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.

VolDiT: Controllable Volumetric Medical Image Synthesis with Diffusion Transformers

Marvin Seyfarth[1,2,3*], Salman Ul Hassan Dar[1,2,3], Yannik Frisch[1,2,3], Philipp Wild[4], Norbert Frey[2,3], Florian André[2,3], and Sandy Engelhardt[1,2,3]

[1] Institute for Artificial Intelligence in Cardiovascular Medicine, Medical Faculty of Heidelberg University, Heidelberg University, Heidelberg, Germany
[2] Department of Cardiology, Angiology, Pneumology, Heidelberg University Hospital, Heidelberg, Germany
[3] DZHK (German Center for Cardiovascular Research), Partner Site Heidelberg/Mannheim, Heidelberg, Germany
[4] University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany
[*]Corresponding author: Marvin.Seyfarth@med.uni-heidelberg.de

Abstract. Diffusion models have become a leading approach for high-fidelity medical image synthesis. However, most existing methods for 3D medical image generation rely on convolutional U-Net backbones within latent diffusion frameworks. While effective, these architectures impose strong locality biases and limited receptive fields, which may constrain scalability, global context integration, and flexible conditioning. In this work, we introduce **VolDiT**, the first purely transformer-based 3D Diffusion Transformer for volumetric medical image synthesis. Our approach extends diffusion transformers to native 3D data via volumetric patch embeddings and global self-attention that operates directly over 3D tokens. To enable structured control, we propose a timestep-gated control adapter that maps segmentation masks to learnable control tokens, which modulate transformer layers during denoising. This token-level conditioning mechanism allows precise spatial guidance while preserving the modeling advantages of transformer architectures. We evaluate our model on high-resolution 3D medical image synthesis tasks and compare it to state-of-the-art 3D latent diffusion models based on U-Nets. Results demonstrate improved global coherence, superior generative fidelity, and enhanced controllability. Our findings suggest that fully transformer-based diffusion models provide a flexible foundation for volumetric medical image synthesis. The code and models trained on public data are available at https://github.com/Cardio-AI/voldit.

Keywords: 3D Medical Image Synthesis · Diffusion Transformers · Conditional Generation

1 Introduction

Diffusion models [7,20] have become the leading approach for high-fidelity image synthesis, often relying on convolutional U-Net backbones [21] that in-

terleave residual blocks with attention layers. In medical imaging, such models have shown promise for modality translation [10,12], data augmentation [9,13,6,26], reconstruction [26], and segmentation-conditioned synthesis [11,6]. However, nearly all 3D medical diffusion models still rely on convolutional architectures. While effective, 3D U-Nets impose strong locality biases that may limit their applicability to global anatomical modeling. Volumetric data requires a consistent structure across slices, yet large receptive fields are costly, and repeated down- and up-sampling can attenuate fine-grained details. In contrast, vision transformers [3] use global self-attention to capture long-range dependencies at every stage, thereby scaling effectively in generative tasks. Diffusion Transformers (DiTs), which replace U-Net backbones with purely transformer-based models, demonstrate strong correlations between model capacity and sample quality [18,17], raising the question: *Can a fully transformer-based backbone replace 3D U-Nets for volumetric medical synthesis?* Prior work has explored hybrid transformer-convolutional designs [26], where transformers operate locally within sub-volumes. While effective, global volumetric modeling remains dominated by convolution, leaving the potential of fully transformer-based 3D diffusion largely unexplored. Beyond scalability, transformer architectures align naturally with emerging foundation-model paradigms in medical AI, enabling unified token-based representations and flexible conditioning on masks [11,6], anatomical priors [8], multimodal inputs [10,12,17], or clinical variables [19]. Here, we introduce **VolDiT**, the first purely 3D Diffusion Transformer for volumetric medical image synthesis. Unlike hybrid approaches, VolDiT operates on the entire latent volume, replacing the 3D U-Net backbone with transformer blocks. Using a 3D patchification strategy, we tokenize the latent volume to enable global self-attention across all spatial locations. We further propose a timestep-gated control adapter *TGCA* [14], which maps spatial conditioning inputs (e.g., segmentation masks) into learnable tokens that modulate the transformer layers. This enables precise, token-level control over generated volumes while maintaining scalability. Our contributions are threefold: (1) formulation of volumetric medical image synthesis using a fully transformer-based diffusion backbone operating on 3D latent volumes, (2) introduction of a timestep-gated control adapter (TGCA) for stable and anatomically precise spatial conditioning, compatible with pretrained transformer backbones. (3) comprehensive evaluation including comparisons to U-Net-based latent diffusion and non-diffusion generative models, as well as ablation studies to assess fidelity, diversity, and scalability on medical datasets.

