Multi-Modal Medical Image Augmentation for Controlled Heterogeneity and Fair Outcomes

Soo Yong Kim * [1] **Seunghyeok Hong** [1]

Abstract

Limited data in medical imaging exacerbate class imbalance and fairness gaps, undermining deep-learning across diverse patient subgroups. GAN- and diffusion-based augmenters can expand datasets but often lack precise control over multiple clinical attributes and fail to cover the full range of real-world variability. We introduce a four-step augmentation pipeline. First, an automated scoring function identifies which classes or regions most urgently need synthetic examples. Second, we construct sketch–image–text triplets from real scans, embedding age, sex, and disease labels. Third, we fine-tune a sketch-conditioned diffusion network for reliable sketch-to-image synthesis and boost variability by generating multiple, similarity-penalized sketches per case. Fourth, we propose a novel diversity metric that simultaneously measures semantic feature-space coverage and pixel-level dispersion—unlike FID or IS, it captures intra-class spread and boundary sharpness without human annotations. Experiments on chest X-rays show our pipeline delivers high-fidelity, diverse images aligned with user-specified conditions, substantially improving fairness and generalizability.

1. Introduction

Deep-learning techniques have revolutionized medical imaging analysis, achieving expert-level accuracy in tasks such as disease classification, lesion detection, and segmentation. However, these models typically require large, well-balanced datasets to generalize effectively across diverse patient populations. In practice, clinical data collection is constrained by privacy regulations, annotation costs, and unequal disease prevalence, resulting in small sample sizes and pronounced class imbalance (3; 19; 20). As a consequence, models often underperform on minority subgroups—such as rare diseases or demographic categories—posing risks to diagnostic equity and patient safety.

Traditional countermeasures to imbalance include undersampling majority classes, oversampling minority classes, and loss re-weighting. While these methods can alleviate skewed label distributions, they either discard potentially informative examples or amplify noise in scarce classes, leading to unstable training and overfitting (18; 21). Data augmentation via generative models offers a promising alternative by synthesizing realistic samples to bolster underrepresented categories without discarding real data.

Generative Adversarial Networks (GANs) have been the de facto choice for synthetic data generation, demonstrating success in various medical imaging modalities (4; 5). Yet GAN-based augmenters often suffer from mode collapse and struggle to capture fine-grained anatomical details, limiting diversity and clinical plausibility (24; 25). To overcome these issues, diffusion models (7) have recently emerged, producing higher-fidelity images with more stable training dynamics (23; 26). Despite these advances, most diffusion-based approaches rely on coarse, text-only prompts and fail to guarantee structural control or demographic consistency.

Recent breakthroughs in controllable diffusion—exemplified by Stable Diffusion (6) and ControlNet (2)—enable conditioning on both textual descriptions and structural sketches. Sketch-guided generators such as DiffSketcher (1) can extract detailed anatomical outlines from real scans, offering a scaffold for precise image synthesis. However, naïvely applying these modules does not ensure sufficient coverage of underrepresented classes, nor does it provide a mechanism to quantify and maximize intra-class variability. Moreover, standard evaluation metrics like Fréchet Inception Distance (FID) (11) and Inception Score (IS) (12) measure global distribution alignment but overlook the pixel-level dispersion and boundary sharpness that are critical for clinical realism.

In this work, we introduce a structured, four-stage augmentation workflow specifically tailored for medical imaging under data-scarce, fairness-critical conditions. First, an automated scoring function identifies classes or anatomical

*Equal contribution [1]Solar Energy Research Lab, MODULABS. Correspondence to: Seunghyeok Hong <shongdr@gmail.com>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

regions that most urgently require synthetic augmentation. Second, we construct sketch–image–text triplets from authentic scans, embedding multiple demographic and disease attributes in a unified representation. Third, we fine-tune a sketch-conditioned diffusion network and enhance diversity by generating multiple sketches per case with a similarity-penalized loss that enforces both high-level semantic variation and low-level textural differences. Fourth, we propose a novel diversity metric that concurrently quantifies feature-space coverage and pixel-level dispersion, distinguishing itself from FID and IS by explicitly measuring intra-class spread and boundary sharpness without human annotations.

