Department of Advanced Computing Sciences
Maastricht University

Explaining Artificial Intelligence Guided Adaptive Radiotherapy Using Counterfactual Explanations in Latent Space

MSc Data Science For Decision Making

Author : Wasti Murad
Student ID : i6348497
Supervisors : Cecile Wolfs (Maastro)
: Mirela Popa (DACS)
: Visara Urovi
: Luca Heising
Submission date : 21 August 2025

Abstract

Adaptive radiotherapy demands more than accurate predictions, it requires transparent, case specific reasoning about how and why a plan should change for a given patient on a given day. This thesis addresses this need by framing the explainability around latent space counterfactuals: minimal, realistic edits to the patient's imaging that reveal how anatomy and predicted segmentations would differ if specific factors were altered. We develop an intrinsically explainable volumetric generative model that couples a variational autoencoder with residual and attention mechanisms, and shape its representation using a decomposed Kullback–Leibler objective with cyclic beta scheduling to encourage disentangled, clinically meaningful factors. Counterfactuals are produced by targeted traversals of individual latent dimensions, yielding smooth, anatomy-preserving edits without per-case optimization. Evaluation focuses on the plausibility, controllability, specificity, and faithfulness of the counterfactual quality using quantitative probes. Ablations show that conventional skip connections can boost overlap scores at the expense of a collapsible bottleneck that hinders counterfactual control, removing these shortcuts and enforcing the proposed regularization restores sparse, semantically organized factors that clinicians can directly steer. While tumor delineation remains challenging, the framework consistently produces fast, anatomically coherent what-if explanations that surface model assumptions and support day to day decision making in adaptive radiotherapy, laying a path toward trustworthy, counterfactual-driven quality control.

Acknowledgements

This thesis was conducted at Maastricht University in collaboration with the Clinical Data Science group at Maastro. I am grateful to Visara Urovi, Cecile Wolfs, Luca Heising, and Mirela Popa for insightful discussions, constructive feedback, and steady encouragement. I also thank the Data Science Research Infrastructure (DSRI) for computing support and The Cancer Imaging Archive (TCIA) for providing public imaging data used in this work. Finally, my heartfelt thanks go to my family and friends for their unwavering support.

Contents

List of Figures

List of Tables

1 Introduction

Adaptive radiotherapy (ART) is reshaping radiation oncology by allowing treatment plans to evolve with patient anatomy over the course of therapy. Anatomical changes between fractions, organ motion, and tumor regression can all erode the accuracy of a fixed plan. In response, ART makes timely adjustments with the aim of maintaining target coverage while sparing healthy tissue [1]. In parallel, ART workflows are increasingly adopting artificial intelligence (AI) for decision support, image guidance, and segmentation [2, 3]. Yet transparency has not kept pace with raw performance: high-capacity models often operate as *black boxes*, offering little justification for their recommendations [4]. This opacity undermines clinician trust, complicates regulatory evaluation, and ultimately limits clinical impact in a safety-critical setting [5, 6, 7].

Explainability is therefore not a luxury in medical imaging, it is a clinical requirement [8]. Post hoc attribution techniques (e.g., saliency maps and class activation variants) can highlight pixels associated with a model output [9, 10], but rarely align with the counterfactual reasoning that clinicians use in practice: *What if the tumor were smaller? What if an organ shifted? What if edema resolved?* [11, 12] ART adds a longitudinal dimension, raising the bar further. Explanations should be fast enough for time limited decisions, consistent across fractions, and coherent with expected anatomical evolution. Existing explainability approaches rarely satisfy fidelity, faithfulness, and timeliness simultaneously in clinical settings. Perturbation methods such as LIME and SHAP are instance-bound and computationally demanding [13]; saliency and attention heatmaps (e.g., Grad-CAM/Attention U-Net) are fast but often lack semantic specificity for treatment decisions [14]. Recent GAN-based counterfactuals can be visually compelling yet remain resource-intensive, dataset-specific, or static at inference [15]. This mismatch is acute in multi fraction ART, where on couch decisions must be made within minutes.

