CoCoLIT: ControlNet-Conditioned Latent Image Translation for MRI to Amyloid PET Synthesis

Alec Sargood[1] *, Lemuel Puglisi[2] *, James H. Cole[1], Neil P. Oxtoby[1], Daniele Ravì[3] †, Daniel C. Alexander[1] †

[1]University College London
[2]University of Catania
[3]University of Messina

Abstract

Synthesizing amyloid PET scans from the more widely available and accessible structural MRI modality offers a promising, cost-effective approach for large-scale Alzheimer's Disease (AD) screening. This is motivated by evidence that, while MRI does not directly detect amyloid pathology, it may nonetheless encode information correlated with amyloid deposition that can be uncovered through advanced modeling. However, the high dimensionality and structural complexity of 3D neuroimaging data pose significant challenges for existing MRI-to-PET translation methods. Modeling the cross-modality relationship in a lower-dimensional latent space can simplify the learning task and enable more effective translation. As such, we present CoCoLIT (ControlNet-Conditioned Latent Image Translation), a diffusion-based latent generative framework that incorporates three main innovations: (1) a novel Weighted Image Space Loss (WISL) that improves latent representation learning and synthesis quality; (2) a theoretical and empirical analysis of Latent Average Stabilization (LAS), an existing technique used in similar generative models to enhance inference consistency; and (3) the introduction of ControlNet-based conditioning for MRI-to-PET translation. We evaluate CoCoLIT's performance on publicly available datasets and find that our model significantly outperforms state-of-the-art methods on both image-based and amyloid-related metrics. Notably, in amyloid-positivity classification, CoCoLIT outperforms the second-best method with improvements of +10.5% on the internal dataset and +23.7% on the external dataset.

Code — https://github.com/brAIn-science/CoCoLIT
Extended version — https://arxiv.org/abs/2508.01292

1 Introduction

Alzheimer's Disease (AD) places a substantial burden on patients, their families, and healthcare systems globally. As the population continues to age, both the human and economic costs associated with AD are steadily increasing (Tay et al. 2024). Early and accurate diagnosis is critical for effective intervention in AD. Among the available neuroimaging techniques, amyloid (Aβ) Positron Emission Tomography (PET) plays a key role by detecting Aβ plaque accumulation—an early hallmark of AD—often years before cognitive symptoms appear (Nordberg 2004). This makes Aβ PET a vital tool for both research and clinical applications where early and reliable diagnosis is essential. However, the high cost and limited availability of Aβ PET (Lee et al. 2021), as well as the radiation exposure, hinder its widespread use as a routine diagnostic tool. In contrast, structural Magnetic Resonance Imaging (MRI) is a more affordable, non-invasive, and widely used modality; however, it is less effective for early AD diagnosis, as it is not designed to highlight Aβ plaques. Despite this limitation, MRI can still capture hidden information related to Aβ pathology (Kerbler et al. 2015). While not a direct replacement for the biochemical accuracy of PET imaging, synthesizing Aβ PET scans from structural MRI is a promising method for enabling large-scale, cost-effective AD screening, especially in resource-limited or low-income countries (Chapleau et al. 2022).

Extensive research has explored translating structural MRI data into PET images, with many approaches leveraging Generative Adversarial Networks (GANs) for their ability to synthesize realistic outputs. Notably, in (Pan et al. 2018), the authors propose 3D-cGAN by extending the CycleGAN architecture (Zhu et al. 2017) to 3D, enabling unpaired PET synthesis from MRI. Similarly, in (Shin et al. 2020), the authors build on the well-known pix2pix framework (Isola et al. 2017) for paired MRI-to-PET translation. While GAN-based methods have shown promising results in generating visually plausible images, they remain prone to training instabilities and mode collapse. Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) have recently emerged as powerful generative models capable of synthesizing high-fidelity, diverse images through a learned denoising process. Consequently, they have been adopted in State-of-the-Art (SOTA) methods for MRI-to-PET synthesis. For instance, FICD (Yu et al. 2024) employs a conditional diffusion model that integrates an additional imaging constraint during training to enhance the fidelity and clinical relevance of the generated PET images. However, performing the diffusion process in the 3D image space imposes significant computational demands. Another recent method, PASTA (Li et al. 2024), introduces a pathology-aware conditional diffusion model with an additional cycle exchange consistency loss. While PASTA ad-

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
* These authors contributed equally as joint first authors.
†These authors contributed equally as joint senior authors.

dresses the challenges of 3D data by operating on sets of 2D slices, this limits the model's ability to fully capture inter-slice dependencies. To mitigate the challenges of image-space modeling, the authors of (Ou et al. 2024) propose IL-CLDM, a diffusion-based MRI-to-PET translation model that operates in a learned latent space. During training, the model is conditioned on the Aβ-positivity label by adding a learned label embedding to the time-step embedding. However, at inference time, the label is unavailable, and only the time-step embedding is used. This mismatch between training and inference conditions may lead to out-of-distribution behavior during generation.

Contributions To address the limitations of prior work, we present CoCoLIT (**Co**ntrolNet-**Co**nditioned **L**atent **I**mage **T**ranslation), a diffusion-based model for conditional 3D medical image synthesis, focused on MRI-to-PET translation. CoCoLIT builds on recent advances in generative modeling, including latent diffusion (Rombach et al. 2022) and ControlNets (Zhang, Rao, and Agrawala 2023). Our key contributions are threefold: **(1)** we introduce a novel Weighted Image Space Loss (WISL), which improves latent representation learning and enhances the fidelity of synthesized images; **(2)** we provide the first formal justification and empirical evaluation of Latent Average Stabilization (LAS)—originally proposed in (Puglisi, Alexander, and Ravì 2025)—showing that while LAS is asymptotically biased, its bias becomes negligible in sufficiently well-trained models; and **(3)** we are the first to successfully apply a ControlNet-based model to the task of MRI-to-PET translation. Our results show that CoCoLIT achieves SOTA performance on both image-based and clinical metrics, significantly outperforming existing methods on internal and external test sets.

2 Preliminaries

In this section, we introduce the background on which CoCoLIT is built, including Latent Diffusion Models (LDMs), the ControlNet conditioning mechanism, and the LAS technique.

