UNICTPET: UNIFYING 3D CT AND PET GENERATION THROUGH ADAPTIVE STRUCTURAL EXCHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative modeling offers a promising way to alleviate the scarcity of paired CT/PET data, yet existing methods typically address different generation tasks with separate models, limiting knowledge reuse across tasks. Unifying these tasks requires accommodating heterogeneous structural constraints and cross-modal dependencies that differ in direction and evolve throughout generation. We propose a unified 3D latent flow-matching framework that supports single-modality, structurally conditioned, CT-to-PET, and joint CT/PET generation within one model, without task-specific retraining. At its core, Time Task Structure Exchange (TTSE) adaptively regulates cross-modal structural interactions using modality feature states and flow time, accommodating one-way conditioning in CT-to-PET synthesis and bidirectional coordination in joint generation without explicit task representations. Complementarily, Skeletal Lesion Guidance (SLG) provides spatial control by injecting multi-scale skeletal and lesion conditions through lightweight encoders, without replicating the backbone encoder. Experiments on three whole-body PET/CT datasets comprising 1,599 paired scans demonstrate improved generation fidelity and downstream lesion segmentation over competing generative methods. Averaged across datasets, our framework improves PET PSNR in joint generation by 3.10 dB over the strongest competing baseline, while TTSE improves CT-to-PET PSNR by 2.15 dB over standard cross-attention. These findings highlight adaptive structural exchange as an effective mechanism for unifying generation tasks with heterogeneous conditions and modality dependencies. Our source code is available at https://anonymous.4open.science/r/UniCTPET-043E/.

1 INTRODUCTION

Integrated positron emission tomography(PET)/computed tomography(CT) acquires spatially registered functional and anatomical images during a single examination. PET characterizes tissue metabolism and molecular processes, whereas CT provides high-resolution anatomical information (Schwenck et al., 2023). Together, they support tumor detection, staging, and treatment response assessment. However, PET depends on dedicated equipment and radioactive tracers, and its high acquisition cost further limits the availability of large-scale, high-quality registered data (Lapi et al., 2024; Dayarathna et al., 2024). Generative models offer a new approach for learning the image distributions of CT and PET and their associations, thereby expanding the capacity for medical image generation (Kazerouni et al., 2023).

Existing CT/PET generation tasks include single-modality generation, generation guided by skeletal structures or lesion masks, cross-modal synthesis from a known modality, and joint generation of registered CT/PET pairs (Yi et al., 2019; Yoon et al., 2025). Existing methods typically use task-specific models for different input states, conditions, and targets, limiting knowledge reuse across tasks (Osuala et al., 2023). As shown in Figure 1(a), multi-task methods mainly span modalities but retain the same input-output configuration, such as generating clean images from differently degraded images (Yang et al., 2024; Chartsias et al., 2019). They therefore cannot unify conditional constraints and modality dependencies across input states.

The difficulty of unified modeling lies in the mismatch of conditional constraints and modality dependencies across tasks (Xiao et al., 2022). In conditional generation guided by skeletal structures

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

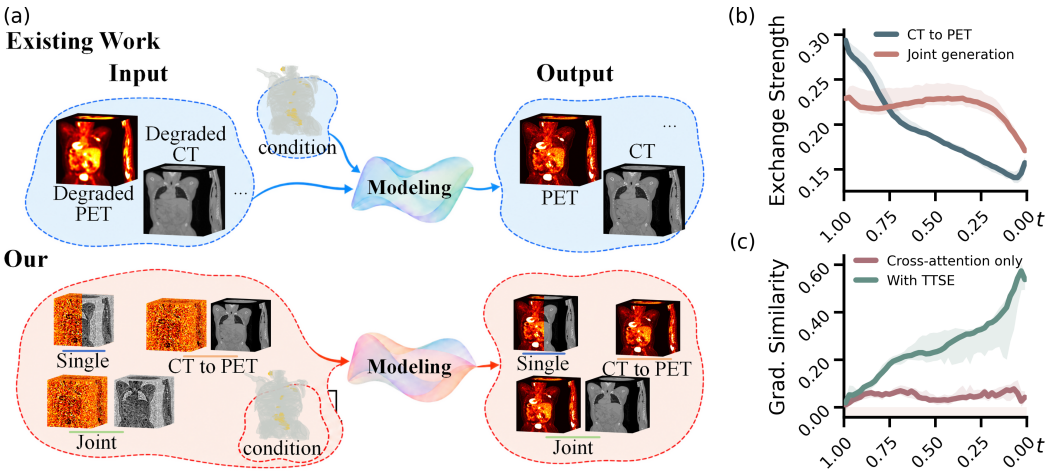

Figure 1: Unified CT/PET generation: task overview and cross-modal interaction analysis. (a) Existing multi-task methods retain fixed input-output configurations; our framework unifies single-modality, conditional, CT-to-PET, and bi-modal joint generation. (b) Distinct exchange strength trajectories in CT-to-PET and joint generation motivate time-dependent interactions. (c) TTSE yields more aligned task gradients than standard cross-attention, especially late in generation.

or lesion masks, generated images must satisfy both the external structural conditions and the target-image distribution (Xu et al., 2024). CT-to-PET generation further requires one-way information transfer (Chartsias et al., 2017) from a clean CT branch to a noisy PET branch, whereas bi-modal joint generation requires bidirectional coordination between two noisy branches. As shown in Figure 1(b), the cross-modal exchange strength follows distinct temporal patterns across the two tasks. When task interactions rely solely on attention, the task gradients with respect to the interaction parameters are nearly orthogonal, as shown in Figure 1(c), indicating limited overlap in their local update directions. Thus, tasks differ in their conditional constraints, information-transfer directions, and temporal exchange dynamics, motivating the joint modeling of modality states and generation time to enable the shared model to adapt to task-specific information dependencies.

To this end, we propose a unified three dimensional CT and PET flow generation framework. The framework uses a shared generative backbone to learn the fundamental image distributions of the two modalities and introduces Time Task Structure Exchange (TTSE) to regulate cross-modal information exchange. TTSE combines the structural features of each modality with the generation time and adaptively adjusts the exchange strength without introducing an explicit task representation, thereby accommodating the one way conditional dependency in cross-modal generation and the bidirectional coordination in bi-modal joint generation. To further support structural conditional guidance, the framework introduces a lightweight module, Skeletal Lesion Guidance (SLG), which separately encodes skeletal structures and lesion masks and injects the conditional features into the generative backbone without replicating the complete U-Net encoder. The framework uses a single network to perform the four generation tasks described above after unified training, without constructing or retraining a dedicated model for each task.

In summary, our main contributions are as follows:

- We propose a unified multi-task framework that supports diverse input-output states and enables multiple tasks within a single framework without task-specific retraining.
- We propose TTSE, which adaptively regulates cross-task information exchange based on flow time and feature states, alleviating exchange-pattern conflicts.
- We propose SLG, a conditional guidance module designed for multi-task settings that provides effective guidance without replicating the full U-Net encoder.

2 Related Work

2.1 Three Dimensional Medical Image Generation and Structural Control

Three dimensional medical image generation must balance the computational cost of volumetric data with the preservation of global structures and local details. Medical Diffusion (Khader et al., 2023) separately models image generation distributions in a three dimensional VQGAN latent space. MAISI (Guo et al., 2025) combines volumetric data compression with latent diffusion to support high-resolution generation under variable volume sizes, voxel spacings, and anatomical conditions. 3D MedDiffusion (Wang et al., 2025) uses block-level encoding and volume-level decoding to balance encoding efficiency with global structure recovery. For structural control, conditional guidance commonly adopts ControlNet (Zhang et al., 2023), which injects spatial conditions into a pretrained generative backbone through an additional trainable branch and provides a general mechanism for structurally controllable generation. These studies lay the foundation for learning three dimensional image distributions and injecting external conditions.

2.2 cross-modal and Joint Medical Image Generation

Cross-modal generation synthesizes target images by exploiting subject-specific information from a known modality. IL-CLDM (Ou et al., 2024) combines image conditions with diagnostic labels available during training to learn a latent mapping that captures both structural information and diagnostic relevance. CoCoLIT (Sargood et al., 2026) combines ControlNet, weighted image-space supervision, and latent mean stabilization to improve synthesis quality and inference consistency. Joint generation simultaneously models multiple spatially associated outputs. MedSyn (Xu et al., 2024) jointly generates three dimensional CT images and anatomical segmentation masks through hierarchical diffusion. TotalGen (Yoon et al., 2025) uses demographic conditions and cascaded three dimensional diffusion to generate registered whole-body PET and CT image pairs. These two types of tasks involve dependence on source modality conditions and coordination among multiple outputs, respectively, and therefore impose different requirements on information interaction within a shared model for CT/PET generation.

2.3 Unified multi-Task Restoration and Generation

Existing research on unified multi-task learning has focused primarily on medical image restoration and multi-modal generation. For unified restoration, AMIR (Yang et al., 2024) uses task adaptive modules to jointly address low-dose PET restoration, CT denoising, and magnetic resonance imaging(MRI) super resolution. Restore-RWKV (Yang et al., 2025) adopts an RWKV architecture with linear complexity to improve the efficiency of multi-task medical image restoration. TAT (Yang et al., 2026) dynamically generates task-specific network-weights to reduce gradient conflicts among shared parameters. It also adjusts the loss weights of individual tasks according to sample-level learning difficulty, preventing the optimization process from being dominated by any particular task. DaPT (Wei et al., 2025) uses degradation-aware prompts and mixture-of-experts routing to handle different modalities and degradation types. For unified restoration, UniMedIR (Cui et al., 2026) combines text prompts with spatial-domain and frequency-domain modeling, extending unified restoration to CT, MRI, and PET. These methods broaden the imaging modalities and degradation types covered by a single model, but they have not yet unified the modeling of differences in input states and modality dependencies associated with different tasks.

