WHERE TO REFINE AND WHAT TO RECOVER: DISENTANGLED CROSS-MODAL SEMANTIC ALIGNMENT FOR CT-TO-PET GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Positron Emission Tomography (PET) and Computed Tomography (CT) play complementary and essential roles in advanced healthcare, particularly for cancer diagnosis. However, CT is far more widely available than PET, primarily due to its substantially lower acquisition and operational costs. To bridge this gap, many CT-to-PET generation methods are developed based on structural alignment while overlooking the intrinsic characteristics of PET, such as metabolic activity, which limits the clinical utility of the synthesized PET images. To address this challenge, we propose a disentangled alignment framework that formulates CT-to-PET generation as a controlled residual generation problem, explicitly guiding both the spatial location ("where") and semantic content ("what") of PET signals. Specifically, a frozen backbone first generates a stable PET prediction, while a Scalable Interpolant Transformer (SiT) models the remaining residual refinement. Then, to characterize what should be recovered, intermediate SiT representations are adopted to align the semantic residual via deep supervision, thereby reinforcing the learning of missing semantics. Furthermore, an activation gate is introduced to identify regions requiring refinement while suppressing redundant residual noise in well-reconstructed areas. Extensive experiments on the popular Lung-PET-CT-Dx and RIDER-Lung-PET-CT datasets demonstrate that our method significantly outperforms state-of-the-art baselines, highlighting its strong potential for CT-to-PET image generation.

1 INTRODUCTION

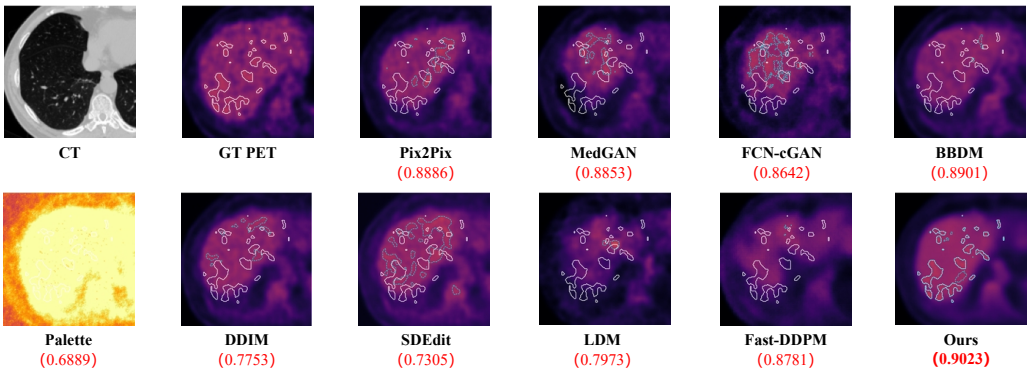


Figure 1: Motivation of our method. Rather than regenerating the entire target image, we selectively recover *what* semantics are missing and refine *where* they are needed, resulting in more faithful target structures. SSIM scores are shown in red numbers within parentheses.

Computed tomography (CT) and positron emission tomography (PET) provide fundamentally different views of disease, and they are both essential for disease diagnosis. Due to the high acquisition cost of PET imaging, synthesizing PET from CT has attracted increasing attention in the machine

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.

learning and computer vision community(Garg et al., 2025; Rathod et al., 2025; Zoghi et al., 2025). However, it is far more challenging than translating between modalities with high structure overlap. The generated PET images should not only perserve spatial structures but also represent semantics that are weakly coupled with structural appearance.

Recently, cross-modality translation in medical imaging has been significantly advanced by generative models. While diffusion models demonstrate remarkable fidelity in image synthesis (Chen et al., 2025; Huang et al., 2025; Kidder, 2024), specialized variants like Brownian Bridge Diffusion Models explicitly constrain the transport path between source and target domains (Li et al., 2023a). Meanwhile, Scalable Interpolant Transformers (SiT) offer a unified probabilistic framework via stochastic interpolants and learned velocity fields (Ma et al., 2024), providing enhanced expressiveness and trajectory controllability for cross-domain translation. Although these methods can generate high-quality PET images, they mainly focus on the structure similarity, ignoring the intrinsic characteristics of PET for semantic representation, thereby limiting the availability of the generated PET images. We show the visualization results of different generative models and ours in Fig. 1.

In this paper, we propose a residual generative modeling framework that explicitly addresses *what* to recover and *where* to refine for CT-to-PET generation. Specifically, a pretrained ResUNet first produces a Generated Anchor PET and is subsequently frozen, providing a globally stable target prediction. Instead of regenerating the complete PET image, a conditional Scalable Interpolant Transformer (SiT) is employed to model only the remaining residual correction. To determine *what* remains missing from the Generated Anchor PET, we introduce Residual Alignment, a target-specific residual representation alignment strategy inspired by Representation Alignment (Yu et al., 2025). A frozen PET-DINO encoder extracts representations from the ground-truth PET and the Generated Anchor PET, and their difference is used as an explicit semantic residual to supervise the intermediate representations of Residual SiT. To further determine *where* the residual correction should be applied, we introduce an Activation Gate conditioned on the CT image and the Generated Anchor PET, which identifies spatial locations requiring additional refinement while suppressing unnecessary residual responses in well-reconstructed regions. Finally, the gated residual is added to the Generated Anchor PET to obtain the final prediction.

Our contributions can be summarized as follows: **Residual generative modeling.** We formulate CT-to-PET synthesis as a residual generation problem, where a powerful, frozen backbone is retained to reconstruct globally stable PET structures, while a conditional SiT is dedicated to capturing semantic residuals. **Recovering semantics.** We introduce Residual Alignment, which supervises the residual generator using the differences between ground-truth PET and the generated anchor PET image, encouraging the generative branch to recover target semantics that are absent in the generated anchor PET image. **Refining targets.** We propose an activation gate strategy, which can guide the target spatially relevant regions to be refined with the residual semantics.

2 Related Work

2.1 Medical Image-to-Image Translation

Medical image synthesis has been widely studied as a paired image translation problem (Nie et al., 2018; Pan et al., 2023; Koetzier et al., 2024). Early approaches were largely based on convolutional encoder-decoder networks and conditional generative adversarial models. Pix2Pix established a general framework for paired image-to-image translation (Isola et al., 2017), while MedGAN extended adversarial learning to medical image synthesis using perceptual and style-based supervision (Armanious et al., 2020) . These methods demonstrated that cross-modality mappings can recover globally plausible target images from paired source data.

CT-to-PET synthesis is more challenging than many anatomical translation tasks because the two modalities encode different information. CT mainly describes anatomical structure and tissue density, whereas PET reflects radiotracer uptake and functional activity. Existing studies have shown that synthetic PET can be estimated from CT and may preserve clinically relevant information in lung cancer cohorts (Salehjahromi et al., 2024). However, deterministic and adversarial models typically optimize the complete target image, which can make spatially sparse target-specific signals difficult to recover accurately.

More recently, diffusion-based models have provided a more flexible framework for image-to-image translation(Kim et al., 2024; Moser et al., 2024; Jiang et al., 2025). Latent Diffusion Models improve efficiency by performing generation in a learned latent space (Rombach et al., 2022), while SDEdit and Brownian Bridge Diffusion Models demonstrate different ways of preserving source structure during stochastic generation and direct source-to-target transport (Meng et al., 2022; Li et al., 2023a). Palette further shows that conditional diffusion can be applied within a unified image translation framework (Saharia et al., 2022). In CT-to-PET synthesis, CPDM incorporates PET- and CT-specific priors into a conditional diffusion model and demonstrates the potential of generative modeling for large-scale paired translation (Nguyen et al., 2025).

Scalable Interpolant Transformers (SiT) combine stochastic interpolants with transformer architectures, providing a unified framework for diffusion- and flow-based generative modeling (Ma et al., 2024). This makes SiT well suited for cross-modal synthesis where deterministic regression may be insufficient.

