Deep Generative Models Unveil Patterns in Medical Images Through Vision- "Language" Conditioning

Xiaodan Xing*
Bioengineering Department and Imperial-X
Imperial College London
London, United Kingdom
xiaodan.xing@imperial.ac.uk

Junzhi Ning*
Bioengineering Department and Imperial-X
Imperial College London
London, United Kingdom

Yang Nan
Bioengineering Department and Imperial-X
Imperial College London
London, United Kingdom

Guang Yang
Bioengineering Department and Imperial-X
Imperial College London
London, United Kingdom
guang.yang@imperial.ac.uk

Abstract

Deep generative models have significantly advanced medical imaging analysis by enhancing dataset size and quality. Beyond mere data augmentation, our research in this paper highlights an additional, significant capacity of deep generative models: their ability to reveal and demonstrate patterns in medical images. We employ a generative structure with hybrid conditions, combining clinical data and segmentation masks to guide the image synthesis process. Furthermore, we innovatively transformed the tabular clinical data into textual descriptions. This approach simplifies the handling of missing values and also enables us to leverage large pre-trained vision-language models that investigate the relations between independent clinical entries and comprehend general terms, such as gender and smoking status. Our approach differs from and presents a more challenging task than traditional medical report-guided synthesis due to the less visual correlation of our clinical information with the images. To overcome this, we introduce a text-visual embedding mechanism that strengthens the conditions, ensuring the network effectively utilizes the provided information. Our pipeline is generalizable to both GAN-based and diffusion models. Experiments on chest CT, particularly focusing on the smoking status, demonstrated a consistent intensity shift in the lungs which is in agreement with clinical observations, indicating the effectiveness of our method in capturing and visualizing the impact of specific attributes on medical image patterns. Our methods offer a new avenue for the early detection and precise visualization of complex clinical conditions with deep generative models. All codes are https://github.com/junzhin/DGM-VLC.

1 Introduction

Deep generative models have traditionally served as vital tools for data augmentation in medical image analysis, enhancing the volume and quality of datasets for downstream tasks. However, the ever-increasing volume of real medical data during routine scanning and advancements in image acquisition algorithms challenge the necessity of using these models merely for data augmentation,

*contributed equally

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

especially when synthetic data may not match the quality of real observations. This evolution prompts a critical reassessment of the broader applications of deep generative models beyond simple data augmentation.

The application of generative models for anomaly detection through reconstruction techniques marked a significant shift [12, 4]. By training on healthy data and inferencing on patient data, these models can highlight differences as anomalies. However, the binary nature of this patient v.s. control strategy limits its application on comparisons across multiple classes, such as categorizing patients into different age groups.

Addressing these challenges, our research proposes a novel aspect of utilizing generative models to identify pattern that correlates with clinical attributes. Our method features by a hybrid condition, including both clinical information, such as gender, age, and diagnosis results, and segmentation masks, to guide the image synthesis process. The clinical information guidance enables the generation of diverse medical image patterns, and the segmentation masks offer structural guidance to minimize bias and highlight distinctive patterns.

The pervasive issue of missing data represents a significant obstacle to our concept's implementation. Our solution involves transforming tabular clinical data into detailed textual descriptions, allowing us to bypass the challenges posed by missing values. This conversion also exploits the potential of pre-trained vision-language models to understand clinical information expressed in simple terms, such as gender and smoking status.

Another challenge in our implementation is that, unlike conventional medical report-guided synthesis, our algorithm is conditioned on clinical information with no direct visual correlation to the images. Medical reports narrate observable patterns in medical images, while clinical parameters—like age, gender, and smoking status—lack established visual representations. Thus, we explore two approaches of text fusion unit including cross-attention module and Affine transformation fusion unit to enhance the conditions, aiming to signify the conditions on generated images.

Our experiments, conducted on a publicly available chest CT dataset, not only showcase the superior synthesis performance of our proposed framework but also highlight its effectiveness in capturing and visualizing the impacts of clinical status in medical image patterns, matching with clinical observations.