Extensive experiments on chest X-ray datasets demonstrate that our pipeline generates high-fidelity images aligned with user-specified conditions while substantially improving fairness and downstream task performance on minority cohorts. We summarize our key contributions as follows:

- A novel scoring mechanism to pinpoint underrepresented classes and guide targeted synthetic augmentation.
- An integrated sketch–image–text triplet construction process that encodes multimodal clinical attributes for controlled diffusion synthesis.
- A similarity-penalized sketch generator and a new diversity metric that together maximize intra-class variability and evaluate distributional coverage beyond standard metrics.

2. Related Work

Medical Image Generation. Various generative models have been developed for text-to-image synthesis (35; 36; 37; 38). (34) describes an approach where a diffusion model is fine-tuned on medical images, subsequently enhancing cancer classification models. Nonetheless, this model exhibits limitations in generating images under complex text conditions that include variables such as sex and age. (33) presents a framework for text-conditional magnetic resonance (MR) imaging generation using canny edge maps of brain, capable of producing realistic MR images that correspond with medical text prompts. However, in the medical field, the diversity and alignment with text semantics of image generation under complex conditions are not fully exploited simultaneously. To solve this problem, we propose a framework leveraging various large-scale pre-trained models.

Quality Diversity. Quality Diversity (QD) is a promising concept which optimizes generative model to produce high-quality diverse outputs (9). Quality Diversity through human feedback (QDHF) (9) employs cosine similarity to compare two feature maps derived from the latents generated from identical input texts. However, this method fails to account for complex prompts that involve attribute binding, spatial reasoning, numeracy, and inherent background details. When processing repeated samples of such intricate prompts, QDHF tends to produce a collection of images that are nearly indistinguishable from one another. Moreover, the introduction of conditional images (8) can significantly reduce the diversity of the generated images. To overcome this problem, we introduce similarity loss function combining high-level and low-level similarity metrics for the high-quality diverse sketch generation.

Similarity Metric. There are wide range of metrics (13; 10; 14; 15) which can be emploied to measure similarities of two images. Among them, LPIPS (10) and DreamSim (14) are the most closely related to our work. LPIPS implemented two-alternative forced choice (2AFC) and just noticeable difference (JND) experiments, including a human survey, to incorporate the human perspective on diversity among images. However, despite human involvement, LPIPS struggles to accurately measure semantic similarity between images neglecting high-level features. To address this issue, Dreamsim employed an ensemble model combining Dinov2 (15) and CLIP to better align with human preferences. However, DreamSim cannot fully cover low-level features and has not been trained with medical images as well as LPIPS. To overcome these limitations, we jointly leveraged both LPIPS and DreamSim to compliment each other.

3. Method

We aim to develop a comprehensive medical image augmentation framework designed to ensure high and uniform model performance across a wide spectrum of patient groups. To begin, we systematically select specific patient subpopulations for augmentation by computing a majority score that quantifies each group's representation within the original dataset. In addition, our framework leverages two complementary generative models. First, DiffSketcher produces preliminary sketch representations, having been optimized with a tailored similarity loss that explicitly enhances the diversity of generated sketch variants. Second, we integrate ControlNet to synthesize realistic medical images under explicitly controlled clinical conditions, including age categories, sex, and disease-specific attributes. These control parameters are intentionally chosen to reflect the unique characteristics of minor or underrepresented patient groups, thereby promoting fairness in data augmentation. To quantitatively assess the variety of outputs, we propose a novel diversity metric based on convolutional neural network saliency maps, which evaluates differences in salient

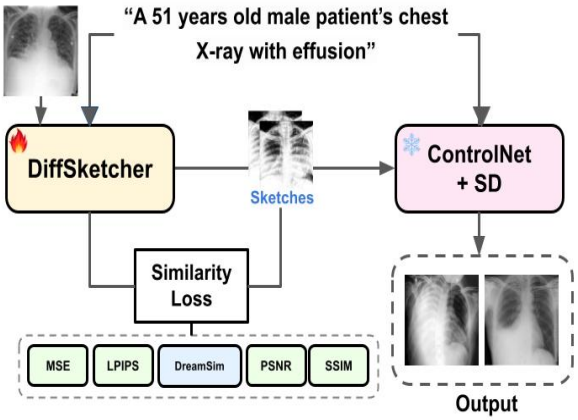

Figure 1. The procedure of inference step in the EDM. We freeze ControlNet and SD while keeping DiffSketcher trainable. We generate multiple sketches from minor patient groups and input them as conditions to the ControlNet.

anatomical regions across generated samples. Overall, our method enables the production of diverse, high-quality, and condition-specific synthetic medical images that effectively address the needs of underrepresented patient cohorts and improve equity and robustness in downstream medical image analysis.