This thesis addresses these gaps by developing an explainable deep learning framework for ART that unifies three properties often treated in isolation: (i) clinical interpretability, (ii) counterfactual reasoning, and (iii) temporal consistency. The central idea is to model anatomy in a generative latent space so that the very mechanism that produces outputs can also answer *why* and *what-if*. Concretely, we couple two specialized three-dimensional variational autoencoders (VAEs): one trained on CT volumes to capture intensity structure, and one trained on segmentation masks to capture anatomical topology. Working in latent space enables fast and plausible counterfactuals, such as organ displacement or tumor shrinkage, without per-case optimization at inference time. The probabilistic formulation also supports uncertainty estimates that are useful for risk aware decisions.

A key challenge is preserving reconstruction fidelity without sacrificing the interpretability and controllability of the latent space. Architectures that emphasize pixel accuracy, especially those with extensive U-Net-style skip connections can bypass the latent bottleneck, harming disentanglement, and limiting controllability. Our design instead favors an informative bottleneck. We restore representational power with residual blocks and attention mechanisms, and we regulate the latent space with a decomposed KL objective (separating mutual information, total correlation, and dimension-wise terms) combined with cyclical annealing. This combination aims to maintain strong reconstructions while preserving active, semantically organized latent dimensions, prerequisites for generating anatomically plausible, temporally coherent counterfactuals in ART.

Research Questions

- **Main question.** How can a deep learning framework that couples dual-stream VAEs with latent-space counterfactuals enhance explainability and decision-making in AI-guided adaptive radiotherapy?
  - **Sub-questions.**
  - How does replacing the standard β-VAE objective with a decomposed β-TCVAE objective (separating mutual information, total correlation, and dimension-wise KL) affect latent disentanglement, reconstruction fidelity, and controllability in 3D medical image segmentation?
  - What is the quantitative and qualitative impact of integrating a CBAM attention mechanism into the segmentation VAE on reconstruction accuracy, attention alignment with clinically relevant anatomy, and the interpretability of latent traversals, compared to an otherwise identical architecture without attention?
  - How does replacing a constant-β schedule with cyclic β-scheduling alter the trade-off between reconstruction fidelity and latent-space disentanglement in a segmentation VAE, and what is the impact on the controllability and anatomical plausibility of counterfactual generations?
  - How can we learn disentangled latent representations from CT and segmentation that capture clinically meaningful factors while preserving reconstruction fidelity?

Contributions

This thesis makes four contributions. First, it proposes a dual VAE architecture that separates intensity structure (CT) from anatomical topology (masks) and supports joint reasoning via coordinated latent traversals. Second, it introduces a training strategy that balances fidelity and disentanglement combining residual learning, attention, a decomposed KL regularizer (MI/TC/DW), and cyclical annealing to maintain an informative bottleneck without skip connections. Third, it develops a fast latent space counterfactual mechanism that exposes uncertainty for risk-sensitive interpretation and enables explicit control over clinically meaningful factors without per-case optimization. Finally, it designs an evaluation protocol that measures not only segmentation accuracy but also latent activity, factor interpretability, attentional alignment with anatomy, and temporal consistency.

Empirical Scope and Limitations

Experiments focus on thoracic cases from the NSCLC-Radiomics collection, using CT images and expert segmentation masks relevant to radiotherapy planning. Preprocessing follows radiotherapy conventions: intensities are normalized within a clinically meaningful Hounsfield range, volumes are resampled to isotropic spacing, and three-dimensional tensors are standardized for fair architectural comparison. Test time evaluation uses fixed per-class thresholds aligned with the codebase and reports cohort means alongside case-level analyses. We do not integrate task specific classifiers, the emphasis is an intrinsic, generative explanation mechanism. This scope enables a detailed study of representation quality and counterfactual behavior. Broader multi-institutional validation and prospective clinical studies lie beyond the present work and are highlighted as directions for future research.

Summary

In summary, the thesis advances a counterfactual, intrinsically explainable approach to ART designed for computational feasibility, clinical plausibility, and temporal stability. By centering interpretable latent spaces, instead of relying on post hoc explanations, the work aims to make AI outputs in ART not only accurate but also transparent and actionable, clarifying how anatomy shapes decisions and how hypothetical changes would alter them.

