Latent Diffusion Modeling for Photoacoustic Computed Tomography

Thomas Wynn
Advised by Evan Scope Crafts and Dr. Umberto Villa

April 2025

1 Introduction

Photoacoustic computed tomography (PACT) is an optical imaging modality that exploits the physical phenomena that conjugated π-systems (such as those found in hemoglobin or melanin) can be excited using light to emit a tiny, rapid temperature rise, which causes mechanical expansion and a resulting acoustic wave. These waves can be detected externally and then used to compute an image by indirectly measuring light absorption. PACT is desirable over other imaging modalities because it is noninvasive and does not require contrast agents; instead, it relies on endogenous chromophores. However, the inherent scattering and attenuation of light and sound waves in biological tissue make a high-fidelity reconstruction of an image difficult due to the introduction of noise and complexity.

Let us denote by PACT the problem of reconstructing initial pressure maps from acoustic data, and by qPACT the associated quantitative inversion where we recover absolute optical absorption coefficients. Recent advances tackle qPACT by embedding a learned prior, $p(x)$, into its Bayesian formulation, regularizing the estimation of absorption coefficients directly in a probabilistic framework. $p(x)$ into the Bayesian formulation. In particular, score-based diffusion models have emerged as flexible priors: they learn to "denoise" from pure noise back to realistic images, and can be used within score-based or posterior sampling methods to regularize the Bayesian inversion [1, 2].

Training a Variational Autoencoder (VAE) in concert with a diffusion model yields a **latent diffusion model (LDM)**: the VAE encoder projects high-dimensional images x into a compact latent representation z, and diffusion training in z provides a tractable learned prior $p(z)$. This prior can be directly incorporated into the Bayesian inverse problem, dramatically reducing computational cost when solving in three-dimensional (3D) or higher-dimensional settings.

As a concrete application, this LDM-based Bayesian framework enables efficient estimation of the Cramér-Rao bound (CRB) on reconstruction error, which

is essential for optimal experimental design (OED) [3]. Instead of computing the Fisher information matrix in the full image space, we compute it in the low-dimensional latent space. Thus, we retain the statistical properties of the original image distribution while lowering the computational complexity CRB calculation.

In this project, I first describe the PACT dataset, then detail the architecture and training of a variational autoencoder—pretrained on natural images and fine-tuned on PACT absorption coefficients—and its accompanying latent diffusion model. In particular, I introduce an SVD-based method to reduce the autoencoder's latent space and train the LDM within this SVD-reduced latent representation.

2 Dataset

Our dataset consists of 2D absorption-coefficient maps of size 1024×1024 pixels, each capturing the spatial distribution of optical absorption at 800 nm over a 2 mm thickness of tissue. The dataset is sourced from 2683 three-dimensional numerical breast phantoms, where each phantom represents the tissue types of a simulated breast embedded in a voxel grid of size $1024 \times 1024 \times 680$, with a voxel resolution of $0.125\,\mathrm{mm}^3$.

We partitioned the 2683 phantoms into 1879 for training, 268 for validation, and 537 for testing, following a 70/10/20 split. To generate training, validation, and testing samples, we randomly selected phantoms from their respective splits, extracted a 16-voxel-thick (2 mm) slice, assigned optical properties to each slice based on tissue type (following this method at a wavelength of 800 nm), and computed an average intensity projection along the thickness dimension. This process yielded 20,000 training samples, 10,000 validation samples, and 10,000 testing samples of dimension 1024×1024 pixels.

However, biological tissues exhibit several orders of magnitude variation in μ_a values. Consequently, tissue types with small but similar μ_a values —such as fat and glandular tissue— are difficult to distinguish in μ_a space. As a result, machine learning loss functions would be insensitive to these differences, potentially degrading model performance. Therefore, we work with $\log(\mu_a)$ maps in future steps.

3 Variational Autoencoder

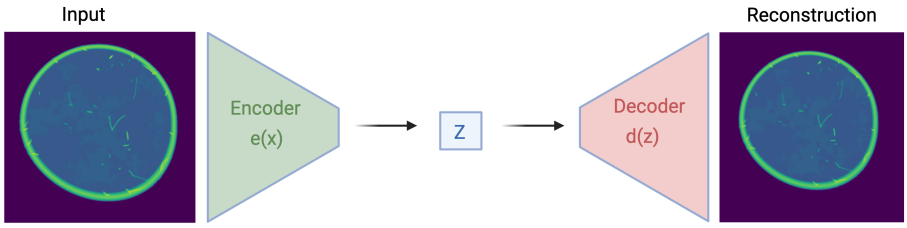

Figure 1: Autoencoder schematic for PACT data using $\log(\mu_a)$ maps.