2.2 Representation Learning and Representation Alignment

Self-supervised visual representation learning has produced feature encoders with strong transferability across a wide range of downstream tasks(Gui et al., 2024; Krishnan et al., 2022). DINO introduced self-distillation without labels for vision transformers and demonstrated that self-supervised transformer features can capture meaningful semantic structure and dense spatial correspondence (Caron et al., 2021). DINOv2 further scaled self-supervised visual pretraining and produced robust general-purpose visual features that transfer effectively across classification, retrieval, and dense prediction tasks (Oquab et al., 2023). These developments have encouraged the use of pretrained visual encoders as external sources of representation-level supervision.

Representation guidance has also been incorporated into generative models (Li et al., 2023b). Perceptual losses use pretrained networks to compare generated and reference images in feature space rather than relying only on pixel-wise differences, providing higher-level structural or semantic supervision. More recently, Representation Alignment (REPA) directly aligns intermediate hidden representations of diffusion- or flow-based transformers with features extracted from frozen pretrained visual encoders (Yu et al., 2025; Lu et al., 2025). REPA demonstrates that the representation quality of generative transformers can be improved through external semantic supervision, leading to more efficient optimization and improved generation quality.

These studies motivate the use of representation-level supervision for generative modeling. Different from aligning with the complete target representation, our Residual Alignment focuses on the representation difference between the ground-truth PET and the Generated Anchor PET, guiding the residual generator toward target semantics that remain missing from the anchor prediction.

3 METHOD

3.1 Overall Architecture

Fig. 2 illustrates the overall architecture of our proposed framework. We formulate cross-modal PET synthesis as a residual generative problem. Specifically, a pretrained ResUNet is first deployed as a baseline backbone to generate an anchor PET image from the CT image, establishing a coarse global distribution. To effectively recover the semantics, we exploit the generative capacity of the Scalable Interpolant Transformer (SiT) for residual modeling. To this end, we introduce two modules for residual refinement: Residual Alignment and the Activation Gate, which address what to recover and where to refine, respectively. Residual Alignment explicitly aligns intermediate SiT feature representations with the semantic residual between the ground-truth PET and the anchor image, thereby capturing the unmodeled semantic discrepancies. Concurrently, the Activation Gate dynamically localizes target refinement regions conditioned on both the input CT image and the anchor PET image. The final synthesized PET image is obtained by adding the gated semantic residual to the anchor PET image. Notably, Residual Alignment is used only during training and introduces no additional computational overhead at inference.

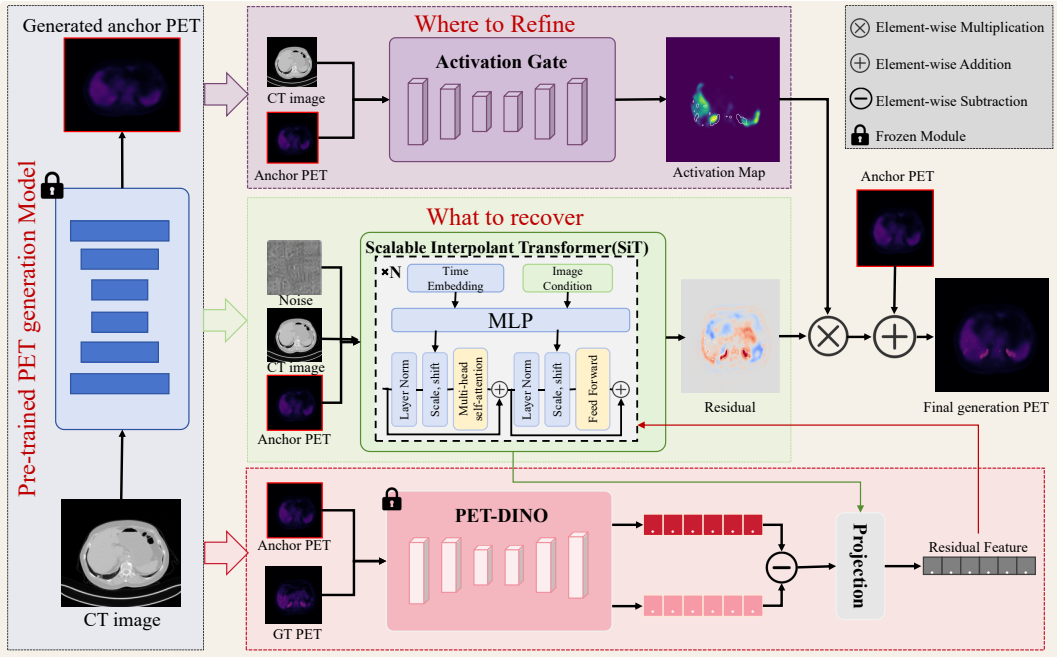


Figure 2: Overview of the proposed CT-to-PET residual generative framework. A frozen ResUNet provides anchor PET, Residual SiT models the correction, Residual Alignment guides the recovery of *what target semantics are missing*, and the Activation Gate determines *where refinement should be applied*. Residual Alignment is used only during training.

3.2 Residual Alignment

While the Residual SiT augments generative capacity beyond deterministic baselines, standard generative supervision fails to explicitly identify the semantics that are insufficient in the generated anchor PET image. To bridge this gap, we introduce Residual Alignment, an explicit feature level guidance mechanism. Residual Alignment enforces the intermediate hidden representations of Residual SiT to align directly with the semantic discrepancy between the ground-truth PET and the generated anchor PET image, thereby providing semantic residual representation for the insufficient semantics.

Semantic residual representation. Let E_{PET} denote the frozen PET-DINO encoder. Given the ground-truth PET image Y and the generated anchor PET image Y_{anchor}, we extract their PET representations as:

$$z_{\mathrm{GT}} = E_{\mathrm{PET}}(Y), \qquad z_{\mathrm{anchor}} = E_{\mathrm{PET}}(Y_{\mathrm{anchor}}). \tag{1}$$

Rather than using the complete target representation z_{GT} as supervision, we construct the semantic residual as follows:

$$z_{\mathrm{res}} = z_{\mathrm{GT}} - z_{\mathrm{anchor}}. \tag{2}$$

The resulting z_{res} explicitly represents target semantics that are present in the target PET but insufficiently captured in the generated anchor image. This formulation is naturally consistent with the role of residual SiT, which is responsible for modeling the remaining correction rather than reconstructing the entire target representation.

Why align residual representations? The key motivation of Residual Alignment is to supervise only what remains missing from the anchor PET, rather than repeatedly matching complete target representation. Let

$$R^* = Y - Y_{\mathrm{anchor}} \tag{3}$$

denote the image space correction from the anchor PET to ground-truth PET. In other words, R^* directly describes what still needs to be recovered after deterministic generation. To understand how this missing correction is reflected in the representation space, we consider the behavior of the frozen PET encoder around Y_{anchor}. According to its first-order Taylor expansion:

$$E_{\mathrm{PET}}(Y) \approx E_{\mathrm{PET}}(Y_{\mathrm{anchor}}) + J_E(Y_{\mathrm{anchor}})R^*, \tag{4}$$

where $J_E(Y_{\text{anchor}})$ denotes the local Jacobian of the PET encoder. The term $J_E(Y_{\text{anchor}})R^*$ describes the corresponding representation change induced by the remaining image-space correction.

Combining Eq. 2 and Eq. 4 yields:

$$z_{\text{res}} \approx J_E(Y_{\text{anchor}})R^*. \tag{5}$$

Eq. 5 explains why z_{res} is an appropriate supervision target for Residual SiT. While R^* specifies the desired correction in image space, z_{res} provides its corresponding direction in the PET representation space. Aligning the intermediate SiT representation with z_{res} therefore explicitly guides the residual generator toward *what* target semantics remain missing from the generated anchor PET, instead of spending generative capacity on information that has already been recovered by the anchor.