2 Background and Related Work

In medical imaging tasks, artificial intelligence (AI) systems have demonstrated excellent performance, frequently matching or even exceeding expert-level accuracy. However, clinical trust and regulatory approval are seriously hampered by their opaque decision making procedures. In high-stakes fields like medical imaging, where physicians need to comprehend how and why AI algorithms make particular predictions, this ”black box” problem is particularly important.

It is crucial to comprehend how anatomical changes impact AI predictions in adaptive radiotherapy (ART). A key component of clinical decision making[5]. is counterfactual reasoning, which is not supported by conventional explanation techniques like saliency maps or feature attributions: ”What would happen if this structure changed?” ART further complicates this need by introducing a temporal dimension, where treatment adaptation depends on consistent understanding of anatomical evolution over time.

Explaining medical images is more complicated than in tabular domains because they are high-dimensional and require anatomical realism.Post-hoc saliency approaches (e.g., Grad-CAM or attention gating) provide pixel-level highlights but offer limited causal or semantic grounding for treatment adaptation [14]. Perturbation-based explanations such as LIME/SHAP are instance-specific and computationally heavy, which challenges on couch use [13]. Counterfactuals generated with adversarial image-to-image models can be compelling, yet they remain resource-intensive, fragile across domains, and are rarely evaluated for clinical faithfulness [15]. These gaps motivate generative, latent-space approaches that target single, interpretable factors. In the framework of ART, this thesis explores the development of realistic, chronologically consistent, and clinically useful counterfactual explanations.

To address these challenges, this thesis introduces a novel VAE-based approach explicitly designed for adaptive radiotherapy, ensuring clinically meaningful disentanglement, temporal consistency, real-time counterfactual generation, hierarchical multiscale modeling, and robust uncertainty quantification.

Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) emerged in response to increasingly complex and opaque machine learning models. While early models like linear regression and decision trees offered transparency, they lacked the capacity to handle complex tasks. Deep learning models improved performance but introduced opacity, resulting in a trade-off between accuracy and interpretability.

Explainable Artificial Intelligence (XAI) has become increasingly important as deep learning models grow in complexity and opacity. Interpretability is essential to ensuring clinical trust and well-informed decision making in high-stakes areas like medical imaging, where choices have a direct impact on patient care. In general, XAI techniques can be divided into two groups: post-hoc explainability methods, which use tools like LIME and SHAP to interpret the

behavior of an already trained model, and intrinsic interpretability, where model transparency is incorporated into the architecture itself (e.g., decision trees, attention mechanisms [16, 14]). Even though these techniques can yield insightful information, they frequently only offer local or instance-specific explanations, which may not adequately convey the model's actual reasoning tendencies, especially when dealing with complicated, high-dimensional medical data.

Explainable Artificial Intelligence in Imaging

Medical imaging presents unique challenges for XAI due to the complexity, scale, and contextual richness of image data [17]. Conventional techniques, such as saliency maps and CAMs, emphasize significant pixels but frequently miss subtle spatial or anatomical correlations that are essential to clinical reasoning.

The majority of existing methods produce static, instance level attributions, even though transformer-based models and attention mechanisms offer richer visual explanations [14]. Their usefulness in applications like ART is limited because they don't explain how certain alterations affect results or encourage the investigation of alternate anatomical conditions.

Counterfactual Explanations

Counterfactual explanations offer an alternative to feature attribution by asking: "What changes would yield a different prediction?" [18]. They are contrastive, actionable, and intuitive, aligning with how humans reason about causality.

Wachter et al. [18] formalized counterfactuals as minimal changes to input features that alter the model's prediction:

$$x' = \arg\min_{x'} d(x, x') \quad \text{subject to} \quad f(x') = y', \quad x' \in \mathcal{X}$$

There are several ways to create counterfactuals, such as model-agnostic and optimization-based approaches, which frequently strike a balance between several goals: diversity, actionability, plausibility, and proximity. However, these approaches are computationally costly and prone to producing irrational results when dealing with high-dimensional inputs, such as photographs.