2.1 Latent Diffusion Models

An LDM (Rombach et al. 2022) is a deep generative model used to learn a target data distribution in a compressed latent space, comprising a forward and reverse Markovian diffusion process. Input images, x, are first encoded into a latent representation z using an encoder $\mathcal{E}$. Gaussian noise is incrementally added to the latent vector in the forward process over T steps, starting from $z_0 = z$. At each step t, noise is added to z_{t-1} by sampling from the Gaussian transition probability $q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}\, z_{t-1}, \beta_t I)$, where β_t follows a predefined variance schedule. This ensures that z_T asymptotically approaches pure Gaussian noise. The reverse process aims to revert each diffusion step, allowing the generation of a latent embedding from the target distribution starting from pure noise z_T. The reverse transition probability has a Gaussian closed form, $q(z_{t-1}|z_t, z_0) = \mathcal{N}(z_{t-1}|\tilde{\mu}(z_0, z_t), \tilde{\beta}_t)$, conditioned on the ground-truth latent z_0. The mean $\tilde{\mu}(z_0, z_t)$ can be reparameterized in terms of z_t and a noise term ϵ. Therefore, a neural network, $\epsilon_\theta(z_t, t)$, is trained to predict the noise by optimizing the following objective (Ho, Jain, and Abbeel 2020):

$$\mathcal{L}_{LDM} := \mathbb{E}_{t, z_t, \epsilon\sim\mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|^2 \right]. \quad (1)$$

For ϵ_θ, we employ a U-Net-based denoising network, which allows the use of a ControlNet mechanism (Zhang, Rao, and Agrawala 2023) for conditional generation, as described in the next section.

2.2 Conditioning ControlNet

A ControlNet mechanism (Zhang, Rao, and Agrawala 2023) enables a pre-trained diffusion model to be conditioned on an additional signal by injecting information into the intermediate layers of its U-Net. Specifically, each U-Net encoder layer $\mathcal{F}(\cdot\,;\Theta)$ is kept frozen, and a trainable copy $\mathcal{F}(\cdot\,;\Theta_c)$ is introduced to learn the conditioning signal. The outputs of this trainable copy are integrated back into the corresponding frozen block via zero-initialized convolutional layers, denoted as $\mathcal{Z}$. These zero-convolution layers ensure that at the start of training, the ControlNet does not alter the original model's behavior. As training progresses, the parameters of $\mathcal{Z}$ adapt to effectively inject the conditioning information. Formally, for a given layer with input v, frozen layer output $w = \mathcal{F}(v;\Theta)$, and conditioning signal z_c, the modified output w_{CN} is computed as:

$$w_{CN} = w + \mathcal{Z}\big(\mathcal{F}(v + \mathcal{Z}(z_c;\Theta_{z1});\Theta_c);\Theta_{z2}\big), \quad (2)$$

The entire denoising network is denoted as $\epsilon_{\theta,\phi}(z_t, t; z_c)$, where θ are the frozen U-Net weights and $\phi = \{\Theta_c, \Theta_{z1}, \Theta_{z2}\}$ is the set of learnable ControlNet parameters. The ControlNet is trained by minimizing the standard diffusion loss:

$$\mathcal{L}_{CN} := \mathbb{E}_{t, z_t, \epsilon\sim\mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_{\theta,\phi}(z_t, t; z_c)\|^2 \right]. \quad (3)$$

2.3 Latent Average Stabilization

A latent conditional generative model aims to learn a conditional distribution $p(z^{(y)}|z^{(x)})$, where $z^{(x)}$ and $z^{(y)}$ are latent variables from paired input and output images x and y, respectively. These latents are computed via encoders, $z^{(x)} = \mathcal{E}^{(x)}(x)$ and $z^{(y)} = \mathcal{E}^{(y)}(y)$, and are recovered back to image space via decoders $x = \mathcal{D}^{(x)}(z^{(x)})$ and $y = \mathcal{D}^{(y)}(z^{(y)})$. Generation is performed by sampling from the learned distribution $p(z^{(y)}|z^{(x)})$, and decoding the samples via $\mathcal{D}^{(y)}$. As such, the process of inferring y is inherently stochastic, with randomness introduced by the sampling procedure. Therefore, for a given input x, we aim to compute the expectation over the decoded samples. This expectation is estimated using the sample mean, an unbiased estimator computed over N samples:

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{N} \mathcal{D}^{(y)} \left(z^{(y,j)} \right),$$

where each $z^{(y,j)}$ is a sample from $p(z^{(y)}|z^{(x)})$. A practical drawback of this method is its computational cost, requiring N forward passes through the decoder. To resolve

this issue, in (Puglisi, Alexander, and Ravì 2024, 2025), the authors propose LAS, which involves taking m samples from the learned latent distribution, $z^{(y,1)}, \ldots, z^{(y,m)} \sim p(z^{(y)}|z^{(x)})$, and decoding their sample mean, $\bar{z}^{(y)}$:

$$\hat{y} = \mathcal{D}^{(y)}\left(\bar{z}^{(y)}\right), \quad \text{for} \quad \bar{z}^{(y)} = \frac{1}{m}\sum_{j=1}^{m} z^{(y,j)},$$

requiring only one forward pass of the decoder. It is shown in (Puglisi, Alexander, and Ravì 2025) that LAS, when applied to a spatiotemporal disease progression modeling task, substantially improves results across a wide range of metrics. As the authors in (Puglisi, Alexander, and Ravì 2025) do not examine its statistical properties, we aim to fill this gap by providing a theoretical analysis of LAS and justifying its use as a reliable estimator in conditional generative tasks.

3 Methods

In this section, we describe the CoCoLIT framework, outlining the staged training process, our novel loss term (WISL), and a theoretical analysis of LAS.

3.1 Proposed Framework: CoCoLIT

The overall pipeline of CoCoLIT, illustrated in Figure 1, comprises five main blocks: (A–B) independent VAEs for MRI and PET representation learning, (C) an LDM for modeling latent PET distributions, (D) a ControlNet for conditional generation, and (E) the inference process incorporating LAS. Blocks A–D correspond to the training stages, while block E details inference. For further clarity, the inference process is also schematically described in Algorithm 1.

