LLM-Based Image Caption Augmentation for Text-Guided Radiology Image Generation Using Diffusion Models with Pretrained Encoders

Ahmed Abul Hasanaath, Hamzah Luqman
College of Information and Computer Science
King Fahd University of Petroleum and Minerals, Khobar, KSA

***Abstract*—The landscape of medical data analysis is evolving with advancements in natural language processing (NLP) and generative models. This research aims to enhance the interpretability of radiology images by integrating textual descriptions and medical concepts. We propose a baseline approach for text-to-image generation in radiology, focusing on two key contributions. First, we leverage large language models (LLMs) to transform image captions, enhancing their descriptive power and aligning them more closely with radiology images. Second, we experiment with both a standard UNet model and a custom VGG16-based UNet model to improve feature extraction and image generation quality. Our approach begins with extensive dataset pre-processing, narrowing down to the most relevant CUIs and filtering low-quality images using entropy measures. We then employ RadBERT as a text encoder to process textual inputs, aiming to capture diverse modalities and enrich feature representations. The diffusion model, central to our methodology, uses a linear noise schedule and classifier-free guidance to generate high-quality, text-aligned images. Experimental results demonstrate the effectiveness of our methods, with the transformed captions and VGG16-based UNet model achieving the best performance in terms of Fréchet Inception Distance (FID). Despite limitations due to hardware constraints, our approach sets a robust baseline for future research in this domain. The findings suggest that integrating advanced NLP techniques with generative models holds significant promise for improving the analysis and interpretation of radiology images, paving the way for more accurate and informed healthcare decisions.**

***Index Terms*—Diffusion Models, Text to Image, Radiology**

I. Introduction

The landscape of medical data analysis and interpretation is continually evolving, driven by advancements in natural language processing (NLP), generative models, and the integration of multi-modal datasets. This research aims to address a pivotal challenge in radiology by enhancing the interpretability of medical images through the integration of textual descriptions and medical concepts. The selected problem revolves around leveraging the capabilities of large language models (LLMs) for improved understanding of radiology images, encapsulating both textual and visual information.

Radiology images, while rich in visual information, often lack a standardized and comprehensive means of representation. Bridging the semantic gap between textual descriptions and medical concepts associated with these images is crucial for enabling more effective analysis and interpretation [1], [2], [3]. Our focus is on predicting Unique Identifiers (CUIs) using image captions, ultimately forming a graph-based representation that integrates information from diverse modalities [4], [5].

The motivation behind solving this problem lies in the potential to revolutionize medical image analysis. By enhancing the synergy between textual descriptions and visual data, our approach aims to provide a more holistic understanding of radiology images. This holistic representation has the potential to empower healthcare professionals with richer insights, aiding in more accurate diagnoses and treatment planning.

Recent works in the literature have explored the intersection of NLP, generative models, and medical data processing. Tang et al. evaluated the performance of LLMs in medical evidence summarization, emphasizing the nuanced challenges in generating consistent and informative summaries [1]. In the realm of medical data processing, Mithun et al. refined automated cohort building from radiology reports using NLP, showcasing the effectiveness of deep learning models [2]. Additionally, the synthesis of medical images through generative models has gained attention, with innovative approaches like TGEDiff and PathLDM addressing segmentation challenges and achieving state-of-the-art results in text-conditioned image generation.

This research makes a significant contribution by proposing a baseline approach to text-to-image generation in radiology. Our contributions are twofold. Firstly, we explore the effect of leveraging LLMs to transform captions for an enhanced description of the image, which when provided to a diffusion model, is able to capture the relationship between the caption and the image. Secondly, we propose a UNet model with a VGG16 model based pretrained encoder as an approach to enhance the standard UNet's feature extraction capabilities. Our approach fills a gap in existing literature regarding generative modeling for medical data, and aims to revolutionize the generation and interpretation of radiology images, paving the way for more informed and accurate healthcare decisions.

The rest of the paper is divided as follows: Section II provides an overview of related literature, Section III details our approach and experimental setup, Section IV presents our results, Section V presents our limitations, Section VI proposes ideas for future work and finally, Section VII concludes the paper.

TABLE I
EXAMPLES OF CAPTION TRANSFORMATION USING LLM

Caption	CUIs	CUI Description	Transformed Caption
Head CT demonstrating left parotiditis.	C0040405	X-Ray Computed Tomography	X-ray, Parotid gland, Side view, Left parotitis.
5.1 cm x 3.4 cm x 4 cm multiloculated hepatic abscess in the inferior posterior aspect of the right lobe	C0040405	X-Ray Computed Tomography	X-ray, Liver, Side, Multiloculated abscess.
Preop CT showing left orbital floor fracture	C0040405	X-Ray Computed Tomography	X-ray, Orbit, Floor of Orbit from below, Fracture.
Postop 22-month CT scan (sagittal): Posteriorly the graft seated in a sound bone	C0040405	X-Ray Computed Tomography	X-ray, Spine, Sagittal View, Normal.
CT demonstrating partially obstructed airway. CT: computed tomography.	C0040405	X-Ray Computed Tomography	Computed Tomography, Airway, Oblique View of, Partially Obstructed Airway.

