Cascaded 3D Diffusion Models for Whole-body 3D 18-F FDG PET/CT synthesis from Demographics

Siyeop Yoon*, Sifan Song*, Pengfei Jin, Matthew Tivnan, Yujin Oh, Sekeun Kim, Dufan Wu, Xiang Li, and Quanzheng Li†

Department of Radiology & Center for Advanced Medical Computing and Analysis,
Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
li.quanzheng@mgh.harvard.edu

Abstract. We propose a cascaded 3D diffusion model framework to synthesize high-fidelity 3D PET/CT volume directly from demographic variables, addressing the growing need for realistic digital twins in oncologic imaging, virtual trials, and AI-driven data augmentation. Unlike deterministic phantoms, which rely on predefined anatomical and metabolic templates, our method employs a two-stage generative process: an initial score-based diffusion model synthesizes low-resolution PET/CT volumes from the demographic variables only, providing global anatomical structures and approximate metabolic activity, followed by a super-resolution residual diffusion model refining spatial resolution. Our framework was trained on 18-F FDG PET/CT scans from the AutoPET dataset and evaluated using organ-wise volume and standardized uptake value (SUV) distributions, comparing synthetic and real data between demographic subgroups. The organ-wise comparison demonstrated strong concordance between synthetic and real images. In particular, most of the deviations in metabolic uptake values remained within 3–5% of the ground truth in sub-group analysis. These findings highlight the potential of cascaded 3D diffusion models to generate anatomically and metabolically accurate PET/CT images, offering a robust alternative to traditional phantoms and enabling scalable, population-informed synthetic imaging for clinical and research applications. Codes can be found at https://github.com/siyeopyoon/TotalGen.

Keywords: Data synthesis · Diffusion Models · PET · CT

1 Introduction

Medical image synthesis has become an essential tool in healthcare, facilitating data augmentation [1], modality translation [17,29], and digital twin simulations [11]. Generative models enable the creation of realistic medical images, addressing challenges related to data scarcity and privacy concerns [15,22]. While early

* Equally Contributed.

† Corresponding Author.

deep generative models such as VAEs [18] and GANs [6] demonstrated promise in medical image synthesis [9,23], diffusion models have recently emerged as the superior alternative, offering enhanced image fidelity and greater control over the generation process [27,8,12]. However, their high computational cost and slow inference speed remain significant challenges. To mitigate these computational demands, Latent diffusion models (LDMs) alleviate memory constraints by performing synthesis in a compressed latent space [25], though the final image quality is heavily dependent on the encoder-decoder network used for latent mapping [16]. In 3D volumetric image synthesis, patch-wise training has gained popularity [29,31]. By breaking a volume into smaller patches, models can operate more efficiently while maintaining high-resolution outputs[32]. Most previous studies on diffusion model-based image synthesis have focused on reconstructing images from undersampled measurements [31,4,10] or performing modality conversion using acquired images [28,21]. However, when models are weakly conditioned or operate unconditionally—without complementary imaging data, the iterative noise removal process often fails to maintain anatomical consistency, resulting in spatial incoherence [24].

The spatial contexts and alignment become even more challenging when models trained on different imaging modalities—such as anatomical and functional images—operate separately. Text-based image generation techniques have been explored as an alternative, using natural language descriptions to synthesize diverse medical images [7]. While such models can produce visually plausible results, they often lack anatomical precision due to the absence of explicit spatial constraints, leading to outputs that may not accurately reflect human anatomy.

To overcome these limitations, we propose an alternative framework that integrates stepwise 3D PET/CT volume synthesis with resolution enhancement, thereby overcoming challenges inherent in weakly conditional synthesis methods. Our approach first generates a coarse anatomical representation using only demographic attributes, establishing spatial relationships and organ layouts in a low-resolution 3D PET/CT volume. This blueprint is then progressively refined using a separated super-resolution residual diffusion model. By decoupling structural generation from visual enhancement, our method reduces dependence on large imaging datasets while streamlining the synthesis pipeline. The proposed framework is evaluated through task-based metrics, including organ volume accuracy and standardized uptake value (SUV) distributions from the AutoPET dataset. Our results confirm that the synthetic images closely align with real demographic-matched data, demonstrating high anatomical fidelity.

