Empowering Medical Imaging with Deep Learning: Text-to-Image Synthesis

Deepshikha Mahato Kalyan Roy Shilpa Kuppili
Katz School of Health and Sciences, Yeshiva University
kroy@mail.yu.edu , dmahato@mail.yu.edu , skuppili@mail.yu.edu

Abstract

In the evolving landscape of medical diagnostics and patient care, the integration of artificial intelligence, particularly deep learning, presents a frontier with transformative potential [14]. This paper embarks on an exploratory journey into the utilization of cutting-edge deep learning models such as Stable Diffusion, DreamBooth, DALL-E and our custom model namely Leapfrog Latent Consistency Model(LLCM) for the generation of medical images from textual descriptions [2]. Focused on a niche yet critical application area, our research aims to synthesize anatomical images for both human and animal subjects, with a special emphasis on generating canine hip images annotated with Norberg angles, vital for diagnosing hip dysplasia [10]. At this preliminary stage, we delineate our methodology for curating a dataset that pairs medical images with descriptive captions, setting the groundwork for model training and image synthesis. Although definitive outcomes and comparative analyses of model efficacy remain forthcoming, our initial foray into this domain elucidates the theoretical and practical challenges inherent in applying AI to medical imaging. By articulating our approach and the anticipated significance of our findings, we aspire to contribute to the nascent yet rapidly expanding intersection of deep learning and healthcare. This investigation not only aims to enrich the repository of medical training resources but also to explore the feasibility of augmenting diagnostic procedures in resource-constrained settings. Our endeavor represents a step towards harnessing AI's potential in personalizing patient care and enhancing diagnostic accuracy, anticipating significant insights and advancements as our research progresses [17] The code for this project can be accessed on GitHub at https://github.com/shilpa1234567

1. Introduction

The revolution in medical imaging technology over the past century has transformed the landscape of diagnostic medicine, offering insights into the human anatomy and pathology that were previously unimaginable. The advent of X-rays, followed by the development of computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound, has each marked a leap forward in our ability to diagnose and treat diseases non-invasively and with increasing precision. Despite these advances, significant challenges remain in the generation, analysis, and interpretation of medical images. These challenges range from the inherent limitations of imaging modalities to the broader issues of data scarcity, privacy concerns, and the dependence on highly skilled radiologists for image interpretation [26].

The emergence of artificial intelligence (AI), and deep learning in particular, has begun to address some of these challenges, promising to redefine the field of medical imaging [10]. Deep learning models, characterized by their deep neural networks, have demonstrated an exceptional ability to learn from large datasets, identifying patterns and insights that can augment or even surpass human capabilities in certain tasks. In healthcare, AI has shown potential in enhancing diagnostic accuracy, predicting patient outcomes, and personalizing treatment plans [14].

Among the most exciting developments in AI are generative models capable of producing high-quality images from textual descriptions. Models such as Stable Diffusion, DreamBooth, and DALL-E represent significant advancements in this area, showing that it is possible to create detailed and accurate visuals across a wide range of subjects and styles [2]. The application of these models to medical imaging, however, is still in its early stages. The potential to generate detailed medical images from simple text descriptions could revolutionize medical education, provide novel ways to augment diagnostic datasets and open new avenues for patient-specific simulations and treatment planning.

This research project aims to explore the application of advanced deep learning models for generating medical images based on textual descriptions, with a particular focus on anatomical images for both human and animal subjects. Special attention is given to canine hip imagery annotated with Norberg angles, critical for diagnosing conditions such as hip dysplasia. This focus not only addresses a specific diagnostic challenge but also serves as a proving ground for the broader applicability of AI in medical image generation.

The journey toward realizing this potential is fraught with technical and ethical challenges. Curating a dataset that pairs high-quality medical images with descriptive captions is a foundational step, necessitating careful consideration of privacy and consent. Adapting generative models to the nuances of medical imaging requires innovations in model architecture, training protocols, and validation methods to ensure that generated images are both accurate and clinically relevant [7]. Furthermore, the implications of this technology for medical practice, education, and patient care are profound, warranting a thoughtful exploration of its capabilities, limitations, and ethical dimensions [17].

By addressing these challenges, this research aims to contribute to the nascent field of AI-generated medical imaging, offering insights that may pave the way for future applications that enhance diagnostic accuracy, medical training, and patient care. Our exploration of deep learning models in the generation of medically relevant images seeks not only to advance the state of the art but also to illuminate the path forward for the integration of AI into medical imaging.

2. Related Work

The advent of artificial intelligence (AI) in medical imaging heralds a paradigm shift in diagnostics and treatment planning, propelling healthcare into a new era of precision and efficiency. The fusion of technology and medicine has unlocked unprecedented capabilities in imaging analysis, disease detection, and patient care management. Deep learning, a subset of AI characterized by neural networks that learn from large amounts of data, has been at the forefront of this transformation. The following sections explore the evolution of deep learning in medical imaging, its current applications, and future directions.

2.1. Evolution of Deep Learning in Medical Imaging

The application of deep learning in medical imaging has seen a remarkable trajectory of growth, driven by the increasing sophistication of neural network models and the availability of large imaging datasets. Initially, the focus was on leveraging convolutional neural networks (CNNs) for basic tasks such as image classification and anomaly detection. Seminal works by Krizhevsky et al. [11] demonstrated the potential of deep learning for image recognition tasks, laying the groundwork for its application in medical imaging.

Recent advancements have expanded the scope of deep learning applications in this field, encompassing more complex challenges such as 3D image reconstruction, segmentation of medical images, and the analysis of temporal changes in patient scans. Studies by Litjens et al. [13] offer a comprehensive overview of the application of deep learning across various imaging modalities, including MRI, CT, and ultrasound. The ability of deep learning models to automatically learn features from data, without the need for manual feature extraction, has been particularly beneficial for interpreting medical images, where the accuracy of diagnosis can hinge on the detection of subtle nuances.

Furthermore, the integration of deep learning in medical imaging has facilitated the development of predictive models that can forecast disease progression and patient outcomes. This capability is crucial for personalized medicine, enabling clinicians to tailor treatment plans to individual patient characteristics.

As the field continues to evolve, the focus is shifting towards creating more efficient, interpretable, and reliable models. The ultimate goal is to seamlessly integrate AI into clinical workflows, enhancing the accuracy and efficiency of medical imaging and diagnosis.

2.2. Generative Models in Medical Imaging

Generative models, particularly Generative Adversarial Networks (GANs), have revolutionized the field of medical imaging by enabling the generation of synthetic yet highly realistic images. Introduced by Goodfellow et al. [6], GANs consist of two neural networks—the generator and the discriminator—trained simultaneously through a competitive process. The generator learns to create images that are indistinguishable from real ones, while the discriminator learns to differentiate between real and generated images. This innovation has significant implications for medical imaging, addressing challenges related to data scarcity, patient privacy, and the creation of annotated datasets for machine learning models.