2 Methodology

This work serves as the methodological foundation for our concurrently submitted 4D spatiotemporal extension (see Supp. 1), in which the proposed architecture is generalized to model dynamic cardiac MR sequences. In contrast, the present paper focuses exclusively on native 3D volumetric medical image synthesis and introduces VolDiT, the first purely transformer-based 3D Diffusion

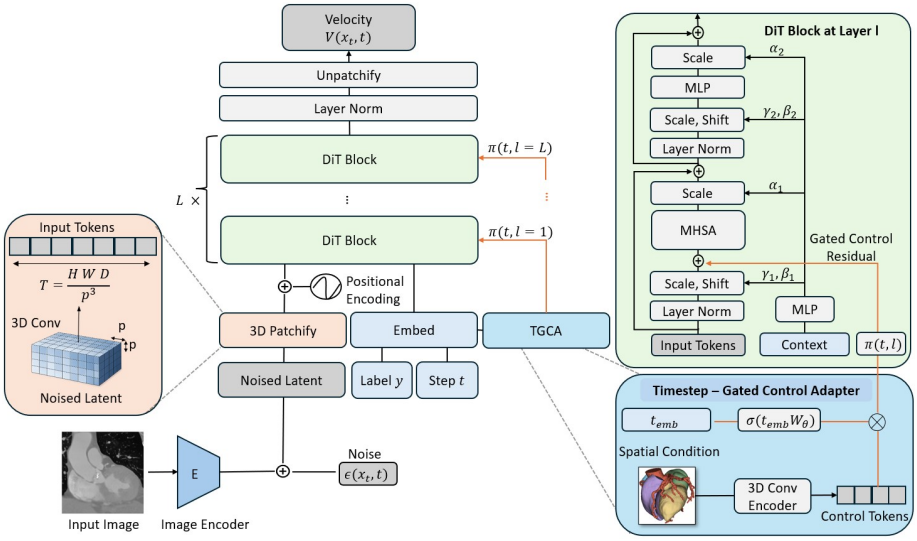

Fig. 1. Overview of the VolDiT framework. Input volumes are encoded into latents, patchified into tokens, and processed by the 3D Diffusion Transformer for denoising, with spatial conditioning injected through the timestep-gated control adapter. The transformer outputs are then unpatchified and decoded to reconstruct volumetric images.

Transformer operating directly on volumetric tokens. We replace the conventional U-Net backbones used in 3D latent diffusion models with global self-attention applied to 3D patch embeddings. Furthermore, we apply a timestep-gated control adapter that injects segmentation-derived control tokens into the denoising transformer, enabling structured, spatially precise conditional generation. The present study establishes the core 3D transformer diffusion framework and systematically evaluates its advantages over convolutional latent diffusion baselines in both unconditional and conditional volumetric synthesis settings.

2.1 Latent Volumetric Representation via VQ-GAN

We adopt a latent diffusion strategy and first learn a compact volumetric representation using a 3D Vector-Quantized GAN (VQ-GAN) [4]. Given an input medical volume $x \in \mathbb{R}^{C \times D \times H \times W}$, an encoder $\mathcal{E}$ maps it to a compressed latent tensor $z_0 = \mathcal{E}(x) \in \mathbb{R}^{C_z \times D' \times H' \times W'}$, which is subsequently reconstructed by a decoder $\mathcal{D}$ as $\hat{x} = \mathcal{D}(z_0)$. The VQ-GAN is trained with a reconstruction objective, along with adversarial and perceptual losses, to ensure high-fidelity reconstructions, achieving a compression rate of 8 in each spatial dimension. All diffusion modeling is performed on this latent representation z_0.

