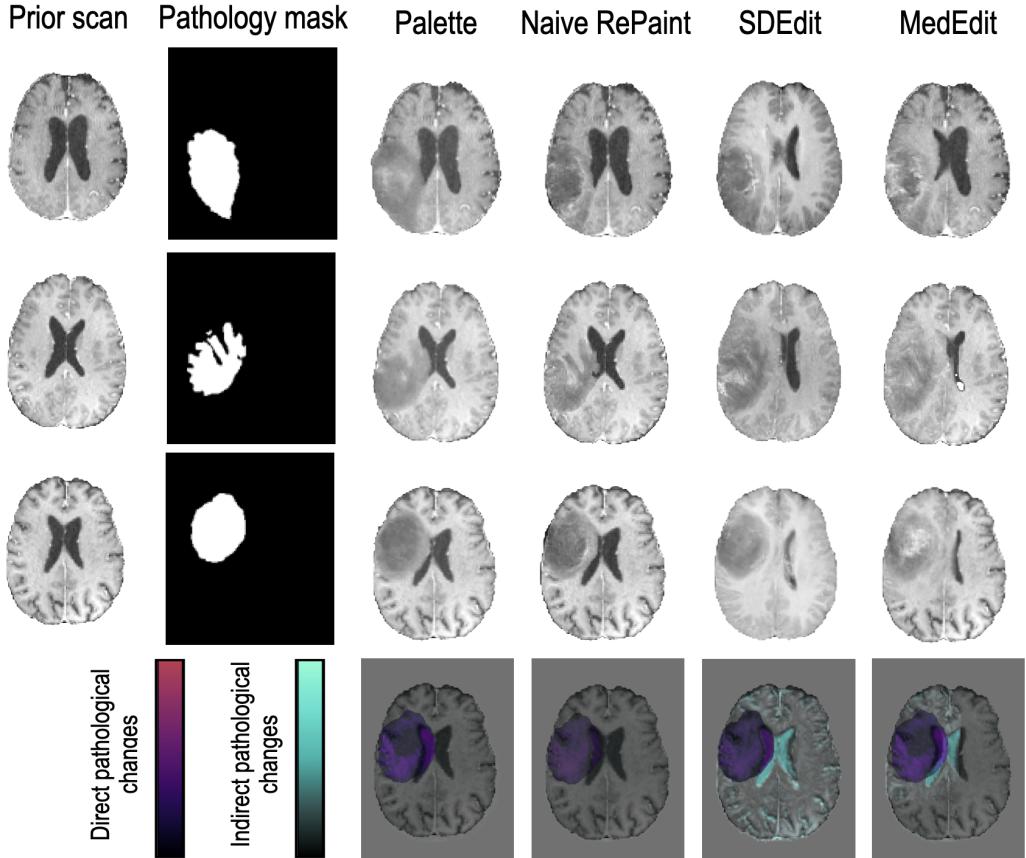


Figure 5.4: Examples of counterfactuals generated by *Palette*, naïve *RePaint*, *SDEdit*, and *MedEdit* using the BraTS 2023 dataset are displayed. While all methods successfully model the pathology in the final case (bottom row, shown in purple difference maps), *MedEdit* stands out by accurately capturing the indirect pathological changes induced by the tumor, highlighted in turquoise. In this instance, *MedEdit* effectively models the tumor-induced ventricular effacement, demonstrating its superior capability to reflect secondary anatomical effects of the pathology.

5.2.1 Ablation study

Ablating the effect of dilation kernel. Similar to strokes, we ablate the effect of the dilation kernel size in *MedEdit* for a fixed number of resample steps for the case of tumors. Results are shown in Table 5.6. The FID scores show minor fluctuations as the kernel size increases. The lowest FID (6.80) is achieved at a kernel size of 25,