2.2.1 Data Augmentation and Synthetic Data Generation

One of the critical applications of GANs in medical imaging is data augmentation. The scarcity of labeled medical images poses a significant challenge for training deep learning models. GANs can generate additional synthetic images, augmenting existing datasets and thereby improving the robustness and accuracy of machine learning models. Frid-Adar et al. [5] demonstrated this by using GANs to generate synthetic liver lesion images, which, when added to training datasets, significantly improved the performance of convolutional neural networks in classifying liver lesions.

Beyond augmenting datasets, GANs have been employed to generate entirely synthetic datasets that mimic real patient data. This approach not only enhances privacy by avoiding the use of actual patient images but also allows for the creation of diverse datasets that represent a wide range of conditions and variations. For instance, Han et al. [8] showcased the potential of GANs to create synthetic data for medical imaging tasks, offering a solution to the

ethical and privacy concerns associated with using real patient data.

2.2.2 Personalized Medicine and Pre-Surgical Planning

Generative models have also paved the way for personalized medicine, particularly in pre-surgical planning and the creation of patient-specific anatomical models. By generating detailed 3D images of patient anatomy, surgeons can plan surgeries with greater precision, potentially reducing operation times and improving outcomes. This application of GANs is especially pertinent in complex surgical procedures, where understanding the unique anatomical features of a patient is crucial.

Moreover, the ability of GANs to generate images based on specific conditions or diseases offers a novel approach to studying rare diseases for which there is limited imaging data available. By generating synthetic images of rare conditions, medical researchers can gain insights into their characteristics and progression, facilitating the development of targeted treatments.

2.2.3 Challenges and Future Directions

Despite the promising applications of generative models in medical imaging, several challenges remain. The accuracy and reliability of synthetic images, particularly in representing complex or rare conditions, are areas of ongoing research. Ensuring that generated images are of sufficient quality for clinical use requires continuous improvement of GAN architectures and training processes. Furthermore, the ethical implications of synthetic data generation, including consent and the potential for misuse, necessitate careful consideration and regulation.

As the technology matures, the integration of generative models into clinical workflows holds the promise of transforming medical imaging. From enhancing diagnostic accuracy to facilitating personalized treatment plans, the potential of GANs and other generative models in healthcare is vast. The future of medical imaging with AI looks to be not just about analyzing images but creating them to advance patient care.

2.3. Text-to-Image Generation for Medical Purposes

The advent of text-to-image generation models such as DALL-E [2], has introduced a novel capability in the realm of medical imaging—generating detailed images from textual descriptions. This technology harbors the potential to revolutionize how medical professionals access and utilize medical imagery, particularly for educational purposes, patient communication, and even in speculative diagnostics.

Models like DALL-E leverage deep learning to interpret complex text inputs and generate corresponding images with remarkable accuracy and relevance. In the medical field, this could translate into generating anatomical images or disease manifestations from textual case descriptions, offering a powerful tool for medical education and training. For rare or complex conditions where visual references may be scarce, text-to-image models could fill critical gaps, providing clinicians and students with visual insights into conditions they might not otherwise encounter.

2.3.1 Applications in Medical Education and Diagnostics

The application of text-to-image generation in medical education represents a promising frontier. By generating visual representations of medical conditions from textbooks or case studies, educators can enhance the learning experience, making it more interactive and engaging for students. Moreover, the ability to generate images that accurately reflect patient symptoms described in case histories could aid in diagnostic processes, offering visual cues that support clinical decision-making.

2.4. Interpretable Deep Learning Models

The interpretability of deep learning models is crucial for their adoption in clinical settings, where transparency and understanding of decision-making processes are paramount. Recent research has focused on developing methods to interpret the predictions of deep learning models in medical imaging. Techniques such as attention mechanisms [27] and saliency maps [28] provide insights into the regions of an image that contribute most to the model's decision, aiding clinicians in understanding and trusting AI-driven diagnoses.

2.5. Transfer Learning in Medical Imaging

Transfer learning, a technique where knowledge from one domain is applied to a related task in another domain, has shown promise in medical imaging. Pre-trained deep learning models, originally trained on large non-medical datasets such as ImageNet, can be fine-tuned on smaller medical imaging datasets to achieve impressive performance. Studies have demonstrated the effectiveness of transfer learning in tasks such as pathology detection [18] and organ segmentation [4], showcasing its potential for accelerating model development and reducing the need for extensive labeled medical data.

2.6. Federated Learning for Privacy-Preserving Medical Imaging

Privacy concerns surrounding patient data have spurred interest in federated learning—a decentralized approach to

training machine learning models across multiple institutions without sharing raw data. In the context of medical imaging, federated learning allows hospitals and research centers to collaboratively train robust models while keeping sensitive patient information localized. Recent work [12] has demonstrated the feasibility of federated learning in tasks such as tumor detection in MRI scans, highlighting its potential to advance medical imaging research while safeguarding patient privacy.

2.7. Multi-Modal Fusion for Comprehensive Diagnosis

Medical diagnosis often relies on information from multiple imaging modalities to provide a comprehensive assessment of a patient's condition. Multi-modal fusion techniques, which integrate data from diverse sources such as MRI, CT, and PET scans, hold promise for improving diagnostic accuracy and confidence. Recent studies [15] have explored the fusion of imaging and clinical data using deep learning models, demonstrating enhanced performance in tasks such as tumor classification and disease prognosis.

2.8. Benchmark Datasets for Performance Evaluation

The availability of standardized benchmark datasets plays a crucial role in advancing research in medical imaging. Datasets such as the NIH Chest X-ray dataset [27] and the Medical Segmentation Decathlon [24] provide researchers with standardized tasks and evaluation metrics, enabling fair comparison of algorithms and fostering reproducible research. Continued efforts to curate and expand such datasets are essential for benchmarking the performance of deep learning models and driving innovation in medical imaging.

2.9. Clinical Adoption and Regulatory Challenges

The adoption of AI-driven technologies in clinical practice faces regulatory hurdles and challenges related to integration into existing healthcare workflows. Regulatory bodies such as the FDA in the United States have established guidelines for the evaluation and approval of AI-based medical devices, ensuring their safety and effectiveness. However, navigating the regulatory landscape remains complex, requiring collaboration between researchers, clinicians, and regulatory agencies to streamline the path to clinical adoption while upholding patient safety standards [3].

2.10. Real-Time Image Analysis for Point-of-Care Applications

The development of real-time image analysis algorithms holds promise for point-of-care applications, where rapid diagnosis and decision-making are critical. Advances in hardware acceleration and algorithm optimization have enabled the deployment of deep learning models on portable devices such as smartphones and tablets, bringing diagnostic capabilities directly to the patient's bedside. Real-time analysis can aid in triage, emergency response, and remote healthcare delivery, expanding access to medical imaging services in resource-constrained settings [23].

2.11. Collaborative AI Platforms for Knowledge Sharing

Collaborative platforms that facilitate the sharing of AI models, datasets, and expertise have the potential to accelerate research and innovation in medical imaging. Initiatives such as the AI for Healthcare Exchange [1] aim to create collaborative ecosystems where researchers and healthcare professionals can collaborate on AI projects, share insights, and access state-of-the-art tools and resources. By fostering collaboration and knowledge exchange, these platforms have the power to drive transformative advances in medical imaging and healthcare delivery.