3 Method

3.1 Unified CT and PET Flow Generation

As shown in Figure 2, the framework first maps spatially registered CT and PET volumes to the latent representations z_0^c and z_0^p using pretrained VQGAN encoders (Esser et al., 2021), respectively. The latent representations z^c, $z^{p_{\mathrm{FDG}}}$, $z^{p_{\mathrm{PSMA}}}$, and $z^{p_{\mathrm{DeepPSMA}}}$ are then jointly modeled by a unified flow-matching model (Lipman et al., 2023) built on a shared 3D U-Net backbone, with different modalities distinguished by modality embeddings. Here, c denotes CT, while p_{FDG}, p_{PSMA}, and p_{DeepPSMA} denote PET from the FDG, PSMA, and DeepPSMA datasets, respectively. For the

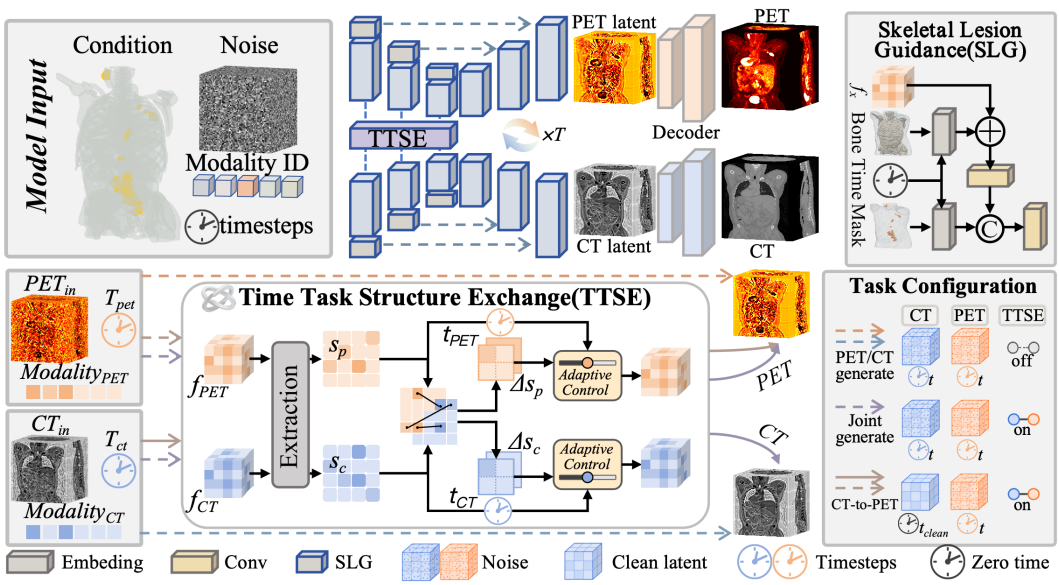

Figure 2: Overview of the unified CT/PET flow generation framework. The shared generative backbone supports single-modality, structurally guided, CT-to-PET, and bi-modal joint generation. Skeletal Lesion Guidance (SLG) injects skeletal and lesion conditions, while Time Task Structure Exchange (TTSE) adaptively regulates structural information exchange between the CT and PET branches based on modality features and flow time.

different generation tasks, the following four input states are used during training:

$$\begin{aligned} \text{CT generation}: \quad & (z_t^c, \varnothing \mid [C]) \qquad \text{PET generation}: \quad (\varnothing, z_t^p \mid [C]) \\ \text{CT to PET}: \quad & (z_0^c, z_t^p) \qquad \text{Joint generation}: \quad (z_t^c, z_t^p \mid [C]), \end{aligned} \tag{1}$$

where C denotes the anatomical condition comprising the skeletal structure and lesion mask, $[\cdot]$ indicates that the condition is optional, and p denotes the PET modality corresponding to the current tracer. CT-to-PET generation keeps the CT latent variable z_0^c fixed and constructs a flow path only for the target PET image. For all other generation objectives, linear interpolation is performed between the data latent variable z_0^m and Gaussian noise ϵ. For $m \in \{c, p_{\text{FDG}}, p_{\text{PSMA}}, p_{\text{DeepPSMA}}\}$, the flow path and target velocity are defined as:

$$z_t^m = (1-t)z_0^m + t\epsilon, \qquad v^m = z_0^m - \epsilon, \qquad t \sim \mathcal{U}(0,1). \tag{2}$$

The U-Net predicts the velocity $\hat{v}^m$ from the noisy latent variable z_t^m and time step t, and is trained using the following velocity matching loss:

$$\mathcal{L}_{\text{flow}}^m = \|v^m - \hat{v}^m\|_2^2 . \tag{3}$$

CT generation is supervised using CT, whereas PET generation and CT-to-PET generation are supervised using PET of the corresponding tracer type. The training loss for joint generation is the sum of the CT and PET velocity matching losses.

3.2 Skeletal Lesion Guidance

Existing conditional generation methods commonly adopt ControlNet, which introduces external conditions by replicating the encoder and middle module of the U-Net. For three dimensional medical images, this design substantially increases the parameter count, GPU memory consumption, and training cost. We therefore employ a lightweight anatomical condition encoder that injects skeletal and lesion information directly into the U-Net backbone.

Specifically, the skeletal structure c_b and lesion mask c_l are first processed by separate three dimensional condition encoders to obtain the multi-scale features f_b^k and f_l^k, respectively. The skeletal

structure and lesion mask are downsampled using convolution and max pooling, and the corresponding conditional features are injected after the initial convolution of the U-Net:

$$h_0 = \mathrm{UNet}_{\mathrm{in}}\left(z_t\right) + f_b^0 + f_l^0. \tag{4}$$

The conditional features are subsequently downsampled stage by stage together with the backbone features and injected into all four encoder stages and the middle block. At the k-th injection location, the skeletal condition is first fused with the backbone feature. The lesion condition is then processed further on the basis of the skeletal conditioned feature:

$$\begin{aligned} h_k &= h_k + \mathrm{ZeroConv}_k^b\left(\mathrm{FiLM}\left[\mathrm{GN}\left(h_k + f_b^k\right), t\right]\right), \\ h_k' &= \mathrm{ZeroConv}_k^l\left(\mathrm{FiLM}\left[\mathrm{GN}\left(\mathrm{Cat}\left(h_k, f_l^k\right)\right), t\right]\right), \end{aligned} \tag{5}$$

where $\mathrm{Cat}(\cdot)$ denotes channel concatenation, and $\mathrm{ZeroConv}(\cdot)$ uses zero initialization. The output h_k' of the lesion branch is directly used as the input to the subsequent backbone.

3.3 Time Task Structure Exchange

TTSE is placed in all four encoder stages and middle module of the U-Net to exchange CT and PET structural information at the same scale. Given the CT feature h_c, PET feature h_p, and their corresponding time embeddings t_c and t_p, respectively, the shared structure encoder uses depthwise convolutions with dilation rates of 1 and 2 to extract structural features under different receptive fields. Flow time information is then incorporated through FiLM (Perez et al., 2018):

$$s_m = \mathrm{FiLM}\left(E_s\left(h_m\right), t_m\right), \qquad m \in \{c, p\}, \tag{6}$$

where E_s is the shared structure encoder, s_m the time-modulated structural feature of modality m.

The features s_c and s_p interact bidirectionally through local-window cross-attention. The outputs are combined with the original structures to derive shared and differential structural information:

$$\begin{aligned} a_c &= \mathrm{CrossAttention}\left(s_c, s_p\right), \qquad a_p = \mathrm{CrossAttention}\left(s_p, s_c\right), \\ s_* &= \mathrm{Norm}\left(\frac{s_c + a_c + s_p + a_p}{4}\right), \qquad \Delta s_m = s_* - s_m, \end{aligned} \tag{7}$$

where a_m denotes the interaction information obtained through $\mathrm{CrossAttention}(\cdot)$, s_* denotes the shared information, and Δs_m denotes the structural difference information.

To adaptively regulate the exchange strength, the gate combines the modality-specific structure s_m, shared structure s_*, and difference magnitude $|\Delta s_m|$ with the modality's time embedding:

$$g_m = \sigma\left(\mathrm{Gate}_m\left(\mathrm{Cat}\left(s_m, s_*, |\Delta s_m|\right)\right) + \mathrm{MLP}\left(t_m\right)\right), \tag{8}$$

where Gate_m denotes the gating layer corresponding to modality m (either c for the CT branch or p for the PET branch). The structural difference undergoes lightweight residual processing and is then used to adjust the original feature:

$$h_m' = h_m + \alpha\,\mathrm{ZeroConv}_m\left(g_m \odot \Delta s_m\right), \qquad m \in \{c, p\}, \tag{9}$$

where α denotes a learnable residual strength. TTSE does not introduce an explicit task representation. The interaction gate g_m uses the structural information and time embedding t_m under different generation states to adjust the exchange strength.

4 Experiments

4.1 Datasets and Data Processing

This study uses three whole-body PET/CT datasets. The FDG dataset (Gatidis et al., 2022) was collected at the University Hospital of Tübingen and LMU University Hospital Munich. It includes patients with malignant melanoma, lymphoma, and lung cancer, as well as control cases without

Table 1: Data retained after preprocessing for the generation and downstream tasks, including sample counts, lesion-positive sample counts, and cancer types in each dataset. For each dataset, the original samples for the downstream training and test sets are drawn from the generation test set.

Full Dataset (1599 Paired PET-CT Scans)				
Tasks	**Dataset**	**AutoPET-FDG (901 pairs)** 165 lung cancer, 143 melanoma 135 lymphoma, 458 negatives	**AutoPET-PSMA (600 pairs)** 531 prostate cancer 69 negatives	**DeepPSMA (98 pairs)** 98 prostate cancer 0 negatives
Multi-Task Generation	Train Set	720 pairs / 360 positive	480 pairs / 423 positive	78 pairs / 78 positive
	Test Set	181 pairs / **83 positive**	120 pairs / **108 positive**	20 pairs / **20 positive**
Downstream Segmentation	Train Set	41 positive	54 positive	10 positive
	Test Set	42 positive	54 positive	10 positive

positive lesions, and uses ^{18}F-FDG as the radioactive tracer. The PSMA dataset (Jeblick et al., 2026) was collected at LMU University Hospital Munich. It includes patients with suspected or confirmed prostate cancer and used either ^{68}Ga-PSMA-11 or ^{18}F-PSMA-1007 as the radioactive tracer. The DeepPSMA dataset (Jackson et al., 2025) was provided by the Peter MacCallum Cancer Centre team. It includes two PET/CT examinations of patients with metastatic castration-resistant prostate cancer, and only the PSMA examination is used in this study. All three datasets provide spatially aligned three dimensional whole-body PET/CT images and the corresponding lesion mask annotations for conditional guidance and segmentation.

During data processing, all PET and CT volumes are first center-cropped and padded to a spatial size of $256 \times 256 \times 256$ voxels. TotalSegmentator (Wasserthal et al., 2023) is then used to extract anatomical structure labels from the CT images. These labels serve as auxiliary conditions for conditionally guided generation. As shown in Table 1, we split the entire dataset into training and test sets at a ratio of $4 : 1$ for the generation task. For the downstream segmentation task, we separately split the original samples containing at least one lesion from each dataset's generation test set at a ratio of $1 : 1$ to construct the downstream training and test sets. For each generation method, we use the conditioning information corresponding to the real training samples to generate synthetic images, which are then used to construct the corresponding downstream training set.

