This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Mask-Guided Attention Regulation for Anatomically Consistent Counterfactual CXR Synthesis

Zichun Zhang, Weizhi Nie*, Honglin Guo, and Yuting Su

School of Electrical Automation and Information Engineering, Tianjin University, Tianjin 300072, China
weizhinie@tju.edu.cn

Abstract. Counterfactual generation for chest X-rays (CXR) aims to simulate plausible pathological changes while preserving patient-specific anatomy. However, diffusion-based editing methods often suffer from structural drift, where stable anatomical semantics propagate globally through attention and distort non-target regions, and unstable pathology expression, since subtle and localized lesions induce weak and noisy conditioning signals. We present an inference-time attention regulation framework for reliable counterfactual CXR synthesis. An anatomy-aware attention regularization module gates self-attention with organ masks, confining structural interactions to anatomical ROIs and reducing unintended distortions. A pathology-guided module enhances pathology-token cross-attention within target anatomical regions during early denoising and performs lightweight latent corrections driven by an attention-concentration energy, enabling controllable lesion localization and extent. Extensive evaluations on CXR datasets show improved anatomical consistency and more precise, controllable pathological edits compared with standard diffusion editing, supporting localized counterfactual analysis and data augmentation for downstream tasks.

Keywords: Counterfactual generation · Diffusion model · Attention regulation.

1 Introduction

Counterfactual medical image generation aims to simulate plausible imaging outcomes under hypothetical changes in pathological conditions while preserving the underlying anatomical structure of the patient, thereby constructing anatomy-preserving "what-if" image pairs for controlled pathological analysis [8, 22]. In chest X-ray (CXR) imaging, counterfactual generation allows localized pulmonary abnormalities to be introduced, removed, or progressively altered without disrupting stable anatomical components, including lung shape, rib structures, and cardiac contours [3]. By explicitly disentangling pathological variations from anatomical consistency, counterfactual medical image generation provides a useful framework for auditing diagnostic models, supporting

controlled data augmentation, and analyzing localized pathology-related visual changes [7].

With the rapid development of diffusion-based generative models [13], a growing body of work has explored counterfactual medical image generation by leveraging image inpainting [16], instruction-driven editing [2] [10], or conditional synthesis [27] [25] frameworks. However, achieving stable and controllable editing often requires domain-specific retraining or learnable control branches [25], adding recurring tuning and data-governance costs under cross-institution variability and continuously evolving data [23], which hinders deployment at scale.

Therefore, we impose constraints on the diffusion sampling process at inference time to improve controllability without additional training overhead. Under this setting, we still need to address the following key challenges:

- **Structural instability.** In diffusion models, global anatomical structures tend to stabilize early and dominate subsequent generation through self-attention [11]. When pathology-related prompts are introduced, this global propagation can cause localized changes to spread to non-target regions [18], leading to unintended structural distortions and reduced anatomical consistency [24].
- **Pathological expression instability.** Pathological features in medical images are often subtle, spatially confined, and heterogeneous, resulting in weak attention responses during generation [8]. Consequently, such features may be suppressed or unintentionally diffused, leading to inaccurate localization or uncontrolled lesion extent [6].

Motivated by these challenges, we propose an inference-time attention regulation framework for counterfactual medical image generation. Specifically, we first introduce an anatomy-aware attention regularization strategy that constrains self-attention responses [21] within valid anatomical regions, preventing structural semantics from excessively propagating into pathology-sensitive areas. Furthermore, to enable precise control over localized pathological changes, we design a pathology-guided attention regulation mechanism that enhances pathology-related cross-attention responses within target anatomical ROIs via a mask-derived spatial prior [5] and applies lightweight corrections to intermediate noise representations during denoising, allowing accurate lesion localization and extent control.

In summary, our contributions are threefold:

- We propose an inference-time attention regulation framework for counterfactual medical image generation, which improves controllability across target pathologies without additional control-branch training.
- We jointly regularize anatomy-aware self-attention and pathology-guided cross-attention to preserve structure while enabling reliable, localized pathological edits.
- Extensive experiments on chest X-ray datasets demonstrate the effectiveness of the proposed framework in producing anatomically consistent and pathology-controllable counterfactual images.