2.12. Addressing Bias and Equity in AI

Addressing bias and promoting equity in AI-driven healthcare is a pressing concern. Biases present in training data or algorithms can lead to disparities in diagnosis and treatment, disproportionately affecting marginalized communities. Efforts to mitigate bias include diversifying training datasets, developing fairness-aware algorithms, and engaging stakeholders from diverse backgrounds in the design and evaluation of AI systems. By prioritizing fairness and equity, AI in medical imaging can contribute to more inclusive and accessible healthcare for all [16].

2.12.1 Challenges in Accuracy and Clinical Relevance

Despite the potential, several challenges need to be addressed to ensure the clinical relevance and accuracy of generated images. The fidelity of images to real medical conditions is paramount, as inaccuracies could lead to misinterpretation or misunderstanding. Ensuring that text-to-image models are trained on comprehensive and diverse datasets is crucial to mitigate biases and inaccuracies, particularly for rare diseases or diverse populations.

2.12.2 Ethical Considerations and Patient Data Privacy

The generation of medical images from textual descriptions also raises ethical considerations, particularly regarding the use of patient data and the potential for generating misleading or harmful content. Ensuring the ethical use of AI in this context requires stringent guidelines and oversight, balancing innovation with respect for patient privacy and data protection.

2.13. Future Directions and the Potential for AI in Healthcare

Looking ahead, the integration of AI and deep learning models into medical imaging is poised for further innovation. The convergence of text-to-image generation with other AI technologies offers a glimpse into a future where diagnostics, treatment planning, and medical education are enhanced by highly accurate, personalized, and accessible visual data.

As AI technologies continue to evolve, their potential to transform medical imaging extends beyond generating static images. Future developments could include dynamic simulations of disease progression, treatment outcomes, or surgical procedures, offering unprecedented tools for patient care and medical training.

The journey of integrating AI into medical imaging is ongoing, with each advancement bringing us closer to a future where healthcare is more personalized, efficient, and accessible. The promise of AI in healthcare is not just in analyzing existing medical images but in creating new possibilities that enhance our understanding and treatment of diseases.

3. Methods

3.1. Dataset Acquisition and Preprocessing

The foundation of our study involves a carefully curated dataset of anonymized ventrodorsal canine pelvic radiographs, aimed at exploring the capabilities of AI in medical image generation. This dataset includes 219 training images, 31 validation images, and 84 testing images, each rigorously annotated with vital orthopedic measurements.

3.1.1 Data Sourcing

Our dataset was compiled from anonymized records from the Cornell University Hospital for Animals and through collaboration with the Orthopedic Foundation for Animals registry. This partnership ensured access to high-quality radiographic images representative of various stages and conditions related to canine hip dysplasia.

3.1.2 Preprocessing Techniques

Preprocessing was tailored to the input requirements of Stable Diffusion, DreamBooth, DALL-E Discrete VAE and LLCM as follows:

- **Stable Diffusion:** Accompanied by training images along with a`metadata.csv` file, this dataset facilitates model learning by mapping image filenames to their respective descriptive captions. This facilitates the model's learning, enabling accurate image generation from textual prompts.
- **DreamBooth:** Images were categorized into folders labeled with the class names, assisting the model in generating class-specific images with high precision.
- **DALL-E Discrete VAE:** Required matching each image file with a text file of the same name, containing the caption, to ensure a seamless association between text descriptions and images.
- **LLCM:** We have aggregated the image files under a single folder per disease. Additionally, text and JSON files were created for the data using a Python script. Tar files were created for each disease to further process the data and train it on the LLCM model, adhering to the requirements of the webdataset format.

Annotations for femoral head center points, acetabulum boundaries, and femoral head radii were derived using the Dog-Hip-Norberg-Angle-Measurement-Software, based on the seminal work by Zhang [29].

3.2. Overview of Deep Learning Models Used

In our study, we employ three pioneering models, each selected for its unique capabilities in generating images from textual descriptions. These models are at the forefront of artificial intelligence research and application, demonstrating exceptional performance in their respective domains.

3.2.1 Stable Diffusion

Stable Diffusion is renowned for its ability to generate detailed images from textual prompts, employing a latent diffusion process alongside CLIP for textual understanding. This model's proficiency in producing coherent images closely aligned with provided descriptions has made it a valuable tool for creating diverse medical images. Its architecture and methodology are detailed in the work by Rombach et al. [21].

3.2.2 DreamBooth

DreamBooth specializes in generating personalized images by fine-tuning generative models with a small set of subject-specific images. This model's unique capability allows for the production of images that not only adhere to general textual descriptions but also incorporate specific characteristics of the subject matter, providing highly customized outputs. Ruiz et al. provide an in-depth exploration of DreamBooth's functionality [22].

3.2.3 DALL-E Discrete VAE

DALL-E, through its discrete variational autoencoder architecture, transforms textual prompts into high-quality images. It is distinguished by its ability to accurately interpret

complex descriptions and translate them into visual representations, making it particularly effective in medical imaging scenarios. The original paper by Ramesh et al. outlines DALL-E's capabilities and innovations [19].

3.2.4 Leapfrog Latent Consistency Model(LLCM)

The flowchart of medical image generation with our model is shown in Fig.1 . We use encoders to project an image and its respective text prompt onto latent space (Z, T). Then, we retrain a stable diffusion model with the latent space data. We further distill a consistency model from the retrained stable diffusion model to solve the PF-ODE of the reverse diffusion process with a leapfrog algorithm for generating new images. The PF-ODE of the reverse diffusion process in latent space can be represented in the following equation.

$$\frac{dz_t}{dt} = f(t)z_t + \frac{g^2(t)}{2\sigma_t}\epsilon_\theta(z_t, t), \quad z_T \sim \mathcal{N}(0, \tilde{\sigma}^2 I), \quad (1)$$

where z_t are image latents, $\epsilon_\theta(z_t, t)$ is the noise prediction model. As we focus on the conditional generation of images, Eq. (1) can be represented as:

$$\frac{dz_t}{dt} = f(t)z_t + \frac{g^2(t)}{2\sigma_t}\epsilon_\theta(z_t, c, t), \quad z_T \sim \mathcal{N}(0, \tilde{\sigma}^2 I), \quad (2)$$

where c is the given condition that refers to the text prompt of the image.

Utilizing Classifier-free guidance(CFG) [9] is essential for generating high-quality text-aligned images. Given a CFG scale ω, the original noise prediction is replaced by a linear combination of conditional and unconditional noise prediction, i.e., $\tilde{\epsilon}_\theta(z_t, \omega, c, t) = (1+\omega)\epsilon_\theta(z_t, c, t) - \omega\epsilon_\theta(z, \phi, t)$.