2.2 3D Diffusion Transformer

We train a diffusion model in the latent space of the pretrained VQ-GAN using a cosine noise schedule [16]. The forward process gradually corrupts the latent variables over $T = 300$ timesteps. Instead of predicting noise directly, we adopt velocity prediction [22]. The model is optimized with a Smooth L1 (Huber) loss between the predicted and target velocity. To parameterize v_θ, we adapt the design principles of the 2D Diffusion Transformer proposed in computer vision [18] to volumetric medical imaging. Specifically, we introduce a purely transformer-based 3D denoising backbone operating on volumetric latent tokens. Given a latent tensor $z_t \in \mathbb{R}^{C_z \times D' \times H' \times W'}$, we partition it into non-overlapping cubic patches of size $p \times p \times p$ (Figure 1). A 3D convolution with kernel size and stride p produces token embeddings $\mathbf{x}_0 \in \mathbb{R}^{T \times d}$, where $T = D'H'W'/p^3$ is the number of tokens and d is the embedding dimension. Fixed 3D sine-cosine positional encodings are added to preserve spatial structure. The backbone consists of a series of L DiT blocks, following the implementation of [18]. The final transformer output is projected back to patch-sized latent predictions and reshaped via an unpatchify operation to recover $\hat{v}\theta(z_t, t) \in \mathbb{R}^{C_z \times D' \times H' \times W'}$.

2.3 Timestep-Gated Control Adapter for Spatial Conditioning

After unconditional pretraining, we apply spatial conditioning through a lightweight adapter (portrayed in the blue part of Figure 1) while keeping the base DiT backbone frozen [30,14]. Given a conditioning volume $s \in \mathbb{R}^{C_s \times D \times H \times W}$ (e.g., a segmentation mask), we encode it with a 3D convolutional network to produce a latent feature map $\mathbf{h}c \in \mathbb{R}^{d \times D' \times H' \times W'}$, explicitly mapping the conditioning input into the DiT token embedding dimension d. The feature map is then flattened into a sequence of control tokens $\mathbf{c}\text{tok} \in \mathbb{R}^{T \times d}$, where $T = D'H'W'$ is the number of volumetric tokens. The final convolution layer is zero-initialized to ensure stable conditioning training.
To adapt the control tokens across diffusion steps, we propose a timestep-dependent gating $\gamma(t)$ computed as the sigmoid of an MLP applied to a time embedding $\mathbf{t}$, i.e., $\gamma(t) = \sigma(\text{MLP}(\mathbf{t}))$. The resulting gating modulates the control tokens as $\tilde{\mathbf{c}}_{\text{tok}} = \gamma(t) \cdot \mathbf{c}_{\text{tok}}$. For selected transformer layers l, control is injected as $\mathbf{x}^{(l)} \leftarrow \mathbf{x}^{(l)} + \lambda_l \tilde{\mathbf{c}}_{\text{tok}}$, where λ_l is a learnable per-layer scaling parameter. This design enables layer-wise control strength, temporal adaptation across diffusion steps, and stable finetuning without disrupting pretrained weights.

2.4 Datasets

Our evaluation datasets include:
Lung CT (LUNA16): This dataset comprised of 888 lung volumes (3D) from the publicly available Luna16 dataset[5]. Among these, 800 volumes were used for training, 40 volumes for validation, and 48 volumes for testing. The volumes