II. RELATED WORK

A. Medical Data and Language modeling

In the realm of natural language processing (NLP), recent studies have highlighted the significant impact of large language models (LLMs) such as GPT-3.5 and ChatGPT in medical evidence summarization [1]. Tang et al. systematically evaluate the zero- and few-shot performance of these models across six clinical domains, employing both automatic metrics and human evaluations. The findings underscore a notable disparity between automatic metrics and the quality of summaries, emphasizing the nuanced challenges of LLMs in medical evidence summarization. The study introduces a comprehensive terminology for error types, revealing potential vulnerabilities in generating inconsistent or misleading summaries. Additionally, it sheds light on LLMs' struggles in identifying salient information, particularly in longer textual contexts, providing crucial insights into the intricacies of their performance. In another exploration, Clusmann et al. address the challenges in the clinical application of LLMs, presenting MultiMedQA as an innovative benchmark for medical question answering [6]. Recognizing the limitations of automated evaluations, the authors propose a human evaluation framework, assessing dimensions such as factuality, comprehension, reasoning, potential harm, and bias. Despite the success of models like Flan-PaLM2 in achieving state-of-the-art accuracy, human evaluations reveal critical gaps, prompting the introduction of instruction prompt tuning.

In the domain of medical data processing, Mithun et al. contribute to the refinement of automated cohort building from radiology reports using Natural Language Processing (NLP) [2]. Their study leverages deep learning (DL) models, including Bi-LSTM and BERT, to predict lung cancer report relevance in a thoracic disease management group cohort. The authors showcase the effectiveness of DL models, particularly the Pre-trained BERT model, in accurately selecting lung cancer reports from radiology datasets. Moving into the realm of tailoring transformer-based language models for radiology NLP applications, Yan et al. present RadBERT, a family of bidirectional encoder representations from transformers fine-tuned for radiology [7]. The study demonstrates that RadBERT variants consistently outperform baselines across abnormal sentence classification, report coding, and report summarization tasks. Dessi et al. explore the application of Deep Learning and Word Embeddings in identifying morbidity types within clinical records [8]. Their preliminary findings suggest that traditional machine learning methods outperform the combination of Deep Learning approaches using word embeddings.

B. Generative Models for Medical Data

In the realm of medical image generation, researchers have been actively exploring innovative approaches to enhance segmentation and generate high-quality images. Dong et al. introduced TGEDiff [3], a groundbreaking end-to-end framework designed for medical image segmentation using denoising diffusion models. TGEDiff integrates a textual attention mechanism into the diffusion model, addressing the limitations of perceptual fields by incorporating a multi-kernel excitation module. Yellapragada et al. contribute to the evolving landscape of histopathology image generation by proposing PathLDM, the first text-conditioned Latent Diffusion Model [9]. Leveraging histopathology reports as guidance, the authors fuse image and textual data to enhance the generation process. Incorporating GPT's capabilities for contextual richness, PathLDM achieves a state-of-the-art FID score for text-to-image generation. Ali et al. explore the synthesis of medical images, particularly employing neural diffusion models, representing a novel avenue in the medical domain [10]. Their approach utilizes a pre-trained DALLE2 model for lungs X-Ray and CT image generation and a stable diffusion model trained on X-Ray images.

Toda et al. delve into the applications of GANs in medical imaging, proposing a GAN architecture for enhancing the resolution of MRI scans [11]. Their approach involves a dual neural network engaging in a competitive process for super-resolution, showcasing promise in significantly enhancing the resolution of medical images, particularly in tuberculosis detection using MRI scans. Gupta et al. provide a comprehensive review of the transformative impact of GANs in radiology,

emphasizing their versatility in image synthesis, accelerated acquisitions, artifact reduction, and abnormality identification [12]. The authors illustrate various GAN applications in radiologic image analysis and discuss the clinical potential, prospective applications, and considerations for radiologists. This review serves as a valuable resource for understanding the evolving landscape of GAN applications in radiology.

C. *Multi-Modal Datasets for Radiology*

The exploration of Natural Language Processing (NLP) techniques coupled with image generation in radiology relies on two key multi-modal datasets: ROCOv1 and ROCOv2 [4], [5]. ROCOv1 consists of diverse medical and multimodal imaging data identified from publications on the PubMed Central Open Access FTP mirror. Automatically categorized into radiology or non-radiology, the dataset includes captions, keywords, UMLS Semantic Types (SemTypes), and UMLS Concept Unique Identifiers (CUIs), offering a foundation for generative models in image captioning, categorization, and retrieval systems.

ROCOv2 represents an evolution, addressing the demand for high-quality labeled data for advanced medical image analysis. It features radiological images, curated medical concepts, and captions from the PubMed Open Access subset. Expert manual curation ensures precision in annotations related to clinical modality, anatomy (X-ray), and directionality (X-ray). With 80,080 images, including 35,852 new additions since 2018, ROCOv2 is a comprehensive resource. It has proven instrumental in tasks like concept detection and caption prediction in ImageCLEFmedical 2023, showcasing its utility in training image annotation models and multi-label image classification. ROCOv2 also paves the way for pre-training models in the medical domain and evaluating deep learning models in multi-task learning scenarios. Together, these datasets form a robust foundation for exploring NLP-image generation synergies in radiology. The dataset is pre-split into train, test and validation. A total of 1935 CUIs are used to describe the images in the dataset.

III. Approach

A. *Dataset Pre-processing*

Our approach started off with an analysis of the dataset. It is important that the dataset can fully represent the domain; analyzing the data distribution not only allows us to remove data that may not be represented adequately but also remove noisy data. This section elaborates on our pre-processing steps.