4.2 Experimental Setup

The proposed method was implemented using PyTorch 2.5.1 and CUDA 12.4, and its training and evaluation were conducted on two NVIDIA A800 GPUs. The total training batch size was 4. The AdamW optimizer was used with an initial learning rate of 0.0001. The values of (β_1, β_2) were set to $(0.9, 0.999)$, and the weight decay was set to 0.01. The learning rate was kept constant throughout training. Separate VQGANs with four-channel latent representations were trained for CT and PET using three dimensional image patches of size $64 \times 64 \times 64$ for 200 epochs each.

The main model was trained in two stages. In the first stage, the condition encoder and the U-Net flow generation backbone were trained jointly. Samples with and without conditional guidance were sampled at a ratio of $1 : 4$, and training proceeded for a total of $200{,}000$ steps. In the second stage, all remaining model parameters were frozen and only the TTSE module was trained. CT-to-PET generation and bi joint generation were sampled at a ratio of $3 : 7$, and training proceeded for a total of $5{,}000$ steps. For a fair comparison, the backbones of all competing methods were trained for $200{,}000$ steps, and the ControlNet modules were trained for $5{,}000$ steps.

4.3 Comparative Experiments

As shown in Table 2, we compare the image quality of the proposed method with seven advanced generative methods, including DDPM (Ho et al., 2020) and DDIM (Song et al., 2021), using five evaluation metrics. For CT generation, our method achieves PSNR values of 28.85, 27.10, and $30.91\,\mathrm{dB}$ on the DeepPSMA, FDG, and PSMA datasets, outperforming the second best results by 1.71, 1.60, and $1.89\,\mathrm{dB}$, respectively. The corresponding SSIM values are 0.885, 0.807, and 0.920, exceeding the second best results by 0.026, 0.040, and 0.023. For PET generation, our method achieves PSNR values of 42.67, 37.78, and $41.27\,\mathrm{dB}$ on the three datasets, outperforming the second best results by 1.64, 1.21, and $1.44\,\mathrm{dB}$, respectively. The corresponding SSIM values are 0.986,

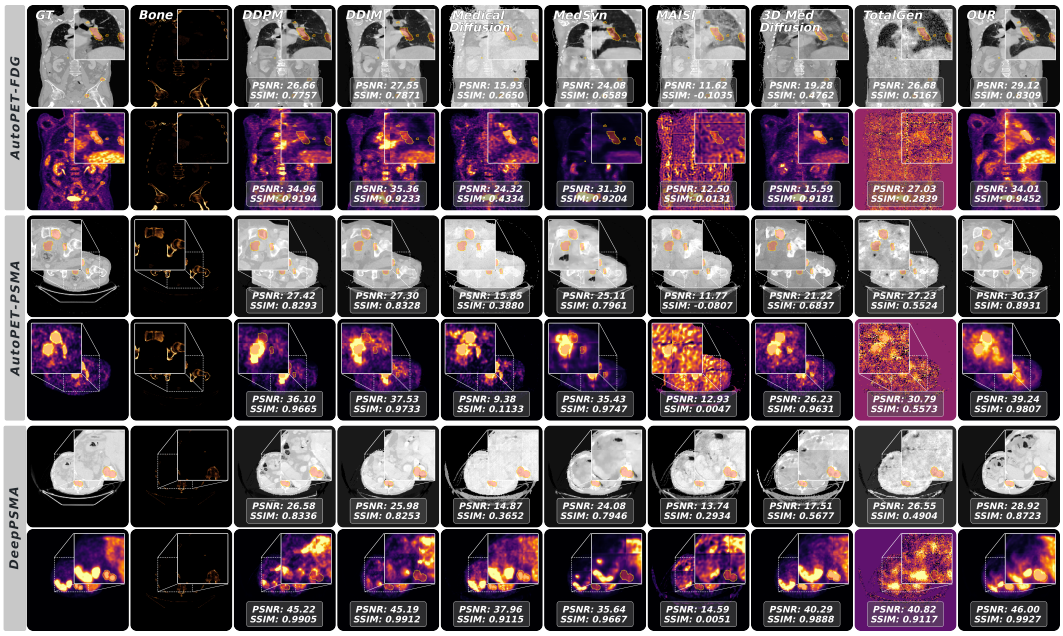

Figure 3: Qualitative comparison of CT and PET generation across three datasets. CT images appear above PET images; columns show the ground truth, bone image, and method results. Magnified views and orange contours highlight differences in anatomical structure and tracer uptake.

Table 2: Quantitative comparison of single-modality CT and PET generation on the FDG, PSMA, and DeepPSMA datasets. The best and second-best results are indicated by boldface in gray-shaded cells and underlining in light-gray-shaded cells, respectively. Means are followed by 95% confidence intervals. The p-values for comparisons with the second-best method are reported alongside the corresponding results in the Ours row of each dataset block.