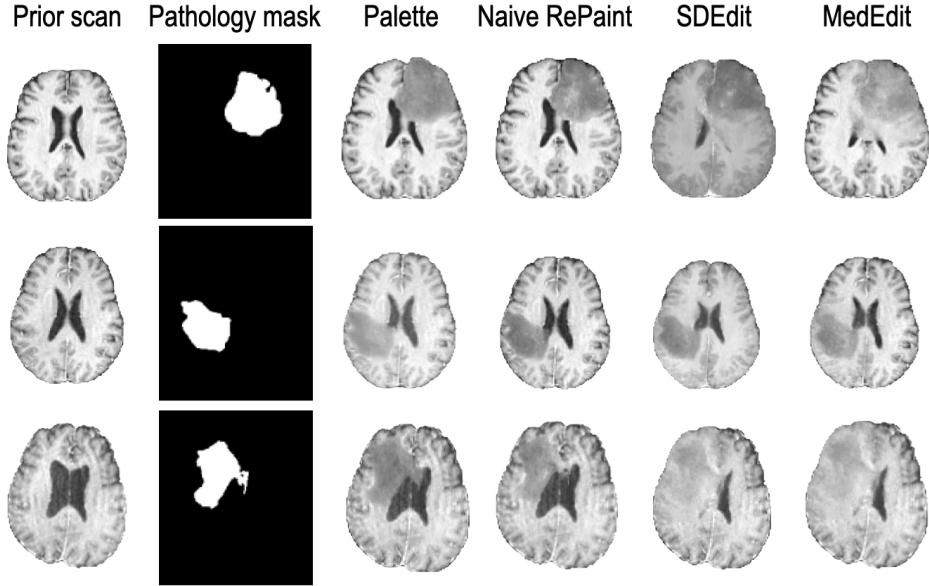


Figure 5.5: Examples of counterfactuals obtained with *Palette*, *Naïve RePaint*, *SDEdit* and *MedEdit* using the BraTS 2023 dataset are displayed. The generated tumors only contain the edematous component and lack a necrotic core.

Table 5.5: We evaluated the computational metrics for the BraTS 2023 dataset. Best results are shown in **bold** and second-best are underlined. \blacktriangle and \blacktriangledown show performance changes relative to the best method and to the best baseline method, respectively.

Method	Computational Metrics		
	(1-Dice) * FID \downarrow	FID \downarrow	Dice \uparrow
SDEdit [Men+22]	<u>2.06</u> \blacktriangle 25%	8.24	0.75
Palette [Sah+22]	2.26 \blacktriangle 37%	6.29	0.64
Naïve RePaint	2.08 \blacktriangle 26%	<u>6.33</u>	<u>0.67</u>
<i>MedEdit</i> (ours)	1.65 \blacktriangledown 20%	6.63	0.75

while the highest FID (6.85) occurs at a kernel size of 20. However, the changes are minimal, indicating that the dilation kernel size has a negligible impact on the quality of generated images. Similar to strokes, the Dice coefficient shows a more noticeable improvement as the kernel size increases. It starts at 0.70 for a kernel size of 10 and

gradually improves to 0.74 at a kernel size of 25. This indicates that larger dilation kernels result in better adherence to desired pathology. In general, The results suggest that increasing the dilation kernel size improves the Dice score while having little impact on the FID. The largest kernel size tested ($k = 25$) results in the best balance between image quality (lower FID) and adherence to desired pathology (higher Dice coefficient). The improvements in Dice are more pronounced compared to the changes in FID, implying that the dilation kernel size has a more significant effect on adherence to desired pathology than on perceptual quality.

Table 5.6: Ablating the effect of dilation kernel size in MedEdit for brain tumors: increasing the dilation kernel size marginally improves both FID and Dice scores. The largest kernel size tested ($k = 25$) results in the best balance between image quality (lower FID) and adherence to desired pathology.

$k, r = 4$	FID	Dice
10	6.81	0.70
15	6.83	0.72
20	6.85	0.73
25	6.80	0.74

Ablating the effect of resample steps. We also ablate the effect of resampling steps for a fixed dilation kernel size. Results are shown in Table 5.7. The results suggest that increasing the number of resample steps does not affect FID or Dice score, implying that the resample steps do not significantly impact the perceptual quality of the generated counterfactuals nor their adherence to desired pathology.

Table 5.7: Ablating the effect of resample steps in MedEdit for brain tumors. Dice and FID are stable as the number of resample steps increases, which suggests the resample steps do not visibly affect neither the realism nor the adherence to the desired pathology.