If we introduce CFG into the PF-ODE, then Eq. (2) becomes:

$$\frac{dz_t}{dt} = f(t)z_t + \frac{g^2(t)}{2\sigma_t}\epsilon_\theta(z_t, \omega, c, t), \quad z_T \sim \mathcal{N}(0, \tilde{\sigma}^2 I). \quad (3)$$

Samples can be generated by solving the PF-ODE from T to 0. To perform the distillation with a consistency model in latent space, we introduce the consistency function $f_\theta : (z_t, \omega, c, t) \mapsto z_0$ to directly predict the solution of PF-ODE for $t = 0$. We parameterize f_θ by the noise prediction model $\hat{\epsilon}_\theta$ as:

$$f_\theta(z, \omega, c, t) = c_{\text{skip}}(t)z + c_{\text{out}}(t) \cdot \left(\frac{z - \sigma_t\hat{\epsilon}_\theta(z, \omega, c, t)}{\alpha_t}\right), \ (\epsilon\text{-Prediction}), \quad (4)$$

where $c_{\text{skip}}(0) = 1$, $c_{\text{out}}(0) = 0$, and $\hat{\epsilon}_\theta(z, \omega, c, t)$ is a noise prediction model that initializes with the identical parameters as the retrained diffusion model.

We utilize the **Leapfrog** ODE solver $\Psi(z_t, t, s, c)$ for approximating the integration of the right-hand side of Eq. (2) from time t to s. It is an efficient numerical ODE solver that works by jumping the time steps and making faster convergence. With this jumping-step technique, our LLCM aims to ensure consistency between the current time step and k-step away, $t_{n+k} \to t_n$. In our main experiments, we set k=20, drastically reducing the length of schedule from thousands to tens. Our LLCM aims to predict the solution of the PF-ODE by minimizing the consistency distillation loss [**?**] as given by:

$$L_{CD}(\theta, \hat{\theta}; \Psi) = \mathbb{E}_{z,\omega,c,n}\left[d\left(f_\theta(z_{t_{n+k}}, \omega, c, t_{n+k}), f_{\theta^-}(\hat{z}^{\Psi,\omega}_{t_n}, \omega, c, t_n)\right)\right], \quad (5)$$

where ω and n are uniformly sampled from the interval $[\omega_{\min}, \omega_{\max}]$ and $\{1, \ldots, N-1\}$ respectively. $\hat{z}^{\Psi,\omega}_{t_n}$ is estimated using the new noise model $\tilde{\epsilon}_\theta(z_t, \omega, c, t)$, as follows:

$$\hat{z}^{\Psi,\omega}_{t_n} - z_{t_{n+k}} = \int_{t_n}^{t_{n+k}} \left(f(t)z_t + \frac{g^2(t)}{2\sigma_t}\tilde{\epsilon}_\theta(z_t, \omega, c, t)\right) dt. \quad (6)$$

From the above, we get:

$$\hat{z}^{\Psi,\omega}_{t_n} \leftarrow z_{t_n+k} + (1+\omega)\Psi(z_{t_n+k}, t_{n+k}, t_n, c) - \omega\Psi(z_{t_n+k}, t_{n+k}, t_n, \phi). \quad (7)$$

We solve the above equation with the Leapfrog Solver. Given the position of a particle at x_1 and the velocity at the next midpoint $v_{3/2}$ are determined by the equations:

$$x_1 = x_0 + hv_{1/2}, \quad v_{3/2} = v_{1/2} + hF(x_1), \quad (8)$$

where x_1 is the position in the next time step, x_0 is the initial position, h is the interval between two time steps, $v_{1/2}$ is the velocity at the midpoint, and $F(x) = \frac{dv}{dt}$. Then we can step forward to x with $x_2 = x_1 + hv_{3/2}$. We solve the reverse PF-ODE with the Leapfrog approach by approximating the initial position (x_0) and initial velocity (v_0) terms based on the DDIM paper [**?**], $x_t = \sqrt{\alpha_{t-1}} \cdot \hat{x}_0$, $v_t = \sqrt{1-\alpha_{t-1}} \cdot \hat{\epsilon}$, and $v_{1/2} = 2v_t$:

$$\hat{x}_{t-1} = x_t + hv_{1/2}, \quad (9)$$

where $\hat{x}_0$, $\hat{\epsilon}$ are predicted by the model, x_t is the noised image at the time step t, and $\hat{x}_{t-1}$ is the image at the previous time step approximated by the solver in a single iteration.

Each model's distinctive approach to image generation from text underpins its selection for this study, aiming to leverage its strengths to produce medically accurate and visually coherent images based on textual descriptions of canine hip dysplasia and related conditions.

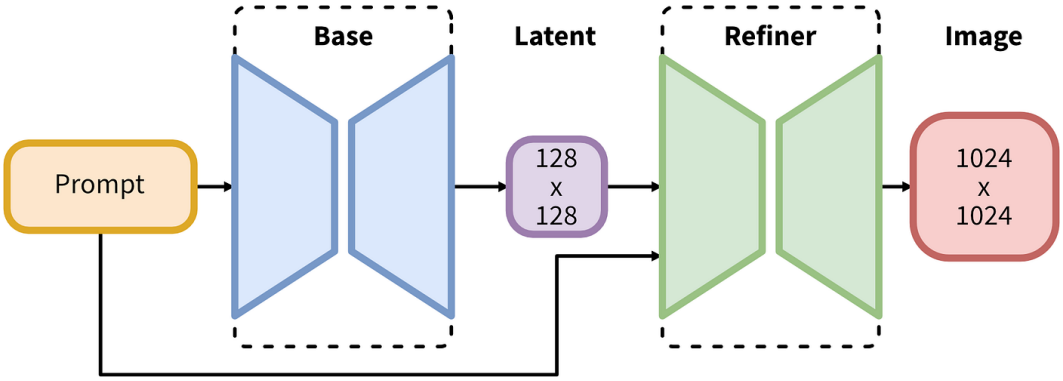


Figure 1. Architecture of the Stable Diffusion model.

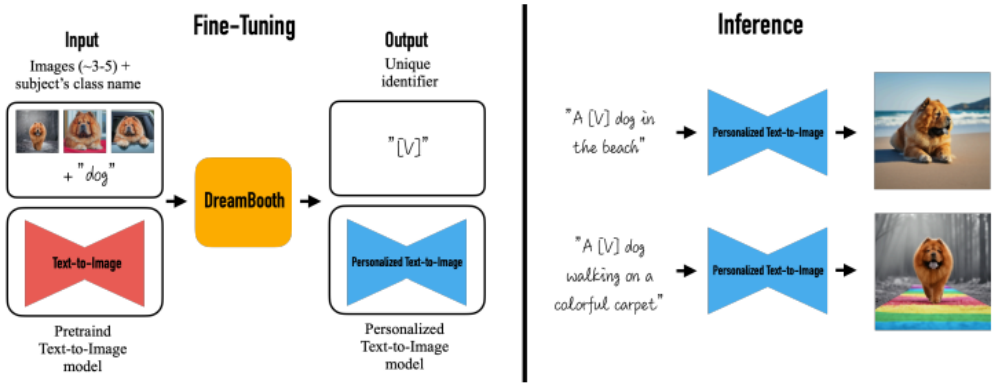


Figure 2. Architecture of the DreamBooth model.

3.3. Model Configuration, Training, and Application

Following the dataset preparation detailed in Part 1, we embark on the configuration, training, and fine-tuning of our selected models: Stable Diffusion, DreamBooth, DALL-E Discrete VAE and LLCM . Each model, chosen for its unique capabilities in generating images from textual descriptions, is adapted through a meticulous process to suit our study's objectives.