In summary, the primary contribution of our study is a novel method that employs generative models to detect medical image patterns that are associated with clinical attributes like age, gender, and smoking history. Our technical advancements include 1) **Conversion of tabular data into text**, which addresses missing data issues and utilizes the capabilities of pre-trained vision-language models to decode clinical information; 2) **Advanced text fusion techniques** including a cross-attention module and an Affine transformation fusion unit, to refine the conditioning process in cases where clinical information does not directly correspond to visual cues in images; and 3) **General implementation for GAN and diffusion models**. This research opens new avenues for employing deep generative models, surpassing traditional applications in data augmentation.

2 Method

The general procedure, shown in Figure 1 of our model pipeline proceeds as follows: First, we utilize any available tabular data related to lung CT scan masks, and use a transformation rule to normalize tabular data to text descriptions. We then employ a pre-trained Bert model [1] specialized in the healthcare domain, to transform this tabular data into clinically relevant text descriptions. These text descriptions are fed into the frozen text encoder to obtain text embeddings. Next, the text embeddings are fused with the generative models using text-vision affine transformation fusion units.

2.1 Transformation of Tabular Data into Textual Representations

Electronic Health Record (EHR) data are predominantly stored in a tabular format. However, utilizing tabular data presents several challenges. The first issue is data missingness, leading to a reduction in the available data. The second issue is that tabular data cannot represent relationships between different classes. For instance, in diagnosing lung fibrosis, both Connective Tissue Disease-Interstitial Lung Disease (CTD-ILD) and Idiopathic Pulmonary Fibrosis (IPF) exhibit an Usual Interstitial

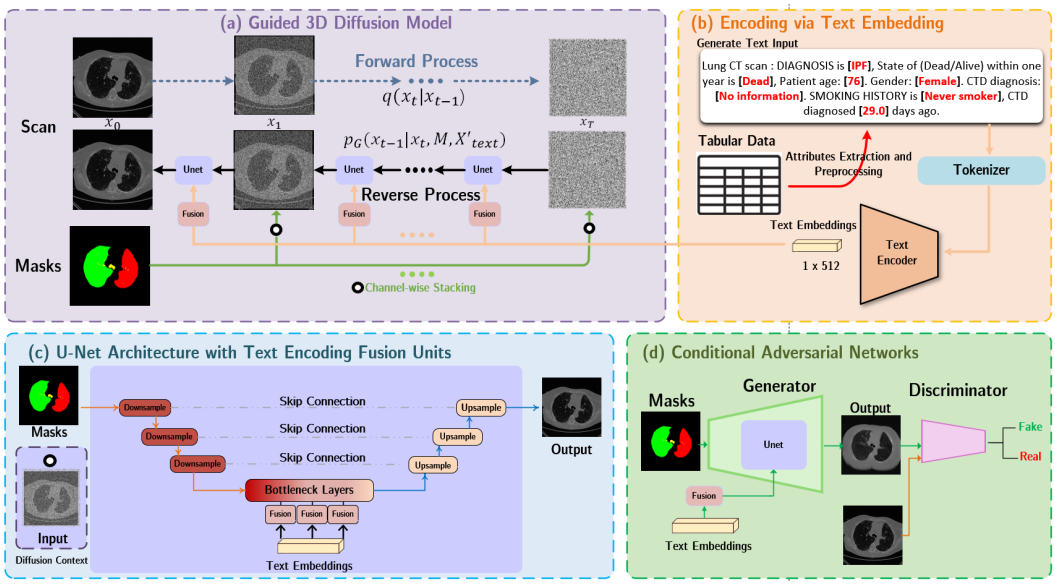

Figure 1: Overview of the proposed method. **Top left:** Illustration of the mechanism by which the text encoding embeddings are incorporated into the conditional diffusion model. **Top right:** Description of the process for utilizing tabular clinical data to obtain text embeddings from the pre-trained text encoder. **Bottom left:** A zoomed-in view of the fusion point where text embeddings integrate with the backbone of the models. **Bottom right:** Depiction of the compatibility of text fusion with a visual generator within the GAN framework.