Variational Autoencoders are a class of neural networks used for unsupervised learning. They encode data into a compressed latent representation and then reconstruct the original image from the latent representation, typically using convolutional layers. Autoencoders consist of two parts: the encoder $e(x)$ and the decoder $d(z)$, where x is the input and z is the latent distribution. Specifically, z is a multivariate Gaussian with a diagonal covariance matrix, from which the decoder will sample. Then, a pass through the autoencoder could be represented with the function composition:

$$(d \circ e)(x) = \tilde{x} \tag{1}$$

Where $\tilde{x}$ denotes the reconstruction. To leverage recent advancements in computer vision, the autoencoder architecture and pretrained weights were adapted from the CompVis group's Stable Diffusion repository (citation). However, the original Stable Diffusion autoencoder was trained on datasets such as the Google Open Images dataset, which consists of $3 \times 256 \times 256$ RGB images.

To fine-tune the pretrained autoencoder on the PACT dataset, additional convolutional and transposed convolutional layers were inserted before the encoder and after the decoder, respectively, to accommodate the different input and output formats. Such that

$$\text{prelayer}(x) \in \mathbb{R}^{3\times256\times256}$$
$$\text{postlayer}(d(z)) \in \mathbb{R}^{1024\times1024}$$

An $8\times$ autoencoder gives us $z \in \mathbb{R}^{4\times32\times32}$. This allows for the fine-tuning of Stable-Diffusion's autoencoder on $\log(\mu_a)$ maps.

In early experiments, we evaluated the included loss function, which combines ℓ_1 pixel reconstruction loss, LPIPS perceptual loss, KL divergence regularization, and an adversarial hinge loss, following the `LPIPSWithDiscriminator` formulation (adapted from the CompVis Stable Diffusion repository). We found that the default loss function provided by Stable Diffusion often omitted finer structural details in the reconstructions. Therefore, to simplify training and better preserve structural features in PACT images, we instead adopted multi-scale

structural similarity (MS-SSIM) as our primary reconstruction loss. Specifically, our loss combines a weighted MS-SSIM loss, pixel-wise ℓ_1 loss, and KL divergence.

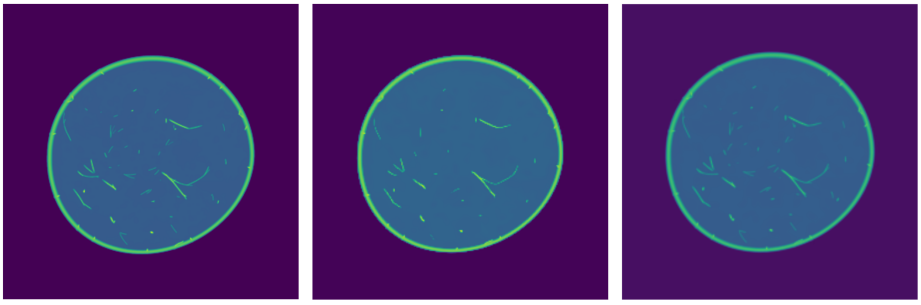

Figure 2: Comparison of the original image (left), reconstruction with the baseline loss (middle), and reconstruction with the custom loss (right). Note the improved preservation of small blood vessels in the reconstruction with the custom loss.

4 Latent Dimensionality Reduction Using SVD

A latent space of $z \in \mathbb{R}^{3\times32\times32}$ corresponds to a dimensionality of $d = 4096$. To further simplify the inverse problem and CRB estimation, we seek to reduce the latent dimensionality while preserving the structure necessary for accurate reconstruction. To achieve this, we construct a latent data matrix by encoding $N = 10{,}000$ samples, flattening each latent code to form the columns of a matrix $Z \in \mathbb{R}^{d\times N}$, and compute its singular value decomposition:

$$Z = U\,\Sigma\,V^T.$$

We then truncate to the first k left singular vectors, yielding

$$U_k \in \mathbb{R}^{d\times k}.$$

We construct our new autoencoder as such:

$$(d \circ U_k^T \circ U_k \circ e)(x) = \tilde{x}$$

With reshaping such that $e(x) \in \mathbb{R}^d$ and $(U_k^T \circ U_k \circ e)(x)$ has the dimensions of our original latent space. If we set $k = 1024$ then our reduced latent is $(U_k \circ e)(x) \in \mathbb{R}^{1024}$ instead of $\mathbb{R}^{4096}$.