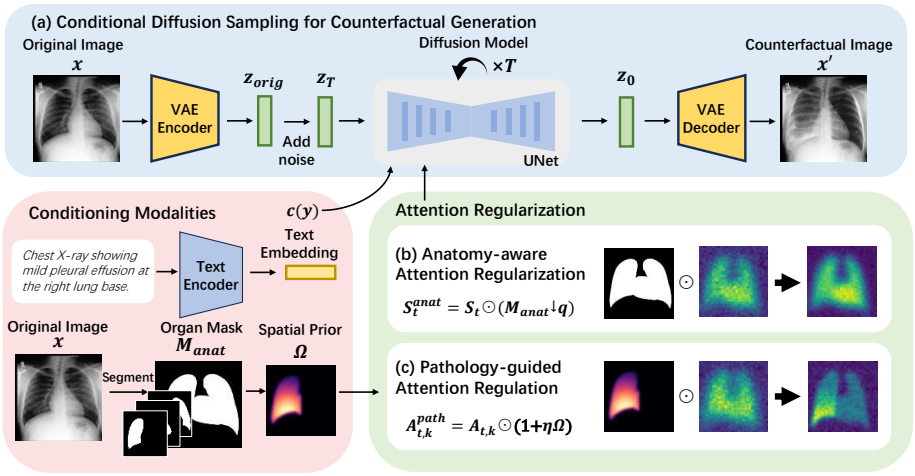

Fig. 1. Overview of our inference-time attention regulation framework for counterfactual CXR generation. (a) The input image is encoded by VAE [15], noised, and denoised by a conditional diffusion model to generate a counterfactual image. (b) Anatomy-aware self-attention gating with M_{anat} preserves structural consistency, while (c) pathology-guided cross-attention reweighting with a mask-derived prior Ω localizes pathological edits.

2 Method

2.1 Problem Definition

Given an input medical image x, an organ mask M_{anat}, and a counterfactual condition y describing the target pathological state, our goal is to generate a counterfactual image x' that preserves anatomical consistency while realizing the desired pathological variation. We denote the counterfactual generator as

$$x' = \mathcal{G}(x, M_{\text{anat}}, c(y); \xi), \tag{1}$$

where ξ denotes the stochasticity.

As shown in Fig.1 (a), in diffusion sampling, we first encode the input image x into a latent $z_{\text{orig}} = \text{Enc}(x)$ and obtain a noisy initialization by forward diffusion:

$$z_T = \sqrt{\bar{\alpha}_T}\, z_{\text{orig}} + \sqrt{1-\bar{\alpha}_T}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

We then perform conditional iterative denoising:

$$z_{t-1} = F_\theta(z_t;\, x, M_{\text{anat}}, c(y), t)\,, \quad t = T, \ldots, 1, \qquad x' = \text{Dec}(z_0), \tag{3}$$

where F_θ is the conditional denoising update implemented by a UNet [20] with parameters θ, and $\text{Dec}(\cdot)$ is the decoder. In this work, we introduce lightweight inference-time regulation on the denoising updates in Eq. (3); the following sections describe our anatomy and pathology regulation strategies.

2.2 Anatomy-aware Attention Regularization

We consider a diffusion UNet where, at each denoising step t, intermediate latents are updated via attention blocks. Let S_t denote the self-attention map capturing spatial interactions, and let A_t denote the cross-attention map aligning conditioning tokens to spatial locations:

$$S_t = \mathrm{Softmax}\left(\frac{Q_{\mathrm{self}}K_{\mathrm{self}}^{\top}}{\sqrt{d}}\right), \qquad A_t = \mathrm{Softmax}\left(\frac{Q_{\mathrm{cross}}K^{\top}}{\sqrt{d}}\right), \tag{4}$$

where $Q_{\mathrm{self}}, K_{\mathrm{self}}$ are projected from current features, and K is projected from the condition embedding $c(y)$. Given an organ mask M_{anat}, we denote by $(M_{\mathrm{anat}} \downarrow q)$ its downsampled version to match the spatial resolution q of the attention map at a specific UNet layer.

Stable anatomical structures tend to become dominant early in denoising and can be propagated through self-attention across the image [11]. When pathological conditions are introduced, such global propagation may induce unintended structural drifts in non-target regions, degrading anatomical consistency. We therefore constrain anatomy-related attention responses to the corresponding anatomical regions using the available organ masks.

As shown in Fig. 1 (b), we apply inference-time region gating to restrict attention-driven information flow by M_{anat}. For self-attention, we gate the attention response at each layer:

$$S_t^{\mathrm{anat}} = S_t \odot (M_{\mathrm{anat}} \downarrow q), \tag{5}$$

which suppresses anatomy-driven interactions outside the anatomical ROI and reduces non-target structural perturbations during denoising.