[5] https://luna16.grand-challenge.org/

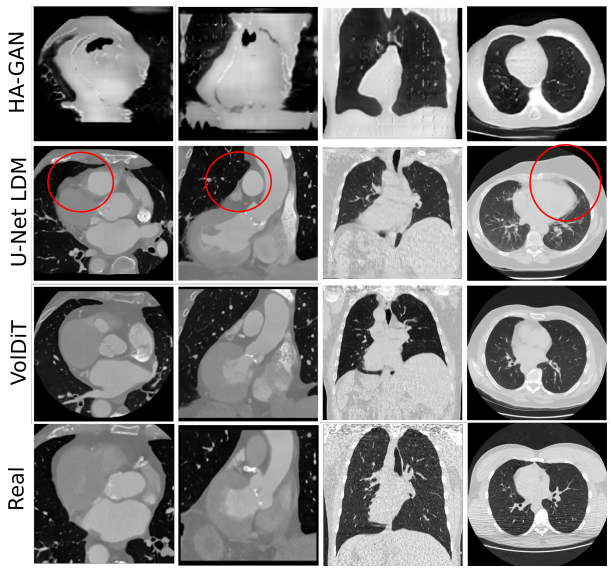

Fig. 2. Synthetic examples on LUNA16 (columns 3 & 4) and TaviCT (columns 1 & 2). HA-GAN produces anatomically implausible structures, while the U-Net LDM yields high-quality images but struggles with global anatomical consistency (red circles). VolDiT generates anatomically coherent volumes with sharper structural detail, reflecting the benefit of global self-attention.

were cropped and padded to $512 \times 512 \times 256$, and their intensities were clipped to the range $[-1200, 300]$.

TaviCT: This in-house dataset contains 1002 cardiac CTA scans of TAVI patients [25], which we split into 600 volumes for training, 201 volumes for validation, and 201 volumes for testing. The volumes were cropped and padded to size $192 \times 192 \times 192$, and their intensities were clipped to the range $[-1000, 1000]$.

2.5 Experimental Setup

Models with minimal validation loss were selected for evaluation. We evaluate generative fidelity using the Fréchet Inception Distance (FID) metric, comparing 100 generated examples to the respective datasets' test splits (FID_{test}). Additionally, precision, recall, coverage, and density were calculated on the extracted features, where the neighborhood parameter k is selected such that a real–real comparison yields scores greater than 0.95, ensuring a stable and meaningful operating point [15] (LUNA16: $k = 7$, TaviCT: $k = 3$). Due to the absence of publicly available feature extractors specifically trained for cardiac CTA data, we employ ImageNet-pretrained feature extractors for the TaviCT dataset and report a 2.5D FID, computed as the average across the three anatomical axes, which has been shown to yield reliable and stable estimates of generative fidelity in volumetric medical imaging [29]. For the LUNA16 dataset, we instead use

a 3D ResNet-50 pretrained on MedicalNet, a large corpus of chest CT scans, which aligns well with the lung CT domain of LUNA16 [1]. Sample diversity is additionally quantified by computing the Multi-Scale SSIM (MS-SSIM) metric across 100 random pairs of the synthetic examples [27]. Lastly, we study the effect of model capacity and patch size on generative quality by evaluating DiT configurations from XS (depth 6, hidden size 384, 6 attention heads) to L, with patch sizes $p = 2$ and $p = 4$, following [18]. All baseline models can be trained and sampled on a single GPU with 24 GB of VRAM. Detailed model training configurations, hardware requirements, and inference times are provided at https://github.com/Cardio-AI/voldit. We compare VolDiT to a state-of-the-art U-Net-based LDM [23] that is trained with the same training strategy, i.e., in the same VQ-GAN latent space, and employs the same noise schedule and time steps. In addition, we include HA-GAN as a non-diffusion-based baseline to contrast diffusion-based generative modeling with adversarial training paradigms [24]. The training procedure and hyperparameters were adopted directly from https://github.com/batmanlab/HA-GAN. For conditional synthesis, we use held-out anatomy test masks to synthesize guided volumes by utilizing our smallest model, *VolDiT-XS*. We then evaluate segmentation-conditioned synthesis using Dice similarity, as well as 95th percentile Hausdorff distance between input segmentation masks and segmentations predicted from generated volumes using TotalSegmentator [28]. To isolate the contribution of timestep-dependent gating, we compare the full model against variants with constant conditioning strengths $\pi(t, l) = 0.1$ and $\pi(t, l) = 1.0$ across diffusion steps.

