MedSoft-Diffusion: Medical Semantic-Guided Diffusion Model with Soft Mask Conditioning for Vertebral Disease Diagnosis

Shidan He[1,2], Enyuan Hu[1,2], Zixuan Tang[1,2], Bin Chen[3], Dongdong Yu[3], Yuan Hong[3], Zhenzhong Liu[4], Mengtang Li[1,2], Lei Liu[5,6], and Shen Zhao[1,2]✉

[1]School of Intelligent Systems Engineering, Sun Yat-Sen University
[2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou
[3]Orthopedics Department, The First Affiliated Hospital of Zhejiang University
[4]Tianjin Key Laboratory for Advanced Mechatronic System Design and Intelligent Control, School of Mechanical Engineering, Tianjin University of Technology
[5]The Chinese University of Hong Kong, Shenzhen
[6]Ant Group

Abstract. Accurate diagnosis of vertebral diseases is vital for preventing severe complications, but data imbalance between abundant normal and rare pathological cases poses a substantial challenge to diagnostic performance. Medical image generation offers a promising solution by synthesizing pathological samples. However, existing diffusion-based methods, pre-trained on natural images, often fall short in capturing complex pathological features due to the pre-training knowledge gap, as well as struggling to obtain precise lesion masks and ensure seamless integration between lesions and the background. To overcome these challenges, we propose a novel diffusion-based medical image generation framework called MedSoft-Diffusion, which involves leveraging detailed medical knowledge to ensure that generated images are not only semantically consistent with the specified pathological conditions but also anatomically accurate. Our framework includes a Medical Semantic Controller (MSC) designed to enhance the alignment between textual prompts and lesion characteristics, ensuring the synthesis of semantically accurate pathological images. Furthermore, the Soft Mask Inpainting Strategy (SMIS) is proposed to combine soft masks with blurring techniques to improve the realism of synthesized images. Experimental results on two vertebral disease datasets demonstrate notable improvements in both image quality and classification performance using our approach. Code is available at MedSoft-Diffusion.

Keywords: Medical Image Generation · Diffusion model · Vertebral Disease Diagnosis.

✉ Corresponding author: Shen Zhao (zhaosh35@mail.sysu.edu.cn)

1 Introduction

Accurate diagnosis of vertebral diseases is essential in clinical practice due to its potential to avert serious complications, including neural impairment, fractures, and chronic pain [1]. Early detection and treatment can significantly enhance patient outcomes and quality of life while improving the overall prognosis [2]. Nonetheless, a major challenge in diagnosing vertebral conditions lies in the class imbalance issue prevalent in clinical datasets. Specifically, normal samples are disproportionately represented compared to pathological cases, a discrepancy that adversely affects the precision and recall rates of predictive models [3,4].

Addressing the scarcity of rare conditions in medical datasets has become a focal point in contemporary research. One promising approach is to augment existing datasets with synthetic medical images, broadening data distribution and enhancing the efficacy of model training [5]. For this purpose, text-to-image diffusion models have garnered significant attention for their adaptability and precise control over image synthesis processes [6,7,8,9,10]. These models, however, are predominantly pre-trained on vast collections of natural images, which limits their effectiveness in capturing the intricate morphological nuances specific to pathological regions within medical imagery. Unlike natural images, medical images possess unique visual and semantic characteristics, and the diffusion model requires detailed and accurate representations of complex pathologies.

Synthesizing high-quality medical images with pathological features is highly challenging, primarily due to the necessity of producing anatomically accurate backgrounds in conjunction with realistic pathological elements. In vertebral disease diagnosis, where normal samples vastly outnumber pathological ones, this disparity inspires an alternative approach: instead of generating entire medical images, we can leverage inpainting techniques [11] to insert lesions into existing normal samples. Such a strategy could simplify the synthesis process and avoid the high computational costs associated with fully generative approaches. Nevertheless, traditional inpainting approaches suffer from the following two problems: (1) Difficulty of annotating precise lesion masks (hard masks). The main reason lies in the inconsistent annotation standards across different institutions and the extremely high cost of manually delineating lesion areas. (2) Low synthesis quality at mask boundaries. The boundaries of masks tend to be overly abrupt, resulting in the generated lesion regions not blending naturally with surrounding tissues.