As detailed in Section II-C, the dataset consists of images with their corresponding CUIs and UMLS Concept Unique Identifier (CUIs). During this stage, the CUIs served as a tool for understanding how the dataset is distributed. A total of 1935 CUIs were used to describe the images. Figure 1 illustrates the number of of images the top 50 CUIs are used to describe. As apparent from the figure, the number of images a CUI describes falls off steeply after the first 10 CUIs. Table II summarizes the frequency of CUIs over the entire dataset;

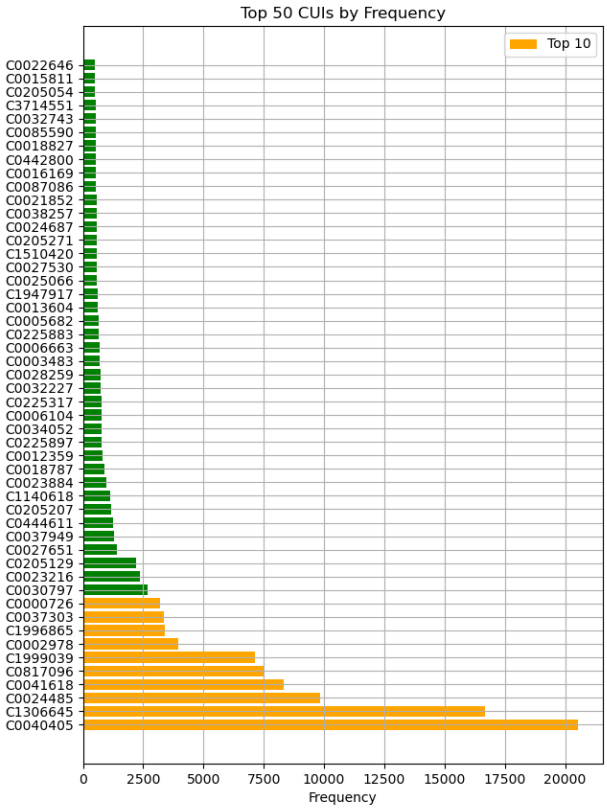

Fig. 1. illustrating the number of images the top 50 CUIs are used to describe.

99% of the CUIs describe less than 874 images. Based on this finding, we chose to limit the dataset to the top 10 CUIs.

Having narrowed the dataset down, we then studied the quality of the images in order to filter out low-quality images; such images would serve as noise during training. The quality of an image is assessed using a measure of entropy, which quantifies the randomness or information content in the image. First, a histogram of the image is computed. The histogram is calculated for intensity values ranging from 0 to 255, representing the distribution of pixel intensities in the image. Next, the entropy is calculated based on this histogram. The formula for entropy calculation is given by:

$$\text{Entropy} = -\sum_{i=0}^{255} p_i \log_2 p_i \tag{1}$$

where p_i represents the probability of occurrence of intensity level i in the image. This probability is computed by dividing the histogram value for intensity level i by the total number of pixels in the grayscale image. The entropy formula quantifies the amount of uncertainty or disorder in the distribution of pixel intensities. It sums over all intensity levels where the histogram value is greater than zero,

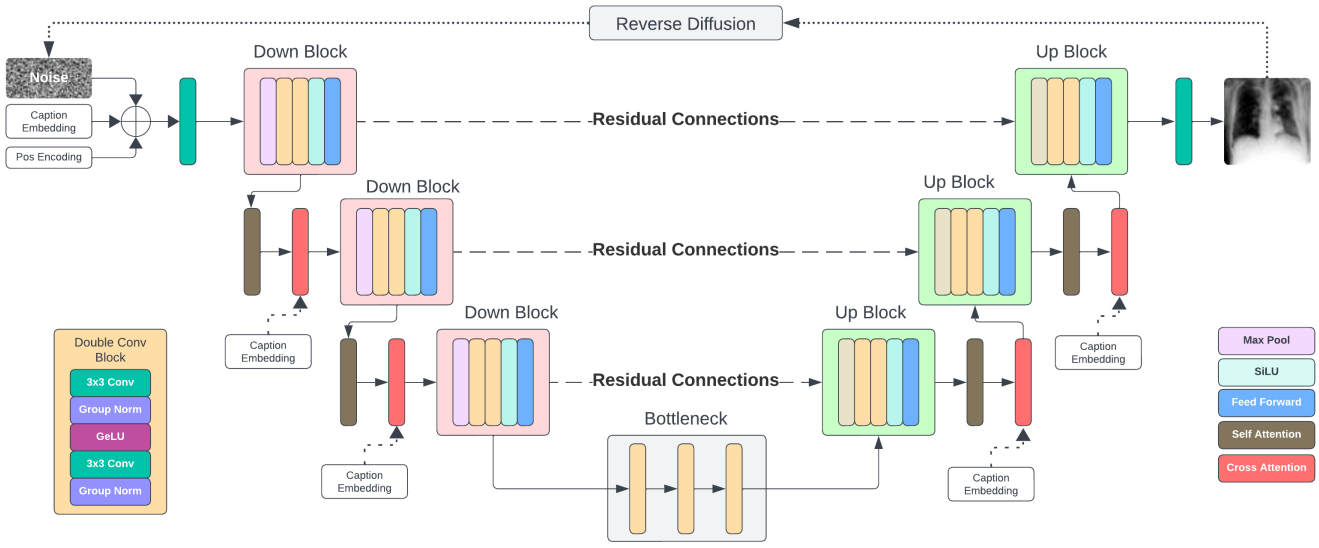

Fig. 2. Our standard UNet model architecture. The VGG UNet architecture follows a similar design; the only difference being that the Down Blocks are replaced with pretrained layers from the VGG16 model.

ensuring that only meaningful intensity levels contribute to the entropy calculation. The logarithm base 2 is used to measure the entropy in bits, providing a clear interpretation of the information content in the image. A higher entropy value indicates greater complexity or variability in the pixel intensities, while a lower entropy suggests a more uniform or predictable distribution. This entropy measure helps in assessing the overall randomness and richness of information present in the image, which is crucial for evaluating image quality and identifying noisy or low-quality images.