3 Results

3.1 Unconditional Volumetric Synthesis

We first compare the proposed 3D DiT against the U-Net-based LDM in the unconditional setting. Across both datasets, the results depicted in Table 1 show that VolDiT achieves a more favorable quality–diversity trade-off. On TaviCT, it substantially improves FID (21.4 vs. 36.8) while maintaining high precision (0.94

Table 1. Unconditional generative metrics as measured by FID, precision (P), recall (R), density (D), coverage (C), and MS-SSIM.

Dataset	Model	$\mathbf{FID_{test}}(\downarrow)$	P(↑)	R(↑)	D(↑)	C(↑)	MS-SSIM(↓)
LUNA16	Real train	0.005	0.96	0.96	0.96	1.0	0.31 ± 0.08
	U-Net LDM [23]	0.031	0.44	**0.94**	0.18	0.76	**0.36 ± 0.09**
	HA-GAN [24]	0.126	**1.00**	0.00	0.32	0.06	0.99 ± 0.00
	VolDiT (L, p=4)	**0.004**	0.91	0.90	**0.82**	**0.92**	0.38 ± 0.08
TaviCT	Real train	3.1	0.95	0.95	1.07	0.90	0.30 ± 0.08
	U-Net LDM [23]	36.8	0.95	0.63	0.89	0.68	0.38 ± 0.07
	HA-GAN [24]	155.5	**1.00**	0.00	**1.25**	0.04	0.99 ± 0.00
	VolDiT (L, p=2)	**21.4**	0.94	**0.73**	1.12	**0.76**	**0.35 ± 0.09**

vs. 0.95) and increasing recall (0.73 vs. 0.63), indicating better mode coverage without sacrificing fidelity. This behavior is further supported by higher density and coverage, suggesting that generated samples both lie closer to the real data manifold and cover a larger portion of it. Additionally, the lower MS-SSIM (0.35 vs. 0.38) reflects increased sample diversity. Similar trends are observed on LUNA16, confirming that VolDiT produces synthetic data that is well aligned with real data across datasets. Qualitative examples are shown in Figure 2.

3.2 Scaling Behavior and Patch Size Ablation

Across our datasets, we do not observe the strictly monotonic improvement with model size reported in 2D computer vision settings [18], as shown in Tab. 2. We attribute this to two main factors: (i) the relatively small size of the medical datasets, and (ii) the fixed number of training iterations used for all models. In the original DiT work, larger models consistently outperform smaller ones when trained longer, but in our setting, larger models may be undertrained and do not always achieve a lower FID. Despite these limitations, smaller DiT models already achieve strong unconditional generation performance, supporting the applicability of transformer-based volumetric diffusion for medical imaging.

3.3 Segmentation-Conditioned Synthesis on TaviCT

As shown in Table 3 and Figure 3, VolDiT with gated control significantly improves anatomical alignment compared to the U-Net-based latent diffusion baseline, without sacrificing generative fidelity. Selecting a fixed conditioning scale requires careful manual tuning; e.g., ($\pi_{0.1}$) leads to weaker anatomical alignment. On the contrary, a high fixed scale ($\pi_{1.0}$) improves Dice and HD95, but risks degrading fidelity due to over-constraining the generative prior. The learnable gating strategy (π_θ) allows the model to adaptively exploit the full conditioning capacity while maintaining the best FID and competitive diversity. This behavior indicates balanced residual modulation rather than structural override and confirms that adaptive temporal scaling stabilizes conditional diffusion in volumetric token space.

Table 2. Effect of model size and patch size on unconditional generation fidelity (FID_{test}, lower is better)

Dataset	Model	Parameter Count	p=2	p=4
LUNA16	VolDiT-XS	17.2M	**0.009**	0.015
	VolDiT-S	33.2M	**0.009**	0.023
	VolDiT-B	131.0M	0.014	0.017
	VolDiT-L	580.0M	–	**0.004**
TaviCT	VolDiT-XS	17.2M	22.0	134.6
	VolDiT-S	33.2M	23.6	131.2
	VolDiT-B	131.0M	24.2	125.8
	VolDiT-L	580.0M	**21.4**	**33.3**