To address the aforementioned issues, we introduce MedSoft-Diffusion, a novel medical image generation framework guided by medical semantic features and soft mask conditioning, which enhances the model's semantic understanding of medical images, mitigates the lesion mask scarcity, and ensures seamless integration between lesions and their backgrounds. Concretely, we propose a Medical Semantic Controller (MSC) to capture detailed pathological features, which leverages a pre-trained medical multimodal model (BiomedCLIP [12] in our experiment) to enhance the alignment between textual prompts and visual structures during the diffusion process. By interpreting masked lesion characteristics from textual prompts, the MSC enhances the diffusion model's capability

to capture nuanced pathological details. Besides, given the limited availability of high-quality lesion masks, we introduce a Soft Mask Inpainting Strategy (SMIS) to substitute hard masks with soft masks (*e.g.*, anatomical segmentation maps or bounding boxes), which provide a flexible and adaptive method to guide the diffusion process and mitigate the issue of insufficient lesion annotations. The overview framework is illustrated in Fig. 1.

Our contributions are as follows:

(1) MedSoft-Diffusion. We propose a diffusion-based medical image generation framework guided by medical semantic features and soft mask conditioning to enhance clinically meaningful pathology synthesis and anatomical consistency while mitigating lesion annotation scarcity.

(2) Medical Semantic Controller. We introduce a medical multimodal-driven controller that aligns textual pathology descriptions with masked images, enabling the diffusion model to capture fine-grained pathological semantics for precise and clinically relevant lesion generation.

(3) Soft Mask Inpainting Strategy. We use soft masks and introduce blurring to smooth lesion boundaries, facilitating flexible lesion inpainting and improving anatomical integration and synthesis realism.

2 Method

2.1 Medical Semantic Controller

The Medical Semantic Controller (MSC) is a transformer-based encoder designed to predict the pathological features of a masked lesion region based on its textual description. It takes as input the textual feature e_t extracted from the lesion description T and the visual feature e_v from the masked image I_m, then outputs a medical semantic feature e_{MSC}. This feature is compared with e_r, the feature of the original image x_0 (which contains the lesion), ensuring that the generated lesion aligns with both the textual description and real pathology. In Fig. 1, the Medical Text Encoder and Medical Visual Encoder are frozen, pre-trained medical multimodal models.

To obtain I_m, we first process the soft mask M by applying a blurring operation. Unlike traditional hard lesion masks, which are rarely available in real-world datasets, we use soft masks that cover the lesion region but are derived from vertebral segmentation masks or bounding box masks of vertebrae. These are more accessible since high-accuracy vertebra segmentation and detection models are already available. The blurring operation is defined as follows:

$$M_b(x, y) = (G_\sigma * M)(x, y), \tag{1}$$

where G_σ is a Gaussian blur kernel. The processed mask M_b is then used to mask the original image:

$$I_m = x_0 \odot (1 - M_b), \tag{2}$$

where $\odot$ denotes element-wise multiplication. This ensures that only the soft-masked region is removed while preserving the background. The MSC takes