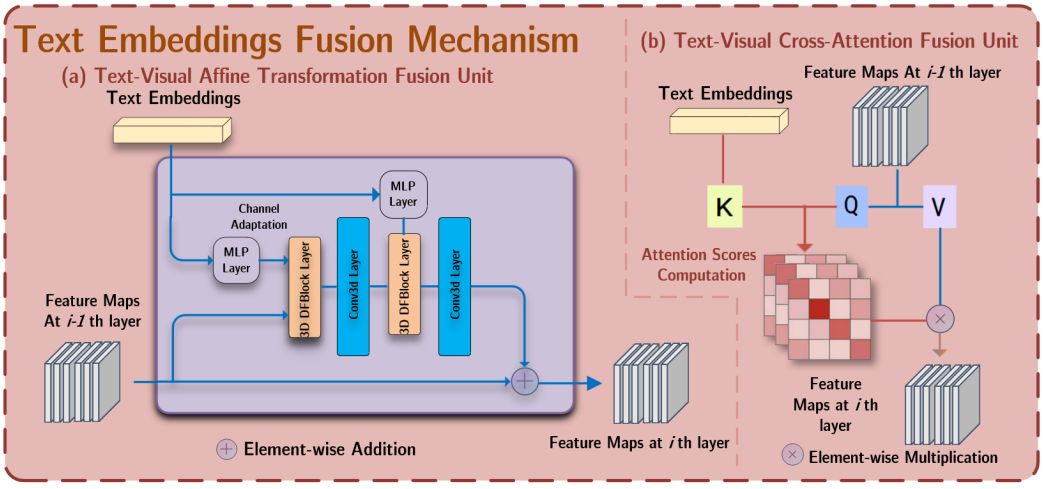

Figure 2: **Left**: Modified DFBlock Fusion Unit **Right**: Modified Cross-attention Fusion Unit

Pneumonia (UIP) pattern. However, tabular data merely categorize these conditions into distinct classes (e.g., 0, 1) without acknowledging their similarities. Lastly, the CLIP model [9, 15] emerges as a notable method for generating text embeddings, offering a training-free approach for feature extraction. Considering these, we design a template that converts tabular data into textual descriptions.

Our framework, outlined in Figure 1 (b), processes each row of tabular data x_i using a function f to generate a textual description x_i', following specific rules R. This involves filtering out attributes with unreasonable or missing values. Unlike other processes, we don't fill in missing values for text encoding with Bert, but **simply omitting them**. The process enriches text encoder context by concisely describing each data entry's attributes and their values. Mathematically, for the set of attributes $\mathcal{J}$, we form a matrix X'_{text} for text embedding, with rows:

$$\forall i \in \{1, \ldots, |X|\}, x_i' = f([k, \mathbf{1}_{\{(x_i)_j \neq \emptyset\}}(x_i)], R). \tag{1}$$

2.2 Incorporating Tabular Data as Text Embeddings in Generative Models

Given that our method of dealing with the text embeddings from tabular data is adaptable across various generative models, we have showcased its effectiveness with two popular generative frameworks pix2pix [7] and 3D diffusion models [3, 6]. We employ the schemes of text embedding fusions in two settings shown in Figure 1 (a) and (d). While these fusion units are technically adaptable to any generative architecture, our choose different units for different generative backbones based on experimental observations.

Text-Visual Affine Transformation Fusion Unit. We enhance the training by incorporating an information mask with random noise during the denoising step and utilize $X^{'}_{\text{text}}$ for data synthesis. Adapting DFBlock from DF-GAN [13] for 3D, we switch 2D convolutions to 3D and use an MLP linear layer for upsampling to maintain channel consistency for integrating text and visual features. This suits the U-shaped network's fusion needs, as shown in Figure 2 (a).

For text embeddings, affine transformations use scaling (γ) and shifting (θ) parameters to transform visual features, optimized via MLPs. Text embeddings are reshaped to align with visual feature channels before affine transformations:

$$AFF(\boldsymbol{e_i} \mid x') = MLP_\gamma(x') \cdot e_i + MLP_\theta(x'), \tag{2}$$

where MLPs match visual feature channels. This is followed by a 3D convolution and another fusion unit. Despite their efficiency in diffusion models, these units face modal collapse in Pix2pix, likely due to the original DFBlock's design for text-to-image synthesis, contrasting Pix2pix's conditional voxel generation.