Method		PSNR ↑	SSIM ↑	NMAE ↓	NMSE ↓	NCC ↑	PSNR ↑	SSIM ↑	NMAE ↓	NMSE ↓	NCC ↑
	Joint	*AutoPET-FDG (CT)*					*AutoPET-FDG (PET)*				
DDPM	×	25.50$_{\pm 0.275}$	0.767$_{\pm 0.005}$	0.109$_{\pm 0.004}$	0.068$_{\pm 0.004}$	0.863$_{\pm 0.005}$	36.37$_{\pm 0.603}$	0.921$_{\pm 0.007}$	0.750$_{\pm 0.018}$	0.681$_{\pm 0.087}$	0.623$_{\pm 0.015}$
DDIM	×	25.43$_{\pm 0.272}$	0.764$_{\pm 0.005}$	0.112$_{\pm 0.004}$	0.069$_{\pm 0.004}$	0.859$_{\pm 0.005}$	36.57$_{\pm 0.597}$	0.921$_{\pm 0.007}$	0.736$_{\pm 0.018}$	0.657$_{\pm 0.121}$	0.622$_{\pm 0.016}$
MedicalDiffusion	×	16.43$_{\pm 0.304}$	0.369$_{\pm 0.015}$	0.740$_{\pm 0.017}$	0.545$_{\pm 0.019}$	0.833$_{\pm 0.006}$	23.19$_{\pm 1.568}$	0.396$_{\pm 0.040}$	7.978$_{\pm 1.534}$	153.5$_{\pm 74.40}$	0.628$_{\pm 0.020}$
MedSyn	✓	23.23$_{\pm 0.265}$	0.671$_{\pm 0.007}$	0.181$_{\pm 0.009}$	0.121$_{\pm 0.008}$	0.860$_{\pm 0.005}$	33.23$_{\pm 0.320}$	0.906$_{\pm 0.005}$	1.455$_{\pm 0.252}$	2.825$_{\pm 0.927}$	0.662$_{\pm 0.014}$
MAISI	×	12.89$_{\pm 0.164}$	0.072$_{\pm 0.023}$	1.194$_{\pm 0.036}$	1.280$_{\pm 0.063}$	0.836$_{\pm 0.007}$	13.93$_{\pm 0.230}$	0.047$_{\pm 0.010}$	38.26$_{\pm 6.018}$	241.3$_{\pm 80.45}$	0.485$_{\pm 0.017}$
3D MedDiffusion	×	19.67$_{\pm 0.300}$	0.508$_{\pm 0.009}$	0.328$_{\pm 0.010}$	0.260$_{\pm 0.011}$	0.743$_{\pm 0.010}$	30.13$_{\pm 0.950}$	0.911$_{\pm 0.007}$	1.043$_{\pm 0.037}$	6.960$_{\pm 2.184}$	0.664$_{\pm 0.021}$
TotalGen	✓	25.01$_{\pm 0.257}$	0.480$_{\pm 0.009}$	0.178$_{\pm 0.004}$	0.076$_{\pm 0.004}$	0.854$_{\pm 0.005}$	28.37$_{\pm 0.792}$	0.376$_{\pm 0.038}$	5.070$_{\pm 0.201}$	4.254$_{\pm 0.284}$	0.298$_{\pm 0.009}$
Ours	✓	**27.10**$^{p<.001}_{\pm 0.277}$	**0.807**$^{p<.001}_{\pm 0.005}$	**0.085**$^{p<.001}_{\pm 0.003}$	**0.047**$^{p<.001}_{\pm 0.003}$	**0.902**$^{p<.001}_{\pm 0.004}$	**37.78**$^{p<.001}_{\pm 0.598}$	**0.942**$^{p<.001}_{\pm 0.005}$	**0.608**$^{p<.001}_{\pm 0.014}$	**0.477**$^{p<.001}_{\pm 0.031}$	**0.731**$^{p<.001}_{\pm 0.014}$
	Joint	*AutoPET-PSMA (CT)*					*AutoPET-PSMA (PET)*				
DDPM	×	29.02$_{\pm 0.314}$	0.897$_{\pm 0.007}$	0.041$_{\pm 0.003}$	0.025$_{\pm 0.002}$	0.864$_{\pm 0.006}$	39.48$_{\pm 0.861}$	0.977$_{\pm 0.003}$	1.181$_{\pm 0.083}$	0.901$_{\pm 0.134}$	0.617$_{\pm 0.022}$
DDIM	×	28.51$_{\pm 0.279}$	0.888$_{\pm 0.007}$	0.046$_{\pm 0.003}$	0.028$_{\pm 0.002}$	0.849$_{\pm 0.009}$	39.83$_{\pm 0.799}$	0.977$_{\pm 0.002}$	1.166$_{\pm 0.091}$	0.805$_{\pm 0.121}$	0.622$_{\pm 0.021}$
MedicalDiffusion	×	14.48$_{\pm 0.358}$	0.337$_{\pm 0.018}$	0.809$_{\pm 0.012}$	0.650$_{\pm 0.020}$	0.769$_{\pm 0.018}$	28.89$_{\pm 2.109}$	0.654$_{\pm 0.048}$	10.78$_{\pm 2.337}$	108.8$_{\pm 45.94}$	0.646$_{\pm 0.022}$
MedSyn	✓	25.26$_{\pm 0.298}$	0.860$_{\pm 0.010}$	0.074$_{\pm 0.006}$	0.061$_{\pm 0.005}$	0.866$_{\pm 0.006}$	36.63$_{\pm 0.536}$	0.974$_{\pm 0.002}$	1.427$_{\pm 0.207}$	2.721$_{\pm 0.929}$	0.616$_{\pm 0.019}$
MAISI	×	11.97$_{\pm 0.132}$	0.024$_{\pm 0.041}$	1.111$_{\pm 0.040}$	1.230$_{\pm 0.078}$	0.807$_{\pm 0.009}$	14.67$_{\pm 0.405}$	0.028$_{\pm 0.011}$	153.5$_{\pm 33.29}$	514.9$_{\pm 196.3}$	0.416$_{\pm 0.018}$
3D MedDiffusion	×	20.97$_{\pm 0.390}$	0.743$_{\pm 0.014}$	0.207$_{\pm 0.012}$	0.156$_{\pm 0.011}$	0.733$_{\pm 0.016}$	31.12$_{\pm 1.173}$	0.971$_{\pm 0.003}$	1.778$_{\pm 0.125}$	9.075$_{\pm 2.104}$	0.588$_{\pm 0.024}$
TotalGen	✓	27.53$_{\pm 0.283}$	0.540$_{\pm 0.013}$	0.119$_{\pm 0.003}$	0.033$_{\pm 0.002}$	0.834$_{\pm 0.006}$	33.24$_{\pm 0.831}$	0.623$_{\pm 0.037}$	11.55$_{\pm 1.207}$	3.798$_{\pm 0.578}$	0.325$_{\pm 0.018}$
Ours	✓	**30.91**$^{p<.001}_{\pm 0.356}$	**0.920**$^{p<.001}_{\pm 0.006}$	**0.028**$^{p<.001}_{\pm 0.002}$	**0.016**$^{p<.001}_{\pm 0.001}$	**0.908**$^{p<.001}_{\pm 0.004}$	**41.27**$^{p<.001}_{\pm 0.863}$	**0.982**$^{p<.001}_{\pm 0.002}$	**1.067**$^{p<.001}_{\pm 0.095}$	**0.572**$^{p<.001}_{\pm 0.082}$	**0.722**$^{p<.001}_{\pm 0.018}$
	Joint	*DeepPSMA (CT)*					*DeepPSMA (PET)*				
DDPM	×	27.14$_{\pm 0.876}$	0.859$_{\pm 0.022}$	0.058$_{\pm 0.010}$	0.036$_{\pm 0.007}$	0.831$_{\pm 0.012}$	41.03$_{\pm 2.414}$	0.981$_{\pm 0.005}$	1.230$_{\pm 0.195}$	0.934$_{\pm 0.139}$	0.483$_{\pm 0.052}$
DDIM	×	26.87$_{\pm 0.744}$	0.853$_{\pm 0.021}$	0.061$_{\pm 0.010}$	0.037$_{\pm 0.006}$	0.823$_{\pm 0.019}$	40.55$_{\pm 2.902}$	0.981$^{p=0.}_{\pm 0.005}$	1.238$_{\pm 0.197}$	1.180$_{\pm 0.424}$	0.469$_{\pm 0.061}$
MedicalDiffusion	×	14.66$_{\pm 0.940}$	0.354$_{\pm 0.050}$	0.772$_{\pm 0.058}$	0.604$_{\pm 0.070}$	0.762$_{\pm 0.046}$	30.21$_{\pm 6.185}$	0.714$_{\pm 0.130}$	15.57$_{\pm 14.58}$	132.7$_{\pm 145.8}$	0.578$_{\pm 0.060}$
MedSyn	✓	22.34$_{\pm 0.669}$	0.821$_{\pm 0.025}$	0.118$_{\pm 0.019}$	0.107$_{\pm 0.019}$	0.844$_{\pm 0.015}$	37.88$_{\pm 1.212}$	0.974$_{\pm 0.004}$	2.084$_{\pm 1.174}$	5.311$_{\pm 6.291}$	0.510$_{\pm 0.055}$
MAISI	×	12.02$_{\pm 0.494}$	0.086$_{\pm 0.076}$	1.062$_{\pm 0.068}$	1.098$_{\pm 0.138}$	0.798$_{\pm 0.020}$	14.72$_{\pm 0.852}$	0.021$_{\pm 0.023}$	217.1$_{\pm 110.0}$	1048$_{\pm 1051}$	0.342$_{\pm 0.026}$
3D MedDiffusion	×	19.18$_{\pm 0.955}$	0.651$_{\pm 0.050}$	0.277$_{\pm 0.049}$	0.225$_{\pm 0.047}$	0.670$_{\pm 0.049}$	32.14$_{\pm 3.642}$	0.977$_{\pm 0.005}$	1.930$_{\pm 0.376}$	14.87$_{\pm 10.04}$	0.485$_{\pm 0.050}$
TotalGen	✓	26.12$_{\pm 0.543}$	0.476$_{\pm 0.021}$	0.137$_{\pm 0.008}$	0.043$_{\pm 0.005}$	0.810$_{\pm 0.017}$	35.30$_{\pm 2.440}$	0.690$_{\pm 0.095}$	13.31$_{\pm 3.444}$	4.007$_{\pm 1.211}$	0.234$_{\pm 0.043}$
Ours	✓	**28.85**$^{p<.001}_{\pm 0.875}$	**0.885**$^{p<.001}_{\pm 0.018}$	**0.043**$^{p<.001}_{\pm 0.008}$	**0.024**$^{p<.001}_{\pm 0.005}$	**0.880**$^{p<.001}_{\pm 0.009}$	**42.67**$^{p<.001}_{\pm 2.489}$	**0.986**$^{p<.001}_{\pm 0.004}$	**1.112**$^{p<.001}_{\pm 0.193}$	**0.633**$^{p<.001}_{\pm 0.080}$	**0.649**$^{p=0.049}_{\pm 0.045}$

0.942, and 0.982, exceeding the second best results by 0.005, 0.021, and 0.005, respectively. In addition, our method achieves the best NMAE, NMSE, and NCC values for both CT and PET generation across all three datasets. Together with the qualitative comparisons shown in Figure 3, these results demonstrate stable generation performance across different modalities and data distributions.

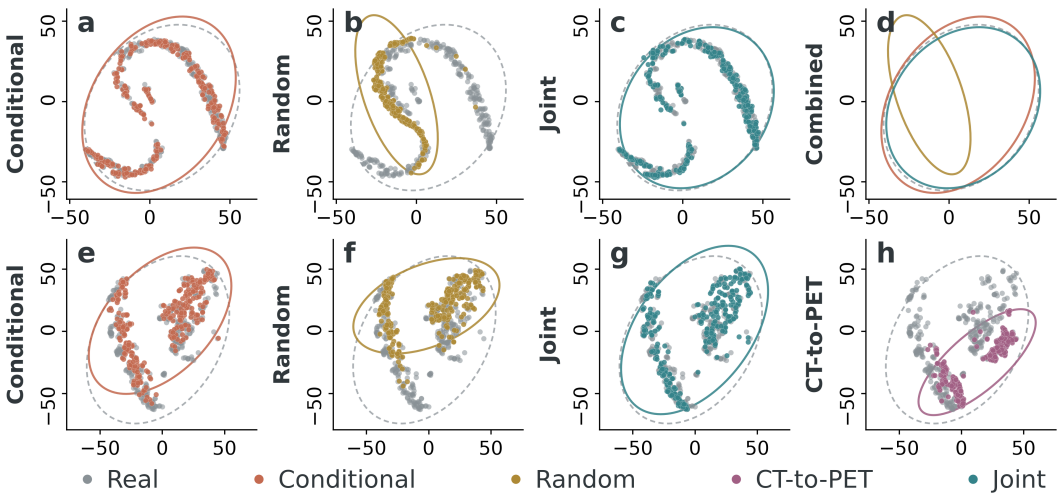

Figure 4: t-SNE of real and generated CT/PET features, globally average-pooled to 2048 dimensions from a Med3D-pretrained 3D ResNet-50 (Chen et al., 2019). Panels (a–c) compare real CT with conditional, random, and joint generation; (e–h) add CT-to-PET generation for real PET; (d) overlays ellipses from (a–c), with x- and y-axes denoting t-SNE dimensions 1 and 2, respectively.

Table 3 reports the average results of bi-modal joint generation and CT-to-PET cross-modal generation across the three datasets. For joint generation, our method achieves PSNR values of $26.10\,\mathrm{dB}$ for CT and $37.89\,\mathrm{dB}$ for PET, outperforming the best competing baselines by 0.08 and $3.10\,\mathrm{dB}$, respectively. The corresponding SSIM values are 0.784 and 0.944, representing improvements of 0.033 and 0.008, respectively. For cross-modal generation, our method achieves a PSNR of $36.99\,\mathrm{dB}$ and an SSIM of 0.943, outperforming the best competing baseline, CoCoLIT, by $19.66\,\mathrm{dB}$ and 0.322, respectively. The t-SNE visualization in Figure 4 further illustrates our framework's ability to capture the feature-distribution patterns of real CT and PET images across multiple generation tasks, with particularly close visual agreement under conditional and joint generation.

4.4 Downstream Segmentation

As shown in Table 3, we evaluate generated images for downstream lesion segmentation. For each dataset and modality, a separate nnU-Net model (Isensee et al., 2021) is trained on each method's generated images and evaluated on a common held-out real-image test set. For DeepPSMA, FDG, and PSMA, our method achieves CT segmentation Dice scores of 0.1680, 0.1675, and 0.0754, respectively. The CT HD95 values are 41.13, 83.19, and 71.72. For PET segmentation, the Dice scores are 0.6540, 0.5166, and 0.3490, respectively. The PET HD95 values are 12.06, 60.86, and 51.73. Compared with the second-best generative methods, the CT Dice scores are higher by 0.0036, 0.0778, and 0.0005, while the PET Dice scores are higher by 0.0055, 0.0119, and 0.0304, respectively. Our method achieves the lowest HD95 among generative methods in five of six settings and the second-lowest HD95 for DeepPSMA CT segmentation. These results indicate that the generated CT and PET images can support lesion segmentation, as further illustrated in Figure 5.

4.5 Ablation Studies

Table 3 presents separate ablation studies of SLG and TTSE. To evaluate SLG, we replace it with ControlNet while keeping all other settings unchanged. In joint generation, SLG improves CT PSNR and SSIM by $0.49\mathrm{dB}$ and 0.017, respectively, and PET PSNR and SSIM by $0.67\mathrm{dB}$ and 0.022, respectively. These results indicate that SLG improves conditional generation performance over ControlNet without replicating the encoding path of the U-Net.