TABLE II
CUI FREQUENCY DISTRIBUTION ACROSS PERCENTILES IN THE DATASET. 75% OF THE CUIS DESCRIBE LESS THAN 44 IMAGES. 99% OF THE CUIS DESCRIBE LESS THAN 874 IMAGES.

25th Percentile	11
50th Percentile	19
75th Percentile	44
99th Percentile	874

B. *Caption Transformation*

In our approach to text-to-image generation for radiology images, we experimented with four different setups for the captions used to guide the image generation process:

1) **Dataset Captions**: We used the original image captions provided in the dataset.
2) **Combined Captions**: We used a combination of the image caption and the corresponding CUI (Concept Unique Identifier) description from the dataset.
3) **LLM Transformed Captions**: We utilized a Large Language Model (LLM) to rewrite the captions in a specific template designed to provide a structured and standardized format. The Llama3 model was used for the caption transformation. Algorithm 1 presents the prompt used to generate the transformed caption.
4) **LLM Transformed Captions with Image Caption and CUI Description**: A simple concatenation of Transformed Captions, Image Caption and CUI Description.

The structured format was provided to the LLM which aimed to ensure that each transformed caption followed the sequence: <Image type>, <Body part>, <View of Image>, <Patient's Condition>. If the patient's condition was not mentioned, it was labeled as "Normal". This transformation was intended to make the captions more consistent and to remove any extraneous information. Table I presents examples of the original captions, CUIs, CUI descriptions, and the transformed captions. This method of caption transformation ensures a uniform and simplified representation of the image captions, enhancing the quality and consistency of the text-to-image generation process.

C. *Model Architecture*

*1) **UNet**:* In our study, we experimented with two distinct UNet-based architectures: a standard UNet and VGG16_UNet. Both models were designed to synthesize radiology images from textual descriptions. Notably, both architectures share the same decoder design, which facilitates a comparative analysis of their encoding mechanisms. Our model architecture is illustrated in Figure 2.

a) Standard UNet: The standard UNet model is constructed using a straightforward yet effective architecture, consisting of convolutional, down-sampling, up-sampling, and attention layers.

- Encoder: The encoder is built from DoubleConv blocks, followed by Down blocks which include max-pooling and