Alignment with Residual SiT. Let h denote an intermediate feature from Residual SiT. Residual Alignment aligns it with the semantic residual z_{res} through

$$\mathcal{L}_{\text{rep}} = \|P_\psi(h) - \text{sg}\left(P_\psi(z_{\text{res}})\right)\|_2^2, \tag{6}$$

where P_ψ is a lightweight projection head and $\text{sg}(\cdot)$ denotes stop-gradient. The resulting loss is back-propagated through the projection head and Residual SiT, while the PET-DINO remains frozen. Thus, residual alignment supervises SiT through representation alignment without injecting z_{res} into its forward generation.

Gradient supervision of SiT. The optimization of residual alignment can be calculated explicitly by considering the gradient with respect to the SiT parameters θ. For Eq. 6, the gradient is

$$\frac{\partial \mathcal{L}_{\text{rep}}}{\partial \theta} = 2\left(\tilde{h} - \tilde{z}_{\text{res}}\right)^\top J_P(h)\frac{\partial h}{\partial \theta}, \tag{7}$$

where $J_P(h)$ denotes the Jacobian of the projection head with respect to h. Eq. 7 shows that Residual Alignment reshapes the internal representation of Residual SiT through auxiliary gradient supervision. When the SiT representation deviates from the semantic residual target, the resulting gradient encourages the SiT blocks to encode the insufficient target semantics in the generated anchor PET image.

3.3 Activation Gate

While residual alignment constrains *what* semantics to recover, it does not determine *where* the residual correction should be applied. Consequently, an unconstrained residual generator risks introducing noise into regions already well-reconstructed in the anchor image. To address this, we introduce the Activation Gate to explicitly constrain the spatial of residual refinements.

Spatial residual modulation. The Activation Gate jointly observes from both CT image and the generated anchor image, which can be represented as:

$$G = \sigma\left(g_\phi\left(X, Y_{\text{anchor}}\right)\right), \qquad G \in [0,1]^{H\times W}, \tag{8}$$

where g_ϕ denotes the activation network and $\sigma(\cdot)$ is the sigmoid function. Their combination allows the gate to estimate whether additional residual refinement is required at each location.

Given the residual ΔY produced by Residual SiT, the final PET prediction is obtained by spatially modulating the residual with the Activation Gate, which can be represented as:

$$\hat{Y} = Y_{\text{anchor}} + G \odot \Delta Y, \tag{9}$$

where $\odot$ denotes element-wise multiplication.

Why activation gating is necessary. Let $R^* = Y - Y_{\text{anchor}}$ denote the ideal residual defined in Eq. 3. Using Eq. 9, the error can be expressed as

$$\hat{Y} - Y = G \odot \Delta Y - R^*. \tag{10}$$

For each location i, the Activation Gate minimizes the residual reconstruction error as follows:

$$e_i(G_i) = \left(G_i \Delta Y_i - R_i^*\right)^2, \quad G_i^* = \text{clip}\left(\frac{R_i^* \Delta Y_i}{\Delta Y_i^2 + \epsilon}, 0, 1\right), \tag{11}$$

where R_i^* denotes the optimal residual and ϵ is a small numerical constant. The clipping operation constrains G_i^* to $[0, 1]$, so the Activation Gate can only preserve or suppress the generated residual at each location, directly controlling *where* to refine.

Eq. 11 provides an intuitive interpretation of the Activation Gate. The activation response is useful when the generated residual is consistent with the correction required by the target prediction. Conversely, if little residual correction is required, or if the generated update is inconsistent with the optimal residual, suppressing the correction better preserves the prediction. The learned gate approximates this spatial selection using only CT and generated anchor PET at inference.

Bounded residual refinement. The gate further imposes an explicit constraint on the magnitude of the generative correction. Since $0 \leq G_i \leq 1$, for every location, we can have

$$\|G \odot \Delta Y\|_2 \leq \|\Delta Y\|_2 . \tag{12}$$

Thus, the Activation Gate cannot amplify the residual magnitude and can only preserve or suppress individual residual responses. This provides a direct mechanism for limiting unnecessary generative modifications and allows the generated anchor PET to remain dominant in regions that are already well reconstructed.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

We conduct experiments on two publicly available CT-PET datasets from The Cancer Imaging Archive (TCIA): Lung-PET-CT-Dx (Li et al., 2020) and RIDER-Lung-PET-CT(Muzi et al., 2015). Both datasets are included in the experimental cohort and split at the patient level into training, validation, and testing sets to avoid subject overlap across different subsets. CT intensities are windowed to a fixed Hounsfield unit range and normalized to $[0, 1]$, while PET is represented using quantitative SUVbw values with a fixed global clipping range followed by normalization. PET volumes are resampled to the corresponding CT geometry, and paired CT–PET images are resized to a unified spatial resolution for training and evaluation.

All compared methods follow the same data split, preprocessing pipeline, and evaluation protocol. The deterministic CT-to-PET backbone is pretrained and frozen during residual generative training. Residual SiT, Residual Alignment, and the Activation Gate are optimized jointly, while the PET-DINO encoder remains frozen and provides representation level supervision only during training. The best model is selected on the validation set, and residuals are generated with a fixed solver during testing. Detailed dataset statistics, preprocessing procedures, and implementation settings are provided in the Appendix.

Table 1: Quantitative comparison on the Lung-PET-CT-Dx and RIDER-Lung-PET-CT datasets. ↓/↑ indicate lower/higher is better. Best results are shown in **bold**, and second-best results are underlined.

	Lung-PET-CT-Dx					RIDER-Lung-PET-CT				
Method	**MAE↓**	**RMSE↓**	**PSNR↑**	**SSIM↑**	**PCC↑**	**MAE↓**	**RMSE↓**	**PSNR↑**	**SSIM↑**	**PCC↑**
Pix2Pix_{2018}	0.009700	0.032696	32.541	0.8886	0.8855	0.017142	0.047110	30.088	0.7401	0.8051
MedGAN_{2020}	0.008740	0.030891	33.539	0.8853	0.8994	0.018869	0.050110	29.087	0.7245	0.7540
UP-GAN_{2021}	0.008381	0.029738	31.950	0.8793	0.8705	0.020589	0.052056	28.407	0.7199	0.7364
FCN-cGAN_{2023}	0.010003	0.029799	33.404	0.8642	0.8789	0.019356	0.050263	28.903	0.7144	0.7584
DDIM_{2020}	0.010510	0.041500	25.900	0.7753	0.7824	0.017161	0.046212	30.355	0.7669	0.8142
Palette_{2022}	0.141200	0.067900	19.056	0.6889	0.6752	0.148308	0.091913	16.233	0.4468	0.6855
SDEdit_{2022}	0.011440	0.048400	22.234	0.7305	0.7093	0.031492	0.051054	28.058	0.4844	0.8163
LDM_{2022}	0.010130	0.035270	29.777	0.7973	0.8051	0.018430	0.050115	29.298	0.7282	0.7666
BBDM_{2023}	0.008989	0.025191	32.945	0.8901	0.9010	0.018261	0.051555	29.103	0.7327	0.7436
Fast-DDPM_{2025}	0.008863	0.030925	32.864	0.8781	0.8915	0.016922	0.046631	30.125	0.7459	0.8152
CoCoLIT_{2026}	0.009349	0.031342	31.652	0.8638	0.8437	0.016808	0.046610	30.280	0.7475	0.8193
Ours	**0.008274**	**0.024928**	**34.316**	**0.9023**	**0.9153**	**0.014231**	**0.040260**	**31.370**	**0.8075**	**0.8237**

4.2 EXPERIMENTAL RESULTS

We compare our method with representative adversarial, diffusion-based, and bridge-based approaches. As shown in Table 1, our method consistently achieves the best results across all five metrics on both datasets. On Lung-PET-CT-Dx, it reaches 34.316 dB PSNR, compared with 33.539 dB for MedGAN, while also improving SSIM from 0.8901 to 0.9023 and PCC from 0.9010 to 0.9153 over BBDM. The advantage is even more evident on RIDER-Lung-PET-CT, where our method improves PSNR from 30.355 to 31.370 dB and SSIM from 0.7669 to 0.8075, while reducing RMSE from 0.046212 to 0.040260.