Text-Visual Cross-Attention Fusion Unit. To address the problem identified earlier, we incorporate the conventional cross-attention mechanism within the Pix2pix method to enhance the integration of text embeddings. This approach is beneficial when there is no direct correlation between the textual information and structural guidance. We employ a tailored strategy that combines text embeddings with visual feature maps, where the text embeddings exclusively act as a "key" to selectively modulate the visual features as "query" and "value". This technique does not require a direct match between textual descriptions and visual conditions. Mathematically, shown in Figure 2 (b), consider an feature map denoted as $X \in \mathbb{R}^{B \times C \times D \times H \times W}$ and a text embedding vector $X^{'}_{text} \in \mathbb{R}^{B \times E}$, the tabular text embeddings cross-attention mechanism is defined as:

$$\begin{aligned} \text{Attention}(X, X^{'}_{text}) &= \text{softmax}\left(\frac{Q(X) \cdot K(X^{'}_{text}))^T}{\sqrt{d_k}}\right) \odot V(X), \\ Q(X) &= \text{Conv}_Q(X),\ K(X^{'}_{text}) = W_K X^{'}_{text},\ V(X) = \text{Conv}_V(X), \end{aligned} \tag{3}$$

where $\odot$ denotes element-wise multiplication, d_k is the scaling factor, typically the dimensionality of the key vectors, softmax is applied over the flattened spatial dimensions of the input tensor X after being projected to the query space, $\text{Conv}_Q(\cdot)$ and $\text{Conv}_V(\cdot)$ are $1 \times 1 \times 1$ convolution operations to generate query and value representations, W_K is a learnable weight matrix for the linear transformation of the text embedding into the key space.

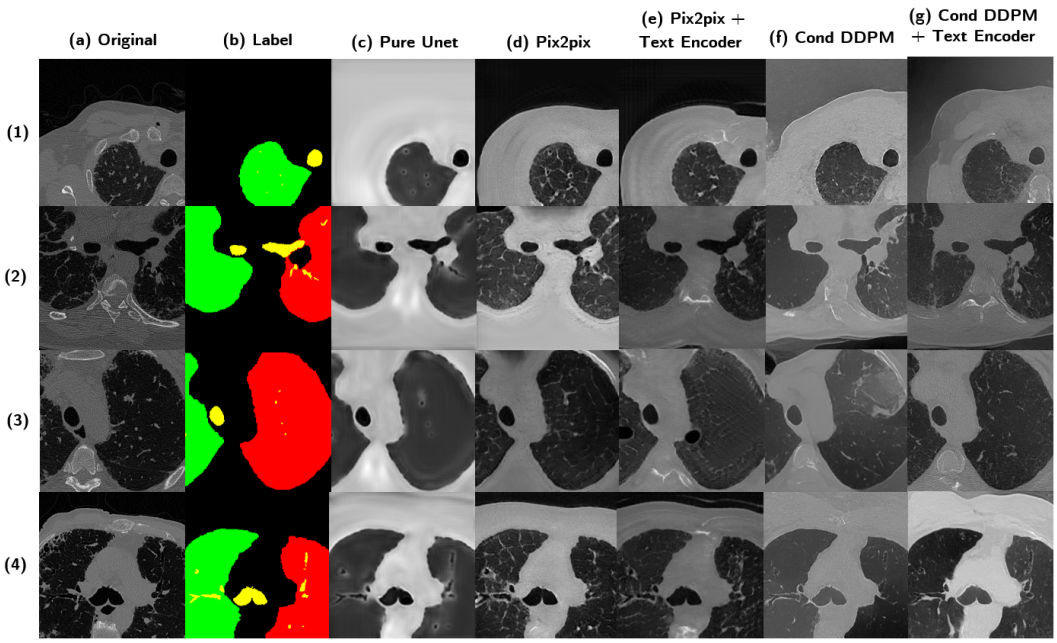

Figure 3: Visual comparison of reconstructed lung CT scans produced by four methods. From left to right: the original scan, followed by the four class labels—right lung, left lung, airway, and background. Subsequent columns present the results of Pure Unet, Pix2pix, Pix2pix with Text Encoder, Cond DDPM and Cond DDPM with Text Encoder.