Counterfactual Explanations in Imaging

Applying counterfactual reasoning to images adds complexity due to the high dimensionality and need for visual realism. Unnatural outputs that deviate from the image manifold are frequently the result of direct pixel manipulation. In order to solve this, GAN-based techniques develop mappings between conditions [19]; however, they are unstable and do not model uncertainty or provide interpretability.

Although StyleGAN provides controlled latent alteration, it is only applicable to natural images and does not generalize to medical fields. Furthermore, human studies show a discrepancy between algorithmic and human judgments of counterfactual usefulness, and common evaluation measures such as SSIM or FID fall short in capturing clinical or semantic relevance.

Counterfactual Explanations in Latent Space

Generating counterfactual explanations in high-dimensional data such as medical images presents unique challenges. Pixel-level changes frequently result in off-manifold or unrealistic

outcomes, which diminishes their clinical utility. Generative adversarial networks (GANs) are capable of producing realistic images, however they usually have limited interpretability, unstable training, and no uncertainty modeling. They are not ideal for safety-critical fields like adaptive radiotherapy because of these drawbacks.

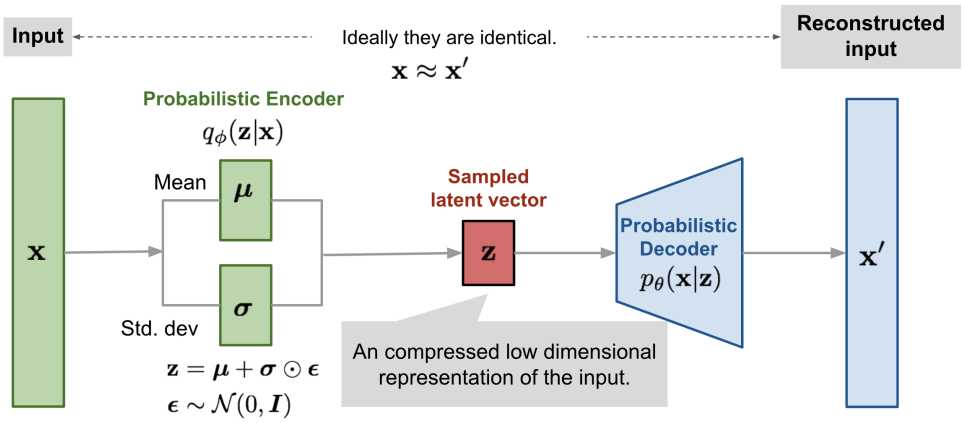


Figure 1: Variational Autoencoder (VAE) architecture. The encoder maps the input to a distribution over latent variables, from which a sample is drawn and passed through the decoder to reconstruct the input. This process encourages the latent space to be continuous, structured, and suitable for semantic manipulation.

Figure 1 illustrates the core components of a Variational Autoencoder. The encoder learns a distribution over latent variables (typically parameterized by a mean and standard deviation), from which a latent vector is sampled and passed through the decoder to reconstruct the input. This probabilistic formulation is essential for generating realistic, low-dimensional representations that can be manipulated for counterfactual reasoning. Many of these issues are addressed by latent space-based methods. Latent space techniques enable the creation of counterfactuals through seamless and semantically significant changes by projecting data into a compact and structured representation. Specifically, the capacity of Variational Autoencoders (VAEs) [20, 21] to represent complex data distributions while maintaining continuity and uncertainty in their latent spaces makes them ideal for medical imaging. Counterfactuals can be generated by perturbing a latent vector z such that its decoded image reflects a desired change in output:

$$z' = \arg\min_{z'} d(z, z') \quad \text{subject to} \quad \text{Decoder}(z') \in \mathcal{Y}'$$

In this thesis, we use a latent-space counterfactual method specifically designed for the detection of anatomical changes. Two distinct 3D VAEs are trained: one for segmentation masks and one for CT images. For structural and anatomical information, respectively, each learns a probabilistic latent representation. Then, without departing from the learnt data manifold, counterfactuals are produced by altering certain latent vector dimensions to mimic believable anatomical alterations, such as tumor shrinking, organ deformation, or positional shifts.