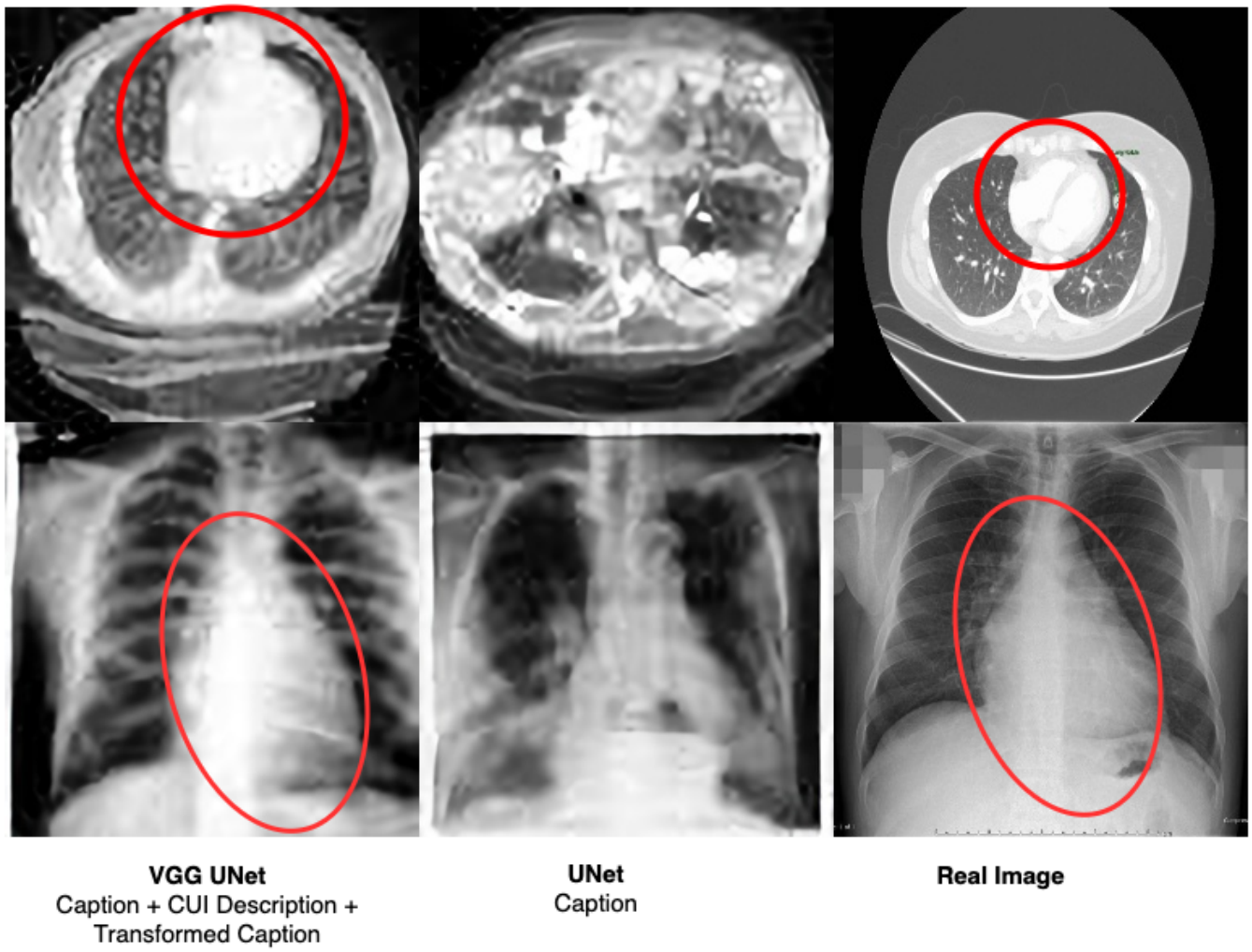

Fig. 3. Comparison between images generated by our best (first column) and worst model (second column). The last column is the real image. The red circles show the ability of our best model to recreate details in the generated image. The worst model is unable to generate such details. Note that the generated images have been upscaled using the super resolution model in [13].

additional DoubleConv layers. These blocks reduce the spatial dimensions while increasing the feature depth.

- Attention Mechanisms: SelfAttention and CrossAttention layers are integrated to allow the model to focus on important parts of the input images and text sequences. The SelfAttention and CrossAttention blocks come after each down sampling and upsampling block. The SelfAttention block focuses on attending to image features, whereas the the CrossAttention blocks focus on attending to important features between the image and its corresponding caption.
- Bottleneck: Intermediate DoubleConv layers provide further feature processing between the encoder and decoder.
- Decoder: The decoder consists of Up blocks, which up-sample the feature maps and concatenate them with corresponding feature maps from the encoder via skip connections from the Encoder. This helps in recovering spatial details lost during down-sampling. Additionally, the embedding layers integrate positional encodings and text embeddings, allowing the model to align textual information with visual features effectively.

b) VGG16_UNet: The VGG16_UNet model leverages the pre-trained VGG16 network as the backbone for its encoder. This architecture takes advantage of the robust feature extraction capabilities of VGG16, followed by a series of convolutional and attention layers to generate high-quality images.

- Encoder: The encoder consists of the initial layers of the VGG16 network, divided into three slices. These slices replace the Down Blocks from the standard UNet model. The slices are as follows:
  - Slice 1: First four layers of VGG16.
  - Slice 2: Layers 5 to 9 of VGG16.
  - Slice 3: Layers 10 to 16 of VGG16.
- Attention Mechanisms: SelfAttention and CrossAttention layers are added after each slice with the same goal as the standard UNet model; attend to important features when it comes to both the image and the text

Algorithm 1 Caption Transformation using LLM

Input: caption

Output: transformed_caption

prompt ← caption + ” ” + CUI description + ”\n====================\n ”

prompt ← prompt + ”Above is a medical caption. Do the following: \n”

prompt ← prompt + ”Rewrite the caption as follows: <Image type>, <Body part>, <View of Image>, <Patient’s Condition>.\n”

prompt ← prompt + ”If the condition is not mentioned, just say Normal in place of Patient’s Condition\’\n”

prompt ← prompt + ”Don’t use abbreviations; example, instead of CT, say computed tomography. Try to be general, for example, instead of saying radiograph, just say X-ray. \n”

prompt ← prompt + ”If the caption does not have the information specified (image type, body part, view), return ’invalid’\n”

prompt ← prompt + ”Do not say anything except the new caption \n”

response ← ollama.generate(model=’llama3’, prompt=prompt)

transformed_caption ← response[’response’]

return transformed_caption

- Bottleneck: A series of DoubleConv layers to further process the features between the encoder and decoder, similar to the standard UNet.
- Decoder: The decoder design is identical for both the standard UNet and VGG16_UNet architectures.

c) Text Encoder: In both UNet models, we experimented with the RadBERT [7] text encoder and the CLIP text encoder [14] to process textual inputs. RadBERT, a pre-trained BERT model fine-tuned on radiology reports, provides robust text embeddings that are projected into a lower-dimensional space suitable for integration with the UNet architecture. CLIP on the other hand excels at learning visual concepts given natural language supervision. The text embeddings are incorporated into the network through CrossAttention layers, allowing the model to align textual information with visual features effectively.

*2) **Diffusion**:* The diffusion process is central to our method, enabling the transformation of noise into meaningful radiology images guided by textual descriptions. This process involves gradually adding noise to an image and then learning to reverse this process using a neural network.

a) Noise Schedule: We employ a linear noise schedule to control the amount of noise added at each timestep. This schedule linearly interpolates between a starting noise level, $\beta_{start} = 1 \times 10^{-4}$, and an ending noise level, $\beta_{end} = 0.02$, over a fixed number of steps, typically 1000. This linear progression ensures a smooth increase in noise, making it easier for the model to learn the denoising process.