Representation Learning Stage This stage involves independently training two VAEs with the same architecture to encode and reconstruct 3D brain images. The MRI VAE (block A) encodes an input MRI volume $x \in \mathbb{R}^D$ into a latent representation $z^{(x)} \in \mathbb{R}^d$ using an encoder $\mathcal{E}^{(x)}$, and reconstructs the original image through a decoder $\mathcal{D}^{(x)}$. Similarly, the PET VAE (block B) maps an Aβ PET scan $y \in \mathbb{R}^D$ into a latent code $z^{(y)} \in \mathbb{R}^d$ via encoder $\mathcal{E}^{(y)}$, and reconstructs it using decoder $\mathcal{D}^{(y)}$. Both VAEs are trained using a composite loss function, $\mathcal{L}_{\text{VAE}}$, which includes reconstruction, perceptual and adversarial losses, and a Kullback-Leibler regularization term, following the formulation in (Guo et al. 2025).

Conditional Generative Modeling Stage The second stage starts by training an LDM (block C), which learns an unconditional distribution over the latents, $z^{(y)}$ (see Section 2.1). The final component (block D) is a ControlNet module, denoted as $\epsilon_{\theta,\phi}$, which operates on top of the trained and frozen LDM backbone (see Section 2.2), thereby learning a conditional distribution, $p(z^{(y)}|z^{(x)})$. To improve image synthesis quality, we propose to incorporate image-space guidance by adding a loss term, which we call WISL, defined as:

$$\mathcal{L}_{WISL} := \mathbb{E}_{t, z_t^{(y)}, \epsilon\sim\mathcal{N}(0,I)}\left[\lambda_t \left\| y - \mathcal{D}^{(y)}(\hat{z}_0^{(y)})\right\|_1\right]. \quad (4)$$

Algorithm 1: CoCoLIT Inference Procedure

Input: MRI volume x, LAS hyperparameter m
Output: Estimated PET scan $\hat{y}$

Encode MRI into latent space: $z^{(x)} = \mathcal{E}^{(x)}(x)$
for $j = 1$ to m **do**
  Sample Gaussian noise: $z_T^{(y,j)} \sim \mathcal{N}(0, I)$
  for $t = T$ to 1 **do**
    Reverse $z_t^{(y,j)} \to z_{t-1}^{(y,j)}$ using $\epsilon_{\theta,\phi}(z_t^{(y,j)}, t; z^{(x)})$
  end for
  Store final latent: $z^{(y,j)} = z_0^{(y,j)}$
end for
Compute $\bar{z}^{(y)} = \frac{1}{m}\sum_{j=1}^{m} z^{(y,j)}$
Decode PET scan: $\hat{y} = \mathcal{D}^{(y)}(\bar{z}^{(y)})$
return $\hat{y}$

Here, we calculate the weighted difference between the ground-truth PET y and the decoded prediction $\mathcal{D}^{(y)}(\hat{z}_0^{(y)})$. The term $\hat{z}_0^{(y)}$ is an estimate of the fully denoised latent, recovered from the noised latent $z_t^{(y)}$ at time-step t (Ho, Jain, and Abbeel 2020), and is given by the formula:

$$\hat{z}_0^{(y)} = \left(z_t^{(y)} - \sqrt{1-\bar{\alpha}_t}\epsilon_{\theta,\phi}(z_t^{(y)}, t; z^{(x)})\right) \cdot \left(\sqrt{\bar{\alpha}_t}\right)^{-1},$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. We adopt a time-step-dependent weighting $\lambda_t \in [0, 1]$ to scale the image-space loss, prioritizing low-frequency synthesis at high t and high-frequency detail reconstruction at low t, in line with the progressive refinement process of the diffusion model (Ho, Jain, and Abbeel 2020). For simplicity, we use a linear schedule defined as $\lambda_t = (T - t)/T$. The final loss term used in ControlNet training is given as:

$$\mathcal{L}_{WCN} = \mathcal{L}_{WISL} + \mathcal{L}_{CN}. \quad (5)$$

Since the loss term $\mathcal{L}_{WISL}$ is dependent on the decoder network, $\mathcal{D}^{(y)}$, we allow the decoder weights to be fine-tuned during ControlNet training (block D).

3.2 Theoretical Analysis of LAS

In this section, we present an analysis of the LAS estimator, $\hat{y}$ (see Section 2.3), to characterize its bias and assess its statistical validity. A full derivation with additional details can be found in the Supplementary Material.

Let $\mu = \mathbb{E}[z^{(y)}|z^{(x)}]$ denote the conditional mean of the learned latent distribution, $p(z^{(y)}|z^{(x)})$. For notational simplicity, we denote $z^{(y)}$ as a sample from the conditional distribution $p(z^{(y)}|z^{(x)})$ throughout the rest of this section. We begin by establishing an approximation, via a second-order Taylor expansion of the decoder $\mathcal{D}^{(y)}$ around μ, assuming that $p(z^{(y)}|z^{(x)})$ has finite second moments and $\mathcal{D}^{(y)}$ is twice continuously differentiable:

$$\begin{aligned}\mathbb{E}[\mathcal{D}^{(y)}(z^{(y)})] \approx \mathbb{E}[\mathcal{D}^{(y)}(\mu)] + \mathbb{E}[\nabla\mathcal{D}^{(y)}(\mu)(z^{(y)} - \mu)] \\ + \frac{1}{2}\mathbb{E}[(z^{(y)} - \mu)^T H_{\mathcal{D}^{(y)}}(z^{(y)} - \mu)].\end{aligned} \quad (6)$$