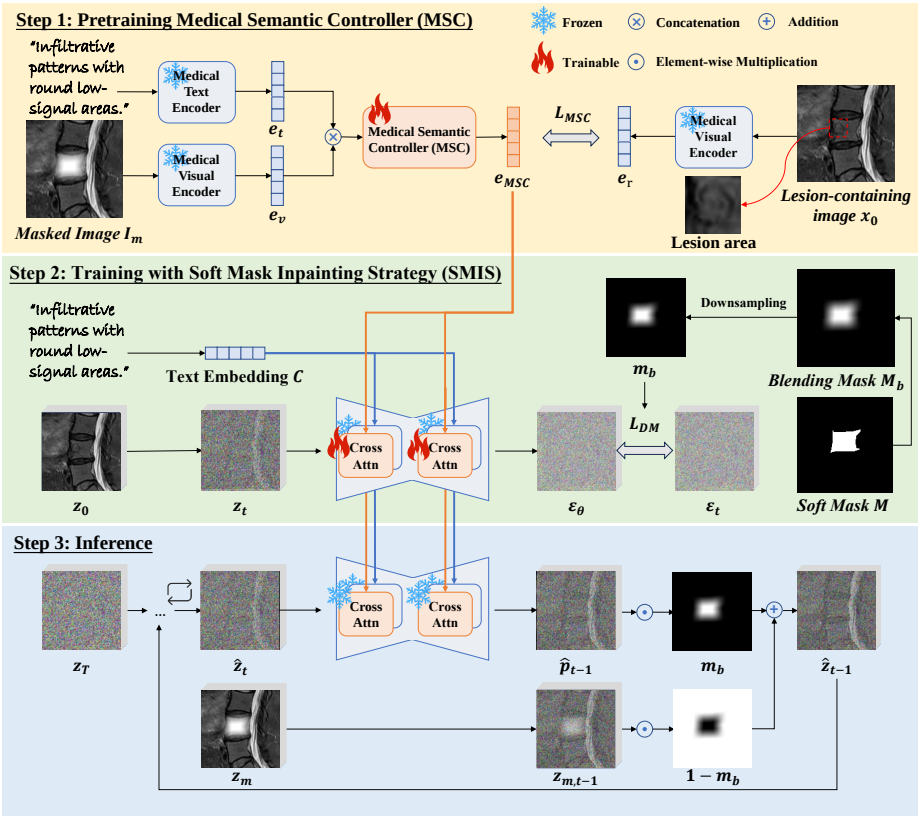

Fig. 1: Overview of our MedSoft-Diffusion method. (1) Pretraining the Medical Semantic Controller (MSC) to predict the medical semantic features of target lesions before the denoising process. (2) Training with the Soft Mask Inpainting Strategy (SMIS), where medical semantic features guide the diffusion model to inpaint lesion regions while the soft mask restricts the learning process to the masked area, ensuring the model focuses on generating realistic lesion textures. (3) During inference, the soft mask confines generation to the lesion region, preserving background integrity and enhancing overall realism.

as input the concatenated textual and visual features and produces a fused representation:

$$e_{MSC} = \text{MSC}(e_t \otimes e_v), \tag{3}$$

where $\otimes$ denotes feature concatenation. The e_{MSC} is then aligned with the full-image feature e_r to ensure that the generated lesion is consistent with real pathological structures. This alignment is enforced by the following loss function:

$$L_{MSC} = \mathbb{E}_{(x_0,M,T)\sim p(\mathcal{X}_0,\mathcal{M},\mathcal{T})}\left[\|e_{MSC} - e_r\|_2^2\right]. \tag{4}$$

By minimizing this loss, MSC learns to project textual pathology descriptions and masked images into a feature space that closely matches real lesion representations, ensuring that the generated pathology characteristics remain both clinically meaningful and visually realistic.

2.2 Training with Soft Mask Inpainting Strategy

The training process is conducted in the latent space [6], where z_0 represents the latent representation of x_0. The mask m_b is the downsampled version of M_b to match the dimensions of the latent space. The forward process of diffusion is defined as: $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where z_t is the noisy image at timestep t, $\epsilon \sim \mathcal{N}(0, I)$ represents Gaussian noise, and $\bar{\alpha}_t$ is the cumulative noise decay factor. To strengthen the interaction between the medical semantic feature e_{MSC} and the text embedding C, we introduce an enhanced cross-attention mechanism in the U-Net structure of the diffusion model, which is formulated as:

$$Z_{\text{new}} = \mathcal{S}\left(\frac{QK^T}{\sqrt{d}}\right)V + \gamma \cdot \mathcal{S}\left(\frac{QK'^T}{\sqrt{d}}\right)V', \tag{5}$$

where γ is a weight balancing the two attention terms, and $\mathcal{S}$ represents the Softmax function. Here, Q, K, and V are the query, key, and value matrices for the attention operation applied to text cross-attention, while K' and V' correspond to medical semantic attention. Given the query features Z and the medical semantic feature e_{MSC}, the query matrix is defined as $Q = ZW_q$, with $K' = e_{MSC}W'_k$ and $V' = e_{MSC}W'_v$. Notably, only W'_k and W'_v are trainable, focusing adaptation on the integration of medical semantics without altering the original text embeddings. This approach enhances the model's ability to synthesize medically meaningful lesion representations. We incorporate the downsampled soft mask m_b to control the inpainting process within the lesion region. The loss function is defined as:

$$L_{DM}(\theta, x) = \mathbb{E}_{z_t \sim q(z_t|z_0), \epsilon \sim \mathcal{N}(0,I)}\left[\|\epsilon_\theta(z_t, t, C, e_{MSC}) - \epsilon\|_2^2 \odot m_b\right], \tag{6}$$

where m_b ensures that the model primarily reconstructs the lesion region. Unlike conventional diffusion models that modify the entire image, our approach localizes generation to the lesion, reducing the learning complexity and enhancing the realism of the generated pathology. Additionally, by leveraging both the medical semantic feature e_{MSC} and the text embedding C, the model generates clinically meaningful pathology features aligned with real medical cases.

2.3 Inference Process

During inference, the trained diffusion model iteratively denoises the input image to synthesize lesions according to the medical text description. The generated image $\hat{z}_{t-1}$ is obtained by blending the predicted image $\hat{p}_{t-1}$ and the noised image $z_{m,t-1}$ using the soft mask m_b:

$$\hat{z}_{t-1} = \hat{p}_{t-1} \odot m_b + z_{m,t-1} \odot (1 - m_b), \tag{7}$$

where $z_{m,t-1}$ follows the noising process: $z_{m,t-1} = \sqrt{\bar{\alpha}_{t-1}} z_m + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon$, and z_m is the latent representation of I_m. The predicted image $\hat{p}_{t-1}$ is estimated using:

$$\hat{p}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_t, t, C, e_{MSC}) \right) + \sigma_t z, \tag{8}$$

where a small noise $\sigma_t z$ is added to enhance diversity in generation. This strategy modifies only the lesion region within the soft mask while preserving anatomical integrity. The blurred soft mask m_b smooths lesion boundaries, ensuring seamless blending and enhancing realism.

3 Experiments

3.1 Experimental Setup

Datasets Two vertebral datasets are used. (1) VerTumor1200 [13] contains 1232 MRI images of vertebral tumors, including benign (hemangiomas) and malignant (metastases, myeloma) cases. Radiologists manually annotate lesion descriptions for clinical accuracy. Soft masks are derived from vertebral segmentation masks in the dataset, offering an accessible alternative to precise lesion masks while preserving anatomical context. (2) RSNA 2024 Lumbar Spine Degenerative Classification [14] includes MRI scans annotated for degenerative conditions. We focus on three sagittal-related diseases: Left Neural Foraminal Narrowing, Right Neural Foraminal Narrowing, and Spinal Canal Stenosis. The dataset provides severity labels (Normal/Mild, Moderate, Severe) across L1/L2 to L5/S1. Lesion descriptions are generated using GPT-4o [15] with prompt engineering for disease types and severity levels, then reviewed by radiologists. Soft masks, created using bounding boxes centered on lesion coordinates, allow flexible localization without requiring pixel-wise segmentation.

Implementations Our method is implemented using the HuggingFace Diffusers library [16] and built on Stable Diffusion 1.5 (SD1.5). The medical text and visual encoder are based on BiomedCLIP [12] for robust medical semantic feature extraction. A new cross-attention layer is added to each of SD1.5's 16 cross-attention layers, with guidance strength $\gamma = 1$. We use the AdamW optimizer [17] with a learning rate of 0.0001 and weight decay of 0.01. For classifier-free guidance, text or medical semantic features are independently dropped with a probability of 0.05, and both together with 0.05 probability. Training is conducted on four NVIDIA 3090 GPUs over three days for 60K iterations. During inference, a 50-step DDIM sampler with a guidance scale of 7.5 ensures a balance between fidelity and generalization in anomaly generation.

3.2 Experimental Results

Quantitative Results To assess the quality of generated medical images, we use Fréchet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), and

Table 1: Comparison Results. The optimal outcomes are highlighted in **bold** font, whereas the next best results are underscored. "w/o" denotes "without".