$r, k = 25$	FID	Dice
1	6.88	0.73
2	6.65	0.74
3	6.70	0.73
4	6.80	0.74

5.2.2 Extending the conditioning process of MedEdit

Here, we investigate the results of our bounding-box conditioning experiments involving tumors (refer to Section 4). Quantitative results are shown in Table 5.8 and qualitative results are presented in Figures 5.6 and 5.7. From Table 5.8, we observe FID scores in the same range as for the type of conditioning that involves strictly defined masks, with *MedEdit* scoring close to the best performing method, namely *Palette*, trailing it by 8%. In the downstream lesion segmentation evaluation, *MedEdit* performs the best, outperforming *Palette* by 40% and 36% in terms of Dice and IoU metrics, respectively. Regarding the combined (1-Dice) * FID metric, *MedEdit* maintains a competitive performance compared to naïve *RePaint*, trailing it by 2%. However, a visual inspection of the generated counterfactuals reveals that for a non negligible amount of cases, these can exhibit a disharmony between the introduced pathology and the rest of the brain in terms of pixel intensity values (except for *SDEdit*). Such examples are shown in Figure 5.7.

Table 5.8: We assess the computational metrics for the bounding-box conditioned versions of *Palette*, Naïve *RePaint*, *SDEdit*, and *MedEdit* using the BraTS 2023 dataset. Best results are shown in **bold** and second-best are underlined. Performance variations relative to the highest-performing method and the second-best method are denoted by \blacktriangle and \blacktriangledown , respectively.

Method	Computational Metrics			
	(1-Dice) * FID \downarrow	FID \downarrow	Dice \uparrow	IoU \uparrow
SDEdit [Men+22]	2.99 \blacktriangle 46%	9.35	0.68	0.58
Palette [Sah+22]	2.90 \blacktriangle 41%	6.45	0.55	0.44
Naïve <i>RePaint</i>	2.05 \blacktriangledown 2%	<u>6.62</u>	<u>0.69</u>	0.60
<i>MedEdit</i> (ours)	<u>2.09</u> \blacktriangle 2%	6.97	0.70	0.60

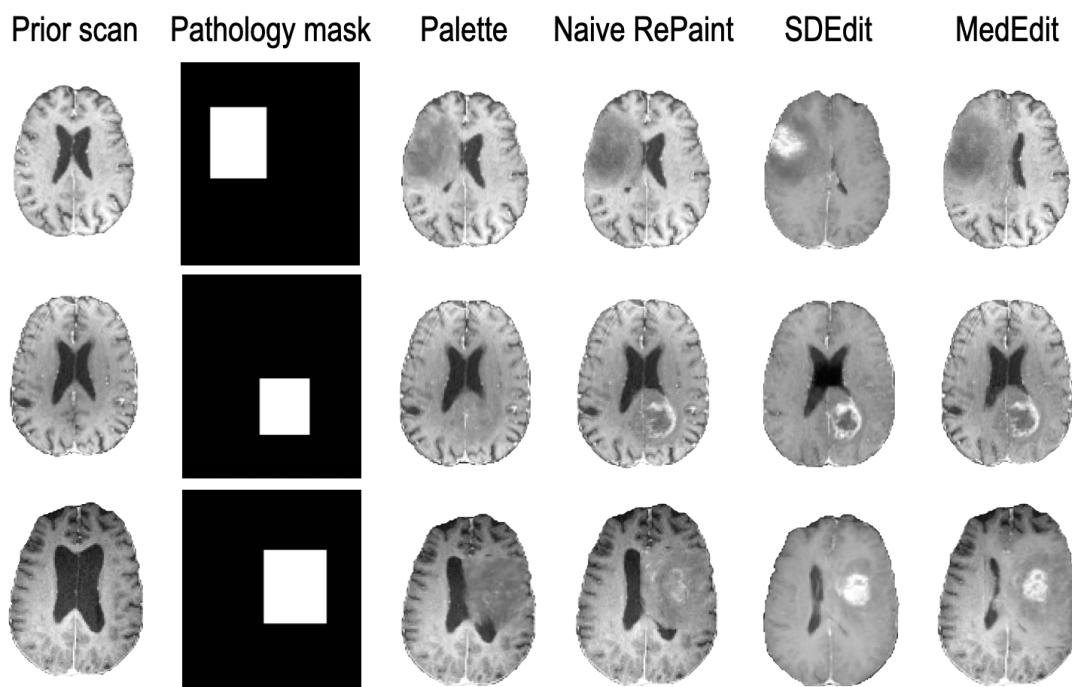


Figure 5.6: Examples of counterfactuals generated by the bounding-box conditioned versions of Palette, Naïve RePaint, SDEdit, and MedEdit using the BraTS 2023 dataset are provided for cases where no visible pixel intensity discrepancies exist between the edited region and the surrounding brain tissue.

