MedTriFlow: Efficient Resolution-Agnostic 3D Medical Image Generation with Implicit Triplane Representation

Chenfan Xu[1], Yulong Dou[1], Tao Luo[1], Qian Wang[1], and Zhiming Cui[1](✉)

School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
cuizhm@shanghaitech.edu.cn

Abstract. 3D medical image generation is crucial for data augmentation and downstream analysis but remains computationally expensive. Existing voxel-based generative models operate on fixed-resolution grids, where the latent representation is coupled with output resolution. This leads to cubic scaling of computational costs and inability to generate volumes at varying resolutions without retraining. Triplane representations offer a more compact and efficient alternative. However, current triplane-based approaches either require costly instance-level optimization or are constrained by discrete voxel supervision, preventing resolution-agnostic generation. To address these challenges, we propose MedTriFlow, an efficient generative framework that bridges a discrete triplane latent space with a continuous anatomical field. The framework employs a triplane-based autoencoder with 3D-aware projection to compress volumetric data into a compact latent space, where generative modeling is performed for efficient generation. A continuous anatomical field, modeled by an implicit neural representation, then reconstructs volumetric anatomy at arbitrary resolutions, effectively enabling resolution-agnostic generation beyond the discrete voxel grid. Extensive experiments on diverse 3D datasets demonstrate that the proposed method, with fast inference (2.6s per volume at 256^3) and resolution-agnostic generation (demonstrated up to 1024^3), achieves superiority over recent 3D generative methods in generation quality, sparse-view CBCT and accelerated MRI reconstruction. Code and pre-trained models are available at https://github.com/ShanghaiTech-IMPACT/MedTriFlow.

Keywords: Resolution-Agnostic Generation · 3D Medical Image Generation · Triplane Representation.

1 Introduction

3D medical image generation plays a vital role in data augmentation and downstream analysis, with recent advances in generative models significantly improving medical image translation, segmentation, and reconstruction [20,1,16,3]. 3D medical images exhibit substantial resolution heterogeneity across modalities,

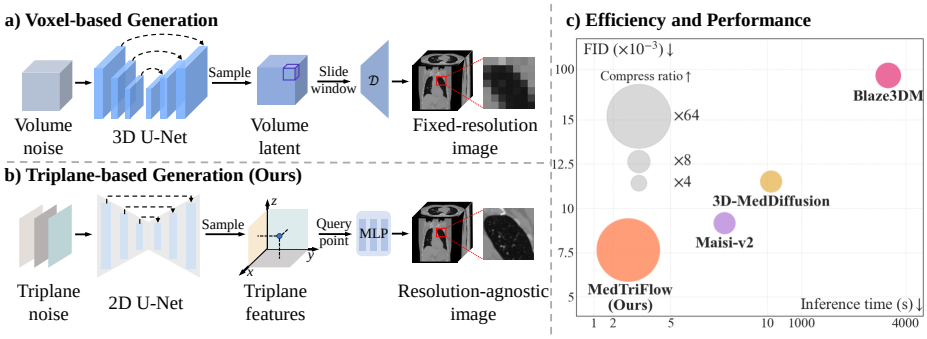

Fig. 1. Workflow and performance comparison. (a) Voxel-based generation with fixed-resolution output. (b) Our triplane-based generation enabling resolution-agnostic output. (c) Efficiency and performance comparison: MedTriFlow achieves the best FID and inference speed with a superior compression ratio (bubble size).

and clinical practice often requires multi-scale analysis of the same anatomy. However, existing generative models cannot flexibly produce outputs at different resolutions without retraining or interpolation. This limitation, compounded by the cubic scaling of computational costs with resolution, makes 3D resolution-agnostic generation a critical yet unsolved challenge.

As shown in Fig. 1(a), the predominant paradigm for 3D medical image generation leverages volumetric latent representations, where images are encoded into voxel grids via VQ-VAE [19] or VAE [14], and generated using 3D UNet backbones [21] with DDPM [12] or Flow Matching [15] (e.g., MedSyn [25], 3D-MedDiff [23], Maisi-v2 [28]). While effective, these voxel-based approaches incur substantial computational overhead due to the cubic scaling of memory and 3D convolution operations with voxel grid resolution. Moreover, the output resolution is inherently coupled with the pre-defined latent grid, limiting clinical flexibility. Thus, triplane-based representations have been introduced to increase efficiency. Existing methods can be broadly categorized into decoder-only frameworks that require per-instance latent optimization [9,27] and encoder-based approaches [17] that directly map volumes to triplane features. However, decoder-only strategies are computationally prohibitive for large-scale generative pre-training, while current encoder-based models remain supervised by voxel-grid reconstruction, thereby inheriting fixed-resolution constraints.