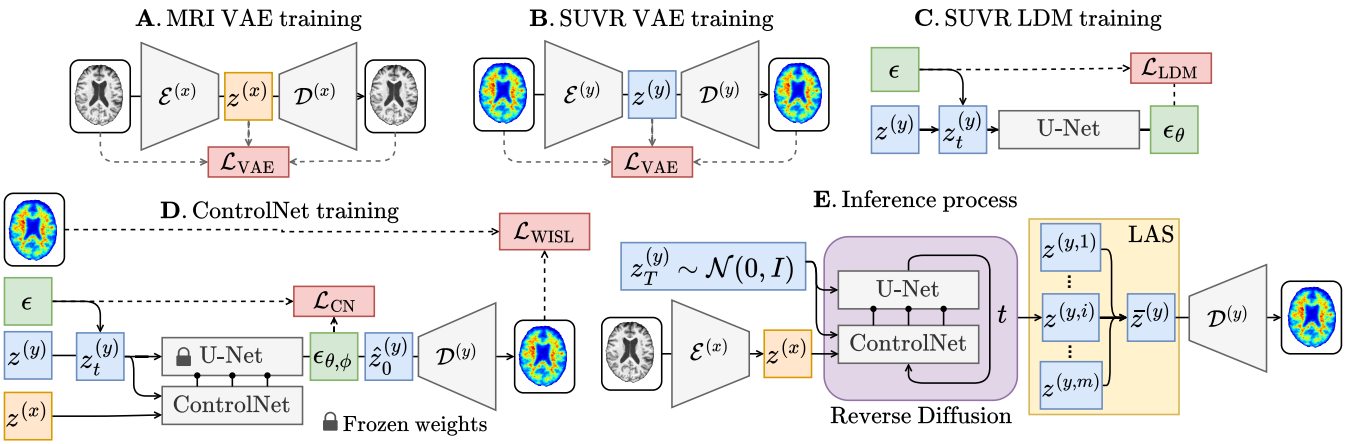

Figure 1: Overview of the CoCoLIT framework. (A–B) Training of the MRI and PET VAEs. (C) Training of the unconditional LDM on PET latents. (D) Training of the ControlNet and fine-tuning of the PET VAE decoder using standard noise loss and WISL. (E) Inference process in CoCoLIT, including the LAS algorithm.

Here, $H_{\mathcal{D}^{(y)}} \in \mathbb{R}^{D\times d\times d}$ denotes the Hessian tensor of the decoder evaluated at μ, comprising one $d\times d$ Hessian matrix per output dimension. Since $\mathbb{E}[z^{(y)} - \mu] = 0$, the first-order term vanishes. Applying the linearity of expectation and the cyclic property of the trace operator yields:

$$\mathbb{E}\left[\mathcal{D}^{(y)}\left(z^{(y)}\right)\right] \approx \mathcal{D}^{(y)}(\mu) + \frac{1}{2}\mathrm{Tr}\left(H_{\mathcal{D}^{(y)}}\Sigma_{z^{(y)}}\right), \quad (7)$$

where $\Sigma_{z^{(y)}} = \mathrm{Cov}(z^{(y)}) = \mathbb{E}\left[(z^{(y)} - \mu)(z^{(y)} - \mu)^T\right]$ is the $d\times d$ covariance matrix. The term $\mathrm{Tr}(H_{\mathcal{D}^{(y)}}\Sigma_{z^{(y)}}) \in \mathbb{R}^D$ represents a vector, containing the trace of each $d\times d$ Hessian and covariance matrix multiplication.

Applying the same approximation to the LAS estimator $\hat{y}$, and noting that for m i.i.d. samples $\mathrm{Cov}(\bar{z}^{(y)}) = \frac{1}{m}\Sigma_{z^{(y)}}$, its expectation is:

$$\mathbb{E}\left[\hat{y}\right] = \mathbb{E}\left[\mathcal{D}^{(y)}\left(\bar{z}^{(y)}\right)\right] \approx \mathcal{D}^{(y)}(\mu) + \frac{1}{2m}\mathrm{Tr}\left(H_{\mathcal{D}^{(y)}}\Sigma_{z^{(y)}}\right). \quad (8)$$

The bias of the LAS estimator is therefore approximated by the difference between Eq. (8) and Eq. (7):

$$\mathrm{Bias}(\hat{y}) \approx \left(\frac{1}{m} - 1\right)\frac{1}{2}\mathrm{Tr}\left(H_{\mathcal{D}^{(y)}}\Sigma_{z^{(y)}}\right). \quad (9)$$

From Eq. (9), we observe that as the number of latent samples $m \to \infty$, the bias does not vanish but instead converges to a constant:

$$\lim_{m\to\infty}\mathrm{Bias}(\hat{y}) = -\frac{1}{2}\mathrm{Tr}\left(H_{\mathcal{D}^{(y)}}\Sigma_{z^{(y)}}\right).$$

This reveals that LAS is an asymptotically biased estimator of the expected output. However, we hypothesize that this bias is negligible in practice for a sufficiently well-trained latent generative model. The practical effectiveness of LAS is justified by the following core assumption about the model's properties:

Assumption 1. *The LAS estimator exhibits negligible bias under the assumption that the latent distribution induced by a well-trained conditional LDM is sufficiently concentrated, such that the decoder $\mathcal{D}^{(y)}$ is approximately linear within the support of the latent samples. This occurs when the covariance $\Sigma_{z^{(y)}}$ is small, restricting samples to a neighborhood where the decoder's curvature is negligible.*

If this condition holds, the asymptotic bias term will be close to zero. In Section 4.6, we empirically show that the decoder behaves linearly within the sampled regions of latent space, and confirm LAS as an effective estimator. Therefore, despite its inherent bias, LAS can serve as a computationally efficient estimator for sufficiently well-trained models.

4 Experiments

This section presents an evaluation of the proposed CoCoLIT framework. We begin by briefly describing the internal and external datasets used in this study, along with the evaluation protocol adopted. We then conduct an ablation study to determine the impact of the LAS hyperparameter m, as well as to quantify the contribution of each component within the CoCoLIT framework. Furthermore, we benchmark CoCoLIT against the SOTA methods in MRI-to-PET translation. Finally, we empirically assess the validity of the theory underpinning LAS.

4.1 Datasets and Pre-processing

For training and evaluation of our framework, we use two publicly available multimodal neuroimaging datasets: ADNI (Petersen et al. 2010) and the A4 Study (including the LEARN substudy) (Sperling et al. 2014). Both datasets contain paired T1-weighted MRI and Florbetapir PET scans. The ADNI dataset includes 1,515 paired scans from 787 subjects (mean age: 74.9 ± 7.6 years; 50.6% female; 85.1% Aβ-positive). We split this dataset into training (80%), validation (5%), and test (15%) sets, ensuring strict subject-level separation to prevent data leakage. To assess generalization,