To assess TTSE, we replace it with standard cross-attention. With TTSE, the CT PSNR and SSIM in joint generation improve by $1.80\,\mathrm{dB}$ and 0.061, respectively. The PET PSNR and SSIM improve by $1.74\,\mathrm{dB}$ and 0.050, respectively. For CT-to-PET generation, the PSNR and SSIM improve by $2.15\,\mathrm{dB}$ and 0.022, respectively. These results support TTSE for joint and cross-modal generation.

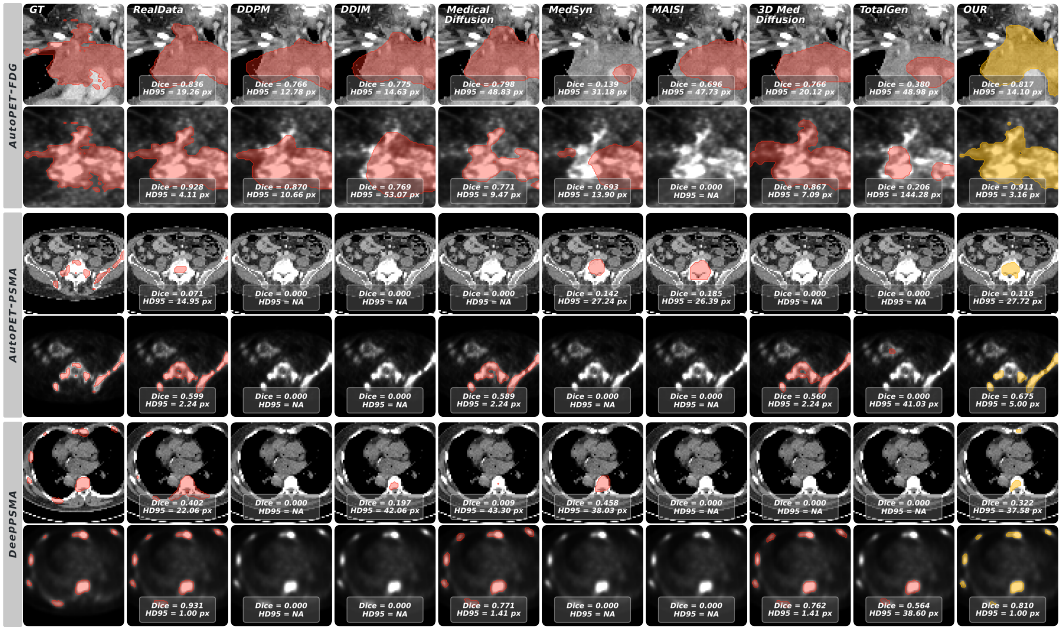

Figure 5: Qualitative comparison of downstream lesion segmentation across three datasets. CT images are shown above PET images; columns show the ground truth, RealData baseline, and segmentation results associated with different generation methods. Colored overlays indicate lesion regions, highlighting differences in localization and boundaries.

Table 3: Downstream lesion segmentation on FDG, PSMA, and DeepPSMA datasets, with pooled sample means for bi-modal joint generation, CT-to-PET generation, and ablations. Results include means, 95% confidence intervals, and p-values. Segmentation p-values accompany the Ours row; generation p-values appear in the final row. Bold gray-shaded and underlined light-gray-shaded cells mark the best and second-best results, respectively, excluding the real-data reference.

Method	*AutoPET-FDG-CT*		*AutoPET-FDG-PET*		*AutoPET-PSMA-CT*		*AutoPET-PSMA-PET*		*DeepPSMA-CT*		*DeepPSMA-PET*	
	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓	Dice ↑	HD95 ↓
GT$_{(\text{Real})}$	$0.2870_{\pm 0.0901}$	$80.72_{\pm 13.66}$	$0.5784_{\pm 0.0792}$	$45.57_{\pm 17.42}$	$0.0823_{\pm 0.0414}$	$81.20_{\pm 14.34}$	$0.4937_{\pm 0.0864}$	$25.19_{\pm 8.97}$	$0.1032_{\pm 0.0624}$	$48.31_{\pm 19.30}$	$0.7098_{\pm 0.1120}$	$15.53_{\pm 8.78}$
DDPM	$\underline{0.0897}_{\pm 0.0408}$	$91.30_{\pm 13.83}$	$0.1398_{\pm 0.0620}$	$86.51_{\pm 15.16}$	$0.0203_{\pm 0.0137}$	$\underline{76.56}_{\pm 11.26}$	$0.0335_{\pm 0.0160}$	$74.04_{\pm 11.24}$	$0.1055_{\pm 0.0438}$	$\mathbf{38.29}_{\pm 10.32}$	$0.0989_{\pm 0.0817}$	$64.57_{\pm 25.33}$
DDIM	$0.0864_{\pm 0.0361}$	$86.69_{\pm 12.61}$	$0.0956_{\pm 0.0524}$	$85.20_{\pm 13.06}$	$0.0179_{\pm 0.0147}$	$77.20_{\pm 11.37}$	$0.0080_{\pm 0.0060}$	$87.10_{\pm 9.04}$	$0.0391_{\pm 0.0252}$	$62.71_{\pm 19.37}$	$0.0356_{\pm 0.0414}$	$73.63_{\pm 18.94}$
MedicalDiffusion	$0.0468_{\pm 0.0293}$	$93.65_{\pm 14.24}$	$0.4686_{\pm 0.0876}$	$78.42_{\pm 18.78}$	$\underline{0.0749}_{\pm 0.0374}$	$80.34_{\pm 14.75}$	$0.2658_{\pm 0.0739}$	$65.51_{\pm 16.68}$	$0.1437_{\pm 0.0733}$	$44.20_{\pm 8.87}$	$\underline{0.6485}_{\pm 0.0686}$	$16.22_{\pm 6.19}$
MedSyn	$0.0416_{\pm 0.0260}$	$92.95_{\pm 13.27}$	$0.1916_{\pm 0.0595}$	$85.28_{\pm 13.01}$	$0.0178_{\pm 0.0131}$	$81.08_{\pm 10.91}$	$0.0025_{\pm 0.0026}$	$79.63_{\pm 10.15}$	$\underline{0.1644}_{\pm 0.0902}$	$41.94_{\pm 15.06}$	$0.0883_{\pm 0.0522}$	$35.00_{\pm 6.41}$
MAISI	$0.0065_{\pm 0.0058}$	$86.62_{\pm 12.18}$	$0.1826_{\pm 0.0648}$	$98.90_{\pm 15.45}$	$0.0542_{\pm 0.0262}$	$92.01_{\pm 17.76}$	$0.0550_{\pm 0.0244}$	$74.08_{\pm 10.47}$	$0.0032_{\pm 0.0028}$	$58.75_{\pm 10.02}$	$0.0325_{\pm 0.0310}$	$82.08_{\pm 30.82}$
3D MedDiffusion	$0.0751_{\pm 0.0480}$	$110.24_{\pm 14.35}$	$\underline{0.5047}_{\pm 0.0758}$	$\underline{67.24}_{\pm 17.58}$	$0.0017_{\pm 0.0024}$	$102.79_{\pm 12.41}$	$\underline{0.3186}_{\pm 0.0747}$	$\underline{63.45}_{\pm 16.31}$	$0.0029_{\pm 0.0036}$	$117.01_{\pm 48.72}$	$0.6286_{\pm 0.0450}$	$\underline{15.39}_{\pm 7.07}$
TotalGen	$0.0444_{\pm 0.0363}$	$\underline{84.94}_{\pm 15.22}$	$0.0875_{\pm 0.0568}$	$101.68_{\pm 14.60}$	$0.0037_{\pm 0.0028}$	$77.96_{\pm 10.60}$	$0.0227_{\pm 0.0168}$	$86.18_{\pm 14.22}$	$0.1134_{\pm 0.0837}$	$43.67_{\pm 6.07}$	$0.0797_{\pm 0.0779}$	$83.22_{\pm 35.22}$
Ours	$\mathbf{0.1675}_{\pm 0.0657}$	$\mathbf{83.19}_{\pm 13.08}$	$\mathbf{0.5166}_{\pm 0.0742}$	$\mathbf{60.86}_{\pm 17.45}$	$\mathbf{0.0754}_{\pm 0.0381}$	$\mathbf{71.72}_{\pm 12.22}$	$\mathbf{0.3490}_{\pm 0.0726}$	$\mathbf{51.73}_{\pm 14.87}$	$\mathbf{0.1680}_{\pm 0.0856}$	$\underline{41.13}_{\pm 5.99}$	$\mathbf{0.6540}_{\pm 0.0664}$	$\mathbf{12.06}_{\pm 5.65}$
p-value	$p = .014$	$p = .301$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$

Method	*Joint Generation-CT*		*Joint Generation-PET*		Method	*CT-to-PET*	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑		PSNR ↑	SSIM ↑
MedSyn	23.93 [23.708, 24.152]	0.751 [0.739, 0.763]	34.79 [34.450, 35.130]	$\underline{0.936}$ [0.931, 0.941]	IL-CLDM	16.38 [15.995, 16.765]	0.173 [0.152, 0.194]
TotalGen	$\underline{26.02}$ [25.796, 26.244]	0.502 [0.494, 0.510]	30.62 [29.993, 31.247]	0.488 [0.459, 0.517]	CoCoLiT	17.33 [16.918, 17.742]	0.621 [0.599, 0.643]
w/o SLG	25.61 [25.295, 25.925]	$\underline{0.767}$ [0.756, 0.778]	$\underline{37.22}$ [36.692, 37.748]	0.922 [0.915, 0.929]	w/o SLG	×	×
w/o TTSE	24.30 [24.069, 24.531]	0.723 [0.713, 0.733]	36.15 [35.598, 36.702]	0.894 [0.887, 0.901]	w/o TTSE	$\underline{34.84}$ [34.307, 35.373]	$\underline{0.921}$ [0.916, 0.926]
Ours	**26.10 [25.752, 26.448]**	**0.784 [0.773, 0.795]**	**37.89 [37.383, 38.397]**	**0.944 [0.939, 0.949]**	**Ours**	**36.99 [36.445, 37.535]**	**0.943 [0.938, 0.948]**
p-value	$p = .465$	$p < .001$	$p < .001$	$p = .117$	p-value	$p < .001$	$p < .001$

5 Conclusion

In this work, we presented UniCTPET that supports single-modality, structurally conditioned, CT-to-PET, and joint CT/PET generation within one model without task-specific retraining. TTSE adapts cross-modal structural interactions to modality feature states and flow time, accommodating both one-way conditional dependencies and bidirectional coordination. Complementing this mechanism, SLG enables spatially controlled generation through lightweight condition encoders without duplicating the backbone encoder. Experiments on three whole-body PET/CT datasets demonstrate improved generation fidelity and downstream lesion segmentation compared with competing generative methods. Our findings highlight adaptive modality interaction as an important complement to parameter sharing when unifying heterogeneous generation tasks, motivating further extensions to a broader range of medical imaging modalities and tasks.