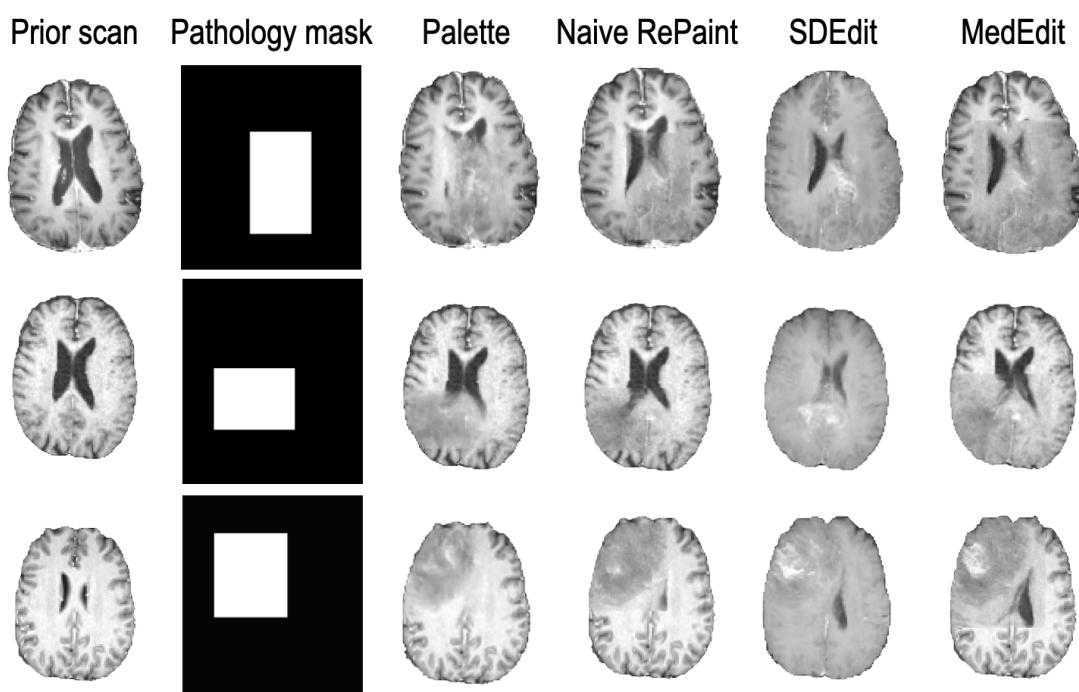


Figure 5.7: Examples of counterfactuals generated by the bounding-box conditioned versions of Palette, Naïve RePaint, SDEdit, and MedEdit using the BraTS 2023 dataset are shown for cases where there is a noticeable visual disharmony in pixel intensity between the edited region and the surrounding brain tissue, with the exception of SDEdit.

6 Discussion

Overall, the conducted experiments demonstrate that *MedEdit* effectively generates counterfactual images representing stroke and tumor lesions, with superior performance observed in stroke cases. More specifically, in comparison to the baseline methods, *MedEdit* can effectively strike a balance between modeling the indirect pathological changes induced by the introduced pathology and maintaining a high fidelity to the prior scan. This capability holds significant potential for advancing the understanding of disease progression and improving data augmentation techniques, which are essential for enhancing diagnostic accuracy, training medical professionals, and supporting personalized treatment planning. Furthermore, in the case of strokes, widely utilized computational metrics such as the Fréchet Inception Distance (FID) and the Dice coefficient, while valuable for fundamental comparisons, are insufficient for capturing the intricate clinical nuances critical in medical diagnostics. FID, for example, evaluates the alignment of overall image distributions but fails to account for essential subtleties, such as the indirect pathological effects associated with stroke lesions, which are crucial for a thorough clinical assessment. This discrepancy is evident in Table ???. Therein, we observe that despite comparable FID scores between Naïve RePaint and *MedEdit*, their clinical assessments differ substantially. *MedEdit* demonstrates a significant advantage in capturing indirect pathological changes, with a 70% improvement over Naïve RePaint. Likewise, while the Dice coefficient concentrates on the accuracy of lesion segmentation, it overlooks other clinically significant features, such as edema and secondary tissue alterations, that are critical for a more comprehensive medical evaluation. The gap between computational assessments and clinical relevance highlights an urgent need for the development of more sophisticated metrics. These new metrics must be capable of comprehensively evaluating both primary and secondary effects of pathological conditions, ensuring that they more accurately reflect clinical observations and align with patient outcomes. Without such advancements, current evaluation methods will continue to fall short of capturing the complexities inherent in medical diagnostics.