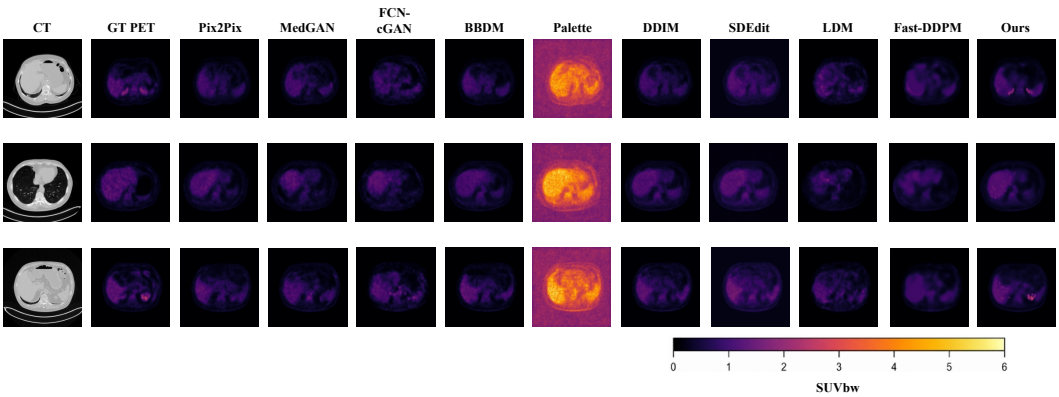


Figure 3: Qualitative comparison on the Lung-PET-CT-Dx dataset under a shared SUVbw scale. Our method better preserves localized target-specific signals while reducing diffuse responses in background regions.

The relative strength of competing methods also changes across datasets. MedGAN and BBDM are more competitive on Lung-PET-CT-Dx, whereas DDIM performs considerably better on RIDER-Lung-PET-CT. In contrast, our method maintains a consistent advantage in both settings. This result is in line with our design motivation: instead of regenerating the entire target image, the framework preserves a reliable anchor prediction and applies generative modeling only to the remaining residual. Residual Alignment further guides the recovery of missing target semantics, while the Activation Gate restricts refinement to spatially relevant regions, leading to more stable improvements across different data distributions.

Fig. 3 compares representative synthesis results under a shared SUVbw scale. Although several baselines reproduce the overall PET appearance, they tend to either attenuate localized target-specific signals or introduce spatially diffuse responses in background regions. Our method better preserves the location, spatial extent, and relative SUV contrast of these localized signals while maintaining stable predictions in surrounding regions. This qualitative behavior is consistent with the SUV-based quantitative results. Residual Alignment guides the recovery of missing target semantics, while the Activation Gate restricts these corrections to spatially relevant regions rather than broadly modifying the entire PET image.

4.3 EXTERNAL EVALUATION

Table 2 reports the external validation results on the Small Cell Carcinoma cohort from TCIA, which is used as an independent external test set. Our method achieves the best performance across all five evaluation metrics, with the lowest MAE of 0.0103 and RMSE of 0.0342, together with the highest PSNR of 31.241 dB, SSIM of 0.8813, and PCC of 0.8486. Compared with the strongest competing results, our method reduces MAE from 0.0106 to 0.0103 and RMSE from 0.0351 to 0.0342, while improving PSNR from 30.910 dB to 31.241 dB. It also consistently improves structural and correlation-based fidelity, surpassing the best baseline SSIM of 0.8776 and PCC of 0.8411. These results indicate that the proposed residual generative framework generalizes well to Small Cell Carcinoma, despite the distribution shift in histopathological subtype, and maintains stable reconstruction quality beyond the primary evaluation setting.

More importantly, this generalization behavior is consistent with our residual formulation. Rather than relearning the complete target image under a shifted distribution, the deterministic backbone provides a stable Generated Anchor PET, leaving the generative branch to model only the remaining discrepancy. Residual Alignment further constrains the missing target semantics, while the Activation Gate limits unnecessary spatial corrections. Such decomposition reduces the burden placed on the generative model and may contribute to the stable refinement observed under the distribution shift.

Table 2: External evaluation on the Small Cell Carcinoma cohort. ↓/↑ indicate lower/higher is better.

Method	MAE↓	RMSE↓	PSNR↑	SSIM↑	PCC↑
Pix2Pix_{2018}	<u>0.010577</u>	<u>0.035118</u>	30.842	0.8749	0.8382
MedGAN_{2020}	0.011149	0.036117	30.674	0.8682	0.8275
UP-GAN_{2021}	0.012666	0.037267	29.925	0.8588	0.8217
FCN-cGAN_{2023}	0.013566	0.038447	29.415	0.8382	0.7952
DDIM_{2020}	0.012662	0.035150	<u>30.910</u>	0.8213	<u>0.8411</u>
Palette_{2022}	0.477712	0.485581	6.338	0.0430	0.7710
SDEdit_{2022}	0.039493	0.049372	26.413	0.3553	0.8407
LDM_{2022}	0.012377	0.037208	30.134	0.8535	0.8158
BBDM_{2023}	0.010692	0.039545	30.580	<u>0.8776</u>	0.8200
Fast-DDPM_{2025}	0.125730	0.035750	29.963	0.8621	0.8291
CoCoLIT_{2026}	0.010856	0.037550	30.355	0.8749	0.8121
Ours	**0.010341**	**0.034231**	**31.241**	**0.8813**	**0.8486**

Table 3: SUV-based reconstruction errors in hypermetabolic regions on Lung-PET-CT-Dx. ↓ indicates lower is better.

Method	MAE@1↓	MAE@2.5↓	RMSE@1↓	RMSE@2.5↓
Pix2Pix_{2018}	0.8552	1.8562	1.1706	2.0788
MedGAN_{2020}	0.8719	1.8571	1.1823	2.0902
UP-GAN_{2021}	0.9322	1.9154	1.2347	2.1416
FCN-cGAN_{2023}	0.9101	1.7086	1.2080	2.0180
DDIM_{2020}	0.8564	1.8701	1.1738	2.0917
Palette_{2022}	4.8041	3.6717	4.9572	3.8472
SDEdit_{2022}	0.7204	<u>1.4618</u>	1.0048	<u>1.7359</u>
LDM_{2022}	0.9686	1.9206	1.2765	2.1750
BBDM_{2023}	0.9052	1.8395	1.2090	2.0935
Fast-DDPM_{2025}	0.8623	1.8623	1.0329	1.9324
CoCoLIT_{2026}	<u>0.6233</u>	2.4911	<u>0.9551</u>	3.0359
Ours	**0.4381**	**0.6379**	**0.5647**	**0.8908**

4.4 Semantic Refinement Analysis

Global reconstruction metrics provide only a coarse view of synthesis quality, as most pixels are already well explained by the deterministic prediction. The more challenging case lies in localized regions with stronger target responses, where the remaining cross-modal discrepancy is more pronounced. We therefore further evaluate reconstruction errors within regions defined by SUV > 1 and SUV > 2.5.

As shown in Table 3, the advantage of our method becomes more evident as the evaluation focuses on these harder regions. At SUV > 2.5, our method reduces MAE from 1.4618 to 0.6379 and RMSE from 1.7359 to 0.8908 compared with the strongest competing results. This larger margin suggests that the gain is not simply caused by improving already well-reconstructed regions. Instead, the residual branch is more effective at recovering target semantics that remain missing from the Anchor PET.

This behavior directly reflects our design. Residual Alignment guides the model toward *what* target semantics remain missing from the Anchor PET, while the Activation Gate controls *where* these corrections should be applied. Their complementary roles allow the residual branch to focus on difficult localized responses without unnecessarily modifying regions that are already well reconstructed.