To address these limitations, we propose MedTriFlow, an efficient framework for resolution-agnostic 3D medical image generation. First, a triplane-based autoencoder learns a compact latent representation and constructs a continuous anatomical field via implicit neural representation. Second, generative modeling is performed in the triplane latent space using flow matching, significantly reducing computational complexity. Benefiting from the ability to reconstruct volumes at arbitrary resolutions (Fig. 1(b)), this design enables flexible, resolution-agnostic 3D generation. Furthermore, a structure-guided adaptation mechanism

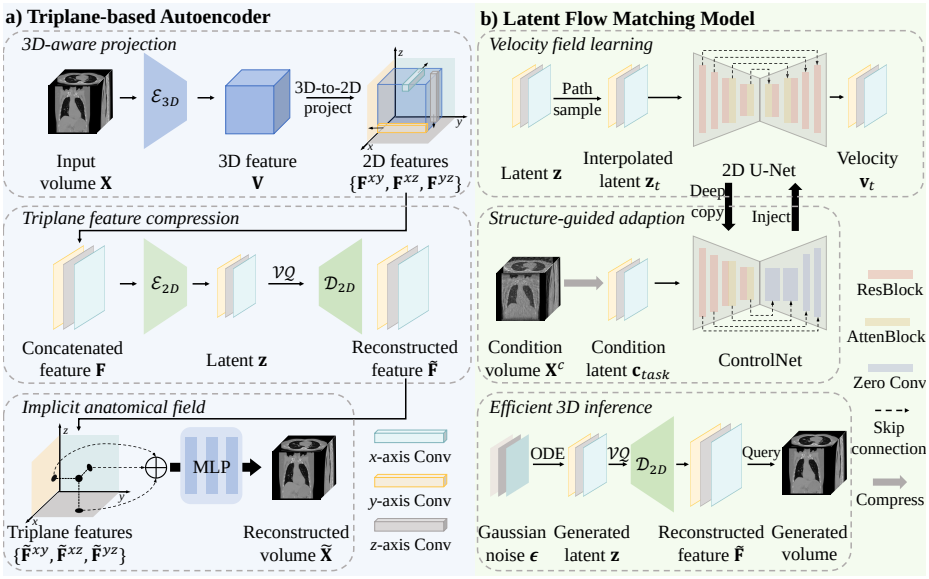

Fig. 2. Overview of the proposed framework. (a) Triplane-based autoencoder for continuous anatomical field modeling and latent compression. (b) Latent flow matching model for efficient, structure-guided 3D generation and reconstruction.

supports conditional generation for diverse clinical reconstruction tasks. Extensive experiments demonstrate consistent advantages over prior voxel- and triplane-based approaches in challenging settings such as sparse-view CT and accelerated MRI reconstruction. As shown in Fig. 1(c), our method also achieves superior compression ratio and inference speed.

2 Method

We propose an efficient framework for resolution-agnostic 3D medical image generation that bridges the discrete triplane latent space and a continuous anatomical field. As shown in Fig. 2, a triplane-based autoencoder (Sec. 2.1) compresses 3D volumes into a compact 2D latent for resolution-agnostic modeling. Subsequently, a latent flow matching model (Sec. 2.2) operates in the triplane latent space to perform efficient 3D generation.

2.1 Triplane-based Autoencoder

Given a medical volume $\mathbf{X} \in \mathbb{R}^{H\times W\times D}$ with spatial dimensions H, W, and D along the x, y, and z axes, we represent the 3D anatomical structures as a continuous anatomical field using a triplane-based implicit neural representation as shown in Fig. 2(a). To facilitate efficient downstream generation, we further

compress the triplane features into a compact latent space while preserving fine structural details.