In the case of tumors, visually inspecting the generated counterfactuals reveals their lack of quality, rendering them unsuitable for clinical evaluation. The frequent failure of *MedEdit*, as well as the baseline methods, to incorporate the necrotic core of a tumor can likely be attributed to the composition of the training set used for *MedEdit*'s

6 Discussion

diffusion model backbone. A significant portion of this dataset comprises images that capture only the edematous component, omitting the necrotic core. This omission typically occurs when the necrotic core is located in the upper or lower regions of the brain, rather than in the central slices, such as the 90th slice in our case. To address this issue, one can either train exclusively on 2D slices containing the necrotic core, or enhance the granularity of the pathology mask by separating it into core and edema components to ensure greater control over the editing process. Alternatively, training on 3D slices would inherently resolve this issue.

In general, we also note that the FID metric used for the evaluation of *MedEdit* is known to be ideally computed on a large sample size, typically around 20,000 samples [Jay+24]. While it is common in the medical imaging domain to use this metric on a much limited number of samples, which is also the case in this work, we acknowledge the limitations of such practice. Additionally, we used a trained nn-Unet model to evaluate the adherence to the desired pathological change of the tried models and we note that are aware of the biases such trained model can express. In addition, we acknowledge the limitations related to the way we compute the mask of the region to be edited by *MedEdit*, i.e. through diluting the pathology mask. We think that the mask of the region to be edited should ideally be automatically computed by the model, in an explicit or implicit form.

7 Conclusion

In this work, we tackle the challenge of counterfactual image editing in brain scans. We assess current image editing and inpainting techniques, revealing their shortcomings in striking a balance between effectively modeling indirect pathological changes and maintaining the integrity of healthy regions in the original scan. To address these limitations, we introduce *MedEdit*, a novel method that is designed to capture this balance. We evaluated *MedEdit* on the Atlas v2.0 and BraTS 2023 datasets to generate counterfactual stroke and tumor scans, respectively. With regards to the strokes dataset, we found that *MedEdit* outperforms state-of-the-art diffusion-based image editing methods such as *Palette* (by 45%) and *SDEdit* (by 61%). We further validated our findings through an anonymized clinical reader study performed by a board-certified neuroradiologist, evaluating the realism, fidelity to the original scan, and the accurate representation of pathological changes, including induced diseases and their indirect effects. Furthermore, our findings reveal significant discrepancies between computational and clinical metrics, emphasizing the need for the development of clinically-relevant metrics to enable automated and more accurate evaluation of generated counterfactuals. Regarding the tumors dataset, although *MedEdit* outperforms competing methods in terms of computational metrics, qualitative assessment of the generated counterfactuals reveals shortcomings. Specifically, the model often inadequately represents the desired pathology, introducing the tumor’s edematous component without its necrotic core. Further investigation reveals this issue likely stems from the training set used for *MedEdit*’s diffusion model backbone. A significant portion of this training set contains scans with only the edematous component, lacking a necrotic core, which is due to the 2D slicing of the original 3D dataset. To address this, we propose two potential solutions that can be explored in future work: either training exclusively on 2D slices containing a necrotic core, or introducing greater granularity to the pathology mask by separating it into core and edema components. Alternatively, utilizing 3D slices for training would eliminate this limitation. Furthermore, we highlight a key limitation of *MedEdit*: it relies on strictly delineated pathology masks, which limits its flexibility and generalisability. To address this, we extended *MedEdit* to generate counterfactuals conditioned on generic bounding boxes instead of strictly defined pathology masks. We evaluated this approach on both strokes and tumors. For tumors, we found that it led to a degradation of the quality of the generated counterfactuals, often resulting in a

7 Conclusion

disharmony between the introduced pathology and the rest of the brain. For strokes, the results are more promising, as such disharmony is not similarly observed. Future work could focus further on mitigating the discrepancy in performance on different datasets, extending the applications of MedEdit to three-dimensional imaging and include modeling of global indirect pathological changes.