3 Experiments

3.1 Dataset and Implementation

Our dataset is sourced from AIIB challenge, where the segmentation masks of lung regions and airways are available [8]. We utilized training data of 50 samples, each with a size of 512×512 along the axial size, which is flexible depending on the individual images. To process the 3D image before training, we randomly crop the 3D image to a size of $256 \times 256 \times 64$ and apply a minimum number of mask criteria to ensure each crop contains some valid values of certain classes in the masks. For validation purposes, we preserve another set of 45 samples of the same size and evaluate the GAN metrics Fréchet Inception Distance (FID)[5], Kernel Inception Distance (KID)[2] , and Inception Score (IS)[11] at a patch-wise level ($256 \times 256 \times 64$), with each test data randomly cropped 5 times.

For the details of the implementation, experiments were conducted on one A100 GPU for 1800 epochs, and the models were optimized through the Adam optimizer with an initial learning rate of 0.0001 and 0.00001 for the Pix2pix and DDpm methods, respectively, with both having a batch size of 2. The learning rate decayed after 800 epochs. The diffusion model is trained and evaluated with the timesteps of 250. The code implementation is based on [14].

3.2 Synthesis Performance Comparison

To evaluate the effectiveness of the tabular data utilization strategies with the existing deep learning framework, we choose the four models, 1) Pure UNet [10] 2) Pix2pix [7], 3) Pix2pix + Text Encoder [15] 4) Cond DDPM, 5) Cond DDPM + Text Encoder to perform the quantitative experiments evaluated on FID, KID and IS metrics.

From the results in Table 1 and Figure 3, we observe the following trends. The Pix2pix method outperforms the other approaches in terms of FID and KID scores, underscoring the potential benefits of integrating tabular text embeddings within the existing conditional GAN framework. However, the performance gain is not replicated with the Conditional 3D Diffusion models when text embedding is introduced; a marginal decrease in performance is noted, with FID scores showing a 2% reduction and KID scores dropping by 0.01. The nature of diffusion models, which rely on multiple iterative steps to

Table 1: Comparative Analysis of Generative Models

Models	Components			Evaluation Metrics		
	Diffusion	GAN	Tabular Data	FID ↓	KID ↓	IS ↑
Pure Unet	✗	✗	✗	221.99	0.207 (0.003)	4.508(0.086)
Pix2pix	✗	✓	✗	143.779	0.098 (0.002)	3.920(0.054)
Pix2pix + Text Encoder	✗	✓	✓	113.097	0.077 (0.002)	3.552(0.054)
Cond DDPM	✓	✗	✗	160.0576	0.114 (0.002)	3.721(0.044)
Cond DDPM + Text Encoder	✓	✗	✓	163.0374	0.120 (0.003)	3.136(0.128)

reconstruct or generate images, may render text embeddings less effective as the incremental denoising process may diminish their impact. Nevertheless, the incorporation of tabular text embeddings still adds value.

Regarding the Inception Score (IS), our interest lies in the model's capacity to leverage the information provided to constrain the generation results. Observations indicate that methods incorporating text embeddings exhibit a decline in IS scores. This suggests that while the diversity of the synthesized output is constrained by the additional information, it is indicative that text embeddings are indeed influencing the behaviour of the models. This finding highlights the complexity of balancing the fidelity and diversity of generated images in generative models.

3.3 Pattern Identification Analysis

We provide visual comparisons to show how clinical data influences image generation. A control experiment examined the impact of slight changes in tabular text descriptions on CT scan synthesis, measuring change by the difference in voxel values, using the Pix2pix and Text Encoder method. As shown in Figure 4, heatmaps highlight differences in test samples with varying "age" and "smoker" descriptions.

Transitioning from "non-smoker" to "smoker" status resulted in a distinct intensity pattern, with dot-like increases linked to the formation of lung nodules or inflammation and dot-like decreases associated with the destruction of lung tissue and the formation of air spaces. This is observed in Figure 4 (c2), aligning with clinical evidence that smoking can lead to a dual effect on lung density: increased in areas of tissue densification and decreased where lung tissue is compromised. Overall, an increase in the intensity is consistently observed.

Conversely, we did not identify a consistent trend in intensity changes with age variations, which we attribute to the limitations of our dataset, starting at a minimum age greater than 30, and the possible inability of language models to discern numerical relationships. Future efforts will focus on independently integrating numerical inputs to overcome this challenge.