3D-aware Projection. To extract both global structures and local textures, a lightweight 3D U-Net $\mathcal{E}_{3D}$ first encodes the volume $\mathbf{X}$ into a feature tensor $\mathbf{V} \in \mathbb{R}^{C\times H\times W\times D}$ where C denotes the channel dimension. To minimize information loss in the 3D-to-2D projection, three anisotropic convolutions with kernel sizes of $(H, 1, 1)$, $(1, W, 1)$ and $(1, 1, D)$ are respectively applied along x, y and z axes, producing triplane features $\{\mathbf{F}^{yz} \in \mathbb{R}^{C\times W\times D}, \mathbf{F}^{xz} \in \mathbb{R}^{C\times H\times D}, \mathbf{F}^{xy} \in \mathbb{R}^{C\times H\times W}\}$. Compared to global axis-wise mean pooling, these learnable projections aggregate cross-slice context and better preserve fine anatomical details during dimensionality reduction.

Triplane Feature Compression. To achieve higher compression efficiency, the triplane features are first padded to $N \times N$ in their spatial dimensions ($N = \max(H, W, D)$), then concatenated along the channel dimension to form $\mathbf{F} \in \mathbb{R}^{3C\times N\times N}$ and mapped into a compact latent space. A 2D encoder $\mathcal{E}_{2D}$ then maps $\mathbf{F}$ to a triplane latent $\mathbf{z} = \mathcal{E}_{2D}(\mathbf{F}) \in \mathbb{R}^{c\times n\times n}$, where c and n denote the channel and spatial dimensions, respectively. This latent representation is regularized via vector quantization $\tilde{\mathbf{z}} = \mathcal{VQ}(\mathbf{z})$ to constrain the latent distribution. A 2D decoder $\mathcal{D}_{2D}$ subsequently reconstructs the triplane features $\tilde{\mathbf{F}} = \mathcal{D}_{2D}(\tilde{\mathbf{z}})$ for anatomical field modeling.

Implicit Anatomical Field. Based on the reconstructed triplane features $\tilde{\mathbf{F}} = \{\tilde{\mathbf{F}}^{yz}, \tilde{\mathbf{F}}^{xz}, \tilde{\mathbf{F}}^{xy}\}$, we define a continuous anatomical field $\mathcal{A}$. For any 3D point $\mathbf{o} = (x, y, z) \in \mathbb{R}^3$, its intensity is decoded via the implicit neural function Ψ_ϕ:

$$\mathcal{A}(\mathbf{o}; \tilde{\mathbf{F}}, \phi) = \Psi_\phi \left(\mathcal{I}(\tilde{\mathbf{F}}^{yz}; (y, z)) \oplus \mathcal{I}(\tilde{\mathbf{F}}^{xz}; (x, z)) \oplus \mathcal{I}(\tilde{\mathbf{F}}^{xy}; (x, y)) \right), \tag{1}$$

where $\mathcal{I}(\cdot;\cdot)$ denotes bilinear interpolation, and $\oplus$ represents channel-wise concatenation. This formulation enables $\mathcal{A}$ to map a discrete triplane representation into a resolution-agnostic continuous anatomical field.

Training Loss. To ensure artifact-free and accurate reconstruction, the framework is optimized using a synergistic loss function:

$$\mathcal{L} = \mathcal{L}_{Rec} + \mathcal{L}_{VQ} + \lambda_1 \mathcal{L}_{LPIPS} + \lambda_2 \mathcal{L}_{Adv}, \tag{2}$$

where $\mathcal{L}_{Rec}$ measures point-wise reconstruction error, $\mathcal{L}_{VQ}$ enforces latent quantization, $\mathcal{L}_{LPIPS}$ with weight λ_1 promotes perceptual similarity in feature space, and $\mathcal{L}_{Adv}$ with weight λ_2 encourages sharp and realistic anatomical textures via adversarial learning.

2.2 Latent Flow Matching Model

To model the complex distribution of 3D anatomy, our generative modeling is performed in the compact triplane latent space before vector quantization (Fig. 2(b)). This 2D latent generation substantially reduces computational cost while preserving rich structural representation.