List of Figures

2.1	Anatomy of the brain	4
2.2	Key requirements Generative models are desired to fulfill	6
2.3	General scheme of DDPMs	7
2.4	Overview of RePaint	10
2.5	Visual interpretation of the Dice score	11
2.6	Visual interpretation of the IoU score	12
2.7	Illustration of possible indirect pathological effects of strokes	14
2.8	Visual lesion distribution of Atlas v2.0	15
2.9	Visual illustration of possible indirect pathological effects that can be caused by tumors	16
3.1	Overview of <i>MedEdit</i>	18
5.1	Examples of counterfactuals obtained with Palette, Naïve RePaint, SDEdit and <i>MedEdit</i>	24
5.2	Visual Ablation of resample steps in <i>MedEdit</i> : larger resample steps result in more refined counterfactuals, showcasing more harmony and reduced artefacts.	27
5.3	Examples of counterfactuals for Atlas V2.0 obtained with bounding-box based methods.	28
5.4	Examples of counterfactuals obtained with Palette, Naïve RePaint, SDEdit and <i>MedEdit</i> for the BraTS 2023 dataset.	29
5.5	Examples of counterfactuals that contain only the edematous component, obtained with Palette, Naïve RePaint, SDEdit and <i>MedEdit</i> for the BraTS 2023 dataset.	30
5.6	Examples of counterfactuals for BraTS 2023 obtained with bounding-box based methods for the case where there is no visual disharmony visible.	33
5.7	Examples of counterfactuals for BraTS 2023 obtained with bounding-box based methods for the case where there is a visible visual disharmony.	34

List of Tables

5.1	Computational and clinical metrics on strokes	25
5.2	Ablating the effect of dilation kernel size in MedEdit for brain strokes. .	25
5.3	Ablating the effect of resample steps in MedEdit for brain strokes . .	26
5.4	Computational metrics for the bounding-box conditioning experiments on strokes	28
5.5	Computational metrics for the BraTS 2023 dataset	30
5.6	Ablating the effect of dilation kernel size in MedEdit for brain tumors .	31
5.7	Ablating the effect of resample steps in MedEdit for brain tumors . .	31
5.8	Computational metrics for the bounding-box conditioning experiments on tumors	32

Bibliography

- [Ala+24] M. B. Alaya, D. M. Lang, B. Wiestler, J. A. Schnabel, and C. I. Bercea. *MedEdit: Counterfactual Diffusion-based Image Editing on Brain MRI*. 2024. arXiv: [2407.15270 \[eess.IV\]](https://arxiv.org/abs/2407.15270).
- [Ana24] T. M. Anatomy. *Anatomy of the human brain*. <https://teachmeanatomy.info/neuroanatomy/vessels/ventricles/>. Accessed: September 22, 2024. 2024.
- [Ata+22] M. Atad, V. Dmytrenko, Y. Li, X. Zhang, M. Keicher, J. Kirschke, B. Wiestler, A. Khakzar, and N. Navab. *CheXplaining in Style: Counterfactual Explanations for Chest X-rays using StyleGAN*. 2022. arXiv: [2207.07553 \[eess.IV\]](https://arxiv.org/abs/2207.07553).
- [Bae24] Baeldung. *IOU metric*. <https://www.baeldung.com/cs/object-detection-intersection-vs-union>. Accessed: September 29, 2024. 2024.
- [Bai+21] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al. “The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radio-genomic classification.” In: *arXiv preprint arXiv:2107.02314* (2021).
- [Bar18] S. A. Barnett. *Convergence Problems with Generative Adversarial Networks (GANs)*. 2018. arXiv: [1806.11382 \[cs.LG\]](https://arxiv.org/abs/1806.11382).
- [Ber+23] C. I. Bercea, M. Neumayr, D. Rueckert, and J. A. Schnabel. *Mask, Stitch, and Re-Sample: Enhancing Robustness and Generalizability in Anomaly Detection through Automatic Diffusion Models*. 2023. arXiv: [2305.19643 \[cs.CV\]](https://arxiv.org/abs/2305.19643).
- [cla24] R. master class. *Mass effect*. https://www.radiologymasterclass.co.uk/tutorials/ct/ct_acute_brain/ct_brain_mass_effect. Accessed: September 25, 2024. 2024.
- [Con11] A. Conley. “Ventricular Enlargement.” In: *Encyclopedia of Clinical Neuropsychology*. Ed. by J. S. Kreutzer, J. DeLuca, and B. Caplan. New York, NY: Springer New York, 2011, pp. 2600–2601. ISBN: 978-0-387-79948-3. doi: [10.1007/978-0-387-79948-3_375](https://doi.org/10.1007/978-0-387-79948-3_375).
- [Cou+22] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord. *DiffEdit: Diffusion-based semantic image editing with mask guidance*. 2022. arXiv: [2210.11427 \[cs.CV\]](https://arxiv.org/abs/2210.11427).