4.5 Downstream Evaluation

Beyond reconstruction fidelity, we further examine whether the synthesized PET preserves representations that are useful for downstream prediction on the Lung-PET-CT-Dx dataset. PET generated by each synthesis method is evaluated under the same downstream setting, and the resulting accuracy is reported in Fig. 4.

Our method achieves the highest accuracy of 0.7857, exceeding BBDM (0.7421) and Fast-DDPM (0.7356) by a clear margin. More importantly, this improvement suggests that the gain is not limited to pixel-level fidelity. By recovering target semantics that remain missing from the generated anchor PET while restricting unnecessary spatial modifications, the proposed residual refinement preserves more task-relevant information for subsequent prediction.

This result provides a complementary view to the reconstruction metrics. Pixel-wise similarity alone does not determine whether the synthesized image retains information that is useful for a subsequent recognition model. The improved downstream accuracy therefore suggests that the residual corrections are not merely low-level modifications, but preserve target-domain cues that remain useful for discrimination.

Table 4: Ablation study of the proposed residual generative framework.

Res. SiT	Res. Align.	Act. Gate	Fusion	PSNR↑	SSIM↑	PCC↑
				30.125	0.8521	0.8762
✓				28.247	0.7935	0.7859
✓	✓			32.562	0.8742	0.8964
✓	✓	✓		33.684	0.8856	0.9058
✓	✓	✓	✓	**34.316**	**0.9023**	**0.9153**

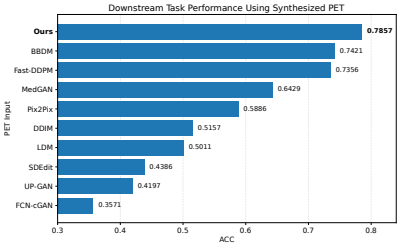


Figure 4: Downstream task accuracy using synthesized PET.

4.6 Ablation Study

We conduct controlled ablations to examine how residual generation, representation alignment, and spatial localization contribute to the final performance. As shown in Table 4, directly adding Residual SiT to the deterministic predictor decreases PSNR from 30.125 to 28.247 dB, with similar drops in SSIM and PCC. This result shows that additional generative capacity alone is not necessarily beneficial, as unconstrained residual generation can introduce unnecessary corrections to an already reliable Generated anchor PET.

Introducing Residual Alignment substantially reverses this degradation, improving PSNR to 32.562 dB. This suggests that guiding the SiT representation toward target semantics missing from the anchor prediction is critical for making the residual branch useful. Adding the Activation Gate further improves PSNR to 33.684 dB, indicating that semantic guidance alone is insufficient without controlling where residual refinement is applied. Finally, the complete model with gated residual fusion achieves the best performance, reaching 34.316 dB PSNR, 0.9023 SSIM, and 0.9153 PCC. Overall, the ablation results support our central design: residual generation provides additional capacity, while Residual Alignment determines what should be recovered and the Activation Gate constrains where the correction should be applied.

Fig. 5 qualitatively illustrates the progressive effect of each component. The Generated anchor PET preserves the coarse target structure but misses several localized responses. Directly adding Residual SiT introduces stronger corrections, yet these responses are still spatially diffuse. Residual Alignment further concentrates the correction on missing target semantics, while the Activation Gate suppresses irrelevant responses and restricts refinement to the appropriate regions. The final prediction therefore better matches the ground-truth PET in both response intensity and spatial localization.

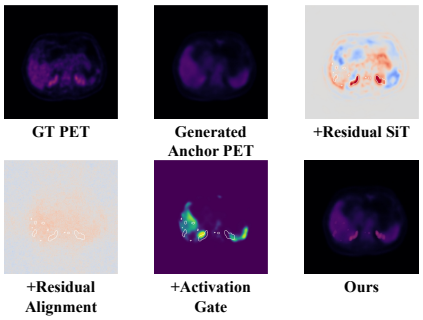


Figure 5: Qualitative ablation of residual refinement. Residual Alignment determines *what* to recover, while the Activation Gate controls *where* to refine.

5 Conclusion

We presented a controlled residual generative framework for CT-to-PET synthesis. Rather than regenerating the complete target image, the proposed method preserves a Generated anchor PET and uses Residual SiT to model only the remaining correction. Residual Alignment guides the recovery of target semantics that remain insufficiently captured by Generated anchor PET, while the Activation Gate controls where residual refinement should be applied. Experiments on Lung-PET-CT-Dx and RIDER-Lung-PET-CT demonstrate consistent improvements in reconstruction quality, together with stronger performance in localized high-response regions, downstream prediction, and external evaluation on Small Cell Carcinoma. These results demonstrate the effectiveness of selective residual refinement for exploiting generative modeling in cross-modal synthesis without regenerating the complete target image.

AI USE STATEMENT

We used AI-assisted tools during the preparation of this paper. Codex was used to assist in generating visualization code for the qualitative and ablation figures directly from our experimental checkpoints. Large language models were also used to support literature search and the organization of related work. In addition, ChatGPT was used for language polishing and improving the clarity of selected parts of the manuscript. All experimental results, analyses, scientific claims, and final manuscript content were reviewed and verified by the authors.

REFERENCES

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized Medical Imaging and Graphics*, 79:101684, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 9630–9640. IEEE, 2021.

Hang Chen, Qian Xiang, Jiaxin Hu, Meilin Ye, Chao Yu, Hao Cheng, and Lei Zhang. Comprehensive exploration of diffusion models in image generation: a survey. *Artificial Intelligence Review*, 58(4):99, 2025.

Pankaj Garg, Gargi Singhal, David Horne, Prakash Kulkarni, Ravi Salgia, and Sharad S Singhal. Molecular pet imaging: Unlocking the secrets of cancer metabolism. *Biochemical Pharmacology*, pp. 117202, 2025.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4409–4437, 2025.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976. IEEE, 2017.

Hongxu Jiang, Muhammad Imran, Teng Zhang, Yuyin Zhou, Muxuan Liang, Kuang Gong, and Wei Shao. Fast-ddpm: Fast denoising diffusion probabilistic models for medical image-to-image generation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Benjamin L Kidder. Advanced image generation for cancer using diffusion models. *Biology Methods and Protocols*, 9(1):62, 2024.

Geonung Kim, Beomsu Kim, Eunhyeok Park, and Sunghyun Cho. Diffusion model compression for image-to-image translation. In *Asian Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471, 2024.

Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1952–1961. IEEE, 2023a.

P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang. A large-scale ct and pet/ct dataset for lung cancer diagnosis (lung-pet-ct-dx), 2020.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023b.

Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. Representation potentials of foundation models for multimodal alignment: A survey. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 16680–16695, 2025.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution, and everything: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):11793–11813, 2024.

P. Muzi, M. Wanner, and P. Kinahan. Data from rider lung pet-ct, 2015.

Dac Thai Nguyen, Trung Thanh Nguyen, Huu Tien Nguyen, Thanh Trung Nguyen, Huy Hieu Pham, Thanh Hung Nguyen, Thao Nguyen Truong, and Phi Le Nguyen. Ct to pet translation: a large-scale dataset and domain-knowledge-guided diffusion approach. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1498–1507. IEEE, 2025.

Dong Nie, Roger Trullo, Jun Lian, Li Wang, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12):2720–2730, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Shaoyan Pan, Tonghe Wang, Richard LJ Qiu, Marian Axente, Chih-Wei Chang, Junbo Peng, Ashish B Patel, Joseph Shelton, Sagar A Patel, Justin Roper, et al. 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine & Biology*, 68(10):105004, 2023.

Narendra Rathod, Warissara Jutidamrongphan, Wolfram Andreas Bosbach, Yizhou Chen, Jan Luca Penner, Hasan Sari, Konstantinos Zeimpekis, Alejandro López Montes, Pawel Moskal, Ewa Stepien, et al. Total body pet/ct: clinical value and future aspects of quantification in static and dynamic imaging. In *Seminars in Nuclear Medicine*, volume 55, pp. 98–106. Elsevier, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVFConferencee on Computer Vision and Pattern Recognition*, pp. 10674–10685. IEEE, 2022.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH Conference Proceedings*, pp. 1–10, 2022.