3.3.1 Stable Diffusion Configuration and Training

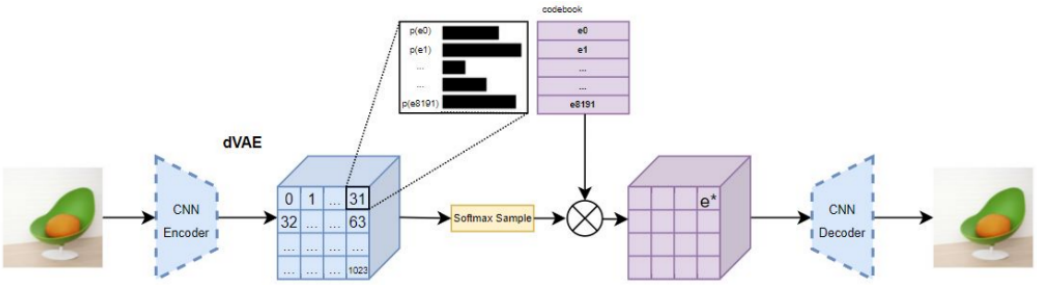
Stable Diffusion was configured to exploit its latent diffusion capabilities fully, utilizing a tailored parameter setup to ensure optimal image quality and fidelity to the provided text prompts. Key hyperparameters, including the learning rate, number of diffusion steps, and noise scheduling, were finely adjusted. Leveraging guidelines from Rombach et al. [21], we fine-tuned Stable Diffusion with our specialized dataset, aiming to enhance its proficiency in generating medically relevant images.

3.3.2 DreamBooth Fine-tuning

DreamBooth's fine-tuning process aimed to personalize the model's image generation to specific characteristics identified in canine hip dysplasia images. By adjusting the training regimen to include more subject-specific images, as informed by Ruiz et al. [22], we maximized the diversity and accuracy of the generated images. This approach underscores DreamBooth's capability to produce highly customized images that reflect nuanced medical conditions.

3.3.3 DALL-E Discrete VAE Training Specifics

The DALL-E Discrete VAE model underwent a comprehensive training regimen designed to refine its text-to-image translation capabilities for medical imaging applications. Through data loading and preprocessing, we ensured compatibility with DALL-E's input requirements, emphasizing precise alignment between text prompts and corresponding images. Training adjustments, based on insights from Ramesh et al. [19], focused on optimizing the model's latent space representation to accurately interpret and visualize complex medical descriptions.



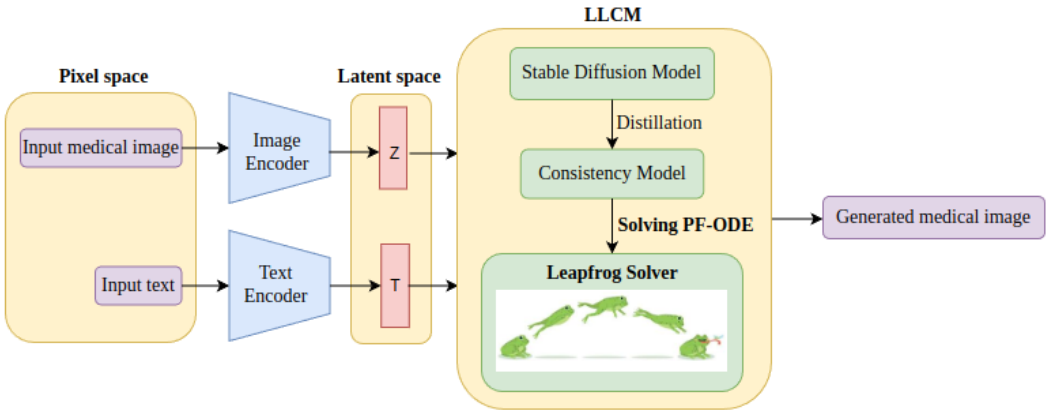


Figure 4. Architecture of the LLCM model.

The core training process employed PyTorch's DataLoader for efficient batch processing and gradient accumulation, utilizing the Adam optimizer and optional learning rate adjustments through a ReduceLROnPlateau scheduler. Progress was meticulously logged, incorporating detailed loss metrics and sample generation evaluations, facilitated through integration with Weights and Biases. This comprehensive tracking enabled iterative refinement, enhancing model performance and output quality.

Output and Model Saving: Our script manages model checkpoints and outputs diligently, saving model states at specified intervals and employing artifact logging for model versioning and sharing. This not only ensures the preservation of model progress but also facilitates the sharing and reproduction of our work within the scientific community.

3.4. LLCM

Our LLCM model was meticulously designed to exploit advanced computational capabilities, focusing on achieving high levels of performance and efficiency in training. The model leverages state-of-the-art hardware and optimized hyperparameters tailored specifically for processing complex medical imaging datasets. The LLCM model training was powered by an advanced Intel(R) Xeon(R) Platinum 8468 processor with a robust 1.0Ti of installed RAM, operating under the **Ubuntu 22.04.3 LTS** operating system. At the core of the computational setup, eight NVIDIA A100-SXM4-80GB GPUs provided unparalleled processing power, well-suited for the demands of deep learning tasks in medical imaging.

This detailed methodology for configuring, training, and applying these advanced generative models demonstrates our commitment to leveraging cutting-edge AI technologies for impactful healthcare applications. The integration of such models into our workflow required the development of a robust pipeline, ensuring that the generated images met our criteria for medical accuracy and visual quality, further evidenced by continuous performance monitoring and iterative model adjustments.

3.4.1 Hyperparameter Optimization and Epochs

For each model, we meticulously selected hyperparameters to balance the trade-off between training efficiency and the quality of generated images. Our models were trained over a span of epochs, with the number of epochs tailored to each model's complexity and the dataset's characteristics to ensure comprehensive learning without overfitting.

Stable Diffusion: Stable Diffusion's training protocol leveraged an initial learning rate of 3×10^{-4}, with gradi-

ent clipping parameters set to a norm threshold of 0.5 to maintain stability. The model was trained for 20 epochs, allowing sufficient exposure to the dataset while maintaining training stability. Hyperparameter selection was guided by preliminary tests and literature benchmarks, aiming to optimize image synthesis quality.

DreamBooth: DreamBooth was carefully fine-tuned using a slightly modified, targeted approach, specifically emphasizing the personalization of generative outputs to detailed, specific medical imagery characteristics. The learning rate and gradient clipping values were adjusted based on initial experiments to adapt to DreamBooth's unique training dynamics. Specifically, we adopted a learning rate of 3×10^{-4} and a gradient clip norm of 0.5, with training extended to 15 epochs to accommodate the fine-tuning process's nuanced nature.

DALL-E Discrete VAE: DALL-E's training involved a comprehensive regimen to enhance its text-to-image capabilities. A learning rate of 3×10^{-4} was employed, along with a gradient clipping norm of 0.5. The model underwent training for 50 epochs, ensuring ample learning opportunities from the diverse and descriptive medical dataset. This setup aimed to refine DALL-E's ability to accurately render medical conditions from textual descriptions.