Velocity Field Learning. Flow matching (FM) is employed to learn a deterministic path between a noise distribution p_{init} and the triplane latent distribution p_{data}. Specifically, we define a linear path $\mathbf{z}_t = t\mathbf{z} + (1-t)\boldsymbol{\epsilon}$, where $\mathbf{z} \sim p_{data}$ and $\boldsymbol{\epsilon} \sim p_{init}$ is Gaussian noise. The generative process is governed by a velocity field $\mathbf{v}_t$ that transforms noise into data. We parametrize this field using a 2D U-Net [6] $\hat{\mathbf{v}}_\theta$ comprising multiple ResBlocks [10] and self-attention blocks (AttenBlock), optimized via the conditional flow matching loss:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{z},\mathbf{c}} \left[\| (\mathbf{z} - \boldsymbol{\epsilon}) - \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t, \mathbf{c}) \|_2^2 \right], \tag{3}$$

where t belongs to the uniform distribution $\mathcal{U}[0,1]$ and $\mathbf{c}$ denotes the class condition. The U-Net is pre-trained on a diverse corpus of 3D medical images spanning different modalities, learning robust generative priors for human anatomy.

Structure-guided Adaptation. To adapt the pre-trained generative prior to downstream tasks, such as sparse-view CT reconstruction and accelerated MRI reconstruction, we incorporate a ControlNet [26] branch. The task-specific input $\mathbf{X}^c$ (e.g., an FDK volume) is projected and compressed into a conditional triplane latent $\mathbf{c}_{task}$ using the triplane-based autoencoder described in Sec. 2.1. This latent is fed into a ControlNet branch, initialized from a deep copy of the U-Net encoder, which extracts and injects structural priors into the main generative backbone through zero-convolution layers. The velocity field $\hat{\mathbf{v}}_\theta$ is thus conditioned on $\mathbf{c}_{task}$, providing structure-guided generation.

Efficient 3D Inference. During inference, generation proceeds in three steps: (1) starting from $t = 0$, the ordinary differential equation (ODE) $\frac{d\mathbf{z}_t}{dt} = \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t, \mathbf{c})$ is solved to obtain the generated latent $\mathbf{z}$; (2) this latent is quantized and decoded via the 2D decoder $\mathcal{D}_{2D}$ to reconstruct the triplane features $\tilde{\mathbf{F}} = \mathcal{D}_{2D}(\mathcal{VQ}(\mathbf{z}))$; (3) finally, the full 3D volume $\tilde{\mathbf{X}}$ is generated by querying the anatomical field $\mathcal{A}$ at the center coordinates of each target voxel. Benefiting from the straight trajectories of FM, 3D volumes with flexible resolutions can be generated within only 20 integration steps, significantly outperforming traditional diffusion models in inference efficiency.

3 Experiments

3.1 Experimental Settings

Dataset. We evaluate our method on five diverse 3D cohorts, including head-neck CT, chest-abdomen CT, lower limb CT, brain MRI, and body MRI. In total, 7088 volumes are collected from public benchmarks [4,2,5,18,22] and in-house datasets. All volumes were cropped to the foreground, resampled to isotropic spacing (<1.25mm), and center-padded to 256^3. Data were split into 6288 training and 800 testing cases with stratified sampling across cohorts. For inverse tasks, 30-view CBCT projections and 8× undersampled MRI were simulated.

Implementation Details. The triplane-based autoencoder employs a lightweight 3D U-Net encoder with channel sizes $[32, 64, 128]$, followed by 2D compression modules with channels $[64, 128, 256]$ and a vector-quantized codebook