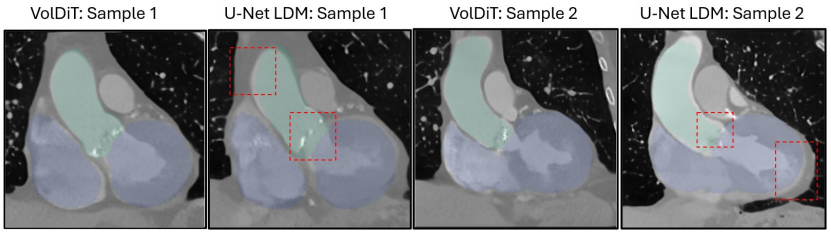

Fig. 3. Conditionally generated TaviCT samples based on held-out test masks of heart structures (blue) and the aorta (green). The U-Net-based model shows less alignment with the input masks and fails to preserve anatomical realism when enforcing the condition (as highlighted by red squares).

4 Conclusion

In this work, we investigated whether a fully transformer-based diffusion backbone can effectively model native 3D volumetric medical data. By directly learning the full 3D latent distribution using global self-attention over volumetric tokens, VolDiT achieves anatomically coherent synthesis within a conceptually simple architecture. Quantitative and qualitative results demonstrate improved fidelity and diversity compared to convolutional U-Net-based latent diffusion baselines. A key design principle of our approach is architectural unification: we rely on a purely token-based backbone operating on the entire 3D latent volume. In this sense, VolDiT provides empirical evidence that diffusion transformers extend naturally beyond 2D image synthesis to higher-dimensional medical data. Although this study focused primarily on 3D volumes, the framework establishes the foundation for high-dimensional generative modeling, as demonstrated in our concurrent extension (see Supp. 1). Additionally, while our evaluations are restricted to a single autoencoder architecture, the approach is not limited to this choice. VolDiT can, in principle, be combined with alternative latent representations, such as autoencoders trained as foundation models [6] or wavelet-based transformations [5], potentially further improving efficiency and generalization. Furthermore, we applied a timestep-gated control adapter *TGCA* for segmentation-conditioned synthesis. By mapping spatial conditions into token space and learning weighted residual modulations of transformer features, the adapter enables stable and anatomically precise control without modifying the

Table 3. Conditional generation performance on TaviCT using held-out test masks. Median (IQR) reported for Dice and HD_{95}. Lower FID and MS-SSIM indicate better fidelity and higher diversity, respectively.

	Heart		Aorta		Gen. Metrics	
Model	Dice ↑	HD_{95} ↓	Dice ↑	HD_{95} ↓	FID_{test} ↓	MS-SSIM ↓
U-Net LDM	0.89 (0.03)	5.39 (2.0)	0.83 (0.10)	4.69 (2.0)	25.8	0.34 ± 0.09
VolDiT ($\pi_{0.1}$)	0.92 (0.04)	4.12 (6.7)	0.90 (0.10)	2.83 (4.1)	25.1	0.36 ± 0.07
VolDiT ($\pi_{1.0}$)	**0.94 (0.02)**	**2.83 (0.8)**	**0.92 (0.02)**	2.00 (0.5)	24.4	**0.33 ± 0.08**
VolDiT (π_θ)	**0.94 (0.02)**	2.83 (0.9)	**0.92 (0.02)**	**2.00 (0.3)**	**22.7**	0.34 ± 0.08

pretrained backbone. Importantly, this design is modality-agnostic: any spatial prior encoded by a pretrained model can, in principle, be projected into token space and integrated via TGCA, facilitating the incorporation of foundation-model linking or multimodal conditioning signals. Finally, by establishing a robust, fully transformer-based baseline for 3D medical data, VolDiT provides the necessary framework for future research on medical image synthesis, such as comparing memorization effects of transformer-based architectures against those observed in traditional U-Net-based LDMs [2].