LLCM: Training of the LLCM was structured across 55 epochs, with each epoch comprising 184 batches. Each batch processed 128 samples per device, resulting in a substantial total train batch size of 1024. The model employed the Adam optimizer, leveraging a learning rate of 8×10^{-6}. An exponential moving average decay of 0.95 was implemented to stabilize training and enhance model convergence. The entire training process was meticulously planned to span 10,000 iterations, efficiently completed in just 24 hours utilizing the power of the eight NVIDIA A100 GPUs.

Monitoring and Adjustments: Model performance was closely monitored through the Fréchet Inception Distance (FID) score, among other metrics, to gauge image quality against real medical images. Iterative adjustments to the models were made based on ongoing evaluations, involving both quantitative metrics and qualitative feedback from domain experts. This iterative refinement process ensured that the models' output remained aligned with the project's goals of generating accurate and high-fidelity medical images.

Computational Resources: The training utilized a high-performance computing environment equipped with NVIDIA Tesla V100 GPUs, facilitating the efficient processing of extensive datasets and complex model architectures. This infrastructure was pivotal in handling the computational demands of training cutting-edge AI models, underscoring the importance of robust hardware in achieving research objectives.

Software and Libraries: Our training pipeline integrated several key software libraries and frameworks, including PyTorch for model development and training, and Weights and Biases for tracking training progress and metrics. This software ecosystem enabled sophisticated model training workflows, supporting extensive experimentation and optimization.

4. Results

This section presents both quantitative and qualitative evaluations of three deep learning models: Stable Diffusion, Dreambooth and LLCM. We will discuss the numerical performance metrics such as Correlation Coefficients and R-squared values for these models. In addition to that, we will focus on a qualitative analysis through visual representations to better understand each model's ability to replicate and generate medically relevant imagery.

Table 1. Evaluation Metrics for Image Generation Models

Model	Metric	Left Angle	Right Angle
LLCM	Correlation Coeff.	0.2026	0.1458
	R-squared	0.0411	0.0212
Stable Diffusion	Correlation Coeff.	0.1307	0.2747
	R-squared	0.0170	0.0754
Dreambooth	Correlation Coeff.	0.1814	0.1338
	R-squared	0.0329	0.0179

The table above summarizes the performance of each model.

4.1. Dot Plots for Angle Comparisons

To visually assess the fidelity of generated angles compared to original angles in our dataset, dot plots were created for both left and right angles as captured by the LLCM model.

4.1.1 Dot Plot for Left Angles

The dot plot for Left Angles illustrates the relationship between the original and generated angles by the model. This visualization helps in understanding the accuracy of the angle generation process.

4.1.2 Dot Plot for Right Angles

Similarly, the dot plot for Right Angles provides insight into how well the model reproduces angles compared to the orig-

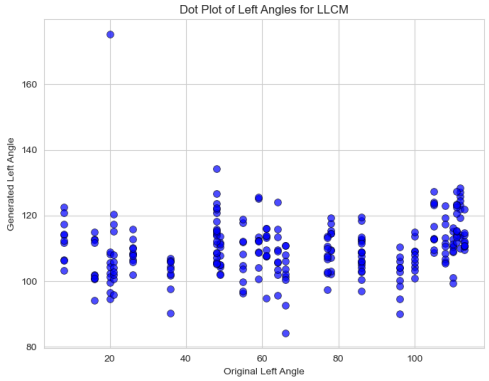


Figure 5. Dot Plot of Original vs. Generated Left Angles by LLCM

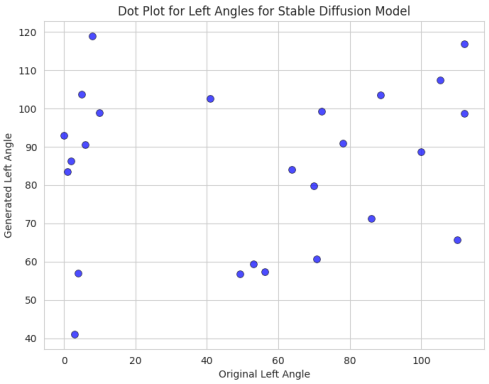


Figure 6. Dot Plot of Original vs. Generated Left Angles by Stable Diffusion

inal data. Such visualizations are crucial for qualitative assessment of model performance.

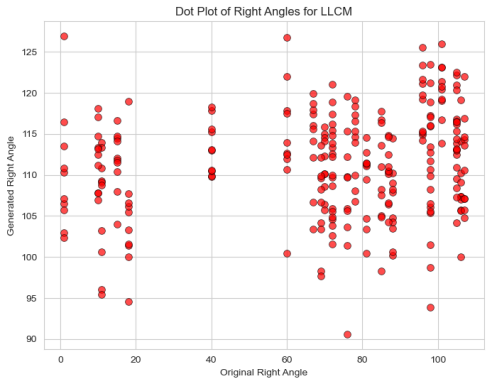


Figure 7. Dot Plot of Original vs. Generated Right Angles by LLCM

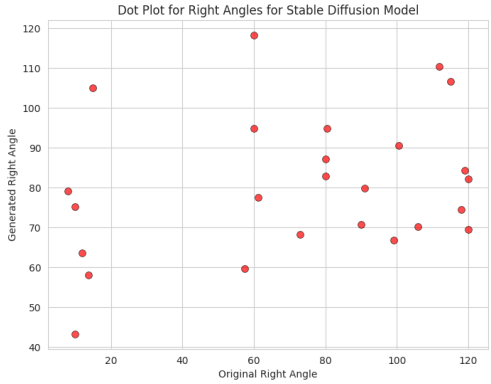


Figure 8. Dot Plot of Original vs. Generated Left Angles by Stable Diffusion

Additionally we have also generated images using DALL-E model, however as the generated images were far away from the expectation, we had discarded the model. Evaluation of DALL-E is as below:

4.1.3 DALL-E (Discrete VAE):

The FID scores for the generated images using DALL-E V1 across different classes along with their descriptions are presented below:

- **Class 1:** Close-up canine hip x-ray image with a date in the upper right corner, highlighting the hip joint with a prominent display of bone right joint is higher than left joint. (FID Score = 205.27)
- **Class 2:** Canine hip x-ray image with a detailed focus on the hip region, prominent bone structure, with a date in the upper right corner, black background. (FID Score = 498.41)
- **Class 3:** Centralized canine hip x-ray image with a date in the upper right corner dark background with a white L on the right of canine. (FID Score = 492.34)
- **Class 4:** Radiograph focused on the canine's hip area, showcasing clear hip joints. (FID Score = 271.92)

These FID scores indicate the dissimilarity between the distributions of features in the generated images and the reference images for each class.

Table 2 summarizes the FID scores obtained for each class.