SETTING	IMAGE-BASED METRICS			Aβ-RELATED METRICS		
	SSIM ↑	PSNR ↑	MSE ↓	CABC ↑	HABC ↑	BA ↑
(A) ABLATION ON m						
$m = 1$	0.865 ± 0.047	22.570 ± 2.427	0.0067 ± 0.0051	0.180 ($p = 0.006$)	0.334 ($p < 0.001$)	57.4%
$m = 2$	0.880 ± 0.047	23.175 ± 2.575	0.0059 ± 0.0049	0.210 ($p = 0.001$)	0.405 ($p < 0.001$)	52.1%
$m = 4$	0.889 ± 0.049	23.735 ± 2.723	0.0054 ± 0.0048	0.300 ($p < 0.001$)	0.369 ($p < 0.001$)	56.4%
$m = 8$	0.892 ± 0.050	23.936 ± 2.738	0.0051 ± 0.0047	0.306 ($p < 0.001$)	0.470 ($p < 0.001$)	57.1%
$m = 16$	0.894 ± 0.050	24.079 ± 2.786	$\mathbf{0.0050 \pm 0.0047}$	0.292 ($p < 0.001$)	0.474 ($p < 0.001$)	60.6%
$m = 32$	0.895 ± 0.050	24.125 ± 2.807	$\mathbf{0.0050 \pm 0.0047}$	0.287 ($p < 0.001$)	0.500 ($p < 0.001$)	56.7%
$m = 64$	$\mathbf{0.896 \pm 0.050}$	$\mathbf{24.135 \pm 2.820}$	$\mathbf{0.0050 \pm 0.0047}$	$\mathbf{0.328}$ ($\mathbf{p < 0.001}$)	$\mathbf{0.522}$ ($\mathbf{p < 0.001}$)	**62.3%**
(B) COMPONENT ABLATION						
Base	0.841 ± 0.054	21.251 ± 2.370	0.0088 ± 0.0053	0.026 ($p = 0.694$)	0.253 ($p < 0.001$)	43.9%
+ ISL	0.870 ± 0.050	22.446 ± 2.767	0.0072 ± 0.0057	0.048 ($p = 0.476$)	0.280 ($p < 0.001$)	58.5%
+ WISL	0.865 ± 0.047	22.570 ± 2.427	0.0067 ± 0.0051	0.180 ($p = 0.006$)	0.334 ($p < 0.001$)	57.4%
+ LAS	0.887 ± 0.040	22.520 ± 2.462	0.0066 ± 0.0038	0.175 ($p = 0.008$)	$\mathbf{0.545}$ ($\mathbf{p < 0.001}$)	56.4%
+ LAS + ISL	$\mathbf{0.896 \pm 0.051}$	24.030 ± 2.681	$\mathbf{0.0050 \pm 0.0043}$	0.281 ($p < 0.001$)	0.422 ($p < 0.001$)	56.7%
+ LAS + WISL	$\mathbf{0.896 \pm 0.050}$	$\mathbf{24.135 \pm 2.820}$	$\mathbf{0.0050 \pm 0.0047}$	$\mathbf{0.328}$ ($\mathbf{p < 0.001}$)	0.522 ($p < 0.001$)	**62.3%**
(C) IMAGE VS. LATENT SPACE AVERAGING						
Unb. Estm. ($\bar{y}$)	0.896 ± 0.050	24.173 ± 2.803	0.0049 ± 0.0047	0.327 ($p < 0.001$)	0.541 ($p < 0.001$)	60.1%
LAS ($\hat{y}$)	0.896 ± 0.050	24.135 ± 2.820	0.0050 ± 0.0047	0.328 ($p < 0.001$)	0.522 ($p < 0.001$)	62.3%

Table 1: (A) Impact of varying LAS hyperparameter m for CoCoLIT on the internal test set. (B) Contribution of different model components (ISL, WISL, LAS) evaluated on the internal test set. (C) Comparison of LAS with the unbiased estimator (Unb. Estm.). Image-based metrics (SSIM, PSNR, MSE) are reported as *mean* ± *std*. Aβ metrics (CABC, HABC, BA) evaluate Aβ-burden correlation and classification performance, with p-values for CABC and HABC reported in parentheses. The best result is highlighted in bold, and the second-best is underlined.

METHOD	IMAGE-BASED METRICS			Aβ-RELATED METRICS		
	SSIM ↑	PSNR ↑	MSE ↓	CABC ↑	HABC ↑	BA ↑
INTERNAL TEST SET						
pix2pix	0.693 ± 0.038	13.968 ± 1.212	0.0416 ± 0.0111	0.178 ($p = 0.007$)	0.363 ($p < 0.001$)	51.8%
FICD	0.678 ± 0.033	12.656 ± 0.659	0.0549 ± 0.0084	0.049 ($p = 0.465$)	0.193 ($p = 0.004$)	48.2%
IL-CLDM	0.718 ± 0.077	18.987 ± 1.153	0.0131 ± 0.0038	-0.062 ($p = 0.357$)	0.280 ($p < 0.001$)	46.0%
PASTA	0.860 ± 0.042	21.630 ± 1.810	0.0076 ± 0.0042	-0.006 ($p = 0.932$)	0.378 ($p < 0.001$)	51.6%
CoCoLIT (*Ours*)	$\mathbf{0.896 \pm 0.050}$	$\mathbf{24.135 \pm 2.820}$	$\mathbf{0.0050 \pm 0.0047}$	$\mathbf{0.328}$ ($p < 0.001$)	$\mathbf{0.522}$ ($p < 0.001$)	**62.3%**
EXTERNAL TEST SET						
pix2pix	0.735 ± 0.035	15.016 ± 1.198	0.0327 ± 0.0088	0.126 ($p = 0.018$)	0.226 ($p < 0.001$)	51.8%
FICD	0.703 ± 0.028	12.876 ± 0.629	0.0521 ± 0.0077	-0.056 ($p = 0.298$)	-0.031 ($p = 0.558$)	49.6%
IL-CLDM	0.744 ± 0.086	19.918 ± 1.253	0.0107 ± 0.0049	0.008 ($p = 0.879$)	0.222 ($p < 0.001$)	50.1%
PASTA	0.882 ± 0.028	22.252 ± 1.795	0.0065 ± 0.0030	0.002 ($p = 0.967$)	0.235 ($p < 0.001$)	56.1%
CoCoLIT (*Ours*)	$\mathbf{0.940 \pm 0.010}$	$\mathbf{26.468 \pm 1.480}$	$\mathbf{0.0024 \pm 0.0011}$	$\mathbf{0.801}$ ($p < 0.001$)	$\mathbf{0.791}$ ($p < 0.001$)	**79.8%**