2 Methods

We propose a cascaded 3D diffusion model framework that employs a global-to-local synthesis strategy to generate anatomically and metabolically consistent PET/CT images only from demographics (Fig. 1).

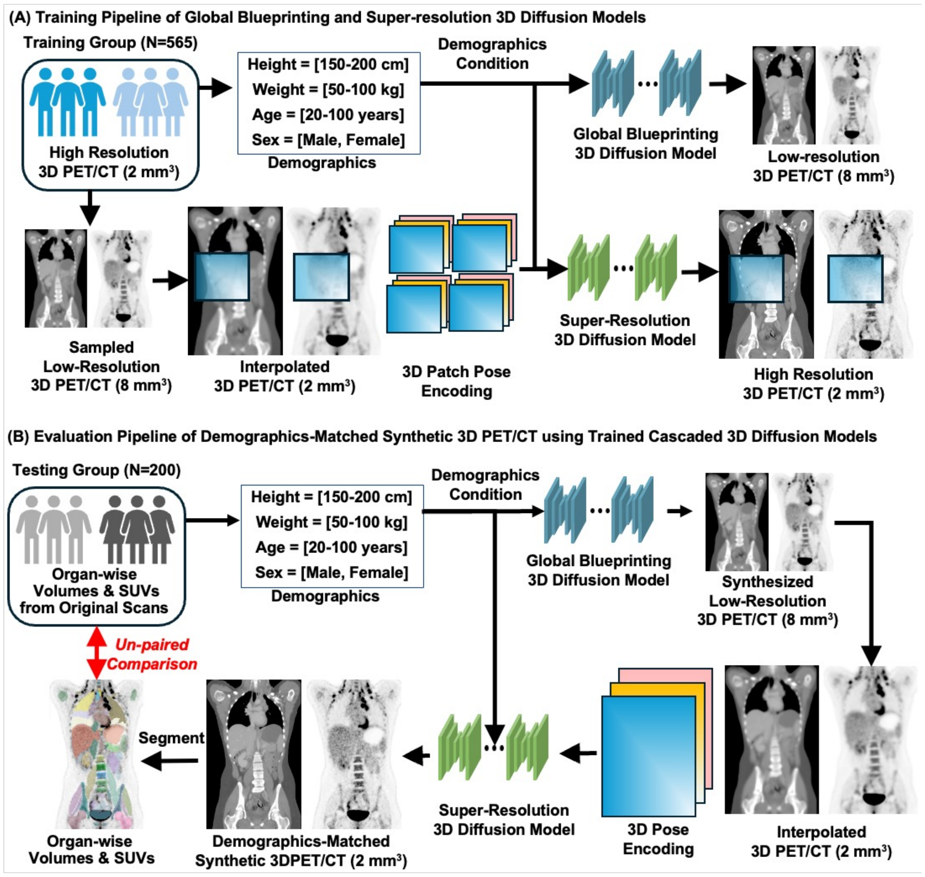
(A) Training Pipeline of Global Blueprinting and Super-resolution 3D Diffusion Models
Training Group (N=565)
High Resolution 3D PET/CT (2 mm³)
Height = [150-200 cm]
Weight = [50-100 kg]
Age = [20-100 years]
Sex = [Male, Female]
Demographics
Demographics Condition
Global Blueprinting 3D Diffusion Model
Low-resolution 3D PET/CT (8 mm³)
Sampled Low-Resolution 3D PET/CT (8 mm³)
Interpolated 3D PET/CT (2 mm³)
3D Patch Pose Encoding
Super-Resolution 3D Diffusion Model
High Resolution 3D PET/CT (2 mm³)
(B) Evaluation Pipeline of Demographics-Matched Synthetic 3D PET/CT using Trained Cascaded 3D Diffusion Models
Testing Group (N=200)
Organ-wise Volumes & SUVs from Original Scans
Height = [150-200 cm]
Weight = [50-100 kg]
Age = [20-100 years]
Sex = [Male, Female]
Demographics
Demographics Condition
Global Blueprinting 3D Diffusion Model
Synthesized Low-Resolution 3D PET/CT (8 mm³)
Un-paired Comparison
Segment
Organ-wise Volumes & SUVs
Demographics-Matched Synthetic 3DPET/CT (2 mm³)
Super-Resolution 3D Diffusion Model
3D Pose Encoding
Interpolated 3D PET/CT (2 mm³)