AI USE STATEMENT

We used generative AI tools solely for manuscript refinement, including language translation, grammar correction, and improving clarity and readability. AI tools were not used for research ideation, hypothesis generation, experimental design, data analysis, result generation, dataset creation, mathematical derivations, proofs, or scientific claims. All AI-assisted suggestions were reviewed and verified by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This study uses existing PET/CT data from the publicly available AutoPET-FDG, AutoPET-PSMA, and DeepPSMA datasets. All data were obtained from the original providers and used in accordance with their applicable access conditions and terms of use. No additional participants were recruited, and no new patient data were collected for this study.

REPRODUCIBILITY STATEMENT

Our source code is available at `https://anonymous.4open.science/r/UniCTPET-043E/`.

REFERENCES

Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Multimodal MR synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, 37(3):803–814, 2017.

Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58:101535, 2019.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer Learning for 3D Medical Image Analysis, 2019. URL `https://arxiv.org/abs/1904.00625`.

Jiaqi Cui, Bo Liu, Xi Wu, Jiliu Zhou, Luping Zhou, Dinggang Shen, and Yan Wang. All-in-one medical image restoration via text-guided prompting and dual-domain modeling. *Information Fusion*, pp. 104558, 2026.

Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. Deep learning based synthesis of MRI, CT and PET: Review and analysis. *Medical Image Analysis*, 92:103046, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, June 2021.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. MAISI: Medical AI for synthetic imaging. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4430–4441. IEEE, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *arXiv preprint arxiv:2006.11239*, 2020.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Price Jackson, Michael Hofman, James Patrick Buteau, Lachlan McIntosh, and Yu Sun. A repository of annotated PSMA and FDG PET/CT images for algorithm development in staging of mCRPC for treament with 177Lu-PSMA therapy, 2025.

Katharina Jeblick, Balthasar Schachtner, Andreas Mittermeier, Jakob Dexl, Philipp Wesp, Thomas Küstner, Sergios Gatidis, Marcel Früh, Matthias P Fabritius, Felix Herr, et al. A whole-body PSMA-PET/CT dataset with manually annotated tumor lesions. *Scientific Data*, 13(1):1023, 2026.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023.

Suzanne E Lapi, Peter JH Scott, Andrew M Scott, Albert D Windhorst, Brian M Zeglis, May Abdel-Wahab, Richard P Baum, John M Buatti, Francesco Giammarile, Ana P Kiess, et al. Recent advances and impending challenges for the radiopharmaceutical sciences in oncology. *The Lancet Oncology*, 25(6):e236–e249, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

Richard Osuala, Kaisar Kushibar, Lidia Garrucho, Akis Linardos, Zuzanna Szafranowska, Stefan Klein, Ben Glocker, Oliver Diaz, and Karim Lekadir. Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *Medical Image Analysis*, 84:102704, 2023.

Zaixin Ou, Yongsheng Pan, Fang Xie, Qihao Guo, and Dinggang Shen. Image-and-Label conditioning latent diffusion model: Synthesizing a β-PET from MRI for detecting amyloid status. *IEEE Journal of Biomedical and Health Informatics*, 29(2):1221–1231, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Alec Sargood, Lemuel Puglisi, James H Cole, Neil P Oxtoby, Daniele Ravì, and Daniel C Alexander. CoCoLIT: ControlNet-conditioned latent image translation for MRI to amyloid PET synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8778–8786, 2026.

Johannes Schwenck, Dominik Sonanini, Jonathan M Cotton, Hans-Georg Rammensee, Christian la Fougère, Lars Zender, and Bernd J Pichler. Advances in PET imaging of cancer. *Nature Reviews Cancer*, 23(7):474–490, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021.

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3D MedDiffusion: A 3D medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Transactions on Medical Imaging*, 44(12):4960–4972, 2025.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Jinbao Wei, Gang Yang, Zhijie Wang, Shimin Tao, Aiping Liu, and Xun Chen. Degradation-aware prompted transformer for unified medical image restoration. *IEEE Transactions on Image Processing*, 34:8583–8598, 2025.

Xiaojiao Xiao, Jianfeng Zhao, and Shuo Li. Task relevance driven adversarial learning for simultaneous detection, size grading, and quantification of hepatocellular carcinoma via integrating multi-modality MRI. *Medical Image Analysis*, 81:102554, 2022.

Yanwu Xu, Li Sun, Wei Peng, Shuyue Jia, Katelyn Morrison, Adam Perer, Afrooz Zandifar, Shyam Visweswaran, Motahhare Eslami, and Kayhan Batmanghelich. MedSyn: text-guided anatomy-aware synthesis of high-fidelity 3D CT images. *IEEE Transactions on Medical Imaging*, 43(10): 3648–3660, 2024.

Zhiwen Yang, Haowei Chen, Ziniu Qian, Yang Yi, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. All-in-one medical image restoration via task-adaptive routing. In *International conference on medical image computing and computer-assisted intervention*, pp. 67–77. Springer, 2024.

Zhiwen Yang, Jiayin Li, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Restore-RWKV: Efficient and effective medical image restoration with RWKV. *IEEE Journal of Biomedical and Health Informatics*, 30(1):513–526, 2025.

Zhiwen Yang, Jiaju Zhang, Yang Yi, Jian Liang, Bingzheng Wei, and Yan Xu. TAT: Task-Adaptive transformer for All-in-One medical image restoration. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025*, pp. 565–575, Cham, 2026. Springer Nature Switzerland.

Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.

Siyeop Yoon, Sifan Song, Pengfei Jin, Matthew Tivnan, Yujin Oh, Sekeun Kim, Dufan Wu, Xiang Li, and Quanzheng Li. Cascaded 3d Diffusion models for whole-body 3D 18-F FDG PET/CT synthesis from demographics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 99–109. Springer, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, pp. 3813–3824. IEEE, 2023.

A Appendix

A.1 Mathematical Background of Flow Matching

This section describes the linear-path flow-matching formulation used in our model. For clarity, we suppress modality indices and fixed conditioning inputs. Let $z_0 \sim p_{\text{data}}$ denote a real image latent, and independently sample $\epsilon \sim \mathcal{N}(0, I)$. The interpolation between data and noise is defined as

$$z_t = (1-t)z_0 + t\epsilon, \qquad t \in [0,1]. \tag{10}$$

Under this convention, $t = 0$ corresponds to data and $t = 1$ to noise. For $t > 0$, the conditional distribution and the induced marginal distribution are

$$\begin{aligned} q_t(z \mid z_0) &= \mathcal{N}\big(z;\, (1-t)z_0,\, t^2 I\big)\,, \\ p_t(z) &= \mathbb{E}_{z_0 \sim p_{\text{data}}}\left[q_t(z \mid z_0)\right]. \end{aligned} \tag{11}$$

These distributions interpolate between the data distribution at $t = 0$ and the standard Gaussian distribution at $t = 1$.

The derivative of the interpolation with respect to increasing t is $\epsilon - z_0$. We define the regression target in the opposite, noise-to-data direction:

$$v = z_0 - \epsilon = -\frac{\mathrm{d}z_t}{\mathrm{d}t}. \tag{12}$$

The velocity network $\hat{v}_\theta(z_t, t)$ is trained by minimizing

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{z_0,\epsilon,t}\left[\left\|\hat{v}_\theta(z_t,t) - (z_0 - \epsilon)\right\|_2^2\right], \qquad t \sim \mathcal{U}(0,1), \tag{13}$$

where t is sampled independently of z_0 and ϵ. The population minimizer of this squared-error objective is the conditional mean

$$v^*(z,t) = \mathbb{E}\left[z_0 - \epsilon \,|\, z_t = z,\, t\right]. \tag{14}$$

Thus, the sample-wise displacement targets provide supervision for a velocity field defined over the intermediate latent distribution.

Let $\widetilde{z}_t$ denote the latent state along a generated trajectory, distinct from the interpolated training sample z_t. With the velocity convention in Eq. equation 12, the corresponding generative ordinary differential equation is

$$\frac{\mathrm{d}\widetilde{z}_t}{\mathrm{d}t} = -\hat{v}_\theta(\widetilde{z}_t, t), \qquad \widetilde{z}_1 \sim \mathcal{N}(0, I). \tag{15}$$

The equation is integrated from $t = 1$ to $t = 0$, and the resulting latent $\widetilde{z}_0$ is mapped to image space by the pretrained decoder. For fixed conditioning inputs, the same formulation applies with the data distribution, regression objective, and velocity field conditioned on those inputs.

A.2 Cross-Modal Exchange and Task-Gradient Alignment

Both analyses use the same test cases, noise realizations, and time points, with skeletal and lesion conditioning disabled.

Relative exchange strength. Figure 1(b) compares two models with the same TTSE architecture, trained separately for CT-to-PET and joint CT–PET generation. For each case, let $h_{p,k}(t)$ and $h'_{p,k}(t)$ denote the input and output PET features of exchange block k at time t, where $k \in \{1, \ldots, 5\}$ indexes the four encoder exchange blocks and the bottleneck exchange block. The overall relative exchange strength $R(t)$ is

$$R(t) = \left[\frac{1}{5}\sum_{k=1}^{5}\left(\frac{\left\|h'_{p,k}(t) - h_{p,k}(t)\right\|_F}{\max\left(\left\|h_{p,k}(t)\right\|_F, 10^{-12}\right)}\right)^2\right]^{1/2}, \tag{16}$$

where $\|\cdot\|_F$ denotes the Frobenius norm over all channel and spatial dimensions, and 10^{-12} ensures numerical stability. This metric measures relative PET feature updates through the equally weighted root mean square of the normalized feature changes across the five blocks.

Task-gradient alignment. For Figure 1(c), we compute the two task gradients within each TTSE or cross-attention model with respect to the structure encoder, temporal FiLM, cross-attention, and PET output projection parameters in all five exchange blocks. Corresponding parameter names and shapes are checked across models.