Bibliography

- [Den+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “Imagenet: A large-scale hierarchical image database.” In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [DN21] P. Dhariwal and A. Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: [2105.05233 \[cs.LG\]](#).
- [Goo+14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. *Generative Adversarial Networks*. 2014. arXiv: [1406.2661 \[stat.ML\]](#).
- [Gu+23] Y. Gu, J. Yang, N. Usuyama, C. Li, S. Zhang, M. P. Lungren, J. Gao, and H. Poon. *BiomedJourney: Counterfactual Biomedical Image Generation by Instruction-Learning from Multimodal Patient Journeys*. 2023. arXiv: [2310.10765 \[cs.CV\]](#).
- [He+15] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: [1512.03385 \[cs.CV\]](#).
- [Hea24] J. H. Healthcae. *Anatomy of the human brain*. <https://johnshopkinshealthcare.staywellsolutionsonline.com/Conditions/Neuroscience/About/85,P00773>. Accessed: September 22, 2024. 2024.
- [hea24] V. well health. *Midline of the brain*. <https://www.verywellhealth.com/the-brain-can-undergo-a-midline-shift-1720044>. Accessed: September 22, 2024. 2024.
- [Heu+17] 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.” In: *Advances in Neural Information Processing Systems* (2017).
- [HJA20] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: [2006.11239 \[cs.LG\]](#).
- [Ise+21] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. “nnUNet: a self-configuring method for deep learning-based biomedical image segmentation.” In: *Nature methods* 18.2 (2021), pp. 203–211.
- [Jay+24] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar. *Rethinking FID: Towards a Better Evaluation Metric for Image Generation*. 2024. arXiv: [2401.09603 \[cs.CV\]](#).
- [Kaw+23] 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*. 2023, pp. 6007–6017.

Bibliography

- [KW22] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: [1312.6114 \[stat.ML\]](#).
- [Kwo+23] K. Kwolek, D. Grzelecki, K. Kwolek, D. Marczak, J. Kowalczewski, and M. Tyrakowski. “Automated patellar height assessment on high-resolution radiographs with a novel deep learning-based approach.” In: *World Journal of Orthopedics* 14.6 (2023), p. 387.
- [Lie+22] S.-L. Liew, B. P. Lo, M. R. Donnelly, A. Zavaliangos-Petropulu, J. N. Jeong, G. Barisano, A. Hutton, J. P. Simon, J. M. Juliano, A. Suri, et al. “A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms.” In: *Scientific data* 9.1 (2022), p. 320.
- [Lug+22] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool. *RePaint: Inpainting using Denoising Diffusion Probabilistic Models*. 2022. arXiv: [2201.09865 \[cs.CV\]](#).
- [Mei+22] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, et al. “RadImageNet: an open radiologic deep learning research dataset for effective transfer learning.” In: *Radiology: Artificial Intelligence* 4.5 (2022), e210315.
- [Men+22] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. *SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations*. 2022. arXiv: [2108.01073 \[cs.CV\]](#).
- [Met+17] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. *Unrolled Generative Adversarial Networks*. 2017. arXiv: [1611.02163 \[cs.LG\]](#).
- [NJ+24] D. NJ, T. SI, L. E, and T. PM. *Counterfactual MRI Generation with Denoising Diffusion Models for Interpretable Alzheimer’s Disease Effect Detection*. 2024.
- [PAR24] L. Puglisi, D. C. Alexander, and D. Ravì. *Enhancing Spatiotemporal Disease Progression Models via Latent Diffusion and Prior Knowledge*. 2024. arXiv: [2405.03328 \[cs.CV\]](#).
- [Pér+24] F. Pérez-García, S. Bond-Taylor, P. P. Sanchez, B. van Breugel, D. C. Castro, H. Sharma, V. Salvatelli, M. T. A. Wetscherek, H. Richardson, M. P. Lungren, A. Nori, J. Alvarez-Valle, O. Oktay, and M. Ilse. *RadEdit: stress-testing biomedical vision models via diffusion image editing*. 2024. arXiv: [2312.12865 \[cs.CV\]](#).
- [Pix24] Pixels. *Axial brain drawing*. <https://pixels.com/featured/brain-transverse-section-evan-oto.html>. Accessed: September 22, 2024. 2024.
- [Red+16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: [1506.02640 \[cs.CV\]](#).