Fig. 1. Overview of the cascaded 3D diffusion framework for demographic-driven PET/CT synthesis. **(A)** During training, demographics guide the global diffusion model to generate a low-resolution 3D PET/CT, which is then interpolated and refined by a super-resolution diffusion model to produce high-resolution outputs. **(B)** For evaluation, the demographics from the testing set are matched and input to the same cascaded process. This yields synthetic PET/CT compared to real data cohorts, focusing on organ-wise volume and SUV metrics.

2.1 Cascaded Synthesis via Conditional Diffusion

Global Anatomical and Functional Synthesis In the first stage of our cascaded framework, a globally coherent low-resolution PET/CT volume is generated through a score-based diffusion process conditioned on demographic attributes. To formalize this, we define a family of distributions $p(I_{\text{LR}}; \sigma, x_{\text{Demo}})$ obtained by perturbing a low-resolution image with i.i.d. Gaussian noise of standard deviation σ with demographics x_{Demo}. At the maximum noise level $\sigma_{\max}$, the distribution $p(I_{\text{LR}}; \sigma_{\max}, x_{\text{cond}})$ approximates pure Gaussian noise, which serves as the initialization for the denoising procedure. The evolution from a noisy sample I_{LR} to a clean, structured, low-resolution volume is modeled by a stochastic differential equation (SDE) that combines deterministic drift with stochastic diffusion by [2]:

$$dI_{\text{LR}} = -\frac{1}{2}\beta(t)\,\nabla_{I_{\text{LR}}} \log p(I_{\text{LR}}; \sigma(t), x_{\text{Demo}})\,dt + \sqrt{\beta(t)}\,dW_t, \tag{1}$$

where $\beta(t)$ is a drift coefficient and dW_t represents a standard Wiener process. And $\nabla_{I_{\text{LR}}} \log p(I_{\text{LR}}; \sigma(t), x_{\text{Demo}})$ denotes the score function that guides I_{LR} toward regions of higher probability density as the noise diminishes under the condition x_{Demo}. Numerical integration of the corresponding reverse-time ODE from Eq. (1)—yields a stable and deterministic trajectory from a highly perturbed state to the final low-resolution PET/CT image.

To approximate the score function $\nabla_{I_{\text{LR}}} \log p(I_{\text{LR}}; \sigma(t), x_{\text{Demo}})$, we train a conditional score network $s_\theta(\cdot, \sigma, x_{\text{Demo}})$ using the following loss [27]:

$$\mathcal{L}_{\text{score, LR}} = \mathbb{E}_{\sigma, I_{\text{LR}}}\Big[\lambda(\sigma)\,\big\|s_\theta(I_{\text{LR}} + \sigma\epsilon, \sigma, x_{\text{Demo}}) - I_{\text{LR}}\big\|^2\Big], \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\lambda(\sigma)$ is a weighting function that balances the contribution of different noise levels [13]. Multiplication σ to ϵ is to normalize the magnitude of the noise. This scaling allows the loss to remain stable across different noise levels and ensures that the network can generalize its denoising capability across the entire noise schedule. The generation of low resolution volume I_{LR} is then evolved by integrating the reverse-time ODE:

$$\frac{dI_{\text{LR}}}{dt} = -\frac{1}{2}\beta(t)\,\nabla_{I_{\text{LR}}} \log p(I_{\text{LR}}; \sigma(t), x_{\text{Demo}}), \tag{3}$$

from $t = T$ (high noise) down to $t = 0$ (no noise). This integration can be performed using standard numerical solvers (EDM2 Solver [12,14]) that provide a stable and accurate approximation of the trajectory. With the trained score network $s_\theta(\cdot, \sigma, x_{\text{Demo}})$, the iterative denoising process reconstructs the complete anatomical and functional features, establishing a structural blueprint for the subsequent high-resolution synthesis stage.

Super-Resolution via Residual Diffusion and Patch-Wise Training In the second stage, the low-resolution image I_{LR} obtained from the global synthesis is refined to recover fine anatomical details. First, an upsampled estimate I_{LU} is computed by linear interpolation to the target high-resolution dimensions. Since simple interpolation does not fully restore the textures and edges, we define a residual term, $R = I_{\text{HR}} - I_{\text{LU}}$, where I_{HR} is the true high-resolution image.

Given that the low-resolution volume serves as a structural blueprint, patch-wise training can be employed effectively. The model is conditioned on the noise level as well as on a spatial prior that specifies the 3D location of each patch within the interpolated PET/CT volume,I_{LU}. The 3D domain Ω is partitioned

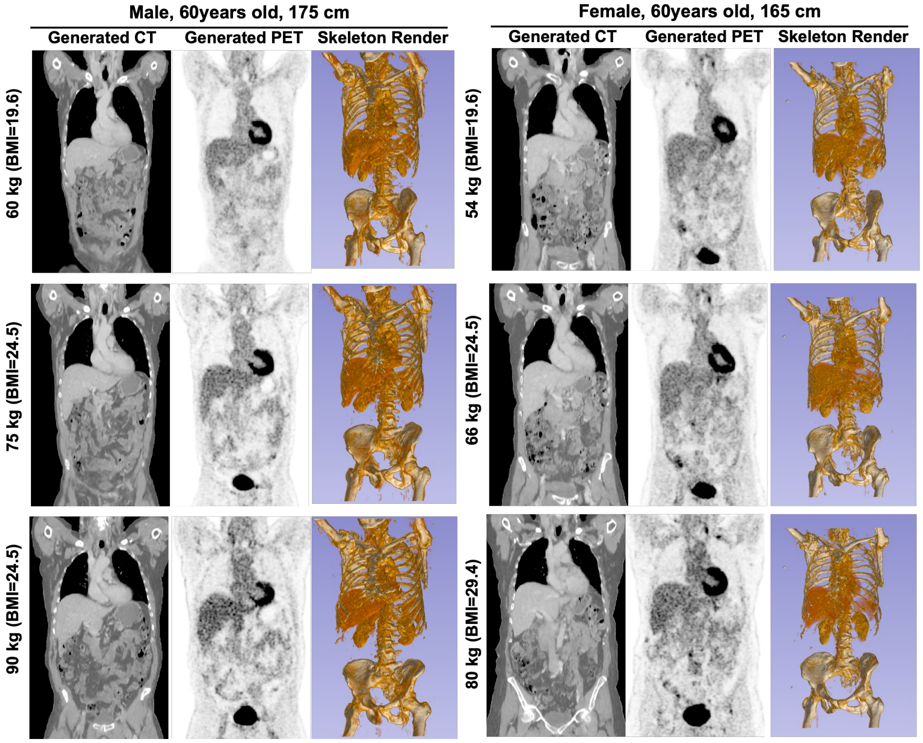

Fig. 2. Representative examples of 18-F FDG PET/CT generated using cascaded 3D diffusion models. The images show CT, 18-F FDG SUV, and 3D renderings for synthetic subjects of the same age (60 years) with different heights (male 175 cm vs. female 165 cm) and BMIs. The results demonstrate plausible anatomical and metabolic differences, including variations in adipose tissue distribution and PET signal heterogeneity.