Morteza Salehjahromi, Tatiana V Karpinets, Sheeba J Sujit, Mohamed Qayati, Pingjun Chen, Muhammad Aminu, Maliazurina B Saad, Rukhmini Bandyopadhyay, Lingzhi Hong, Ajay Sheshadri, et al. Synthetic pet from ct improves diagnosis and prognosis for lung cancer: Proof of concept. *Cell Reports Medicine*, 5(3), 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Shervin Zoghi, Clemens Mingels, Ramsey D Badawi, Benjamin A Spencer, Tracy L Yarbrough, Lorenzo Nardo, and Abhijit J Chaudhari. Role of total body pet/ct in inflammatory disorders. In *Seminars in nuclear medicine*, volume 55, pp. 41–51. Elsevier, 2025.

A Dataset and Preprocessing Details

We conduct experiments on two publicly available CT–PET datasets from The Cancer Imaging Archive (TCIA): Lung-PET-CT-Dx and RIDER-Lung-PET-CT. The two datasets are combined into a unified cohort and split at the patient level into training, validation, and test sets to avoid information leakage across different subsets. Let a paired CT–PET sample be denoted as (X, Y), where X is the CT image and Y is the corresponding quantitative PET image.

CT intensity preprocessing. CT intensities are represented in Hounsfield units (HU). The raw CT image X_{raw} is first clipped to $[-1000, 400]$ HU:

$$X_{\mathrm{clip}} = \mathrm{clip}\left(X_{\mathrm{raw}}, -1000, 400\right). \tag{13}$$

The clipped CT image is then linearly normalized to $[0, 1]$:

$$X = \frac{X_{\mathrm{clip}} + 1000}{1400}. \tag{14}$$

This fixed transformation provides a consistent intensity scale across subjects and avoids introducing patient-specific normalization statistics.

Standardized uptake value. PET intensity is represented using body-weight-normalized standardized uptake value (SUVbw), defined as:

$$\mathrm{SUV}_{\mathrm{bw}}(p) = \frac{C_{\mathrm{PET}}(p)}{A_{\mathrm{inj}}/W}, \tag{15}$$

where $C_{\mathrm{PET}}(p)$ denotes the decay-corrected radiotracer activity concentration at spatial location p, A_{inj} is the injected activity, and W denotes body weight. SUVbw provides a quantitative scale for comparing PET intensity across spatial locations and subjects.

Quantitative PET preprocessing. To preserve a consistent quantitative scale across subjects, we apply a fixed global SUV range rather than patient-specific or slice-wise normalization. Given the raw SUVbw image Y_{raw}, we first clip its values to $[0, 10]$:

$$Y_{\mathrm{clip}} = \mathrm{clip}\left(Y_{\mathrm{raw}}, 0, 10\right), \tag{16}$$

and subsequently normalize it as:

$$Y = \frac{Y_{\mathrm{clip}}}{10}. \tag{17}$$

The inverse transformation used for quantitative evaluation is

$$Y_{\mathrm{SUV}} = 10Y. \tag{18}$$

This global normalization preserves the relative SUV scale across subjects and enables consistent evaluation in the original quantitative PET domain. We do not apply patient-specific percentile normalization, slice-wise min-max normalization, or 8-bit intensity conversion, since these operations may map identical SUV values to different normalized intensities across subjects.

CT–PET spatial alignment. Because CT and PET are acquired at different spatial resolutions, the PET volume is resampled to the geometry of the corresponding CT volume before paired training. Let $\mathcal{R}(\cdot; \mathcal{G}_X)$ denote resampling to the CT spatial grid $\mathcal{G}_X$. The aligned PET image is written as:

$$Y_{\mathrm{align}} = \mathcal{R}\left(Y; \mathcal{G}_X\right). \tag{19}$$

After alignment, corresponding CT and PET slices share the same spatial coordinate system, ensuring that pixel-wise reconstruction and spatial gating operate on corresponding locations.

Spatial resizing. The aligned CT–PET pairs are resized to a common spatial resolution of 256×256:

$$\tilde{X} = \mathcal{S}(X), \qquad \tilde{Y} = \mathcal{S}(Y_{\mathrm{align}}), \tag{20}$$

where $\mathcal{S}(\cdot)$ denotes the resizing operator. The same spatial transformation is applied to both modalities to preserve their correspondence throughout training and evaluation.

Region masks for quantitative evaluation. For the region-level reconstruction analysis, evaluation regions are defined directly in the original SUV domain. Given an SUV threshold τ, we construct the binary reference mask

$$M_\tau(p) = \mathbb{I}\left[Y_{\mathrm{SUV}}(p) > \tau\right], \tag{21}$$

where p indexes spatial locations and $\mathbb{I}[\cdot]$ denotes the indicator function. We use $\tau = 1$ and $\tau = 2.5$ for the reported intensity-based evaluations. Reconstruction errors are then computed only over pixels satisfying $M_\tau(p) = 1$.

The corresponding MAE is

$$\mathrm{MAE}_\tau = \frac{\sum_p M_\tau(p) \left| \hat{Y}_{\mathrm{SUV}}(p) - Y_{\mathrm{SUV}}(p) \right|}{\sum_p M_\tau(p)}, \tag{22}$$

and the RMSE is

$$\mathrm{RMSE}_\tau = \sqrt{\frac{\sum_p M_\tau(p) \left(\hat{Y}_{\mathrm{SUV}}(p) - Y_{\mathrm{SUV}}(p) \right)^2}{\sum_p M_\tau(p)}}. \tag{23}$$

This evaluation focuses on localized target responses that may contribute only weakly to global image-level metrics.

B Implementation Details

Generated anchor PET. We employ a ResUNet as the deterministic CT-to-PET backbone. The backbone is first trained independently to predict PET from CT and subsequently frozen during residual generative training. Given a preprocessed CT image X, the deterministic predictor produces Generated anchor PET,

$$Y_{\mathrm{anchor}} = f_{\mathrm{anchor}}(X), \tag{24}$$

which serves as a fixed anchor for subsequent residual refinement. Freezing the deterministic backbone ensures that the residual generator is optimized relative to a stable target-domain estimate rather than a continuously changing prediction.

Residual SiT. Residual refinement is modeled using a conditional Scalable Interpolant Transformer (SiT). Instead of synthesizing the complete PET image, the SiT learns the correction relative to the anchor PET. At time t, the generator receives the stochastic state x_t together with CT and the anchor PET,

$$\Delta Y = f_\theta \left(x_t, t; X, Y_{\mathrm{anchor}} \right), \tag{25}$$

where ΔY denotes the generated residual. CT provides source-domain structural conditioning, while Generated anchor PET provides the current target-domain estimate, allowing the generative branch to focus on information that remains insufficiently captured by Generated anchor PET.

Residual Alignment supervision. Residual Alignment is implemented as an auxiliary training pathway and does not modify the forward input of Residual SiT. Ground-truth PET and Anchor PET are separately encoded by the frozen PET-DINO encoder to obtain a semantic residual target. An intermediate SiT feature h is then aligned with this target through a lightweight projection head:

$$\mathcal{L}_{\mathrm{rep}} = \left\| P_\psi(h) - \mathrm{sg}\left(P_\psi(z_{\mathrm{res}}) \right) \right\|_2^2. \tag{26}$$

The corresponding gradients are propagated through P_ψ and the preceding SiT blocks, while the PET-DINO encoder remains frozen. Thus, Residual Alignment acts as representation-level supervision rather than an additional conditioning input. At inference, the complete representation-alignment pathway is removed.

Activation Gate. The Activation Gate jointly takes CT and anchor PET as input and predicts a single-channel spatial modulation map,

$$G = \sigma \left(g_\phi \left(X, Y_{\mathrm{anchor}} \right) \right), \qquad G \in [0,1]^{H \times W}. \tag{27}$$

The gate is applied directly to the generated residual, and the final PET prediction is obtained as:

$$\hat{Y} = Y_{\mathrm{anchor}} + G \odot \Delta Y. \tag{28}$$

This formulation preserves anchor PET in regions with weak gate responses while allowing stronger residual refinement at spatial locations requiring additional correction.