3.1. Selecting patient groups for the data augmentation

First of all, we introduce majority score that assists in selecting patient groups for data augmentation based on their low data diversity and scarcity. We can select patient groups with lower majority scores for the data augmentation. In Figure 1, the X-ray image inputted into DiffSketcher represents data from these minor patient groups which have low majority scores.

$$M_K = \frac{1}{2}\left(\frac{|\mathcal{D}_K|}{|\mathcal{D}|} + S_K\right) \tag{1}$$

$$S_K = \frac{2 * \sum_{1 \le i \le j \le |\mathcal{D}_K|} d_{ij}}{|\mathcal{D}_K|(|\mathcal{D}_K| - 1)} \tag{2}$$

$$d_{ij} = \frac{1}{3} * [\frac{1}{C}\sum_{c=1}^{C} f(v_c^i, v_c^j) + \frac{|g^i//10 - g^j//10|}{g^m//10} + \delta(s^i, s^j)] \tag{3}$$

$$f(v_c^i, v_c^j) = \begin{cases} 1 & \text{if } v_c^i \neq v_c^j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$\delta(s^i, s^j) = \begin{cases} 1 & \text{if } s^i \neq s^j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Regarding the parameters of majority score, M_K is the majority score and K is the index of patient groups. $|\mathcal{D}|$ is the total number of data and $|\mathcal{D}_K|$ is the number of data from K^{th} patient group. v is the label vector and c is the class index. g indicates the age of the patient and m is the index of the maximum age. s represents the sex of the patient.

3.2. Preparing data augmentation

The second step is preparing the procedure for the data augmentation. In this step, we obtain image-sketch-text triplets leveraging the real patients' dataset. For this, we employ DiffSketcher. DiffSketcher is a kind of learnable vector graphics which draws sketches using mathematical formulas. This process is executed one-by-one which is different from Figure 2 which generates multiple sketches at the same time. Then we train ControlNet with the triplets from the second step. It takes a text prompt and an image as inputs and then generates a synthetic image as an output.

3.3. Constructing synthetic dataset

Third step is regarded as the inference step to construct synthetic datasets. We obtain multiple sketches for each real patient's image respectively from the minor patient groups. We enable this by introducing similarity loss to the DiffSketcher. For each sketch, we calculate the similarity with the rest of the sketches and train DiffSketcher to decrease this similarity. The similarity loss for DiffSketcher is as below.

$$\begin{aligned} \mathcal{L}_{sim} = & \alpha * \frac{\sum_{1 \le j \le n, j \neq i}(1 - s(x_i, x_j))}{n-1} + \\ & (1-\alpha) * \frac{\sum_{1 \le j \le n, j \neq i}(1 - l(x_i, x_j))}{n-1} \end{aligned} \tag{6}$$

α regulates the contribution of low-level and high-level distance. x is the input image. s represents the loss function to calculate high-level similarity of input images. We used DreamSim to calculate this. l is the function to calculate low-level similarity of input images. We leveraged MSE, PSNR, SSIM, LPIPS and compared the results through the experiments. n is the number of sketches we generate at the same time.

Using the obtained various sketch data, we generate synthetic medical images using fine-tuned ControlNet from step two.

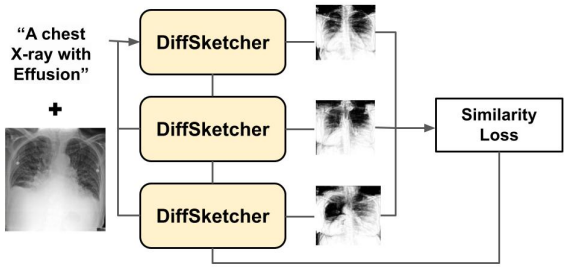

Figure 2. Training schematic of diversity-enhanced DiffSketcher. We improve the diversity of DiffSketcher by training multiple DiffSketchers at the same time with the similarity loss.