For each case, let $\mathbf{g}_{\text{joint}}(t)$ and $\mathbf{g}_{\text{CT}\to\text{PET}}(t)$ denote the gradients of the joint-generation and CT-to-PET losses, respectively, flattened and concatenated over the selected parameters across all five blocks. Their cosine similarity is

$$\text{CosSim}\left(\mathbf{g}_{\text{joint}}(t), \mathbf{g}_{\text{CT}\to\text{PET}}(t)\right) = \frac{\left\langle \mathbf{g}_{\text{joint}}(t), \mathbf{g}_{\text{CT}\to\text{PET}}(t)\right\rangle}{\left\|\mathbf{g}_{\text{joint}}(t)\right\|_2 \left\|\mathbf{g}_{\text{CT}\to\text{PET}}(t)\right\|_2}, \tag{17}$$

where CosSim denotes cosine similarity, and $\langle\cdot,\cdot\rangle$ and $\|\cdot\|_2$ denote the Euclidean inner product and norm, respectively. Higher values indicate greater agreement between the task-gradient directions for the selected parameters, whereas negative values indicate gradient conflict. Unused parameter gradients contribute zero entries.

A.3 Evaluation Metrics

In this section, we define the metrics used to evaluate generated CT and PET images and downstream lesion segmentation. Each metric is summarized below:

- **Peak Signal-to-Noise Ratio (PSNR)**
  Purpose: Measures the fidelity of generated images relative to their reference images.
  Methodology: Computes $10 \log_{10}(R^2/\mathrm{MSE})$, where $R = \max(y) - \min(y)$ is the intensity range of the reference volume y.
  Interpretation: Higher PSNR indicates lower reconstruction error.
- **Structural Similarity Index Measure (SSIM)**
  Purpose: Evaluates the structural similarity between generated and reference images.
  Methodology: Compares local luminance, contrast, and structure in 3D using the reference intensity range R.
  Interpretation: Higher SSIM indicates greater structural similarity.
- **Normalized Mean Absolute Error (NMAE)**
  Purpose: Measures the magnitude of intensity differences between generated and reference images.
  Methodology: Divides the sum of absolute voxel-wise errors by the sum of absolute reference intensities.
  Interpretation: Lower NMAE indicates smaller intensity errors.
- **Normalized Mean Squared Error (NMSE)**
  Purpose: Measures intensity differences while placing greater weight on large errors.
  Methodology: Divides the sum of squared voxel-wise errors by the sum of squared reference intensities.
  Interpretation: Lower NMSE indicates better agreement with the reference image.
- **Normalized Cross-Correlation (NCC)**
  Purpose: Assesses the correspondence between generated and reference image intensities.
  Methodology: Computes the correlation between voxel intensities after centering and scaling them.
  Interpretation: Higher NCC indicates stronger intensity correspondence.
- **Dice Similarity Coefficient (Dice)**
  Purpose: Evaluates the overlap between predicted and reference lesion masks.
  Methodology: Computes twice the volume of their intersection divided by the sum of their volumes.
  Interpretation: Higher Dice indicates better segmentation overlap.
- **95th-Percentile Hausdorff Distance (HD95)**
  Purpose: Evaluates the agreement between predicted and reference lesion boundaries.
  Methodology: Computes the 95th percentile of the distances between the two boundaries.
  Interpretation: Lower HD95 indicates better boundary agreement.

A.4 Supplementary Quantitative Results

Table 4: FID-Med3D comparison of CT and PET feature distributions across generation settings. The scores compare generated and real images in Med3D feature space, complementing the t-SNE visualization in Fig. 4; lower values indicate closer distributional agreement. CT-to-PET generates PET only, so its CT entry is not applicable. Bold gray-shaded and underlined light-gray-shaded cells mark the best and second-best scores in each column, respectively.

Setting	CT FID-Med3D ↓	PET FID-Med3D ↓
Conditional	<u>0.0029201</u>	<u>0.0000842</u>
Random	0.1956792	0.0002544
CT-to-PET	–	0.0001290
Joint	**0.0008950**	**0.0000226**

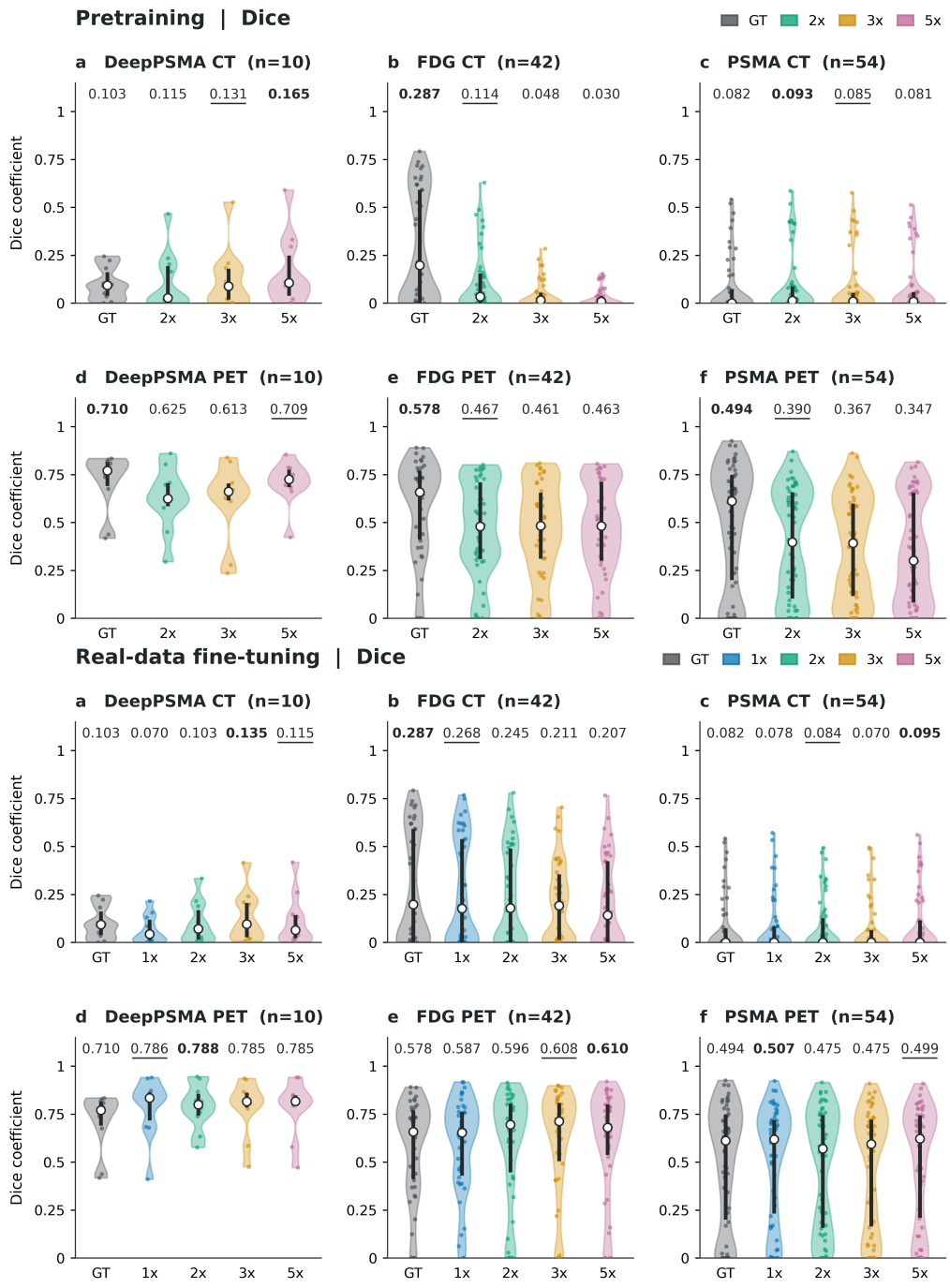

Figure 6: Dice scores for downstream lesion segmentation on DeepPSMA, FDG, and PSMA. Upper panels show results for nnU-Net trained on images generated by our framework; lower panels show results after subsequent fine-tuning on real images for 20 epochs. GT denotes the baseline trained directly on real images; multipliers indicate synthetic training-set sizes relative to the real training set. Violin plots show per-case distributions, with mean Dice above each group; higher values indicate greater overlap. After fine-tuning, PET Dice exceeds the baseline at all evaluated scales on DeepPSMA and FDG and selected scales on PSMA. CT Dice exceeds the baseline in selected DeepPSMA and PSMA settings but remains below it on FDG.