Optimization. During residual generative training, the deterministic ResUNet and PET-DINO encoder remain frozen. Residual SiT, the Residual Alignment projection head, and the Activation Gate are optimized jointly. The overall training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda_{\text{rep}}\mathcal{L}_{\text{rep}}, \tag{29}$$

where $\mathcal{L}_{\text{gen}}$ denotes the primary generative objective and λ_{rep} controls the contribution of residual representation alignment.

Inference. Inference requires only the input CT image. The frozen ResUNet first produces the anchor PET, after which Residual SiT generates the residual correction and the Activation Gate predicts its spatial modulation. The gated residual is then added to the Anchor PET to obtain the final PET prediction. Since Residual Alignment is used only as auxiliary supervision during training, neither ground-truth PET nor the PET-DINO encoder is required at test time.

Training configuration. All experiments are conducted on four NVIDIA GeForce RTX 3090 GPUs using distributed FP16 training. The model is trained for 3,000 optimization steps with an effective batch size of 32, and an EMA model with a decay rate of 0.995 is maintained throughout training. Residual Alignment is applied from SiT Block 4 onward, and validation uses a 10-step Heun solver. The remaining implementation details are summarized in Table 5.

Table 5: Implementation details.

Configuration	**Setting**
Hardware	4× NVIDIA GeForce RTX 3090
Input resolution	256×256
Residual generator	ResidualSiT-S/8
Batch size	8 per GPU / 32 total
Training steps	3,000
Mixed precision	FP16
Weight decay	1×10^{-4}
Max gradient norm	1.0
EMA decay	0.995
Alignment block	4th SiT block
Alignment coefficient	0.50
Alignment weights	$0.45/0.45/0.10$
Residual scale	0.25
Gate-only stage	500 steps
Representation adaptation	1,000 steps
Gate Stage-A LR	1×10^{-4}
Flow Stage-B LR	2×10^{-6}
Gate Stage-B LR	2×10^{-5}
Sampling solver	Heun
Sampling steps	10
Validation interval	250 steps

C Additional Experimental Results

C.1 Additional Qualitative Comparison

We provide additional qualitative comparisons on the Lung-PET-CT-Dx and RIDER-Lung-PET-CT test sets to further examine the reconstruction behavior of different CT-to-PET synthesis methods. All predictions are visualized under the same PET display setting and compared with the corresponding ground-truth images.

Fig. 6 presents a representative case from Lung-PET-CT-Dx under a shared SUVbw scale. The common quantitative range enables direct comparison of predicted intensity patterns across methods. Our method better preserves localized target responses and their relative contrast, whereas several

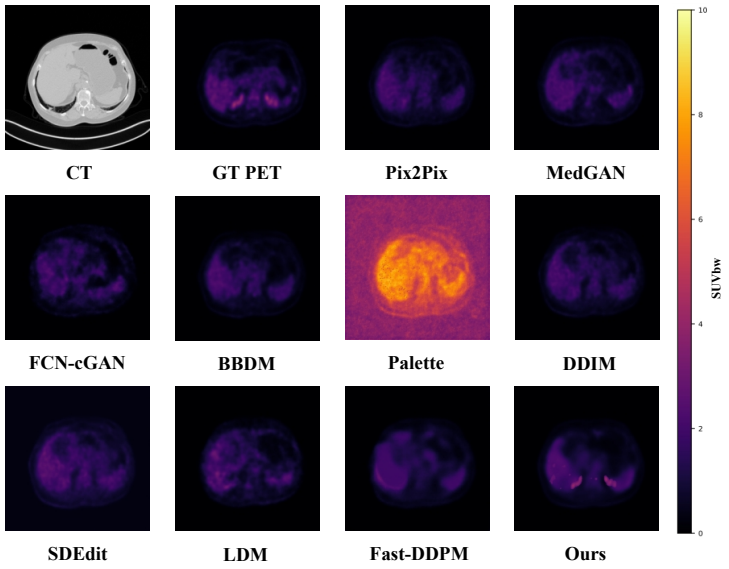


Figure 6: Additional qualitative comparison on Lung-PET-CT-Dx under a shared SUVbw scale.

baselines either attenuate these responses or produce overly diffuse predictions. This behavior is consistent with the SUV-based quantitative results reported in the main paper.

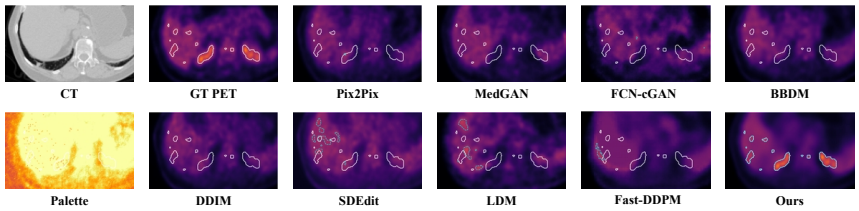


Figure 7: Additional qualitative comparisons on representative Lung-PET-CT-Dx test cases.

As shown in Fig. 7, several competing methods reproduce the overall PET appearance but exhibit different local reconstruction errors. Some methods weaken localized target responses, while others introduce spatially diffuse predictions that are weakly supported by the ground-truth PET. In contrast, our method better preserves the location and spatial extent of these localized responses while maintaining cleaner surrounding regions.

Fig. 8 provides an additional qualitative comparison on RIDER-Lung-PET-CT under a shared SUVbw scale. Although most competing methods recover the coarse PET appearance, noticeable differences remain in the reconstruction of localized target responses. Pix2Pix, MedGAN, BBDM, and DDIM tend to produce smoother predictions with attenuated local variations, while LDM and Fast-DDPM recover stronger responses but also introduce relatively diffuse structures. Palette exhibits a substantial intensity mismatch under the shared quantitative scale. In contrast, our method better preserves the localized high-response patterns observed in the ground-truth PET while maintaining relatively stable surrounding regions. This behavior suggests that the proposed residual formulation does not simply increase global response intensity, but instead concentrates the generative correction on information that remains insufficiently represented by the anchor PET.

Fig. 9 further compares local reconstruction behavior on RIDER-Lung-PET-CT. Several baselines exhibit attenuated responses or mismatched spatial extent in the highlighted regions, while our prediction more closely follows the ground-truth location and shape. This is consistent with our design, where Residual Alignment guides missing target semantics and the Activation Gate restricts refinement to relevant spatial regions.

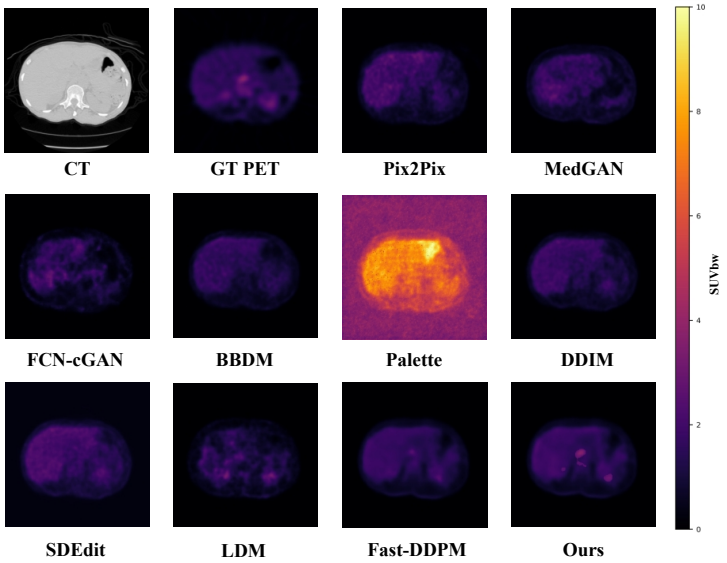


Figure 8: Additional qualitative comparison on RIDER-Lung-PET-CT under a shared SUVbw scale.