2.3 Pathology-guided Attention Regulation

To realize counterfactual edits, we regulate token-wise cross-attention maps and apply a lightweight trajectory correction during inference as shown in Fig. 1 (c). We interpret the token-wise cross-attention weight map $A_{t,k}$ as the spatial attribution of token k over latent locations at denoising step t. We construct a sample-specific spatial prior map Ω at the same resolution as $A_{t,k}$.

Given the organ masks (e.g., left/right lung and heart), we deterministically select or combine an ROI mask M_{ROI} using simple anatomy cues from the text y (e.g., laterality and coarse lung regions), and obtain Ω by downsampling M_{ROI} to the attention resolution followed by optional smoothing and normalization, i.e., $\Omega = \mathrm{Normalize}(\mathrm{Blur}(M_{\mathrm{ROI}} \downarrow q))$.

Let K_{path} denote the index set of pathology-related tokens, and let $A_{t,k}$ be the cross-attention map for token $k \in K_{\mathrm{path}}$ at step t. During early denoising steps, we reweight token-wise cross-attention with a soft multiplier map $(1+\eta\Omega)$:

$$A_{t,k}^{\mathrm{path}} = A_{t,k} \odot \Big(1 + \eta\,\Omega\Big), \quad k \in K_{\mathrm{path}},\ t < \mu T, \tag{6}$$

where η controls the enhancement strength and μ specifies the early-step window.

To mitigate attention diffusion and leakage to non-target areas during editing, we introduce a differentiable concentration metric that quantifies how well pathology-token attribution aligns with the target ROI, which further enables a lightweight inference-time correction. We define a region concentration score:

$$\mathrm{score}_{t,k} = \frac{\langle A_{t,k}^{\mathrm{path}},\, \Omega \rangle}{\langle A_{t,k}^{\mathrm{path}},\, \mathbf{1} \rangle}, \tag{7}$$

and the pathology energy

$$L_{\mathrm{path}}(t) = 1 - \frac{1}{|K_{\mathrm{path}}|} \sum_{k \in K_{\mathrm{path}}} \mathrm{score}_{t,k}. \tag{8}$$

We then apply a single inference-time correction to the intermediate latent within the same early-step window:

$$\hat{z}_t \leftarrow z_t - \alpha_t \nabla_{z_t} L_{\mathrm{path}}(t), \tag{9}$$

where α_t is a step-size schedule. This lightweight update steers the denoising trajectory toward better ROI concentration, improving lesion localization and controlling the affected extent.

2.4 Inference Procedure

Given $(x, M_{\mathrm{anat}}, y)$, we run conditional diffusion sampling for $t = T, \ldots, 1$. At each denoising step, we first apply anatomy-aware self-attention regularization by gating self-attention with the downsampled organ mask in Eq. (5), thereby restricting attention-driven propagation to anatomically valid regions. We then perform pathology-guided cross-attention regulation for pathology tokens by reweighting token-wise cross-attention with the spatial prior Ω in Eq. (6). During early steps ($t < \mu T$), we additionally apply a single latent correction update in Eq. (9) using the pathology energy in Eq. (8) before the denoising update. Finally, we decode the resulting z_0 to obtain the counterfactual image x'.

3 Experiments

Datasets. We use frontal-view (PA) studies from MIMIC-CXR-JPG [14] and ChexpertPlus [4], resizing images to 512×512 as a practical trade-off between computational cost and visual fidelity. Multi-organ anatomical masks are obtained by HybridGNet [9]. We refine the paired report text using a GPT-based [1] cleaning step that removes boilerplate and non-imaging content and compresses the remaining findings into concise prompts. Text conditions are truncated to fit the SD tokenizer limit (77 tokens).