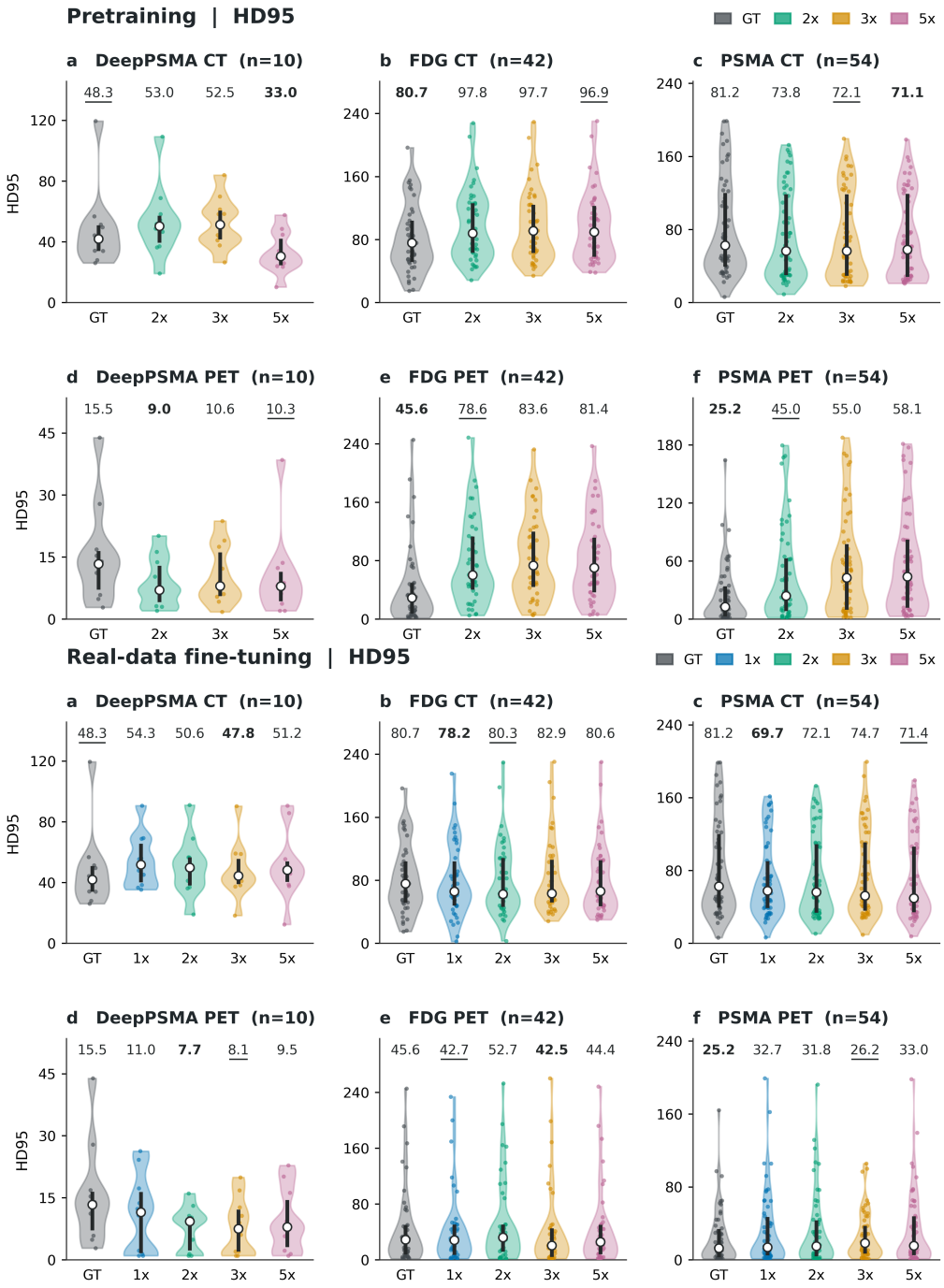

Figure 7: HD95 for downstream lesion segmentation on DeepPSMA, FDG, and PSMA. Upper panels show results for nnU-Net trained on images generated by our framework; lower panels show results after subsequent fine-tuning on real images for 20 epochs. GT denotes the baseline trained directly on real images; multipliers indicate synthetic training-set sizes relative to the real training set. Violin plots show per-case distributions, with mean HD95 above each group; lower values indicate closer boundary agreement. After fine-tuning, HD95 is below the baseline at all evaluated scales for DeepPSMA PET and PSMA CT and selected scales in other settings; PSMA PET remains above the baseline.

A.5 Generation Examples

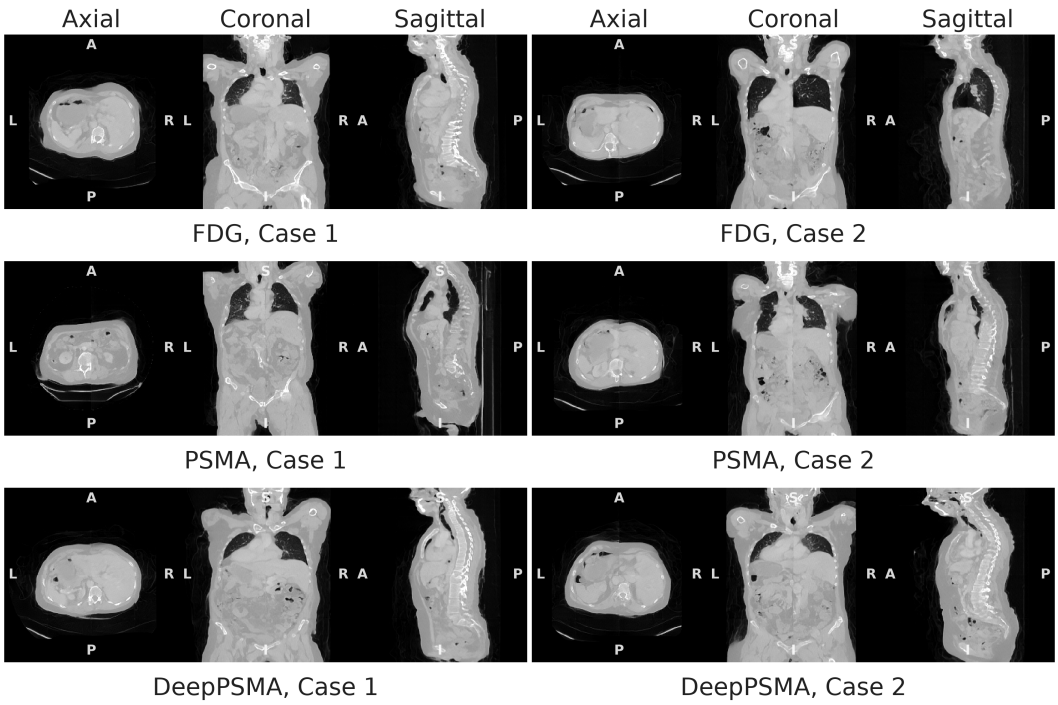

Figure 8: Randomly selected examples from the FDG, PSMA, and DeepPSMA datasets for single-modality CT generation. Each case is shown in axial, coronal, and sagittal views from left to right.

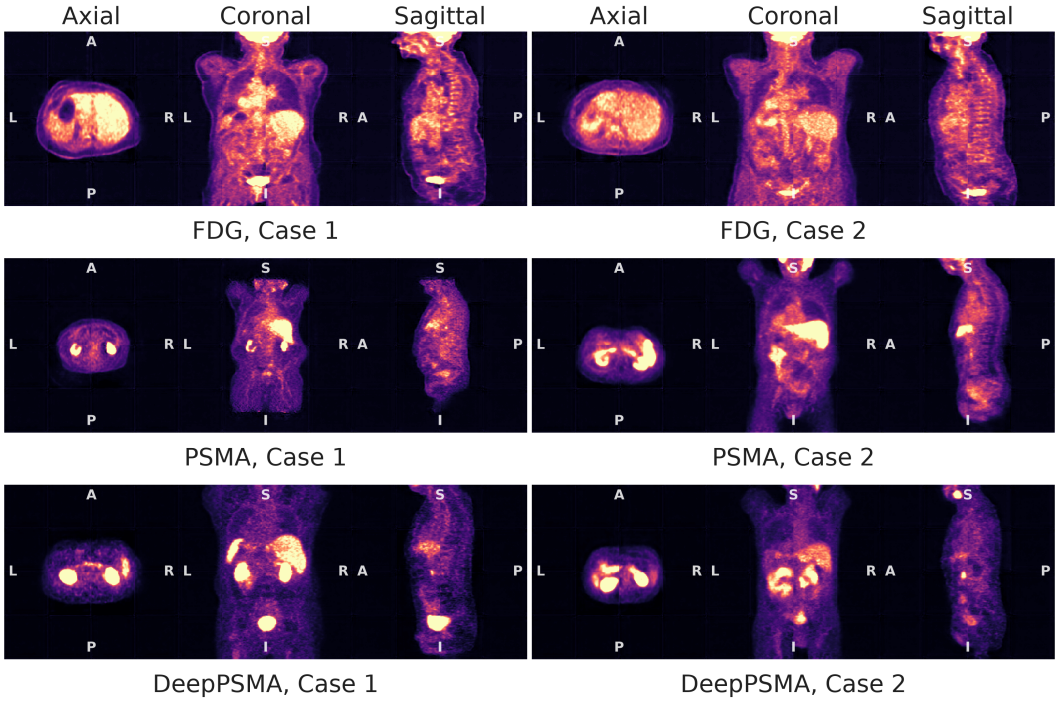

Figure 9: Randomly selected examples from the FDG, PSMA, and DeepPSMA datasets for single-modality PET generation. Each case is shown in axial, coronal, and sagittal views from left to right.

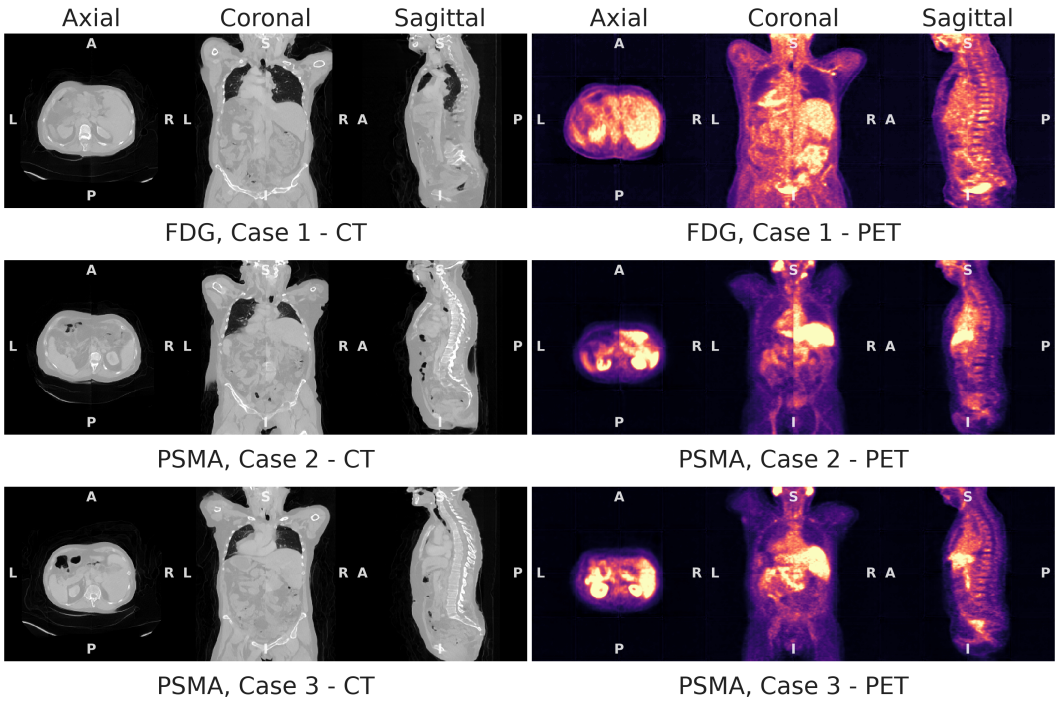

Figure 10: Randomly selected examples for bi-modal joint CT-PET generation. For each case, the generated CT volume is shown in the left triplet and the corresponding PET volume in the right triplet. Axial, coronal, and sagittal views are shown from left to right within each triplet.