Table 2: Quantitative results from the comparison with baseline methods. Image-based metrics (SSIM, PSNR, MSE) are reported as *mean* ± *std*. Aβ metrics (CABC, HABC, BA) evaluate Aβ-burden correlation and classification performance, with p-values for CABC and HABC reported in parentheses. The best result is highlighted in bold, and the second-best is underlined.

we use the A4 cohort as an external test set, drawing a random sample of 350 image pairs from 350 unique subjects (mean age: 63.9 ± 22.8 years; 59.7% female; 83.1% Aβ-positive). Following standard practice (Schreiber et al. 2015; Royse et al. 2021), we convert PET scans to Standardized Uptake Value Ratio (SUVR) maps using the cerebellar gray matter as the reference region. Ground-truth $A\beta$-positivity is defined as the mean SUVR value in the cerebral cortex exceeding the commonly used threshold of 1.11 (Schreiber et al. 2015; Royse et al. 2021). To retain subject-specific patterns and preserve inter-modality relationships, we perform a z-score standardization on both MRI and PET scans independently using statistics computed from the training set. We resample all MRI and SUVR volumes to a uniform spatial resolution of 1.5 mm^3. Full pre-processing details are provided in the Supplementary Material.

4.2 Implementation Details

The MRI and PET VAEs used in Section 3.1 were obtained by fine-tuning separate instances of the MAISI VAE (Guo et al. 2025). This network was chosen for its extensive pre-training on large amounts of 3D medical imaging data. All the blocks in CoCoLIT were implemented using the MONAI framework (Pinaya et al. 2023), and all training and experiments were conducted on an NVIDIA A100 GPU. At inference time, latent samples were generated from the LDM using an implicit sampling strategy (DDIM) (Song, Meng, and Ermon 2020), with 50 inference steps.

4.3 Evaluation Protocol

We assess model performance using six metrics: three image-based measures—Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Mean Squared Error (MSE)—and three $A\beta$-related measures. Specifically, we compute Spearman correlations between predicted and ground-truth mean SUVR values in the cerebral cortex and hippocampus, referred to as Cerebral Amyloid Burden Correlation (CABC) and Hippocampal Amyloid Burden Correlation (HABC), respectively. These regions are selected due to their known association with Aβ accumulation (Hampel et al. 2021). Lastly, we evaluate binary $A\beta$-positivity classification using Balanced Accuracy (BA) to address class imbalance. To account for possible systematic biases in each method, predicted Aβ-positivity is determined by applying a data-driven threshold to the predicted mean cortical SUVR. This threshold is selected on the internal validation set to maximize BA and is held fixed during testing. We provide the threshold values for each method, along with details on the statistical tests performed in our experiments, in the Supplementary Material.

4.4 Ablation Study

In this section, we present an ablation study to assess (i) the effect of the LAS hyperparameter m on model performance, and (ii) the contribution of each individual component within the CoCoLIT framework.

LAS Hyperparameter Analysis We evaluate the effect of the LAS hyperparameter m on model performance, with results presented in Table 1-A. Image-based metrics consistently improve as m increases. While the rate of improvement slows beyond $m = 8$, the overall trend indicates that larger m values enhance structural fidelity. Similarly, Aβ-related metrics show substantial gains over the baseline ($m = 1$), with both correlation measures (CABC, HABC) and Aβ-positivity classification (BA) reaching their highest values at $m = 64$. The results also suggest that the estimated burden increasingly aligns with the ground-truth as m grows. Based on these findings, we select $m = 64$ as the optimal configuration for all subsequent experiments.

Evaluating Individual Components We perform an ablation study to evaluate the contribution of key components in our framework, focusing on: (i) the use of LAS at inference time; (ii) a variant of our proposed WISL with constant weight ($\lambda_t = 1 \; \forall t \in [0, T]$), referred to as ISL; and (iii) the proposed time-step-dependent WISL loss. Comparing the results of ISL and WISL allows us to assess the impact of varying λ_t over time, as defined in Section 3.1. We define the "Base" model as CoCoLIT without LAS ($m = 1$) and the ControlNet trained without either ISL or WISL. We then independently assess the contribution of each component by progressively adding them. Results are summarized in Table 1-B. Adding ISL during training leads to consistent improvements in both image-based and Aβ-related metrics, suggesting that ControlNet and the decoder benefit from image-space guidance during training. Introducing the time-step-dependent weighting (WISL) yields further gains, particularly in Aβ correlation metrics (Base + WISL vs. Base + ISL). These improvements become even more pronounced when LAS is used, with WISL outperforming ISL across all metrics (Base + LAS + WISL vs. Base + LAS + ISL). We hypothesize that ISL, lacking temporal weighting, may prematurely enforce fine detail generation early in the denoising process, potentially disrupting the learned trajectory. In contrast, WISL aligns supervision with the progressive nature of denoising, thereby preserving generative stability. Finally, LAS itself contributes positively across all configurations, enhancing both image fidelity and Aβ-related performance. Based on these findings, we adopt the full CoCoLIT model with LAS and WISL for all subsequent experiments.

4.5 Comparison with State-of-the-Art

In this section, we compare CoCoLIT with several baseline models. We conduct a thorough quantitative evaluation using the metrics described in Section 4.3, and present qualitative examples of predictions on both internal and external test sets.

Baselines We compare CoCoLIT against existing baseline approaches: PASTA (Li et al. 2024), IL-CLDM (Ou et al. 2024), FICD (Yu et al. 2024) and pix2pix (Isola et al. 2017). All baselines were implemented using their official code.