into N patches, $\Omega = \bigcup_{i=1}^{N} \Omega_i$, and the patch-wise loss is formulated as

$$L_{Patch} = \sum_{i=1}^{N} \mathbb{E}_{R, I_{\mathrm{LU}}, \sigma} \left\| S_\theta \Big(R_{\Omega_i} + \sigma\epsilon; I_{\mathrm{LU}, \Omega_i}, \sigma, x_{\mathrm{Demo}}, \Omega_i \Big) - R_{\Omega_i} \right\|_2^2 \qquad (4)$$

, where R_{Ω_i} denotes the residual signal corresponding to the positional encoding of patch Ω_i. I_{LU} and x_{Demo} represent the low-resolution PET/CT and demographics, respectively [29,32]. The inclusion of the spatial Ω_i and the low-resolution priors $I_{(\mathrm{LU},\ \Omega_i)}$ ensures that local reconstructions are consistent with the blueprint. Once the conditional score network S_θ is trained with the above patch-wise loss, it is used to estimate the residuals for high-resolution PET/CT volumes through the reverse-time ODE flow, as defined in Eq. (3). Finally, the super-resolved high-resolution PET/CT volume is obtained by adding the recovered residual R to the upsampled estimate $I_{\mathrm{SR}} = I_{\mathrm{LU}} + R$.

2.2 Dataset and Implementation Details

We utilize the AutoPET dataset [5], which comprises whole-body ^{18}F-FDG PET/CT scans. A total of 765 subjects with complete demographic information (age, sex, height, and weight) were retrospectively selected (565 training and 200 testing subjects). All volumes were resampled to a resolution of 2 mm^3.

To ensure consistent anatomical coverage, including key organs such as the heart, liver, and kidneys, CT volumes were segmented using the TotalSegmentator, and images were cropped to the region extending from the clavicle to the sacrum bones. Each volume was zero-padded to maintain a standardized spatial dimension: $I_{\mathrm{HR}} \in \mathbb{R}^{224\times224\times384}$. The low-resolution volumes were derived from the high-resolution volumes using voxel subsampling. Specifically, each HR volume was sampled by a factor of 4 along each spatial dimension using stratified voxel extraction, defined as $I_{\mathrm{LR}}^{(x,y,z)} = I_{\mathrm{HR}}[x :: 4, y :: 4, z :: 4]$, with Cartesian multiplication of $(x, y, z) \in \{0, 1, 2, 3\}$. This process generates 64 unique low-resolution volumes per subject, significantly augmenting the training set size to low-resolution volumes while preserving anatomical diversity.

To standardize intensity values, CTs were clipped in the range $[-500, 500]$ HU and linearly scaled to $[0, 1]$. PET images were converted to SUV units, clipped to $[0, 25]$, and normalized using $SUV_{\log} = \frac{\log(SUV+1)}{\log(26)}$. During evaluation, normalized images were inverted to the original scale. Position encoding follows PatchDiffusion [29], tagging each patch with normalized (–1 to 1) x,y,z coordinates. Demographics (age, height, weight)/100, sex (0=female, 1=male) are concatenated as additional channels of input volume.

For the model implementation, we extend the EDM2 framework [14,12] to the 3D domain for conditional PET/CT synthesis by integrating 3D convolutions. Our modified 3D U-Net captures spatial dependencies across all axes, ensuring anatomical fidelity while incorporating demographic attributes to enhance realism. The training was performed on 16,252K images with an accumulated batch size of 2048, 64 base channels, and an initial learning rate of 0.017 decayed over 35,000 batches, requiring 72 hours on an NVIDIA DGX A100 system (4×40GB A100 GPUs for each model). A two-stage approach is employed: a global context model using a $56 \times 56 \times 96$ input and a super-resolution model refining patches of the same size cropped from high-resolution volumes. For seams among the patches, advanced artifact solutions can be utilized such as MultiDiffusion [3].