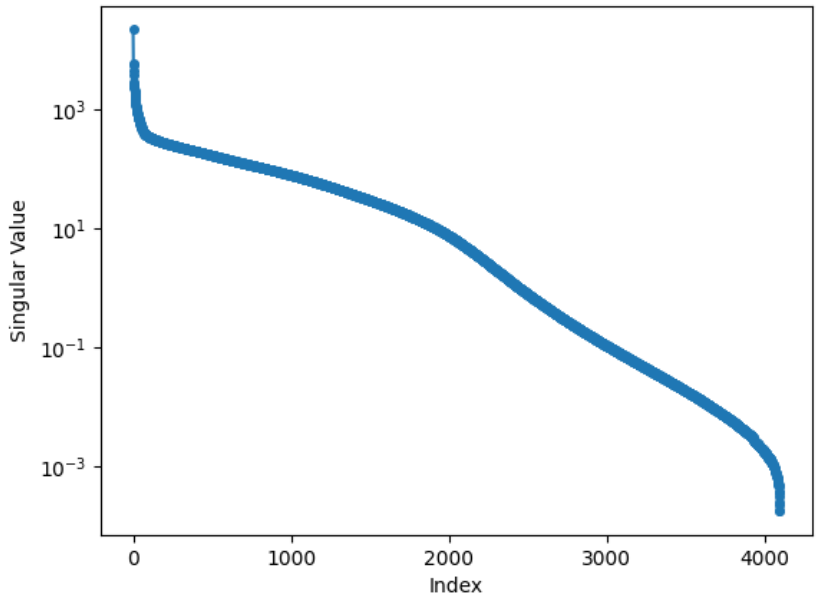

Figure 3: Singular value decay of the latent data matrix Z

5 Latent Diffusion Model

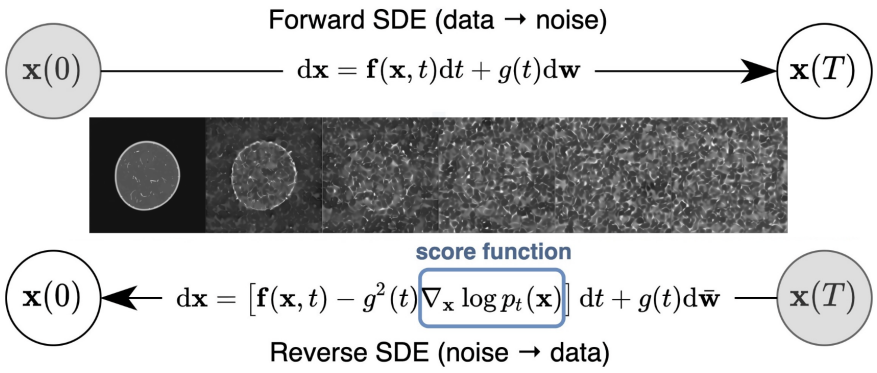

Figure 4: Schematic of a score-based diffusion model operating in image space (adapted from Song et al.).

To train a diffusion model in the latent space, we adapt the Latent Diffusion Model from the Stable Diffusion repository, utilizing OpenAI's DDPM framework. Inputs are first encoded into latent representations, and forward diffusion is simulated directly in this latent space. We then train a U-Net that, for each timestep t, takes the noisy latent z_t and the timestep t as input, and predicts the score $\nabla_{z_t} \log q_t(z_t)$ to guide the reverse denoising process.

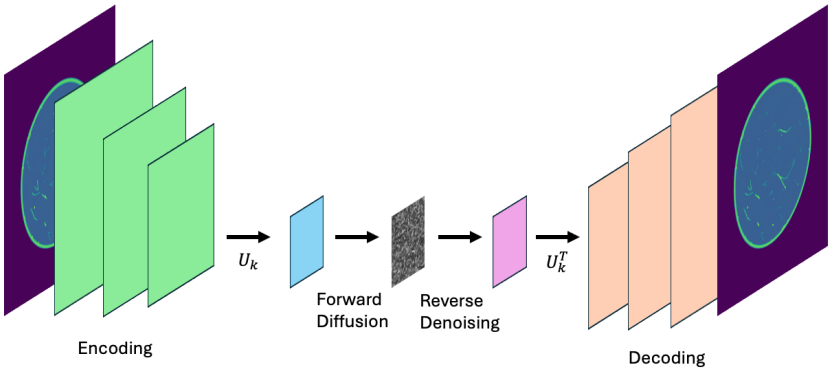

Figure 5: Diffusion model operating in the latent space, with the final denoised latent shown in pink.