(a) Quality of image generation

Metric	SD [6]	T2I-Adapter [7]	ControlNet [8]	IP-Adapter [9]	BrushNet [10]	Ours
FID ↓	3.105	3.087	3.098	3.092	3.062	**2.875**
PSNR ↑	10.214	10.135	10.289	10.367	10.312	**10.845**
SSIM ↑	0.634	0.628	0.637	0.641	0.633	**0.669**

(b) Classification Performance

Method	VerTumor1200					RSNA				
	AUC ↑	Acc ↑	Pre ↑	Rec ↑	F1 ↑	AUC ↑	Acc ↑	Pre ↑	Rec ↑	F1 ↑
Baseline[18]	0.838	0.876	0.640	0.777	0.702	0.823	0.843	0.426	0.797	0.555
SD[6]	0.843	0.881	0.653	0.782	0.712	0.824	0.839	0.419	0.805	0.552
T2I-Adapter[7]	0.847	0.884	0.661	0.787	0.718	0.830	0.849	0.438	0.805	0.567
ControlNet[8]	0.849	0.885	0.662	0.793	0.722	0.831	0.850	0.440	0.805	0.569
IP-Adapter[9]	0.861	0.891	0.674	0.814	0.737	0.845	0.857	0.455	0.829	0.588
BrushNet[10]	0.864	0.899	0.700	0.809	0.751	0.841	0.856	0.453	0.821	0.584
Ours w/o MSC	0.861	0.901	0.711	0.798	0.752	0.835	0.851	0.442	0.813	0.573
Ours w/o SMIS	0.879	0.909	0.726	0.830	0.774	0.871	0.891	0.536	0.846	0.656
Ours	**0.898**	**0.930**	**0.795**	**0.846**	**0.820**	**0.890**	**0.905**	**0.575**	**0.870**	**0.693**

Structural Similarity Index Measure (SSIM). We compare our method with Stable Diffusion (SD) [6], T2I-Adapter [7], ControlNet [8], IP-Adapter [9], and BrushNet [10]. As shown in Table 1 (a), our method achieves the best performance across all three metrics, indicating superior image quality, structural accuracy, and clinical realism compared to existing approaches.

To evaluate the effectiveness of our method in vertebral disease diagnosis, we use the area under the curve (AUC), accuracy (Acc), precision (Pre), recall (Rec), and F1-score as evaluation metrics. We employ a baseline classification model [18] and augment the training dataset with synthetic images generated by different methods. We generate synthetic diseased images using only the training data and train the classifier on a mixture of real and synthetic images, while the evaluation is performed solely on real test images. To mitigate class imbalance, we increase each diseased sample category to match the number of normal samples. The impact of synthetic data on classification performance is assessed by measuring the improvement relative to the baseline. As shown in Table 1 (b), our method achieves the best performance across all metrics. Notably, our approach improves the F1-score compared to the baseline, with an increase of 11.8% on the VerTumor1200 dataset (from 0.702 to 0.820) and 13.8% on the RSNA dataset (from 0.555 to 0.693). These results demonstrate that our method enhances classification performance not only by mitigating data imbalance but also by

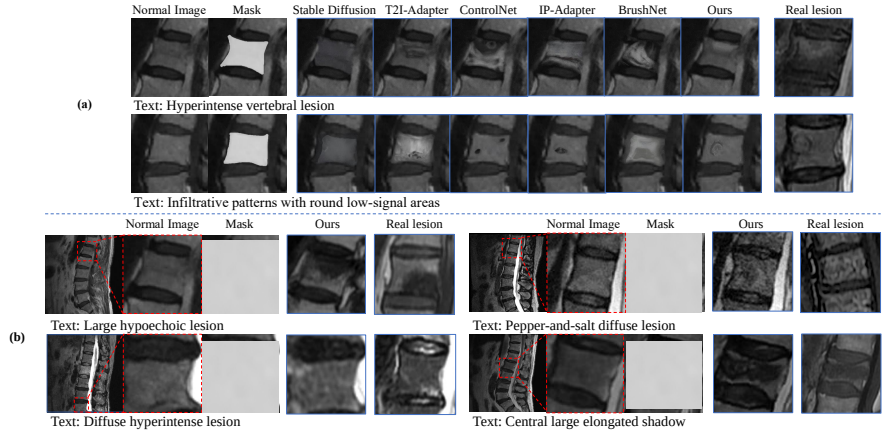

Fig. 2: (a) Lesion synthesis within vertebral-shaped soft masks. Our approach produces more anatomically consistent and realistic pathological features. (b) Lesion-aware vertebra synthesis within bounding box-based soft masks. Our method reconstructs both the vertebral structure and lesions coherently, preserving fine details and structural integrity.