In the model testing, for both diffusion and 3D flow-matching models [19,20], the number of sampling steps was set to 35 for global synthesis and 100 for super-resolution. All models were evaluated on a single 40GB A100 GPU. The global synthesis stage required a peak of 2GB GPU memory and approximately 30 seconds per sample, whereas the super-resolution stage used up to 24GB GPU memory and took about 6 minutes per sample. Note that super-resolution was performed on partial volumes of size $224 \times 224 \times 96$, which were subsequently concatenated along the z-axis to reconstruct the full high-resolution output. Codes can be found at https://github.com/siyeopyoon/TotalGen.

Performance evaluation of the proposed cascaded 3D diffusion framework was conducted through both task-oriented quantitative analyses. Organ volumes

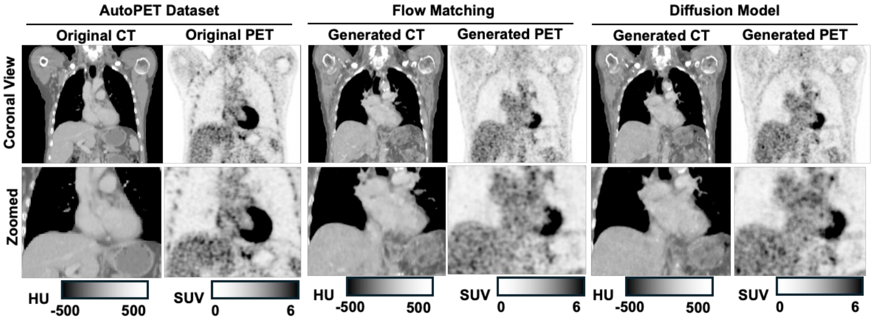

Fig. 3. Representative slices from the AutoPET (left) compared with synthetic CT/PET generated by the flow-matching (middle) and the diffusion model (right).

were measured using segmentation masks obtained from both original and generated CT images via TotalSegmentator [30], while organ SUV_{mean} and SUV_{max} were computed using the corresponding liver, heart, and kidney masks. Notably, failures or poor-quality segmentations were not observed in this study.

3 Results

Figure 2 shows representative CT, 18-F FDG PET, and the corresponding surface volume rendering of synthetic images. The synthetic PET/CT volumes exhibit visually realistic anatomical and metabolic features that are consistent with variations in demographic inputs. In particular, the generated images capture subtle morphological differences such as variations in adipose tissue distribution and heterogeneous PET signals across subjects.

Table 1 presents a quantitative comparison between the real (AutoPET) and synthetic datasets across multiple demographic groups. In general, synthetic PET / CT volumes showed strong concordance with real data at the group level. The mean SUV values in the synthetic PET images were within a close margin of the real data, and most subgroups did not show statistically significant differences ($p>0.05$) in liver and kidney volumes. However, female groups exhibited significant offsets in heart volume (30 mL, 6.9% difference) due to segmentation variability at the boarder of blood pool and myocardium. Similarly, a significant underestimation of heart SUV_{max} was observed in the male group.

In our experiments, the flow-matching model produced synthetic PET/CT images with lower variability in quantitative metrics compared to the AutoPET data, as evidenced by a reduced standard deviation in organ volumes and SUV measurements. Flow matching using ODE formulations tends to yield less diverse samples compared to stochastic diffusion models, mainly because the deterministic integration in ODE approaches can constrain variability[26]. In contrast, diffusion models that rely on stochastic sampling can better capture and preserve the natural variability of the data, while the resolution enhancement stage

Table 1. Results of Organ-wise Volume and Standard Uptake Values (mean, peak) measured in the AutoPET and synthetic datasets of 18-F FDG PET/CT. * indicates p-value <0.05 compared to AutoPET. Data are Mean ± Std (Mean difference %).