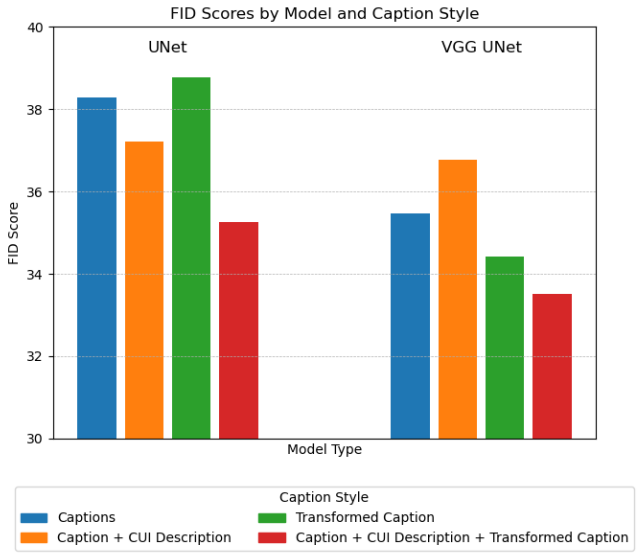

Fig. 4. FID Scores from Table III illustrated.

b) Classifier-Free Guidance: To enhance the quality of generated images and ensure they are well-aligned with the input text, we use classifier-free guidance. This technique involves training the model with both conditional (text) and unconditional (no text) data. During inference, we generate two sets of predictions: one conditioned on the text and one unconditioned. By adjusting the weighting between these predictions, we can guide the generation process more strongly towards the textual description, resulting in more accurate and relevant images.

c) Exponential Moving Average (EMA): To stabilize training and improve the robustness of the model, we implement an Exponential Moving Average (EMA) of the model weights. EMA maintains a moving average of the model's parameters, which are updated as follows:

$$\theta_{\text{EMA}}^{(t)} = \alpha \theta_{\text{EMA}}^{(t-1)} + (1 - \alpha)\theta^{(t)} \tag{2}$$

where α is a decay rate close to 1 (commonly 0.999 or higher), $\theta^{(t)}$ are the current model parameters, and $\theta_{\text{EMA}}^{(t)}$ are the EMA parameters. Using the EMA parameters for inference typically results in more stable and reliable performance compared to using the raw model parameters.

d) Sampling Process: The sampling process starts with a random noise image and iteratively refines it using the trained UNet model. At each timestep, the model predicts the noise present in the current image, which is then subtracted to produce a cleaner image. This iterative denoising continues until the final image is generated. The use of classifier-free guidance and EMA ensures that the generated images are both high-quality and closely aligned with the provided textual descriptions.

D. *Experimental Setup*

First, we experimented with both standard UNet and VGG UNet architectures to determine the best text input for our models. The three types of text inputs tested were the image caption from the dataset, the image caption combined with the CUI description, and the transformed caption generated by a language model. For this step, we used the RadBERT encoder. Next, we experimented to determine the best text encoder by comparing CLIP and RadBERT. In this step, the best text input from the determined in the previous step was used. Finally. we experimented with different image sizes in order to determine the best image size to train the diffusion model on. (These experiments were not carried out due to GPU issues. The RAM was running out leading too ResourceExhaustedError)

All experiments were conducted on an Ubuntu 22 machine equipped with an NVidia A5000 GPU with 24GB VRAM. A batch size of 16 was employed. We used a learning rate of 3×10^{-4} with a linear decay scheduler. The models were trained for 100 epochs, optimizing for the Mean Squared Error (MSE) loss between the actual noise and the noise predicted by the Diffusion Model. The training objective is defined as follows:

$$\mathcal{L}_{\text{denoise}} = \frac{1}{N} \sum_{i=1}^{N} \|\epsilon_i - \hat{\epsilon}_i\|_2^2 \tag{3}$$

where ϵ_i is the actual noise and $\hat{\epsilon}_i$ is the predicted noise.

We evaluated the models using Fréchet Inception Distance (FID). FID measures the statistical similarity between the feature distributions of real and generated images in a pre-trained network. Specifically, it calculates the distance between the means and covariances of the feature representations of real and generated images, thereby assessing how closely the generated images replicate the distribution of real images:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}\right) \tag{4}$$

where μ_r and Σ_r are the mean and covariance of the real images, and μ_g and Σ_g are the mean and covariance of the generated images.

IV. Results and Discussion

The experiments conducted with various text inputs and different models yielded significant insights into the impact of text processing on the FID (Fréchet Inception Distance) scores. Table III summarizes the FID scores for both UNet and VGG UNet models using different combinations of text inputs processed by the RadBERT text encoder. The results clearly indicate that the incorporation of additional contextual information through the CUI (Concept Unique Identifier) descriptions and transformed captions generated by a language model notably enhances the quality of the generated images, as reflected by the lower FID scores. Table III summarizes our results and Figure 4 illustrates the same results.

For the UNet model, using only the image caption as the text input resulted in a baseline FID score of 38.2849. The introduction of the transformed caption slightly improved the FID score to 37.2050, indicating that the language model's ability to refine and contextualize the caption positively influenced the image generation process. However, combining the original caption with the CUI description slightly degraded the performance (FID score of 38.7760). This suggests that while the CUI descriptions add valuable information, they may also introduce some noise or redundancy when not effectively integrated. The most significant improvement was observed when all three text inputs—caption, CUI description, and transformed caption—were used together, resulting in the lowest FID score of 35.2501. This demonstrates that a holistic approach to textual information, leveraging both structured and unstructured data, can substantially improve the generative model's performance.