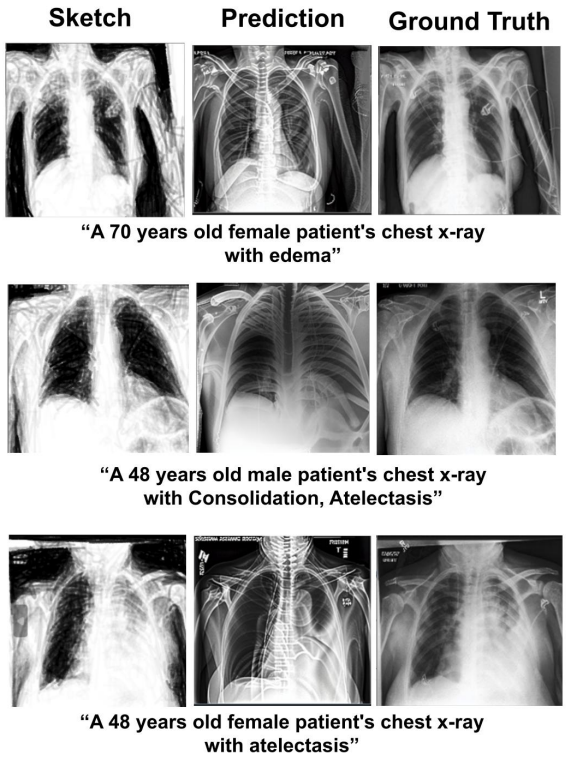

Figure 3. Results while training the ControlNet. We can see that ControlNet learns how to generate synthetic X-ray images from the sketches. A sketch and a prompt are inputted to the ControlNet.

4. Experiments

4.1. Training ControlNet using image-sketch-text triplets

To begin with, we fine-tuned ControlNet using the sketch-image-text triplets obtained from the CheXpert. The number of image-sketch-text triplets used in our experiments was ten. Regarding the hyperparameters, the number of epochs was 255, learning rate was 1e-5, batch size was 1, optimizer was AdamW and we did not use scheduler. We freezed stable diffusion (SD) and only trained ControlNet. Fig 3 shows the fine-tuned ControlNet successfully generated modest synthetic X-ray images with only a few triplets. The result is illustrated in Figure 3.

4.2. Diversity enhanced sketch generation

In our experimental setup, we aim to evaluate the impact of implementing similarity loss in the DiffSketcher by observing the diversity of generated sketches. To conduct this analysis, we employ a stable diffusion model trained specifically on medical image data, ensuring relevance to medical imaging contexts. This choice was made to enhance the model's ability to accurately capture and replicate the distinctive features of medical images.

Input text prompt is "A chest X-ray with pneumonia", "A chest X-ray with effusion", "A chest X-ray with cardiomegaly", with the process running for 500 iterations and utilizing 1000 paths. The comparative results clearly show that the sketches from the DiffSketcher with similarity loss exhibited significantly better diversity compared to those without the similarity loss. This outcome highlights the effectiveness of similarity loss in enhancing the diversity of generated medical sketches.

To optimize the results of our experiments, we conducted various tests with different hyperparameters, particularly the number of paths and timing of applying similarity loss and type of low-level metric. We compared various numbers of paths and found that 1000 paths resulted in the best output images. Too many paths made the sketch too dark, distorting patient's anatomy. Recognizing the importance of preserving anatomical integrity in medical images, we explored the consequences of introducing similarity loss at different stages of the training process. We found that introducing the similarity loss at epoch 100 delivered the highest visual fidelity; applying it significantly earlier or later resulted in reduced diversity and noticeable anatomical distortions. Our qualitative comparisons under different scheduling regimes confirm that this mid-training insertion yields the most accurate and realistic outputs.

In addition, we compared various forms of similarity loss function. We employed MSE, LPIPS, PSNR, SSIM to measure the low-level similarity of given input images. We also

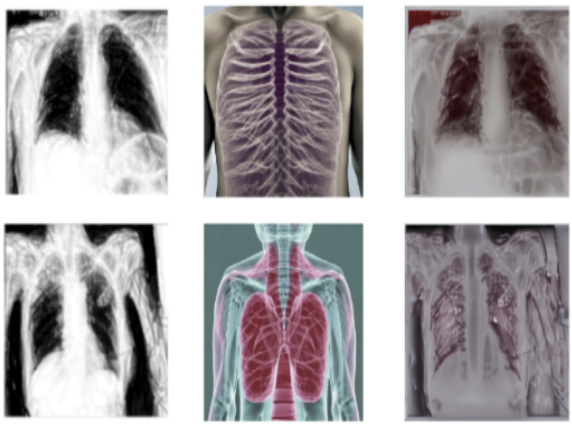

Figure 4. First column is the input sketch, second is result of ControlNet, third is result of ControlNet++. Second and Third show the result without any fine-tuning

conducted experiments using each metric alone for the similarity loss. Through these experiments, we observed that MSE resulted in the best performance for pneumonia and effusion cases, while mixing LPIPS and DreamSim resulted in the best results for cardiomegaly detection.