	Test cohort demographics					
	Male (N = 108)			**Female (N = 92)**		
Age (years)	58 ± 17			59 ± 15		
Heights (cm)	177 ± 7			165 ± 7		
Weights (kg)	83 ± 15			76 ± 19		
	Measurement in Liver					
Method	AutoPET	Flow	Diffusion	AutoPET	Flow	Diffusion
Volume(L)	1.76±0.44 -	1.78±0.31 (1.1%)	1.78±0.36 (1.3%)	1.55±0.37 -	1.67±0.35* (7.6%)	1.62±0.33 (4.7%)
SUV_{mean}	2.26±0.42 -	1.96±0.26* (-13.4%)	2.18±0.34 (-3.9%)	2.34±0.33 -	2.15±0.22* (-8.6%)	2.38±0.40 (1.5%)
SUV_{max}	7.29±4.28 -	5.27±1.83* (-27.7%)	6.58±3.02 (-9.7%)	7.03±4.57 -	5.41±1.67* (-23.1%)	7.22±3.76 (2.8%)
	Measurement in Heart					
Volume(L)	0.68±0.15 -	0.74 ± 0.10* (8.3%)	0.71±0.11 (3.8%)	0.55 0.09 -	0.63±0.09* (15.8%)	0.58±0.09* (6.9%)
SUV_{mean}	2.48±1.03 -	2.68±0.52 (7.8%)	2.85±1.17 (14.5%)	3.07±1.39 -	3.07±0.51 (0.2%)	3.04±1.17 (-0.8%)
SUV_{max}	10.20±5.92 -	13.72±4.71* (34.5%)	13.96±7.25* (36.8%)	12.20±7.46 -	15.26±4.59* (25.%)	13.69±7.51 (12.2%)
	Measurement in Kidneys					
Volume(L)	0.30±0.08 -	0.33±0.05* (-6.9%)	0.35±0.07 (-2.3%)	0.30±0.06 -	0.30±0.05 (3.0%)	0.30±0.06 (3.0%)
SUV_{mean}	2.50±0.41 -	2.20±0.29* (-12.0%)	2.47±0.42 (-0.9%)	2.70±0.45 -	2.48±0.17* (-8.1%)	2.71±0.38 (0.4%)
SUV_{max}	11.20±5.24 -	8.90±2.77* (-19.7%)	10.63±4.04 (-5.0%)	11.83±5.55 -	11.39±3.58 (-3.8%)	11.79±3.63 (-0.4%)

yielded outputs with similar quality (Fig 3). Overall, shape metrics across demographic groups did not differ significantly, confirming that the synthetic images preserve anatomical structure while reflecting realistic metabolic activity.

4 Conclusion

In this work, we introduced a cascaded 3D diffusion framework for the synthesis of PET/CT images directly from demographic variables. Our approach leverages a two-stage process, where an initial conditional diffusion model generates a coarse, low-resolution anatomical framework, and a subsequent super-resolution diffusion model refines this output to recover fine metabolic and structural details. This division of the synthesis task allows our method to effectively capture global anatomical structures while also ensuring that local, high-frequency details are faithfully reproduced. Quantitative evaluation on the AutoPET dataset demonstrates that the synthetic images closely replicate key clinical metrics, including organ volumes and standardized uptake values (SUV), across a range of demographic groups. The strong agreement observed in liver, kidney, and heart measurements indicates that the framework is capable of producing anatomically accurate and metabolically realistic images. The results of our study highlight the potential of this framework to serve as a reliable and scalable alternative to traditional imaging phantoms. By reducing the dependency on large, annotated imaging datasets, our method provides a novel solution for data augmentation, digital twin simulations, and virtual clinical trials. Future work will refine the synthesis pipeline to address residual artifacts in specific cohorts and evaluate its generalizability to other imaging modalities. We will also adapt the framework for multi-sequence MRI datasets and develop a text-conditioned pathology generation module.