Bibliography

- [RFB15a] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: [1505.04597 \[cs.CV\]](#).
- [RFB15b] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation.” In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer. 2015, pp. 234–241.
- [Rom+22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: [2112.10752 \[cs.CV\]](#).
- [Sah+22] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi. *Palette: Image-to-Image Diffusion Models*. 2022. arXiv: [2111.05826 \[cs.CV\]](#).
- [Sal+16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. *Improved Techniques for Training GANs*. 2016. arXiv: [1606.03498 \[cs.LG\]](#).
- [San+22] P. Sanchez, A. Kascenas, X. Liu, A. Q. O’Neil, and S. A. Tsaftaris. *What is Healthy? Generative Counterfactual Diffusion for Lesion Localization*. 2022. arXiv: [2207.12268 \[cs.CV\]](#).
- [SH22] T. Salimans and J. Ho. “Progressive distillation for fast sampling of diffusion models.” In: *arXiv preprint arXiv:2202.00512* (2022).
- [Shi+18] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski. “Medical image synthesis for data augmentation and anonymization using generative adversarial networks.” In: *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings* 3. Springer. 2018, pp. 1–11.
- [SME22] J. Song, C. Meng, and S. Ermon. *Denoising Diffusion Implicit Models*. 2022. arXiv: [2010.02502 \[cs.LG\]](#).
- [Syn24] Synapse. *BraTS 2023 Glioma challenge*. <https://www.synapse.org/Synapse:syn51156910/wiki/622351>. Accessed: September 25, 2024. 2024.
- [Sze+15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. *Rethinking the Inception Architecture for Computer Vision*. 2015. arXiv: [1512.00567 \[cs.CV\]](#).
- [Tia+22] C. Tian, X. Zhang, J. C.-W. Lin, W. Zuo, Y. Zhang, and C.-W. Lin. *Generative Adversarial Networks for Image Super-Resolution: A Survey*. 2022. arXiv: [2204.13620 \[eess.IV\]](#).

Bibliography

- [WHO24] WHO. *The top 10 causes of death*. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>. Accessed: September 25, 2024. 2024.
- [Wik24a] Wikipedia. *Brain atrophy*. https://en.wikipedia.org/wiki/Cerebral_atrophy. Accessed: September 25, 2024. 2024.
- [Wik24b] Wikipedia. *Strokes*. https://en.wikipedia.org/wiki/Stroke#cite_note-HLB2014W-5. Accessed: September 25, 2024. 2024.
- [Wya+22] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks. “Anodddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 650–656.
- [XKV22] Z. Xiao, K. Kreis, and A. Vahdat. *Tackling the Generative Learning Trilemma with Denoising Diffusion GANs*. 2022. arXiv: [2112.07804 \[cs.LG\]](https://arxiv.org/abs/2112.07804).
- [Zho+17] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. “Places: A 10 million image database for scene recognition.” In: *IEEE transactions on pattern analysis and machine intelligence* 40.6 (2017), pp. 1452–1464.