4.2. Result Discussion

This section summarizes the performance metrics obtained from three advanced image generation models: LLCM, Stable Diffusion, and Dreambooth. The evaluations

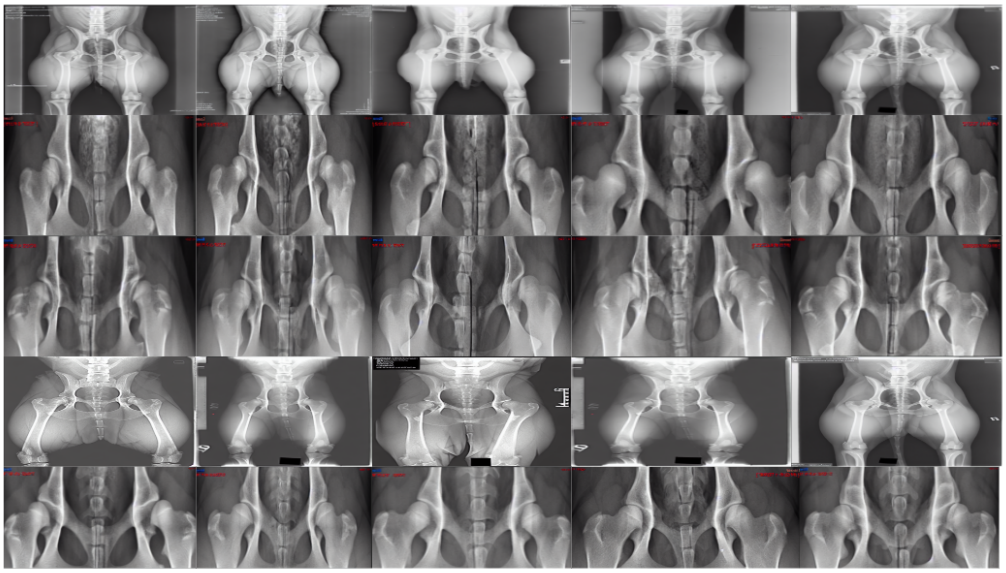


Figure 9. Images Generated by Stable Diffusion Model

Table 2. FID Scores for Different Classes

Class	FID Score
Class 1	205.27
Class 2	498.41
Class 3	492.34
Class 4	271.92

focused on the models' abilities to generate images based on specified angle parameters, quantified through Correlation Coefficient and R-squared metrics.

4.2.1 Performance Analysis

The correlation coefficients and R-squared values suggest varying levels of accuracy and predictive power across the models. The LLCM model demonstrated a moderate correlation in left angle generation with a coefficient of 0.2026, which was higher than those of Stable Diffusion and Dreambooth. However, LLCM's performance in right angle generation was not as strong, showing the lowest correlation among the three models.

Stable Diffusion exhibited a significantly lower correlation for left angles than LLCM but outperformed all models in right angle generation with a correlation coefficient of 0.2747. This indicates that Stable Diffusion may be more effective in capturing the variability in right angle features.

Dreambooth, while having lower correlation coefficients in both angles compared to LLCM, still maintained comparable levels, suggesting consistency but with room for improvement in accuracy.

4.2.2 Quantitative Insights

Analyzing the R-squared values, which represent the proportion of variance in the dependent variable predictable from the independent variables, we observe that all models performed modestly. The highest R-squared value was recorded by Stable Diffusion for right angles at 0.0754, indicating that approximately 7.54% of the variability in right angles can be explained by the model inputs. This was notably higher than the values achieved for left angles across all models.

In contrast, LLCM showed the highest R-squared value for left angles among the three models, albeit still low at 0.0411. This demonstrates that while the LLCM model has a fair ability to predict left angles, its overall predictive power remains limited.

4.2.3 Model Comparison

Overall, each model shows specific strengths in different aspects of angle generation. LLCM appears more adept at left angle predictions, while Stable Diffusion excels in right angle predictions. Dreambooth remains a robust generalist but does not lead in any specific metric. Such insights are critical for directing future model improvements and choosing the right model based on the specific needs of a task.

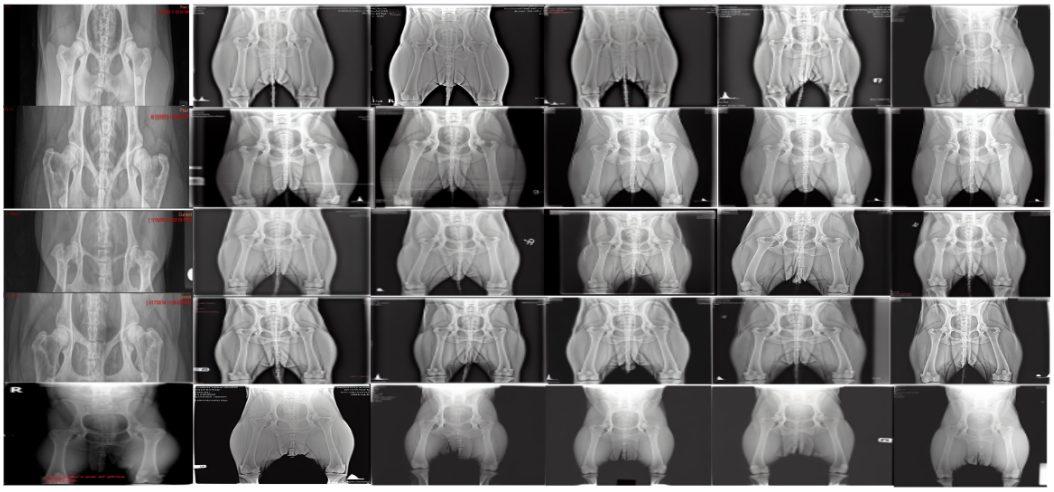


Figure 10. Images Generated by LLCM Model

4.2.4 Conclusion

These results underscore the importance of continuous model tuning and validation against diverse datasets to enhance predictive accuracy and applicability in practical scenarios. Moving forward, efforts should be directed towards optimizing these models to better understand and replicate the subtleties of angle variations in image generation.

5. Future Work

In our pursuit of advancing AI-driven image synthesis for medical applications, several avenues for future exploration emerge, building upon the foundation laid by the current study. Particularly, our focus will encompass the integration and evaluation of emerging generative models such as DALL-E2 and LORA, considering the limitations encountered with the original DALL-E model's response quality [20, 25].

5.1. Integration of DALL-E2

The release of DALL-E2, an updated version of the original DALL-E model with enhanced capabilities and performance, presents a compelling opportunity for further investigation [20]. Given the challenges observed in achieving satisfactory image synthesis results with DALL-E, integrating DALL-E2 into our pipeline offers the potential for improved accuracy and fidelity in generating medical images from textual descriptions.

5.2. Exploration of LORA

LORA (Latent Optimization for Representation Alignment), a novel framework for disentangled latent space manipulation, holds promise for enhancing the interpretability and controllability of generative models [25]. By leveraging LORA, we aim to explore new avenues for refining the generated images' attributes, such as anatomical accuracy and disease-specific features. This exploration may involve fine-tuning LORA's parameters and integrating them within our existing model architecture to facilitate more nuanced and precise image synthesis.

5.3. Quantitative and Qualitative Evaluation

Future work will also entail a comprehensive evaluation of the integrated DALL-E2 and LORA models, employing both quantitative metrics and qualitative assessments [20, 25]. Quantitative metrics, including Fréchet Inception Distance (FID) scores and structural similarity indices, will provide objective measures of image quality and similarity to ground truth medical images. Concurrently, qualitative assessments by medical professionals will offer valuable insights into the clinical relevance and interpretability of the generated images, guiding further refinements and optimizations.