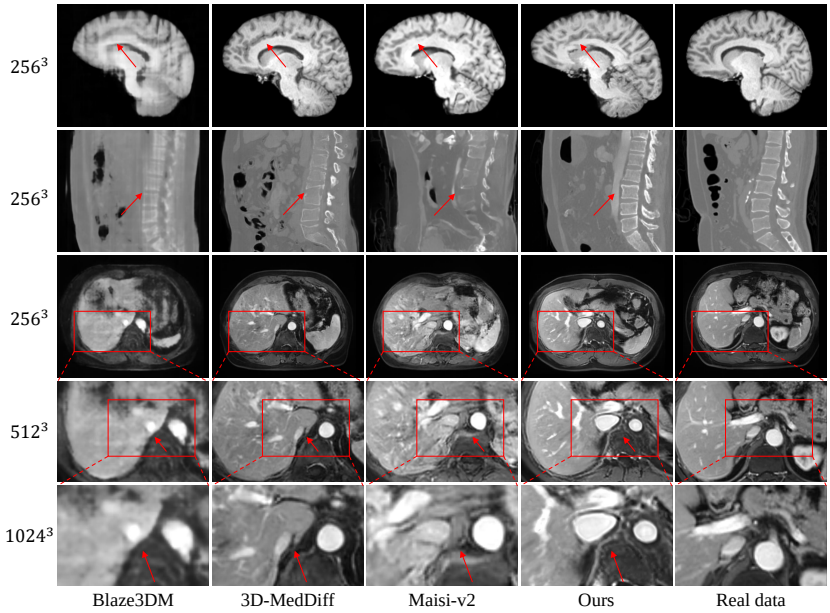

Fig. 3. Qualitative comparison of category-conditioned 3D generation across different resolutions. Our method preserves robust structural consistency without blurring via the continuous anatomical field, while the 512^3 and 1024^3 results for all baselines and real data are obtained through trilinear interpolation.

Table 1. Quantitative comparison of different methods across generation and reconstruction tasks. "Generation" metrics (FID and MMD) are averaged over all cohorts for category-conditioned generation. SV-CT and Fast MRI refer to sparse-view CBCT and accelerated MRI reconstruction, respectively. Best results are in bold.

Method	Efficiency	Generation		SV-CT		Fast MRI	
	Params(M) /Time(s)	FID↓ ($\times 10^{-3}$)	MMD↓ ($\times 10^{-6}$)	PSNR↑ (dB)	SSIM↑	PSNR↑ (dB)	SSIM↑
FDK [7]	0/0.1	-	-	18.20	0.3895	-	-
Zero-filling	0/0.8	-	-	-	-	22.19	0.6131
Blaze3DM [9]	521/3685.9	89.2237	13.3857	24.45	0.6489	23.92	0.6596
3D-MedDiff [23]	158/130.4	11.5096	5.3097	24.93	0.7290	24.94	0.8277
Maisi-v2 [28]	165/7.8	9.1598	2.3858	25.23	0.7331	26.86	0.8565
Ours	**827/2.6**	**7.6425**	**1.8196**	**26.75**	**0.7618**	**28.33**	**0.8742**

of 8192 entries (64-dim). The continuous anatomical field is parameterized by a 4-layer MLP (256 hidden units). The model is trained with reconstruction, perceptual, and adversarial losses ($\lambda_1 = 0.1, \lambda_2 = 0.01$) and a 3D PatchGAN [13] discriminator on 128^3 patches. Optimization was performed using Adam (10^{-4}) on eight A100 GPUs. The 2D latent generative U-Net adopts four stages with