The VGG UNet model exhibited a similar trend, with the lowest FID score of 33.5230 achieved when all text inputs were combined. Notably, this model consistently outperformed the UNet model across all text input variations, with a baseline FID score of 35.4735 using only the caption. The inclusion of the CUI description improved the FID to 34.4119, and the transformed caption alone resulted in a score of 36.7792. The superior performance of the VGG UNet model suggests that its architecture is more effective at leveraging the additional textual information provided by the CUI descriptions and transformed captions. These findings highlight the importance of integrating diverse and enriched textual inputs to enhance the fidelity of generated images, pointing to promising directions for future research in multimodal image generation.

Figure 3 compares our best model (VGG UNet with Caption + CUI Description + Transformed Caption) with the worst model (UNet with Caption + CUI Description) and the ground truth. Note that due to the model generating very small images (64x64), we used the super resolution model proposed in [13] to upscale our generated images. The red ellipses in the figure indicate the ability of our best model in recreating important details in the image simply from a textual input. For example, the image caption for the images in the first row was "Computerized tomography of the chest in 2021 showing an increase in the lung nodule size to 15 mm". The nodule is clearly visible in the ground truth image as indicated by the red circle. This detail is present in the image generated by the best model. The worst model (conditioned on image caption and CUI description) fails to create a discernable image. This can be attributed to the VGG UNet's superior feature extraction ability along with the transformed caption's ability to enhance the image description. A similar case is illustrated in the second row.

V. Limitations

The research faced several limitations primarily due to hardware constraints, specifically the VRAM capacity of the NVIDIA A5000 GPU used, which has 24GB of VRAM. These constraints significantly impacted various aspects of the experimental setup and model training.

TABLE III

FID YIELDED FROM EXPERIMENTS WITH DIFFERENT TEXT INPUT PROCESSED; IMAGE CAPTION IS THE UNALTERED CAPTION FROM THE DATASET, CUI DESCRIPTION IS THE UNALTERED DESCRIPTION OF THE IMAGE CUI AND TRANSFORMED CAPTION IS THE TRANSFORMED TEXT GENERATED BY AN LLM WHEN FED WITH THE IMAGE CAPTION AND THE CUI DESCRIPTION

Model	**Text Input**	**Text Encoder**	**FID**
UNet	Caption	RadBERT	38.2849
UNet	Transformed Caption	RadBERT	37.2050
UNet	Caption + CUI Description	RadBERT	38.7760
UNet	Caption + CUI Description + Transformed Caption	RadBERT	35.2501
VGG UNet	Caption	RadBERT	35.4735
VGG UNet	Transformed Caption	RadBERT	36.7792
VGG UNet	Caption + CUI Description	RadBERT	34.4119
VGG UNet	**Caption + CUI Description + Transformed Caption**	**RadBERT**	**33.5230**

TABLE IV

FID YIELDED FROM EXPERIMENTS WITH DIFFERENT TEXT ENCODERS. TRANSFORMED CAPTIONS ARE USED DUE TO THEIR SUPERIOR PERFORMANCE. THESE EXPERIMEMTS COULD NOT BE CARRIED OUT DUE TO LIMITATIONS. SEE SECTION V

Model	**Text Input**	**Text Encoder**	**FID**
UNet	Caption + CUI Description + Transformed Caption	RadBERT	35.2501
UNet	Caption + CUI Description + Transformed Caption	CLIP	
VGG UNet	Caption + CUI Description + Transformed Caption	RadBERT	33.5230
VGG UNet	Caption + CUI Description + Transformed Caption	CLIP	

TABLE V

EXPERIMENT RESULTS ON THE BEST CAPTION-ENCODER COMBINATION WITH DIFFERENT IMAGE SIZES FOR BOTH UNET ARCHITECTURES. THESE EXPERIMEMTS COULD NOT BE CARRIED OUT DUE TO LIMITATIONS. SEE SECTION V

Model	**Text Input**	**Text Encoder**	**Image Size**	**FID**
UNet	Caption + CUI Description + Transformed Caption	RadBERT	64	35.2501
UNet	Caption + CUI Description + Transformed Caption	RadBERT	128	
UNet	Caption + CUI Description + Transformed Caption	RadBERT	256	
VGG UNet	Caption + CUI Description + Transformed Caption	RadBERT	64	33.5230
VGG UNet	Caption + CUI Description + Transformed Caption	RadBERT	128	
VGG UNet	Caption + CUI Description + Transformed Caption	RadBERT	256	

Firstly, the experiments were conducted using a very small image size of 64x64 pixels. This reduced image resolution was a direct consequence of the VRAM limitations, as larger image sizes led to out-of-memory errors during training. This restriction likely affected the performance of the models, as higher resolution images generally provide more detailed information and can lead to improved generative results.

Moreover, we were unable to experiment with state-of-the-art (SOTA) diffusion model architectures due to the same VRAM limitations. Diffusion models, known for their high computational demand, were not feasible to implement without encountering memory issues. This limitation prevented a direct comparison with more advanced generative models that could potentially offer better performance.