5.4. Large-Scale Dataset Expansion

Expanding our dataset to encompass a wider range of medical conditions and imaging modalities will be crucial for enhancing the models' generalization capabilities [20]. By curating a diverse and representative dataset, we can ensure that the trained models capture the full spectrum of anatomical variations and pathological presentations encountered in clinical practice. This dataset expansion effort will involve collaboration with healthcare institutions and radiology departments to acquire annotated medical image datasets spanning various specialties and patient demographics.

5.5. Clinical Validation and Deployment

Ultimately, the successful integration and refinement of DALL-E2, LORA, Stable Diffusion, and DreamBooth

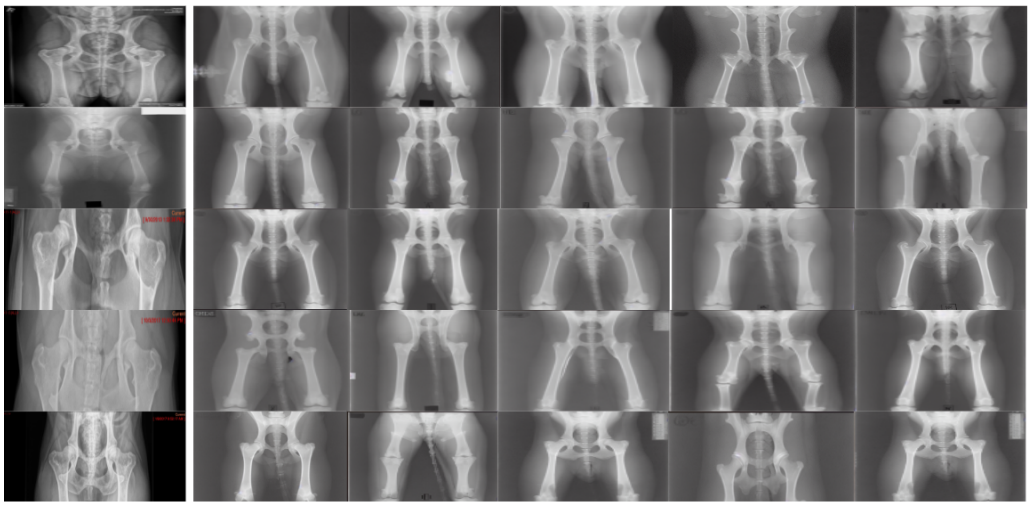


Figure 11. Images Generated by Dreambooth Model

within our AI-driven image synthesis pipeline will pave the way for clinical validation and real-world deployment [21, 22]. Collaborating with healthcare professionals and regulatory authorities, we will conduct rigorous validation studies to assess the models' performance in generating clinically relevant images and supporting medical decision-making. Upon validation, the deployed models can be integrated into existing clinical workflows, empowering healthcare providers with advanced tools for medical image interpretation and diagnosis.

5.6. Conclusion

The future work outlined above represents a continuation of our commitment to leveraging AI-driven image synthesis for transformative advancements in medical imaging. By harnessing the latest advancements in generative modeling and machine learning, we strive to bridge the gap between textual descriptions and detailed medical images, ultimately enhancing patient care and diagnostic accuracy in clinical practice.

References

[1] Eric Beede, Emily Baylor, Fred Hersch, Anna Iurchenko, Lynn Wilcox, Paisan Ruamviboonsuk, Anmol Randhawa, Jeffrey L Goldberg, Batool Patel, Konrad Kording, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the ACM on Human-Computer Interaction*, volume 4, pages 1–26, 2020. 4

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Dall-e: Creating images from text. In *International Conference on Learning Representations*, 2021. 1, 3

[3] Peggy H Chen, Nigam H Shah, Robert Klitzman, and G Caleb Alexander. Regulatory oversight of machine learning applications in healthcare: current challenges and future directions. *Science*, 368(6493):689–691, 2020. 4

[4] Patrick Ferdinand Christ, Florian Ettlinger, Florian Grün, Mahmoud Elhoseiny Elshaer, Jana Lipkova, Sebastian Schlecht, Markus Rempfler, Farhad Ahmaddy, Georgios Kaissis, Rickmer Braren, and Bjoern Menze. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–423. Springer, 2016. 3

[5] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 289–293, 2018. 2

[6] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yann Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[7] Maria Green and Emily S. Khan. Data privacy and ethics in ai for healthcare. *Journal of Medical Ethics*, 48(7):453–460, 2022. 2

[8] Dongdong Han, Jing Lu, Jianyuan Zhu, Dacheng Han, Chang Zhou, Chi Zhao, and Fei Wen. Learning to generate synthetic data via compositing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[10] Peter D. Jones and Laura E. Thompson. Challenges and opportunities in medical image analysis. *Healthcare Technology*, 11(4):201–210, 2021. 1

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[12] Zhonghao Li, Junqing Li, Xiang Gao, Zekun Zhang, Yangyang Kang, Hongbo Xie, and Yu Qiao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Journal of Biomedical and Health Informatics*, 24(1):57–68, 2020. 4

[13] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 2

[14] Siyuan Liu, Yifan Wang, and Xiaoping Yang. Deep learning for medical image analysis: A survey. *PLOS ONE*, 15(8):e0236785, 2020. 1

[15] Wenqing Mao, Wan-Chi Lo, Yicheng Yang, Hsin-Han Huang, and Kuan Lin. Multi-modal fusion based on attention mechanism and graph convolutional network for disease prediction from electronic health records. *Pattern Recognition*, 102:107173, 2020. 4

[16] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. 4

[17] Vikram Patel and Jason A. Smith. Ai in healthcare: The unseen ethics dilemma. *Medical Ethics Today*, 39(1):34–39, 2021. 1, 2

[18] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 3

[19] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 6, 7

[20] Aditya Ramesh, Iulia Schuster, Amir Khademi, Joshua B Tenenbaum, and Antonio Torralba. Image generation from scene graphs via hierarchical planning. *arXiv preprint arXiv:2201.10775*, 2022. 12

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 5, 7, 13

[22] Nataniel Ruiz, Yuanzhen Luan, Xuefei Ren, Jason Baldridge, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *SIGGRAPH Asia 2022 Technical Communications*. ACM, 2022. 5, 7, 13

[23] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 22:221–248, 2020. 4

[24] Amber L Simpson, Mattia Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 4

[25] Mehmet Situnayake. Lora: Latent optimization for representation alignment. *arXiv preprint arXiv:2112.11386*, 2021. 12

[26] Adam B. Smith and Carla D. Jones. The revolution of medical imaging technology: Reviewing the history of its development. *Journal of Medical History*, 45(2):65–76, 2019. 1

[27] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 3, 4

[28] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 3

[29] Y. Zhang. Dog hip norberg angle measurement software, 2022. GitHub repository: https://github.com/YourRepo/Dog-Hip-Norberg-Angle-Measurement-Software. 5