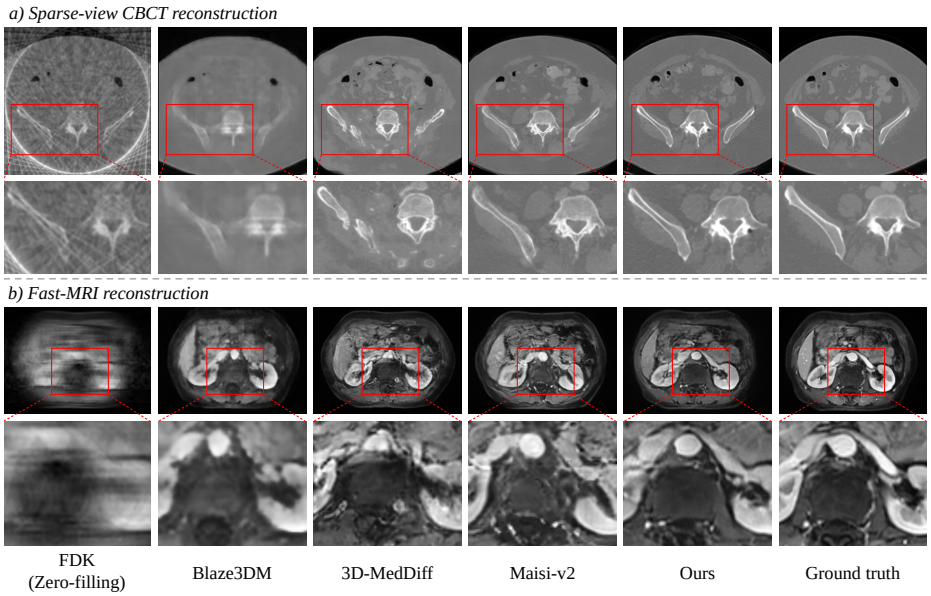

Fig. 4. Qualitative comparison of reconstruction performance on downstream tasks. (a) 30-view sparse-view CBCT reconstruction and (b) 8× accelerated fast MRI reconstruction. Our method recovers finer anatomical structures and suppresses artifacts more effectively than state-of-the-art baselines. (CT window: $[-1000, 1000]$)

channels $[384, 768, 1152, 1536]$ and is pre-trained on the full training dataset. For reconstruction tasks, FDK-reconstructed volumes and zero-filled MRI were employed as structural conditions.

Baselines and Metrics We compare with representative 3D generative models, including Blaze3DM [9], 3D-MedDiff [23], and Maisi-v2 [28]. FDK [7] and zero-filling serve as task-specific baselines. Generation quality is evaluated using FID [11] and MMD [8], while reconstruction fidelity is measured by PSNR and SSIM [24] in 3D space.

3.2 Comparison Results

3D Anatomical Generation. Fig. 3 shows qualitative results of category-conditioned generation across datasets. Blaze3DM exhibits structural artifacts and detail loss, while 3D-MedDiff and Maisi-v2 better preserve global anatomy but produce over-smoothed textures. In contrast, our method generates sharper anatomical structures and more realistic high-frequency details. Notably, the continuous anatomical field enables resolution-agnostic generation up to 1024^3 without interpolation artifacts. These observations are consistent with Table 1, where our method achieves the best FID (7.64×10^{-3}) and MMD (1.82×10^{-6}).

Sparse-view CBCT and Accelerated MRI Reconstruction. On 30-view CBCT and 8× accelerated MRI reconstruction (Fig. 4), conventional task-specific

Fig. 5. Incremental ablation study of the triplane-based autoencoder. "Base model" uses global axis-wise mean pooling and is trained only with $\mathcal{L}_{Rec}$ and $\mathcal{L}_{\mathcal{VQ}}$. Visual and quantitative results demonstrate the incremental contribution of the 3D-aware projector, perceptual loss ($\mathcal{L}_{LPIPS}$), and adversarial loss ($\mathcal{L}_{Adv}$) to structural fidelity and texture sharpness. (CT window: $[-1000, 1000]$)

baselines (FDK and zero-filling) suffer from severe artifacts. Blaze3DM, 3D-MedDiff, and Maisi-v2 partially suppress noise but fail to fully recover fine anatomical details. Our method reconstructs both global structures and fine textures more accurately, achieving the highest PSNR and SSIM across both tasks (Table 1).

Efficiency Ananlysis. Blaze3DM requires instance-level latent optimization during inference (∼1h). In contrast, our model achieves the fastest inference (2.6s per volume), outperforming 3D-MedDiff by approximately 50× (Table 1). This efficiency stems from latent-space generative modeling and the reduced complexity of triplane representations.

3.3 Ablation Study

An incremental ablation was conducted to assess each component of the triplane-based autoencoder (Fig. 5). The base model, which uses global axis-wise mean pooling, achieves limited structural fidelity (0.8051 SSIM). Replacing it with the proposed 3D-aware projector improves structural accuracy (0.8317 SSIM). Introducing perceptual loss $\mathcal{L}_{LPIPS}$ further enhances reconstruction quality (0.8742 SSIM). Finally, adding adversarial supervision $\mathcal{L}_{Adv}$ yields the best performance (0.9051 SSIM) with sharper anatomical details. These results confirm that the 3D-aware projection ensures structural consistency, while perceptual and adversarial losses are essential for high-frequency detail recovery.

4 Conclusion

In this work, we propose an efficient framework for resolution-agnostic 3D medical image generation, achieved by bridging discrete triplane latent modeling with a continuous anatomical field. Experiments show consistent improvements over state-of-the-art methods in generation quality, reconstruction accuracy, and inference speed. The decoupling of latent representation from output resolution

provides a promising paradigm for efficient, clinically adaptable medical image generation. Future work will extend the framework to direct training on volumes with heterogeneous resolutions, enabling learning from native clinical data without resampling.