This observation confirms that text encoders such as BERT or CLIP are adequately sensitive to condition the synthesis through their integration within the generative framework, employing mechanisms like cross-attention or affine transformation fusion.

4 Conclusion

In this study, we developed a versatile framework demonstrating the potential of deep generative models for uncovering invisible patterns in medical images associated with various clinical states. We innovatively transformed tabular data into textual descriptions, enabling the integration of clinical data with image synthesis methods through pre-trained vision-language models. Given the unique optimization challenges of generative models, we designed two distinct units to fuse textual and structural guidance for both GAN and Diffusion Model backbones, ensuring high-quality image synthesis while maintaining clinical relevance.

Moreover, we observed that while language models are good at understanding abstract concepts like life and death, they struggle with numerical understanding, such as recognizing that age 24 is less than age 68, which suggests a need for inputs beyond just text. In our future work, we aim to explore various conditioning to broaden the use of generative models beyond only data augmentation.

Our results are promising, showing that generative models can identify unseen patterns related to specific clinical attributes, such as the smoking status in lungs. These findings underscore the potential

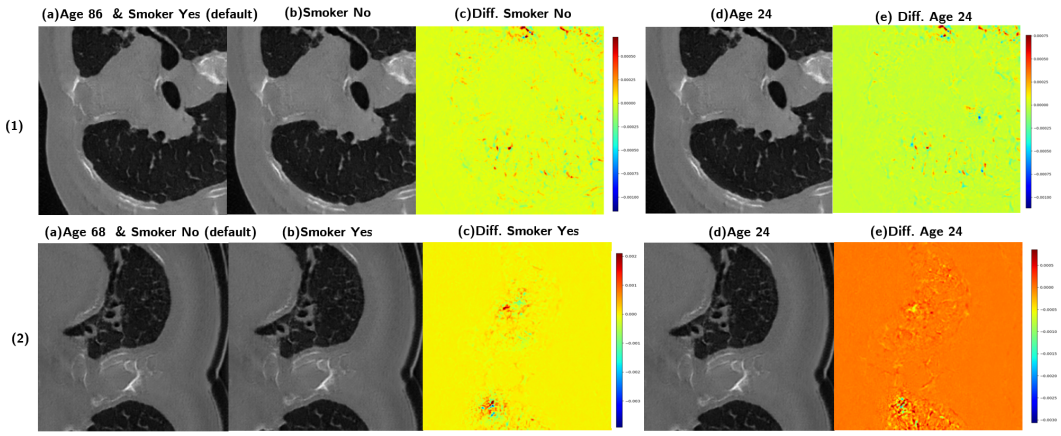

Figure 4: Qualitative Assessment of Pixel-Level Variations Following Prompt Modification: Illustrative Cases Demonstrating the Impact of Altered Prompt Content on Prediction Outcomes.

for significant advancements in the early detection of lung diseases and other complex medical conditions. This research opens new avenues for employing deep generative models, surpassing traditional applications in data augmentation.

5 Acknowledgement

This study was supported in part by the ERC IMI (101005122), the H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC\NSFC\211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, Wellcome Leap Dynamic Resilience, and the UKRI Future Leaders Fellowship (MR/V023799/1).

References

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72, 2019.

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[4] Changhee Han, Leonardo Rundo, Kohei Murao, Tomoyuki Noguchi, Yuki Shimahara, Zoltán Ádám Milacski, Saori Koshino, Evis Sala, Hideki Nakayama, and Shin'ichi Satoh. Madgan: Unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction. *BMC bioinformatics*, 22(2):1–20, 2021.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[8] Yang Nan, Xiaodan Xing, Shiyi Wang, Zeyu Tang, Federico N Felder, Sheng Zhang, Roberta Eufrasia Ledda, Xiaoliu Ding, Ruiqi Yu, Weiping Liu, et al. Hunting imaging biomarkers in pulmonary fibrosis: Benchmarks of the aiib23 challenge. *Medical Image Analysis*, 97:103253, 2024.

[9] 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.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[12] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[13] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022.

[14] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[15] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv e-prints*, pages arXiv–2210, 2022.