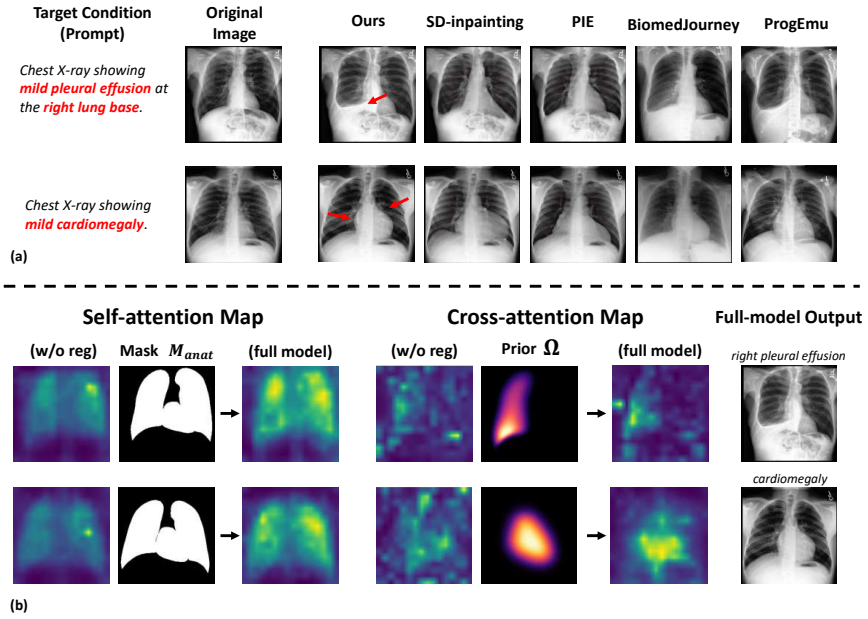

Fig. 2. Qualitative comparison and attention visualization for counterfactual CXR generation. (a) Comparison with state-of-the-art counterfactual CXR generation methods under two target prompts. (b) Attention-map visualization of the proposed regulation strategy. M_{anat} and Ω are used to guide self-attention regularization and pathology-token cross-attention regulation, respectively.

Implementation Details. Our generator is built on Stable Diffusion v1.5 [19] and is fine-tuned on chest X-ray data for domain adaptation. The backbone is fine-tuned for 50 epochs using AdamW with a learning rate of 1×10^{-4} and an effective batch size of 16; we use mixed precision (bf16) and gradient accumulation under GPU memory constraints. We perform DDIM sampling with T=50 steps and classifier-free guidance s=7.5. For attention regulation, we align the masks/prior to the attention resolution and use $q \in \{64, 32, 16\}$ for 512×512 inputs. Pathology-guided attention regulation is applied in an early window $t < \mu T$ with μ=0.4 and enhancement strength η=1.5. Within the same window, latent correction uses one gradient step per denoising step with a linearly decayed step size α_t, where $\alpha_{\max}$=0.05 at the start of the window.

3.1 Comparison with State-of-the-Art

We conduct both quantitative and qualitative comparisons with state-of-the-art methods for counterfactual CXR generation. As summarized in Table 1, our method achieves the best Conf and CLIP-I when evaluated on a selected subset of the MIMIC-CXR-JPG dataset. For pathological accuracy, our approach

Table 1. Performance comparison with state-of-the-art methods on the counterfactual CXR generation task.

Method	Conf↑	CLIP-I↑	LPIPS[26]↓	FID[12]↓
SD-inpainting[19]	0.662	0.827	0.23	37.6
BiomedJourney[10]	0.693	0.868	0.19	30.8
ProgEmu[17]	0.701	0.866	**0.17**	**28.4**
PIE[16]	0.691	0.842	0.18	29.7
Ours	**0.709**	**0.870**	0.18	29.0

Table 2. Ablation study of inference-time regulation modules on the counterfactual CXR generation task.

Variant	Conf↑	CLIP-I↑	SSIM↑
Full (Ours)	**0.71**	**0.87**	**0.80**
w/o anatomy self-attn regularization (Eq. 5)	0.69	0.85	0.76
w/o pathology cross-attn regulation (Eq. 6)	0.66	0.82	0.80
w/o latent correction (Eq. 9)	0.70	0.86	0.79

provides more reliable and better-aligned target edits, indicating that the proposed inference-time regulation improves the intended pathological changes while maintaining global semantic consistency. Although ProgEmu obtains slightly better LPIPS and FID, our results remain competitive in terms of perceptual realism and distribution-level fidelity, suggesting that the improved controllability does not come at the expense of visual plausibility.

Fig. 2 (a) further shows representative visual examples. Compared with instruction-based editing methods, our generations exhibit stronger stability in background and non-target regions. Meanwhile, relative to prior editing/inpainting baselines, the pathological changes are more accurate and more tightly confined to relevant regions, demonstrating better preservation of irrelevant areas and validating the effectiveness of our framework.

3.2 Ablation Study

As shown in Table 2, we conduct an ablation study on the MIMIC-CXR-JPG dataset to evaluate pathological accuracy and the preservation of structural characteristics in counterfactual image generation. Overall, the full configuration consistently yields the best results, indicating that neither anatomy stabilization nor pathology injection alone is sufficient for robust counterfactual generation. In particular, anatomy-aware self-attention gating primarily safeguards structural consistency and global appearance, while pathology-guided cross-attention regulation is the key factor for reliably expressing the intended pathological change. The latent correction step provides an additional layer of stability: although its impact is more modest, it consistently improves the final outcome by refining the

denoising trajectory, leading to more dependable edits without over-amplifying the target signal.