Acknowledgments. This work was supported by Capital's Funds for Health Improvement and Research under Grant No. CFH2024-1-4101, the National Natural Science Foundation of China (NSFC) under Grant Nos. 6230012077 and 62531024, Shanghai Municipal Central Guided Local Science and Technology Development Fund Project under Grant No. YDZX20233100001001, and Beijing Natural Science Foundation under Grant No. L242133. The authors also acknowledge the HPC Platform of ShanghaiTech University for providing computational resources.

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

References

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Armato, S., McLennan, G., Bidaut, L., McNitt-Gray, M., Meyer, C., Reeves, A., Zhao, B., Aberle, D., Henschke, C., Hoffman, E., Kazerooni, E., MacMahon, H., {Van Beek}, E., Yankelevitz, D., Biancardi, A., Bland, P., Brown, M., Engelmann, R., Laderach, G., Max, D., Pais, R., Qing, D., Roberts, R., Smith, A., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G., Jude, C., Munden, R., Petkovska, I., Quint, L., Schwartz, L., Sundaram, B., Dodd, L., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., {Vande Casteele}, A., Gupte, S., Sallam, M., Heath, M., Kuhn, M., Dharaiya, E., Burns, R., Fryd, D., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B., Clarke, L.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (Feb 2011). https://doi.org/10.1118/1.3528204
3. Cao, C., Cui, Z.X., Wang, Y., Liu, S., Chen, T., Zheng, H., Liang, D., Zhu, Y.: High-frequency space diffusion model for accelerated mri. IEEE Transactions on Medical Imaging **43**(5), 1853–1865 (2024)
4. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. The Lancet **392**(10162), 2388–2396 (2018)
5. Deng, Y., Wang, C., Hui, Y., Li, Q., Li, J., Luo, S., Sun, M., Quan, Q., Yang, S., Hao, Y., Liu, P., Xiao, H., Zhao, C., Wu, X., Zhou, S.K.: Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography (2024), https://arxiv.org/abs/2105.14711
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Feldkamp, L.A., Davis, L.C., Kress, J.W.: Practical cone-beam algorithm. Journal of the Optical Society of America A **1**(6), 612–619 (1984)
8. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The journal of machine learning research **13**(1), 723–773 (2012)

9. He, J., Li, B., Yang, G., Liu, Z.: Blaze3dm: Integrating triplane representation with diffusion for solving 3d inverse problems in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 56–66. Springer (2025)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. 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)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
16. Liu, J., Anirudh, R., Thiagarajan, J.J., He, S., Mohan, K.A., Kamilov, U.S., Kim, H.: Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10498–10508 (2023)
17. Liu, X., Li, H., Qiao, Z., Huang, Y., Liu, X., Zhang, J., Qian, Z., Zhen, X., Zhang, B.: D3t: Dual-domain diffusion transformer in triplanar latent space for 3d incomplete-view ct reconstruction. International Journal of Computer Vision **133**(8), 5238–5261 (2025)
18. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2022). https://doi.org/10.1109/TPAMI.2021.3100536
19. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
20. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Cukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (2023)
21. 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)
22. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R.: Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine **12**(3), 1–10 (03 2015). https://doi.org/10.1371/journal.pmed.1001779, https://doi.org/10.1371/journal.pmed.1001779
23. Wang, H., Liu, Z., Sun, K., Wang, X., Shen, D., Cui, Z.: 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. IEEE Transactions on Medical Imaging (2025)

24. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
25. Xu, Y., Sun, L., Peng, W., Jia, S., Morrison, K., Perer, A., Zandifar, A., Visweswaran, S., Eslami, M., Batmanghelich, K.: Medsyn: text-guided anatomy-aware synthesis of high-fidelity 3-d ct images. IEEE Transactions on Medical Imaging **43**(10), 3648–3660 (2024)
26. 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)
27. Zhang, Z., Ehinger, K.A., Drummond, T.: Tcam-diff: Triplane-aware cross-attention medical diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 22732–22740 (2025)
28. Zhao, C., Guo, P., Yang, D., Tang, Y., He, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D.: Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. arXiv preprint arXiv:2508.05772 (2025)