improving the realism and clinical relevance of synthetic diseased images. The ablation study results in Table 1 (b) indicate that MSC plays a more crucial role in directing the model's understanding of pathology, as removing MSC (*Ours w/o MSC*) results in an F1-score drop of 6.8% (from 0.820 to 0.752) on VerTumor1200, whereas removing SMIS (*Ours w/o SMIS*) leads to a smaller decrease of 4.6% (from 0.820 to 0.774), respectively.

Qualitative Results Fig. 2 compares lesion synthesis results under different soft mask constraints, demonstrating that our method better preserves anatomical consistency while generating clear and realistic pathological features. Theoretically, the soft masks can take any shape as long as they cover the intended lesion region. In our experiments, we use vertebral segmentation masks and bounding box masks. The former ensures that lesions are confined within the vertebrae, maintaining anatomical consistency, while the latter allows for shape alterations, enabling the synthesis of lesions that may cause vertebral deformation (as shown in the bottom right example of Fig. 2 (b)).

4 Conclusion

In this work, we propose a novel diffusion-based framework called MedSoft-Diffusion, leveraging detailed medical knowledge to generate high-quality images with specified pathological conditions that are also anatomically accurate. MSC is designed to enhance the alignment between textual prompts and lesion characteristics, ensuring the synthesis of semantically accurate pathological images.

SMIS is proposed to combine soft masks with blurring techniques to improve the realism of synthesized images. Experimental results on two vertebral disease datasets demonstrate notable improvements in both image quality and classification performance using our approach. Moreover, our method can be easily extended to the synthesis of pathological images for other organs.

Acknowledgments. This work is supported by the National Key Research and Development Program Inter-governmental Special Project for International Science and Technology Innovation Cooperation under grant 2022YFE0112500, Foundation for Shenzhen Science and Technology Program under Grant JCYJ20240813151224032, Shenzhen Medical Research Fund under Grant B2402030 Foundation for Shenzhen Science, and Technology Program under Grant JCYJ20240813151102004.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Pelc, M., Vilimkova Kahankova, R., Blaszczyszyn, M., Mikolajewski, D., Konieczny, M., Khoma, V., Bara, G., Zygarlicki, J., Martinek, R., Gupta, M.K., et al.: Initial study on an expert system for spine diseases screening using inertial measurement unit. Scientific Reports **13**(1), 10440 (2023)
2. Dong, Y., Peng, R., Kang, H., Song, K., Guo, Q., Zhao, H., Zhu, M., Zhang, Y., Guan, H., Li, F.: Global incidence, prevalence, and disability of vertebral fractures: a systematic analysis of the global burden of disease study 2019. The Spine Journal **22**(5), 857–868 (2022)
3. Kumar, P., Bhatnagar, R., Gaur, K., Bhatnagar, A.: Classification of imbalanced data: review of methods and applications. In: IOP conference series: materials science and engineering. vol. 1099, p. 012077. IOP Publishing (2021)
4. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10795–10816 (2023)
5. Luo, Y., Yang, Q., Fan, Y., Qi, H., Xia, M.: Measurement guidance in diffusion models: Insight from medical image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
7. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 38, pp. 4296–4304 (2024)
8. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (2023)
9. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)

10. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: European Conference on Computer Vision (ECCV). pp. 150–168. Springer (2024)
11. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 11461–11471 (2022)
12. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2024)
13. Zhao, S., Chen, B., Chang, H., Chen, B., Li, S.: Reasoning discriminative dictionary-embedded network for fully automatic vertebrae tumor diagnosis. Medical Image Analysis **79**, 102456 (2022)
14. Richards, T., Talbott, J., Ball, R., Colak, E., Flanders, A., Kitamura, F., Mongan, J., Prevedello, L., Vazirabad., M.: Rsna 2024 lumbar spine degenerative classification. https://kaggle.com/competitions/rsna-2024-lumbar-spine-degenerative-classification (2024), kaggle
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
16. Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022)
17. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)