Quantitative Comparison Quantitative results are presented in Table 2. On the internal test set, CoCoLIT significantly outperforms all baseline methods across both image-based and Aβ-related metrics. Notably, while the baselines perform near chance level in Aβ-positivity classification, CoCoLIT achieves much better performance, with a BA of 62.3% (+10.5% over the second-best method), along with correlations of 0.328 ($p < 0.001$) and 0.522 ($p < 0.001$) for CABC and HABC, respectively. Unexpectedly, on the external test set, all methods exhibit improved image-based metrics and BA. We suspect this is likely due to differences in acquisition protocols and post-processing of PET scans in the A4 study, which result in smoother SUVR signals that are easier to predict. On this dataset, CoCoLIT achieves an SSIM of 0.94, CABC and HABC scores exceeding 0.79, and a BA of 79.8% (+23.7% over the second-best method). All performance improvements are statistically significant both on the internal test set (image-based metrics: $p < 0.001$, BA: $p < 0.05$, except for PASTA [$p < 0.1$]) and on the external test set (image-based metrics: $p < 0.001$, BA: $p < 0.001$). Our results establish CoCoLIT as the new SOTA for MRI-to-PET synthesis and underscore its generalization capabilities,

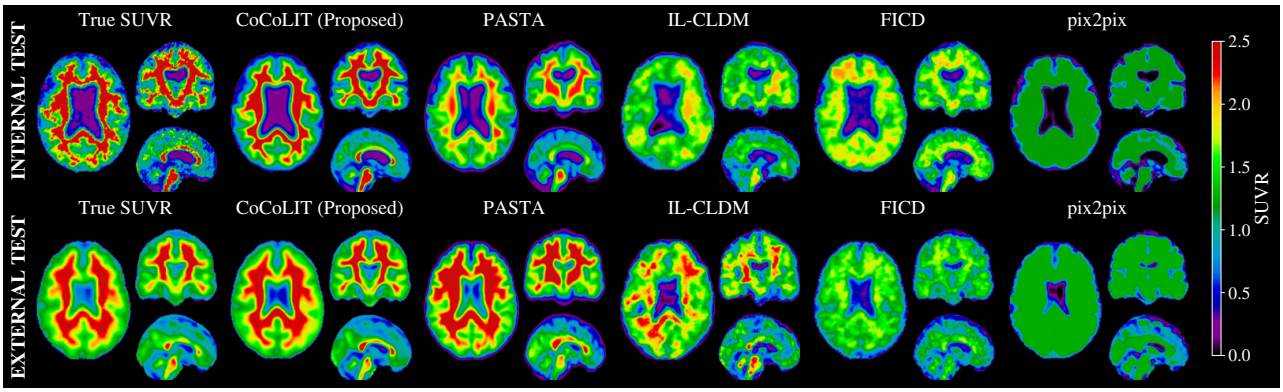

Figure 2: Qualitative comparison of SUVR maps predicted from structural MRIs using CoCoLIT and baseline methods on both internal and external test sets. The color bar on the right indicates SUVR values ranging from 0.0 to 2.5.

supporting its potential for future clinical translation.

Qualitative Comparison Figure 2 presents visual comparisons of predicted SUVR maps using CoCoLIT and baseline models. Across both the internal and external test sets, CoCoLIT (second column) better approximates the ground-truth Aβ accumulation (first column). In the ground-truth volumes, the smoother SUVR signal observed in the A4 dataset is apparent when compared to ADNI.

4.6 Empirical Assessment of LAS Theory

In this section, we empirically evaluate Assumption 1 (see Section 3.2) and compare the performance of the LAS estimator, $\hat{y}$, with the unbiased estimator, $\bar{y}$ (see Section 2.3).

Local Linearity of the Decoder (Assumption 1) As we show in Section 4.4, the LAS bias depends on the decoder's curvature and becomes negligible when the decoder $\mathcal{D}^{(y)}$ is locally linear. We empirically validate this local linearity assumption with two complementary tests. First, we define the experimental setup. For each test subject, we randomly sample five unique pairs of latent vectors, $(z_i^{(y)}, z_j^{(y)})$, from the conditional distribution $p(z^{(y)}|z^{(x)})$. These are decoded to their corresponding outputs, $y_i = \mathcal{D}^{(y)}(z_i^{(y)})$ and $y_j = \mathcal{D}^{(y)}(z_j^{(y)})$. We then construct a linear interpolation path in the latent space using 10 evenly spaced steps $s \in [0, 1]$:

$$z^{(y)}_{\text{interp}}(s) = z_i^{(y)} + s(z_j^{(y)} - z_i^{(y)})$$

The resulting path in the image space is $y_{\text{interp}}(s) = \mathcal{D}^{(y)}(z^{(y)}_{\text{interp}}(s))$. Based on this, we perform two tests.

Test 1: This test assesses whether the distance traveled in the image space increases linearly with the latent interpolation step s. We measure this by computing the Pearson Correlation Coefficient (PCC) between the steps s and the corresponding L1 distances from the start point, $d(s) = \|y_{\text{interp}}(s) - y_i\|_1$.

Test 2: This test directly quantifies how much the output path deviates from a perfect straight line. We compare the actual path $y_{\text{interp}}(s)$ to an ideal linear path $\hat{y}_{\text{interp}}(s) = y_i + s(y_j - y_i)$. The deviation is measured as the MSE between these two paths, averaged over all steps s.

Across all subjects and latent pairs, we find the mean PCC to be 0.9994 ± 0.0015 and the mean MSE to be 0.00045 ± 0.00057. Together, these results provide empirical evidence supporting Assumption 1.

Practical Effectiveness of the LAS Estimator To evaluate the effectiveness of the LAS estimator, $\hat{y}$, we empirically compare it with the unbiased estimator, $\bar{y}$, which decodes all $N = m = 64$ samples before averaging. As shown in Table 1-C, both estimators achieve comparable performance. We therefore conclude that LAS is a practically effective estimator, as its bias does not lead to any meaningful degradation in output quality.

5 Discussion and Limitations

In this study, we present CoCoLIT, a novel ControlNet-conditioned latent diffusion framework that outperforms SOTA methods in 3D MRI-to-PET translation. Our method introduces WISL, an image-space supervision loss, and integrates LAS, whose effectiveness is supported by theoretical and empirical results. While our work focuses on MRI-to-PET translation, the CoCoLIT framework is generalizable to a broader range of conditional generative tasks, such as disease progression modeling (Puglisi, Alexander, and Ravì 2025), image quality transfer (Gao et al. 2023), and translation across other imaging modalities (Moschetto et al. 2025).

Despite these promising results, some limitations remain. Although the model achieves higher Aβ-positivity classification accuracy compared to the SOTA, further improvements may be required to ensure reliable clinical translation. Additionally, although LAS is more efficient than the unbiased estimator at inference time, drawing m samples can still incur computational costs without GPU-parallelization.

In conclusion, future work could explore evaluating CoCoLIT on a broader spectrum of image-to-image tasks, further advancing its synthesis capabilities. Moreover, by leveraging the framework's flexible conditioning mechanism, the integration of clinically relevant covariates may enhance both the predictive power and clinical utility of the model.