To train a diffusion model in the reduced latent space $\mathbb{R}^{1024}$ (reshaped as $\mathbb{R}^{4\times16\times16}$), we wrap the standard diffusion U-Net with projection and back-projection layers. Although the model inputs and outputs are in the reduced

latent space, we perform denoising in the full latent space to better fit the convolutional structure of the U-Net. Thus, we can model diffusion in a compressed latent space while still benefiting from spatial correlations present in the full latent representation. Given a reduced latent input $y \in \mathbb{R}^{4\times16\times16}$, we first flatten and project it up to the full latent dimension:

$$z \;=\; U_k^T \,\mathrm{vec}(y) \;\in\; \mathbb{R}^{4096},$$

which we reshape to $z \in \mathbb{R}^{4\times32\times32}$ and feed into the U-Net, $g(z)$

$$\hat{z} \;=\; g(z) \;\in\; \mathbb{R}^{4\times32\times32}.$$

Finally, we flatten and project back down,

$$\hat{y} \;=\; U_k \,\mathrm{vec}(\hat{z}) \;\in\; \mathbb{R}^{1024},$$

and reshape to $\hat{y} \in \mathbb{R}^{4\times16\times16}$ as the reconstructed latent. In compact form, the wrapper implements

$$f(y) \;=\; \big(\mathrm{reshape}_{4\times16\times16} \circ U_k \circ g \circ \mathrm{reshape}_{4\times32\times32} \circ U_k^T \circ \mathrm{vec}\big)(y) = \hat{y}$$

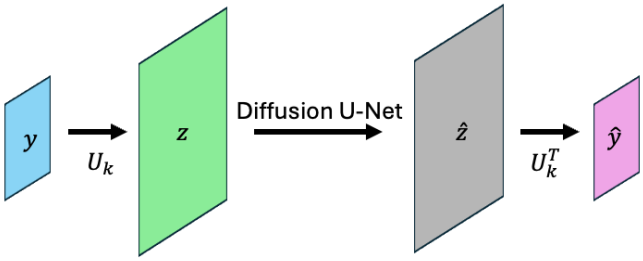

Figure 6: Schematic of the diffusion model wrapper. Inputs in the reduced latent space $\mathbb{R}^{4\times16\times16}$ are projected to the full latent space $\mathbb{R}^{4\times32\times32}$, denoised using the diffusion U-Net, and then projected back down to the reduced space. The overall mapping is represented as $f(y) = \hat{y}$.

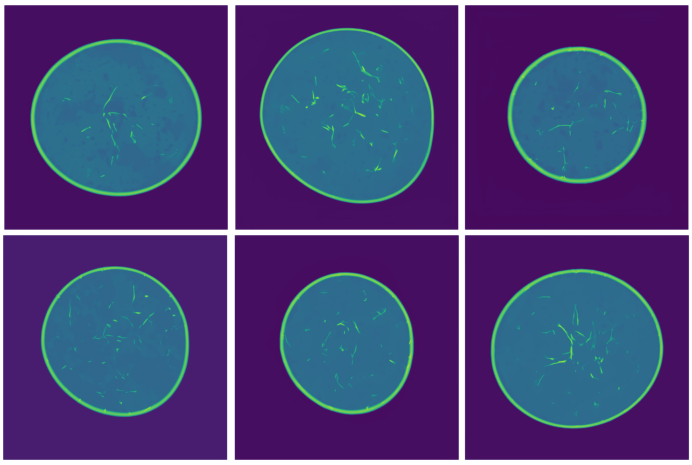

Figure 7: Samples generated by the latent-space diffusion model after denoising and decoding.

6 Discussion

By leveraging Stable Diffusion's pretrained models, we fine-tuned a latent-space diffusion model on $\log(\mu_a)$ maps using a custom loss function and an SVD-based dimensionality reduction of the latent space. Training a VAE in the reduced latent space enables a reformulation of the inverse problem in a lower-dimensional domain. Additionally, modeling a diffusion process in this reduced space constructs a latent prior distribution for the Bayesian inverse problem, paving the way for more efficient Optimal Experimental Design (OED) and Cramér-Rao Bound (CRB) estimation.

7 References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: arXiv preprint arXiv:2006.11239 (2020). URL: https://arxiv.org/abs/2006.11239.

[2] Yang Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: International Conference on Learning Representations (ICLR). 2021. URL: https://openreview.net/forum?id=PxTIG12RRHS.

[3] Evan Scope Crafts and Umberto Villa. "Can Diffusion Models Provide Rigorous Uncertainty Quantification for Bayesian Inverse Problems?" In: arXiv preprint arXiv:2503.03007 (2025). URL: https://arxiv.org/abs/2503.03007.