Acknowledgments. This work was supported by Heidelberg Faculty of Medicine at Heidelberg University, by the Multi-DimensionAI project of the Carl Zeiss Foundation (P2022-08-010), by the EU-Horizon Project *DVPS* (101213369), by the Klaus Tschira Foundation, and by the DZHK (reference number 81X3500151, project number 2098). The authors acknowledge *1-* the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG, *2-* the state of Baden-Württemberg through bwHPC and the DFG through grant INST 35/1597-1 FUGG.

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

References

1. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis (2019), https://arxiv.org/abs/1904.00625
2. Dar, S.U.H., Seyfarth, M., Ayx, I., Papavassiliu, T., Schoenberg, S.O., Siepmann, R.M., Laqua, F.C., Kahmann, J., Frey, N., Baeßler, B., Foersch, S., Truhn, D., Kather, J.N., Engelhardt, S.: Unconditional latent diffusion models memorize patient imaging data. Nature Biomedical Engineering (Aug 2025). https://doi.org/10.1038/s41551-025-01468-8
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
5. Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis. In: MICCAI workshop on deep generative models. pp. 11–21. Springer (2024)
6. Guo, P., Zhao, C., Yang, D., Xu, Z., Nath, V., Tang, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., et al.: Maisi: Medical ai for synthetic imaging. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4430–4441. IEEE (2025)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Kebaili, A., Lapuyade-Lahorgue, J., Vera, P., Ruan, S.: Multi-modal mri synthesis with conditional latent diffusion models for data augmentation in tumor segmentation. Computerized Medical Imaging and Graphics **123**, 102532 (2025)

9. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., et al.: Denoising diffusion probabilistic models for 3d medical image generation. Scientific reports **13**(1), 7303 (2023)
10. Kim, J., Park, H.: Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In: Proceedings of the IEEE/CVF Winter conference on applications of computer Vision. pp. 7604–7613 (2024)
11. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2024)
12. Li, X., Shang, K., Wang, G., Butala, M.D.: Ddmm-synth: A denoising diffusion model for cross-modal medical image synthesis with sparse-view measurement embedding. arXiv preprint arXiv:2303.15770 (2023)
13. Mao, J., Wang, Y., Tang, Y., Xu, D., Wang, K., Yang, Y., Zhou, Z., Zhou, Y.: Medsegfactory: Text-guided generation of medical image-mask pairs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21525–21535 (2025)
14. 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. AAAI'24/IAAI'24/EAAI'24, AAAI Press (2024). https://doi.org/10.1609/aaai.v38i5.28226
15. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)
16. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021)
17. Pan, S., Abouei, E., Wynne, J., Chang, C.W., Wang, T., Qiu, R.L., Li, Y., Peng, J., Roper, J., Patel, P., et al.: Synthetic ct generation from mri using 3d transformer-based denoising diffusion model. Medical Physics **51**(4), 2538–2548 (2024)
18. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
19. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI workshop on deep generative models. pp. 117–126. Springer (2022)
20. 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)
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. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
23. Seyfarth, M., Dar, S.U.H., Ayx, I., Fink, M.A., Schoenberg, S.O., Kauczor, H.U., Engelhardt, S.: MedLoRD: A Medical Low-Resource Diffusion Model for High-Resolution 3D CT Image Synthesis. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 1–12. Springer (2025)

24. Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K.: Hierarchical amortized GAN for 3D high resolution medical image synthesis. IEEE J Biomed Health Inform **26**(8), 3966–3975 (Aug 2022)
25. Tölle, M., Garthe, P., Scherer, C., Seliger, J.M., Leha, A., Krüger, N., Simm, S., Martin, S., Eble, S., Kelm, H., et al.: Real world federated learning with a knowledge distilled transformer for cardiac ct imaging. npj Digital Medicine **8**(1), 88 (2025)
26. 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)
27. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (2003). https://doi.org/10.1109/ACSSC.2003.1292216
28. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024
29. Woodland, M., Castelo, A., Al Taie, M., Albuquerque Marques Silva, J., Eltaher, M., Mohn, F., Shieh, A., Kundu, S., Yung, J.P., Patel, A.B., Brock, K.K.: Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 87–97. Springer Nature Switzerland, Cham (2024)
30. 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)