We predefine interpretable latent directions during training, in contrast to optimization based counterfactual approaches that necessitate inference time search and frequently yield unstable outcomes. Real time "what-if" scenario development is made possible by the

disentanglement of these orientations and their semantic alignment with clinical characteristics. Additionally, this removes the requirement for external classifiers, increasing the method's computational efficiency and modularity. The latent traversal remains consistent across time points, supporting the temporal coherence required in ART workflows.

Overall, latent space counterfactuals offer a clinically meaningful, efficient, and interpretable alternative to traditional post hoc explainability methods in medical imaging. They enable structured exploration of treatment relevant anatomical changes, aligning model outputs with clinician expectations and improving trust in AI-guided adaptive radiotherapy.

VAE for Counterfactual Explanations in Medical Imaging

Several groups have explored VAE-based approaches in medical imaging with promising results. Liu et al. [22] developed a VAE framework for medical image enhancement, focusing on reference based super-resolution. Chartsias et al. proposed factorized representation learning to disentangle anatomical from modality-specific features in cardiac imaging. Pawlowski et al. [23] introduced deep structural causal models that integrate VAEs with causal inference for generating population level counterfactuals in neuroimaging. Sauer and Geiger utilized StyleGAN to generate counterfactual images, achieving high visual quality however, their approach lacks uncertainty quantification and does not preserve temporal coherence [24].

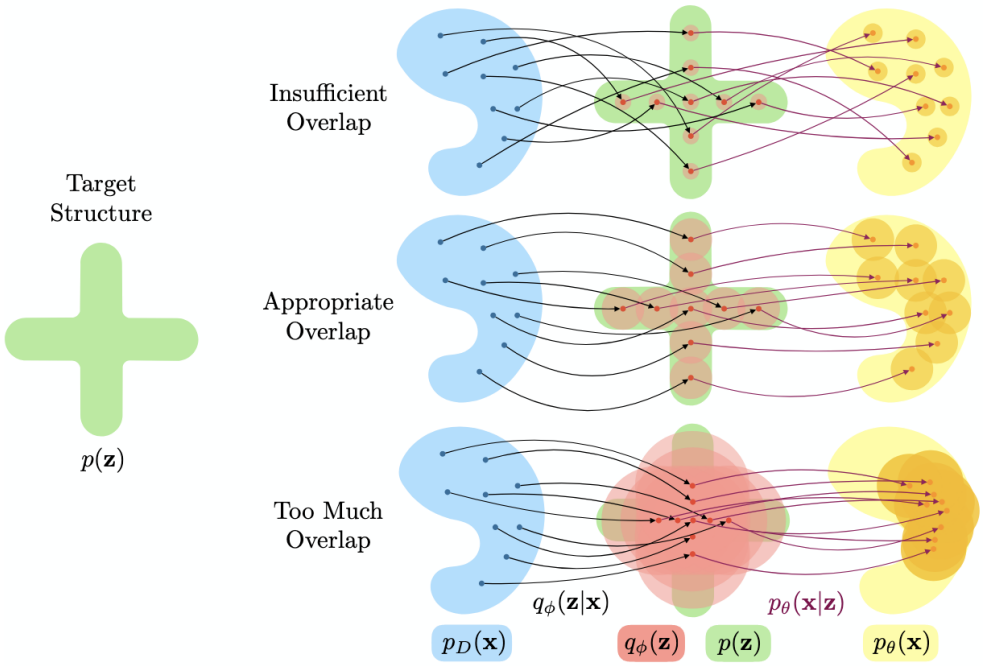


Figure 2: Visualizing latent space overlap in VAEs. When the posterior distribution $q_\phi(z|x)$ has insufficient overlap with the prior $p(z)$ (top), generalization suffers. Too much overlap (bottom) leads to poor latent structure. An appropriate balance (middle) enables semantically meaningful and disentangled latent representations [25].