Additionally, while I aimed to incorporate CLIP for text encoding due to its robust performance in understanding multimodal data, VRAM limitations posed a significant challenge. Although it was marginally usable, the batch size had to be drastically reduced to avoid running out of memory. This reduction in batch size would prevent any comparison with

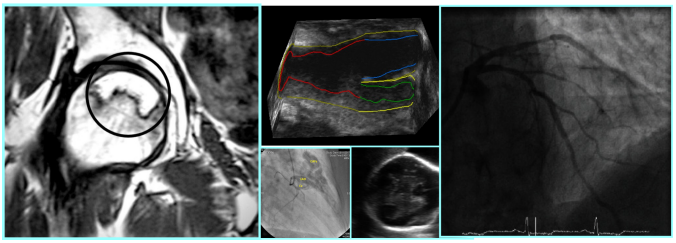

Fig. 5. Collage of noisy and indiscernable images from the dataset; these images were frequent in the dataset despite the filtering process and served as noise during training.

the RadBERT encoder used in our approach.

Lastly, the integration of image captions, CUI descriptions, and transformed captions using an attention module was planned to enhance the textual context provided to the model. However, the high memory requirement for implementing such an attention mechanism led to VRAM exhaustion, and we were unable to proceed with this approach. The inability to use this advanced textual integration method may have limited the potential improvements in the generative quality of the models.

Finally, the dataset also posed a significant challenge. Despite our image filtering process mentioned in Section III-A, the dataset still had a significant amount of noisy and non-descriptive images. Figure 5 illustrates some of these images. Such images occured were frequent in the dataset and hence served as noise when training the diffusion model.

VI. Future Works

Future research should address several limitations encountered in this study to fully harness the potential of text-to-image generation in radiology. Firstly, expanding the computational resources to more powerful GPUs will allow training on higher-resolution images, which is essential for capturing the detailed nuances of radiology images. Utilizing GPUs with greater VRAM capacity would also enable the exploration of state-of-the-art diffusion model architectures, which are known for their superior performance but require significant memory.

Additionally, integrating the CLIP text encoder more effectively should be a priority. Although CLIP showed potential, the necessity to significantly reduce the batch size limited its usability. Enhanced hardware resources would facilitate the seamless incorporation of CLIP, leveraging its strong multimodal learning capabilities to improve image generation quality.

Further advancements could be achieved by implementing an attention module to combine image captions, CUI descriptions, and transformed captions. The attention mechanism would enable the model to weigh different parts of the textual input according to their relevance, potentially enhancing the fidelity and coherence of the generated images. Exploring alternative methods to reduce VRAM usage or employing memory-efficient training techniques could make this feasible.

Overall, these future directions aim to overcome current hardware limitations, enabling more sophisticated models and techniques to significantly improve the quality and applicability of generated radiology images.

VII. Conclusion

In this study, we explored the efficacy of text-to-image generation for radiology images using UNet-based architectures and advanced textual input transformations. By analyzing and pre-processing the dataset, we ensured a more uniform and high-quality input for training. The experiments revealed that combining different textual descriptions significantly impacts the generative performance, with the inclusion of transformed captions and multiple descriptive sources (captions and CUI descriptions) leading to notable improvements in FID scores. Specifically, the integration of image captions, CUI descriptions, and transformed captions in the RadBERT text encoder demonstrated the best performance, highlighting the importance of comprehensive and structured textual input. Despite hardware constraints limiting image resolution and model complexity, our approach underscores the potential of using well-processed textual data and robust model architectures to enhance the generation of radiology images.

References

[1] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158, 2023.

[2] S Mithun, Ashish Kumar Jha, Umesh B Sherkhane, Vinay Jaiswar, Nilendu C Purandare, V Rangarajan, A Dekker, Sander Puts, Inigo Bermejo, and L Wee. Development and validation of deep learning and bert models for classification of lung cancer radiology reports. *Informatics in Medicine Unlocked*, page 101294, 2023.

[3] Zhiwei Dong, Genji Yuan, Zhen Hua, and Jinjiang Li. Diffusion model-based text-guided enhancement network for medical image segmentation. *Expert Systems with Applications*, page 123549, 2024.

[4] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.

[5] ROCOv2: Radiology Objects in COntext Version 2, An Updated Multimodal Image Dataset — zenodo.org. https://zenodo.org/records/8333645. [Accessed 29-02-2024].

[6] J Clusmann, FR Kolbinger, HS Muti, ZI Carrero, JN Eckardt, NG Laleh, CML Löffler, SC Schwarzkopf, M Unger, GP Veldhuizen, et al. The future landscape of large language models in medicine. communications medicine, 3 (1), 141, 2023.

[7] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.

[8] Danilo Dessi, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. Tf-idf vs word embeddings for morbidity identification in clinical notes: An initial study. *arXiv preprint arXiv:2105.09632*, 2021.

[9] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5182–5191, 2024.

[10] Hazrat Ali, Shafaq Murad, and Zubair Shah. Spot the fake lungs: Generating synthetic medical images using neural diffusion models. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 32–39. Springer, 2022.

[11] Ryo Toda, Atsushi Teramoto, Masakazu Tsujimoto, Hiroshi Toyama, Kazuyoshi Imaizumi, Kuniaki Saito, and Hiroshi Fujita. Synthetic ct image generation of shape-controlled lung cancer using semi-conditional infogan and its applicability for type classification. *International Journal of Computer Assisted Radiology and Surgery*, 16:241–251, 2021.

[12] Rohit Gupta, Anurag Sharma, and Anupam Kumar. Super-resolution using gans for medical imaging. *Procedia Computer Science*, 173:28–35, 2020.

[13] Haoyu Chen, Jinjin Gu, and Zhi Zhang. Attention in attention network for image super-resolution, 2021.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.