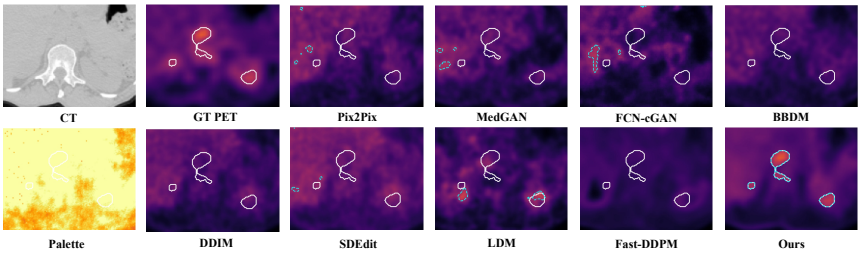


Figure 9: Additional qualitative comparisons on representative RIDER-Lung-PET-CT test cases.

C.2 Additional Spatial Agreement Analysis

Table 6: Region-level spatial agreement on Lung-PET-CT-Dx and RIDER-Lung-PET-CT at SUV $>$ 1. $\uparrow$ indicates higher is better. Best results are shown in **bold**, and second-best results are underlined.

	Lung-PET-CT-Dx			RIDER-Lung-PET-CT		
Method	**Dice↑**	**Precision↑**	**Recall↑**	**Dice↑**	**Precision↑**	**Recall↑**
Pix2Pix_{2018}	0.7121	0.7148	0.7094	0.5889	0.6510	0.5376
MedGAN_{2020}	0.6894	0.7130	0.6673	0.5233	0.5982	0.4651
UP-GAN_{2021}	0.6569	0.6370	0.6781	0.5113	0.5200	0.5030
FCN-cGAN_{2023}	0.6332	0.6214	0.6455	0.5130	0.5786	0.4607
DDIM_{2020}	0.7093	0.7263	0.6931	0.5997	0.6607	0.5490
Palette_{2022}	0.1559	0.1496	0.1627	0.1713	0.0937	0.1623
SDEdit_{2022}	0.5743	0.5475	0.6039	0.6026	0.4808	**0.8070**
LDM_{2022}	0.6669	0.6756	0.6584	0.5180	0.6381	0.4359
BBDM_{2023}	0.7149	0.7175	0.7123	0.5412	**0.7245**	0.4319
Fast-DDPM_{2025}	0.7087	0.7225	0.6954	0.6066	0.6560	0.5642
CoCoLIT_{2026}	0.6220	0.6376	0.6377	0.6060	0.6573	0.5622
Ours	**0.7323**	**0.7357**	**0.7289**	**0.6287**	0.6795	0.6983

To further evaluate whether the synthesized PET preserves the spatial extent of localized target responses, we measure region-level overlap at SUV $>$ 1 on both Lung-PET-CT-Dx and RIDER-Lung-

PET-CT. Specifically, Dice, precision, and recall are computed between the thresholded synthesized PET and the corresponding ground-truth PET.

As shown in Table 6, our method achieves the best Dice score on both datasets, reaching 0.7323 on Lung-PET-CT-Dx and 0.6287 on RIDER-Lung-PET-CT. On Lung-PET-CT-Dx, our method also obtains the highest precision and recall, indicating that the synthesized responses are both spatially accurate and well covered. On RIDER-Lung-PET-CT, although BBDM achieves higher precision and SDEdit achieves higher recall, our method provides the best overall Dice score, suggesting a better balance between localization accuracy and spatial coverage.

These results complement the SUV-based intensity evaluation in the main paper. They indicate that the proposed residual refinement does not simply increase response magnitude, but also improves where the recovered target responses are placed. This observation is consistent with the role of the Activation Gate, which restricts residual updates to spatial regions requiring additional refinement.

C.3 Additional Downstream Evaluation

To evaluate whether the synthesized PET images preserve discriminative information useful for subsequent prediction, we further conduct a downstream classification experiment on the Lung-PET-CT-Dx dataset. For each CT-to-PET synthesis method, the corresponding synthesized PET images are used as input to the same ResNet-50 classifier. The classifier architecture, data split, training protocol, and evaluation procedure are kept identical across all methods, such that the only difference is the source of the synthesized PET input.

Table 7: Additional downstream classification results on Lung-PET-CT-Dx using synthesized PET as input. ↑ indicates higher is better.

PET Input	AUROC↑	AUPRC↑	ACC↑	F1↑
Pix2Pix	0.6042	0.6716	0.5886	0.5021
MedGAN	0.5208	0.5919	0.6429	0.6154
UP-GAN	0.5875	0.6548	0.4197	0.6012
FCN-cGAN	0.5417	0.6483	0.3571	0.5263
SDEdit	0.3842	0.3966	0.4386	0.3422
DDIM	0.4367	0.5295	0.5157	0.5413
LDM	0.5833	0.5093	0.5011	0.6316
BBDM	<u>0.6878</u>	<u>0.7563</u>	<u>0.7421</u>	<u>0.6813</u>
Fast-DDPM	0.6792	0.7412	0.7356	0.6738
Ours	**0.7392**	**0.7784**	**0.7857**	**0.7373**

We evaluate classification performance using AUROC, AUPRC, accuracy (ACC), and F1 score. As shown in Table 7, our synthesized PET achieves the best performance across all four metrics on Lung-PET-CT-Dx, reaching 0.7857 ACC and 0.7373 F1. Compared with BBDM, the strongest competing method, our approach improves AUROC from 0.6878 to 0.7392 and ACC from 0.7421 to 0.7857, while also achieving higher AUPRC and F1.

These consistent gains suggest that the benefit of our method extends beyond pixel-level reconstruction. The selective residual refinement better preserves task-relevant target semantics, resulting in synthesized PET representations that are more useful for downstream classification.

C.4 Residual Alignment at Different SiT Layers

We further investigate where Residual Alignment should be introduced within Residual SiT by varying the starting block of representation supervision. As shown in Table 8, the choice of alignment depth has a clear impact on reconstruction quality. Applying Residual Alignment from Block 4 to Block 7 achieves the best performance across all three metrics, reaching 34.316 dB PSNR, 0.9023 SSIM, and 0.9153 PCC.

An interesting trend is that introducing alignment too early does not necessarily improve performance. Although early SiT blocks preserve more low-level information, their representations may not yet contain sufficiently developed target-domain semantics. Applying semantic supervision at

Table 8: Effect of applying Residual Alignment at different SiT blocks. ↑ indicates higher is better.

Alignment Layer	PSNR↑	SSIM↑	PCC↑
Block 0–7	34.010	0.8967	0.9086
Block 1–7	34.006	0.8893	0.9015
Block 2–7	33.986	0.8875	0.9003
Block 3–7	33.965	0.8866	0.9000
Block 4–7 (Ours)	**34.316**	**0.9023**	**0.9153**
Block 5–7	33.933	0.8921	0.9056
Block 6–7	33.927	0.8893	0.9082
Block 7	33.929	0.8865	0.9038

these stages can therefore over-constrain features that are still responsible for building the basic generative representation. This is reflected by the lower performance of Block 0–7 to Block 3–7 compared with our setting.

Conversely, applying Residual Alignment only to later blocks also leads to a performance drop. At this stage, the representation is already closer to the final residual prediction, leaving fewer subsequent transformations to absorb and refine the aligned semantic information. Block 4–7 provides a favorable balance between these two cases: the intermediate representation has developed sufficient target semantics for meaningful alignment, while the remaining SiT blocks still retain enough capacity to transform the aligned representation into effective residual corrections.

Overall, these results suggest that Residual Alignment is most effective at an intermediate semantic stage rather than at the earliest or latest layers. This observation is consistent with the role of our alignment objective, which is designed to guide the recovery of missing target semantics without overly restricting the underlying generative dynamics.