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

References

1. Akrout, M., Gyepesi, B., Holló, P., Poór, A., Kincső, B., Solis, S., Cirone, K., Kawahara, J., Slade, D., Abid, L., et al.: Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 99–109. Springer (2023)
2. Anderson, B.D.: Reverse-time diffusion equation models. Stochastic Processes and their Applications **12**(3), 313–326 (1982)
3. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. Proceedings of Machine Learning Research **202**, 1737–1752 (2023)
4. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated mri. Medical image analysis **80**, 102479 (2022)
5. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
7. Hamamci, I.E., Er, S., Sekuboyina, A., Simsar, E., Tezcan, A., Simsek, A.G., Esirgun, S.N., Almas, F., Doğan, I., Dasdelen, M.F., et al.: Generatect: Text-conditional generation of 3d chest ct volumes. In: European Conference on Computer Vision. pp. 126–143. Springer (2024)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Hu, Q., Li, H., Zhang, J.: Domain-adaptive 3d medical image synthesis: An efficient unsupervised approach. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 495–504. Springer (2022)

10. Jiang, C., Pan, Y., Liu, M., Ma, L., Zhang, X., Liu, J., Xiong, X., Shen, D.: Pet-diffusion: Unsupervised pet enhancement based on the latent diffusion model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–12. Springer (2023)
11. Kadry, K., Gupta, S., Nezami, F.R., Edelman, E.R.: Probing the limits and capabilities of diffusion models for the anatomic editing of digital twins. npj Digital Medicine **7**(1), 1–12 (2024)
12. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)
13. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems **35**, 26565–26577 (2022)
14. Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., Laine, S.: Analyzing and improving the training dynamics of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24174–24184 (2024)
15. Kazeminia, S., Baur, C., Kuijper, A., Van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. Artificial intelligence in medicine **109**, 101938 (2020)
16. 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)
17. 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)
18. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
20. Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R.T., Lopez-Paz, D., Ben-Hamu, H., Gat, I.: Flow matching guide and code. arXiv preprint arXiv:2412.06264 (2024)
21. 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)
22. Pianykh, O.S., Langs, G., Dewey, M., Enzmann, D.R., Herold, C.J., Schoenberg, S.O., Brink, J.A.: Continuous learning ai in radiology: implementation principles and early applications. Radiology **297**(1), 6–14 (2020)
23. Pinaya, W.H., Graham, M.S., Kerfoot, E., Tudosiu, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., Da Costa, P.F., Patel, A., et al.: Generative ai for medical imaging: extending the monai framework. arXiv preprint arXiv:2307.15208 (2023)
24. 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)
25. 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)

26. Schusterbauer, J., Gui, M., Ma, P., Stracke, N., Baumann, S.A., Hu, V.T., Ommer, B.: Fmboost: Boosting latent diffusion with flow matching. In: European Conference on Computer Vision. pp. 338–355. Springer (2024)
27. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
28. Wang, T., Yang, X.: Take ct, get pet free: Ai-powered breakthrough in lung cancer diagnosis and prognosis. Cell Reports Medicine **5**(4) (2024)
29. Wang, Z., Jiang, Y., Zheng, H., Wang, P., He, P., Wang, Z., Chen, W., Zhou, M.: Patch diffusion: Faster and more data-efficient training of diffusion models. arXiv preprint arXiv:2304.12526 (2023)
30. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
31. Yoon, S., Pratap, J.S., Liu, W.C., Tivnan, M., Ren, H., Bhashyam, A., Li, Q., Chen, N., Li, X.: High-resolution 3d ct synthesis from bidirectional x-ray images using 3d diffusion model. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2024)
32. Yoon, S., Tivnan, M., Hu, R., Wang, Y., Son, Y.d., Wu, D., Li, X., Kim, K., Li, Q.: Volumetric conditional score-based residual diffusion model for pet/mr denoising. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 754–763. Springer (2024)