Acknowledgements

AS is supported by the EPSRC-funded UCL Centre for Doctoral Training i4health (EP/S021930/1) and by the Wellcome Trust (221915). LP is supported by the PNRR initiative (DM 118/2023). NPO is a UKRI Future Leaders Fellow (MR/S03546X/1, MR/X024288/1). NPO and DCH acknowledge funding from the E-DADS project (EU-JPND 2019; UKRI MR/T046422/1). DCH acknowledges funding from the Wellcome Trust CU-MONDAI project (WT 221915), the Wellcome Trust Democratizing MRI project (WT 317797), and the NIHR UCLH Biomedical Research Centre. DR acknowledges funding from the "Rete eHealth: AI e strumenti ICT Innovativi orientati alla Diagnostica Digitale (RAIDD)" project (CUP J43C22000380001).

References

Chapleau, M.; Iaccarino, L.; Soleimani-Meigooni, D.; and Rabinovici, G. D. 2022. The role of amyloid PET in imaging neurodegenerative disorders: a review. *Journal of Nuclear Medicine*, 63(Supplement 1): 13S–19S.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10021–10030.

Guo, P.; Zhao, C.; Yang, D.; Xu, Z.; Nath, V.; Tang, Y.; Simon, B.; Belue, M.; Harmon, S.; Turkbey, B.; et al. 2025. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4430–4441. IEEE.

Hampel, H.; Hardy, J.; Blennow, K.; Chen, C.; Perry, G.; Kim, S. H.; Villemagne, V. L.; Aisen, P.; Vendruscolo, M.; Iwatsubo, T.; et al. 2021. The amyloid-β pathway in Alzheimer's disease. *Molecular psychiatry*, 26(10): 5481–5503.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kerbler, G. M.; Fripp, J.; Rowe, C. C.; Villemagne, V. L.; Salvado, O.; Rose, S.; Coulson, E. J.; Initiative, A. D. N.; et al. 2015. Basal forebrain atrophy correlates with amyloid β burden in Alzheimer's disease. *NeuroImage: Clinical*, 7: 105–113.

Lee, Y.-S.; Youn, H.; Jeong, H.-G.; Lee, T.-J.; Han, J. W.; Park, J. H.; and Kim, K. W. 2021. Cost-effectiveness of using amyloid positron emission tomography in individuals with mild cognitive impairment. *Cost Effectiveness and Resource Allocation*, 19(1): 50.

Li, Y.; Yakushev, I.; Hedderich, D. M.; and Wachinger, C. 2024. PASTA: P athology-A ware MRI to PET Cro S s-modal T r A nslation with Diffusion Models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 529–540. Springer.

Moschetto, A.; Puglisi, L.; Sargood, A.; Dell'Acqua, P.; Guarnera, F.; Battiato, S.; and Ravì, D. 2025. Benchmarking GANs, Diffusion Models, and Flow Matching for T1w-to-T2w MRI Translation. *arXiv preprint arXiv:2507.14575*.

Nordberg, A. 2004. PET imaging of amyloid in Alzheimer's disease. *The lancet neurology*, 3(9): 519–527.

Ou, Z.; Pan, Y.; Xie, F.; Guo, Q.; and Shen, D. 2024. Image-and-Label Conditioning Latent Diffusion Model: Synthesizing Aβ-PET from MRI for Detecting Amyloid Status. *IEEE Journal of Biomedical and Health Informatics*.

Pan, Y.; Liu, M.; Lian, C.; Zhou, T.; Xia, Y.; and Shen, D. 2018. Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, 455–463. Springer.

Petersen, R. C.; Aisen, P. S.; Beckett, L. A.; Donohue, M. C.; Gamst, A. C.; Harvey, D. J.; Jack, C. R.; Jagust, W. J.; Shaw, L. M.; Toga, A. W.; et al. 2010. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology*, 74(3): 201–209.

Pinaya, W. H.; Graham, M. S.; Kerfoot, E.; Tudosiu, P.-D.; Dafflon, J.; Fernandez, V.; Sanchez, P.; Wolleb, J.; Da Costa, P. F.; Patel, A.; et al. 2023. Generative ai for medical imaging: extending the monai framework. *arXiv preprint arXiv:2307.15208*.

Puglisi, L.; Alexander, D. C.; and Ravì, D. 2024. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 173–183. Springer.

Puglisi, L.; Alexander, D. C.; and Ravì, D. 2025. Brain Latent Progression: Individual-based spatiotemporal disease progression on 3D Brain MRIs via latent diffusion. *Medical Image Analysis*, 103734.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Royse, S. K.; Minhas, D. S.; Lopresti, B. J.; Murphy, A.; Ward, T.; Koeppe, R. A.; Bullich, S.; DeSanti, S.; Jagust, W. J.; Landau, S. M.; et al. 2021. Validation of amyloid PET positivity thresholds in centiloids: a multisite PET study approach. *Alzheimer's research & therapy*, 13(1): 99.

Schreiber, S.; Landau, S. M.; Fero, A.; Schreiber, F.; Jagust, W. J.; Initiative, A. D. N.; et al. 2015. Comparison of visual and quantitative florbetapir F 18 positron emission tomography analysis in predicting mild cognitive impairment outcomes. *JAMA neurology*, 72(10): 1183–1190.

Shin, H.-C.; Ihsani, A.; Xu, Z.; Mandava, S.; Sreenivas, S. T.; Forster, C.; Cha, J.; and Initiative, A. D. N. 2020. GANDALF: Generative adversarial networks with discriminator-adaptive loss fine-tuning for Alzheimer's disease diagnosis from MRI. In *Medical Image Computing and*

Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23, 688–697. Springer.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Sperling, R. A.; Rentz, D. M.; Johnson, K. A.; Karlawish, J.; Donohue, M.; Salmon, D. P.; and Aisen, P. 2014. The A4 study: stopping AD before symptoms begin? *Science translational medicine*, 6(228): 228fs13–228fs13.

Tay, L. X.; Ong, S. C.; Tay, L. J.; Ng, T.; and Parumasivam, T. 2024. Economic burden of Alzheimer's disease: a systematic review. *Value in health regional issues*, 40: 1–12.

Yu, M.; Wu, M.; Yue, L.; Bozoki, A.; and Liu, M. 2024. Functional Imaging Constrained Diffusion for Brain PET Synthesis from Structural MRI. *arXiv preprint arXiv:2405.02504*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.