4.3. Comparision Study

We tried to compare the results of our Fine-Tuned ControlNet(8) to those of Fine-Tuend ControlNet ++. However, as we could not find the official Train code of ControlNet ++, we could not fine tune. However, we could see the possibility that if we succeed to scratch the train code, ControlNet ++ may show better performance than our Baseline. Because as you can see in Figure 4, although ControlNet ++ was not trained on the data related to Chest X-Ray, it shows great performance, considering the zero-shot inference situation. As you can see the second column in Figure 4, without any training, ControlNet hardly generates the image related to the sketch.

4.4. Ablation Study

We carried out an ablation study to test whether our sketch-conditioned, ControlNet-generated chest-X-ray images improve disease classification across diverse architectures and an extended label set. Three backbones—ResNet-50, EfficientNet-B0 and ViT-Base—were trained on the MIMIC-CXR corpus containing fourteen findings, from common pathologies such as Pneumonia and Atelectasis to rarer labels like Pleural Other. All models first learned from twenty-thousand real scans and were evaluated on an unchanged validation pool of two-thousand real images, providing a no-augmentation baseline.

Adding synthetic data consistently boosted accuracy. Injecting two-hundred generated images per class (total train size = 22 800) raised EfficientNet-B0 from 46.9 percent to 48.1 percent, ResNet-50 from 45.1 percent to 46.6 percent, and ViT-Base from 42.4 percent to 44.1 percent. Increasing to three-hundred synthetic images per class (train size = 24 200) delivered further but tapering gains: 48.7 percent, 47.3 percent and 44.9 percent for EfficientNet-B0, ResNet-50 and ViT-Base respectively.

Rare categories profited most. Labels with limited real examples—Pleural Other, Lung Lesion—saw accuracy jumps of up to four percentage points, showing that sketch-guided augmentation mitigates class imbalance. However, returns diminished beyond two-hundred synthetic images per class, indicating an optimal synthesis budget between two-hundred and three-hundred. Overall, our approach improves generalization across architectures and disease spectrum without altering the evaluation protocol.

5. Conclusion

In conclusion, our research presents a framework for enhancing the diversity of medical image generation by employing ControlNet and DiffSketcher optimized with similarity loss. This methodology improves condition-specific medical image generation and introduces a new diversity assessment metric incorporating both high-level and low-level image features. Our experimental results demonstrate that similarity loss significantly boosts sketch diversity and quality, while fine-tuned ControlNet with image-sketch triplets generates modest quality X-ray data. This work underscores the potential of advanced generative models to revolutionize medical imaging by providing diverse datasets for improved diagnostic accuracy. Future work may explore additional augmentations and collaboration with physicians to align biological science with model performance.

6. ACKNOWLEDGMENTS

This research was supported by Kakao Impact Tech for Impact Lab Program (1st Cohort, 2024)

References

[1] R. Barandela, J. S. Sánchez, V. García, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849–851, 2003.

[2] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint*, arXiv:1801.01973, 2018.

[3] C. Bodnar, "Text to image synthesis using generative adversarial networks," *arXiv preprint*, arXiv:1805.00676, 2018.

[4] W. Cho et al., "Towards enhanced controllability of

diffusion models," *arXiv preprint*, arXiv:2302.14368, 2023.

[5] B. de Wilde, A. Saha, R. P. G. ten Broek, and H. Huisman, "Medical diffusion on a budget: textual inversion for medical image generation," *arXiv preprint*, arXiv:2303.13430, 2023.

[6] L. Ding, J. Zhang, J. Clune, L. Spector, and J. Lehman, "Quality diversity through human feedback," *arXiv preprint*, arXiv:2310.12103, 2023.

[7] F. A. Fardo, V. H. Conforto, F. C. de Oliveira, and P. S. Rodrigues, "A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms," *arXiv preprint*, arXiv:1605.07116, 2016.

[8] S. Frolov et al., "Adversarial text-to-image synthesis: A review," *Neural Networks*, vol. 144, pp. 187–209, 2021.