The attention visualization in Fig. 2 (b) further provides a qualitative explanation for the ablation results. In the representative cases, the full model shows more anatomically constrained self-attention under the guidance of M_{anat}, and pathology-token cross-attention becomes more concentrated around the soft spatial prior Ω compared with the unregulated setting. These observations are consistent with the quantitative improvements in Table 2, suggesting that the proposed regulation helps preserve anatomical structure and guide localized pathology editing.

4 Conclusion

In this work, we propose an inference-time attention regulation framework for counterfactual medical image generation to avoid recurring retraining and tuning costs. It addresses two practical challenges in diffusion-based medical editing: maintaining structural fidelity and producing accurate, localized pathological changes. By injecting organ-mask priors into attention modulation, our method suppresses unintended alterations in irrelevant regions while strengthening pathology-focused attention within target areas to better realize localized pathological edits. Extensive experiments demonstrate improved anatomical consistency and semantic alignment over strong diffusion editing baselines, highlighting its potential for controlled counterfactual analysis and research-oriented data augmentation.

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
3. Çallı, E., Sogancioglu, E., Van Ginneken, B., Van Leeuwen, K.G., Murphy, K.: Deep learning for chest x-ray analysis: A survey. Medical image analysis **72**, 102125 (2021)
4. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats (2024), https://arxiv.org/abs/2405.19538
5. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5343–5353 (2024)

6. Chen, Z., Wei, F., Xu, R., Li, J., Duan, L., Yao, A., Li, W.: The devil is in attention sharing: Improving complex non-rigid image editing faithfulness via attention synergy. arXiv preprint arXiv:2512.14423 (2025)
7. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
8. Cohen, J.P., Brooks, R., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.: Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest x-rays. In: International Conference on Medical Imaging with Deep Learning (MIDL). pp. 74–104 (2021)
9. Gaggion, N., Mansilla, L., Mosquera, C., Milone, D.H., Ferrante, E.: Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. IEEE Transactions on Medical Imaging **42**(2), 546–556 (2022)
10. Gu, Y., Yang, J., Usuyama, N., Li, C., Zhang, S., Lungren, M.P., Gao, J., Poon, H.: Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. arXiv preprint arXiv:2310.10765 (2023)
11. Guo, H., Chen, R., Nie, W., Wang, L., Liu, A.: Compcraft: Foreground-driven image synthesis with customized layouts. IEEE Transactions on Circuits and Systems for Video Technology (2025)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
15. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
16. Liang, K., Cao, X., Liao, K.D., Gao, T., Ye, W., Chen, Z., Cao, J., Nama, T., Sun, J.: Pie: Simulating disease progression via progressive image editing (2023), https://arxiv.org/abs/2309.11745
17. Ma, C., Ji, Y., Ye, J., Zhang, L., Chen, Y., Li, T., Li, M., He, J., Shan, H.: Towards interpretable counterfactual generation via multimodal autoregression (2025), https://arxiv.org/abs/2503.23149
18. Min, H., You, T., Lee, H., Cho, Y., Cho, S.: Instructx2x: An interpretable local editing model for counterfactual medical image generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 279–288. Springer (2025)
19. 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. pp. 10684–10695 (2022)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
22. Wang, T., Wang, W., Giunchiglia, F., Zhao, F., Zhang, Y., Yu, D., Liu, G.: Mbt-polyp: A new multi-branch memory-augmented transformer for polyp segmentation. Image and Vision Computing **163**,

105747 (2025). https://doi.org/https://doi.org/10.1016/j.imavis.2025.105747, https://www.sciencedirect.com/science/article/pii/S026288562500335X
23. Wang, T., Lei, Y., Fu, Y., Wynne, J.F., Curran, W.J., Liu, T., Yang, X.: A review on medical imaging synthesis using deep learning and its clinical applications. Journal of applied clinical medical physics **22**(1), 11–36 (2021)
24. Yang, F., Yang, S., Butt, M.A., van de Weijer, J., et al.: Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. Advances in Neural Information Processing Systems **36**, 26291–26303 (2023)
25. 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. pp. 3836–3847 (2023)
26. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
27. Zhang, Z., Yao, L., Wang, B., Jha, D., Keles, E., Medetalibeyoğlu, A., Bagci, U.: Emit-diff: Enhancing medical image segmentation via text-guided diffusion model. arXiv.org (2023). https://doi.org/10.48550/arxiv.2310.12868