[9] S. Fu et al., "DreamSim: Learning new dimensions of human visual similarity using synthetic data," *arXiv preprint*, arXiv:2306.09344, 2023.

[10] J. Gong and H. Kim, "RHSBoost: Improving classification performance in imbalance data," *Comput. Stat. Data Anal.*, vol. 111, pp. 1–13, 2017.

[11] I. J. Goodfellow et al., "Generative adversarial networks," *arXiv preprint*, arXiv:1406.2661, 2014.

[12] D. Gravina, A. Liapis, and G. N. Yannakakis, "Quality diversity through surprise," *IEEE Trans. Evol. Comput.*, vol. 23, no. 4, pp. 603–616, 2018.

[13] M. Heusel et al., "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *arXiv preprint*, arXiv:1706.08500, 2018.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint*, arXiv:2006.11239, 2020.

[15] K. Kim et al., "Controllable text-to-image synthesis for multi-modality MR images," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pp. 7936–7945, 2024.

[16] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint*, arXiv:1312.6114, 2022.

[17] K. Kobayashi et al., "Sketch-based medical image retrieval," *arXiv preprint*, arXiv:2303.03633, 2023.

[18] S. K. Venu and S. Ravula, "Evaluation of deep convolutional generative adversarial networks for data augmentation of chest X-ray images," *Future Internet*, vol. 13, no. 1, p. 8, 2020.

[19] I. Ktena et al., "Generative models improve fairness of medical classifiers under distribution shifts," *Nature Medicine*, pp. 1–8, 2024.

[20] I. Ktena et al., "Generative models improve fairness of medical classifiers under distribution shifts," *arXiv preprint*, arXiv:2304.09218, 2023.

[21] D.-C. Li, C.-W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets," *Comput. Biol. Med.*, vol. 40, no. 5, pp. 509–518, 2010.

[22] M. Li et al., "ControlNet++: Improving conditional controls with efficient consistency feedback," *arXiv preprint*, arXiv:2404.07987, 2024.

[23] R. K. Mahabadi et al., "TESS: Text-to-text self-conditioned simplex diffusion," *arXiv preprint*, arXiv:2305.08379, 2023.

[24] M. U. Nasir et al., "Llmatic: Neural architecture search via large language models and quality-diversity optimization," *arXiv preprint*, arXiv:2306.01102, 2023.

[25] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *arXiv preprint*, arXiv:2304.07193, 2024.

[26] Y. Qin et al., "Class-balancing diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18434–18443, 2023.

[27] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint*, arXiv:2103.00020, 2021.

[28] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *Int. J. Mach. Learn. Comput.*, vol. 3, no. 2, p. 224, 2013.

[29] R. Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," *arXiv preprint*, arXiv:2112.10752, 2022.

[30] S. Sadat et al., "CADS: Unleashing the Diversity of Diffusion Models through Condition-Annealed Sampling," *arXiv preprint*, arXiv:2310.17347, 2023.

[31] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36479–36494, 2022.

[32] E. Strelcenia and S. Prakoonwit, "A survey on GAN techniques for data augmentation to address the imbalanced data issues in credit card fraud detection," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 1, pp. 304–329, 2023.

[33] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," in *ACM SIGGRAPH Conf. Proc.*, pp. 1–11, 2023.

[34] X. Xing et al., "DiffSketcher: Text Guided Vector Sketch Synthesis through Latent Diffusion Models," *arXiv preprint*, arXiv:2306.14685, 2024.

[35] W. Wu, Y. Zhao, H. Chen, Y. Gu, R. Zhao, Y. He, H. Zhou, M. Z. Shou, and C. Shen, "DatasetDM: Synthesizing data with perception annotations using diffusion models," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[36] X. Xing, C. Wang, H. Zhou, J. Zhang, Q. Yu, and D. Xu, "DiffSketcher: Text Guided Vector Sketch Synthesis through Latent Diffusion Models," *arXiv preprint*, arXiv:2306.14685, 2024.

[37] M. Zameshina, O. Teytaud, and L. Najman, "Diverse diffusion: Enhancing image diversity in text-to-image generation," *arXiv preprint*, arXiv:2310.12583, 2023.

[38] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint*, arXiv:2302.05543, 2023.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *arXiv preprint*, arXiv:1801.03924, 2018.

[40] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models," *arXiv preprint*, arXiv:2305.16322, 2023.

[41] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," *arXiv preprint*, arXiv:1901.07042, 2019.