A well-structured latent space is critical for counterfactual generation. As shown in Figure 2, the degree of overlap between the approximate posterior $q_\phi(z|x)$ and the prior $p(z)$ influences the quality of the learned representation. Techniques like β-VAE explicitly control this overlap via KL divergence weighting, encouraging disentanglement and interpretability in the latent space. In the context of adaptive radiotherapy, current VAE-based counterfactual techniques

are inadequate not withstanding these developments. Due to their lack of temporal consistency between treatment sessions, most techniques produce counterfactuals for isolated images, which violates the biological continuity required for clinical monitoring. Furthermore, rather than separating clinically significant variables like tumor response or morphological deformation, current disentanglement algorithms frequently separate low level technical differences like contrast or scanner noise. In addition to being computationally demanding, optimization-based methods for creating counterfactuals are not feasible for real-time clinical procedures where treatment choices must be modified under tight time limitations. Furthermore, these models frequently overlook the necessity for multi scale representations that capture changes at both the local tissue level and the global organ level, failing to take into consideration the hierarchical nature of medical imaging.

While related works demonstrate promising generative capabilities, they are not directly sufficient for our ART setting. Liu et al. focus on reference-based super-resolution rather than clinically constrained counterfactual editing [26]; Pawlowski et al. propose deep structural causal models that are theoretically appealing but not validated on 3D thoracic data or within ART time constraints [27]; and Counterfactual Generative Networks are primarily developed on natural images with limited evaluation of non-target preservation and clinical plausibility [28]. Our work therefore tailors a VAE-based approach to volumetric medical data with explicit objectives for independence, controllability, and speed.Our method introduces temporally consistent latent representations by decomposing the KL divergence to separate stable anatomical structures from dynamic, treatment-induced changes, thereby ensuring longitudinal coherence. In contrast to generic methods like β-VAE, we match latent factors with semantically significant anatomical variances by using structured priors based on clinical knowledge. By pre learning interpretable latent directions during training, real-time counterfactual production is accomplished without requiring inference-time optimization. Our hierarchical framework provides a thorough understanding of everything from fine-grained tissue alterations to organ deformation by capturing anatomical variation across several spatial scales. Finally, the framework includes uncertainty quantification for each counterfactual, enabling clinicians to assess prediction confidence and make informed, risk-aware decisions in real-world treatment planning.

3 Methodology

Overview

For adaptive radiotherapy (ART), we develop an intrinsically explainable framework based on Variational Autoencoders (VAEs) that can both reconstruct anatomy and justify its behavior in clinically meaningful terms [20, 17]. The framework addresses the critical need for interpretability in radiation therapy planning, where understanding model decisions is essential for clinical acceptance and safety. Rather than fixing an architecture *a priori*, we follow an iterative design strategy that increases model complexity step by step while jointly monitoring two axes: reconstruction fidelity and latent-space quality. This approach exposes a fundamental trade-off between preserving fine spatial detail and maintaining an informative, controllable bottleneck. The methodology culminates in counterfactual explanations that emerge naturally via latent traversals, where specific latent directions are varied to produce interpretable anatomical changes such as tumour shrinkage or organ displacement [18].

3.1 Dataset

We utilize the NSCLC–Radiomics dataset from The Cancer Imaging Archive (TCIA), which provides thoracic CT scans and expert-annotated masks for gross tumour volume (GTV)

and lungs [29]. The dataset comprises 422 patients with non-small cell lung cancer and offers clinical diversity that is important for robust model development. Patient ages range from 33.7 to 91.7 years with a mean of 67.1 ± 10.8 years, there is a male predominance, and the cohort includes squamous cell carcinoma, large cell carcinoma, adenocarcinoma, and not otherwise specified histologies across stages I to IIIb. This heterogeneity supports generalization across tumour presentations, anatomical variation, and disease progression typical of ART workflows.

Medical images are provided as DICOM series with RTSTRUCT contour sets. To support efficient 3D learning and consistent metadata handling, we convert CT series and contours to NIfTI volumes and per-class binary labelmaps in the CT reference frame. This DICOM→NIfTI conversion preserves spatial orientation and voxel spacing while reducing I/O overhead during training. We implement conversion and metadata handling with MONAI utilities to ensure robustness and reproducibility [30].

3.2 Architectural Design

We select VAEs over standard autoencoders because VAEs learn a probabilistic latent space regularized toward a simple prior, which enables smooth interpolation between anatomically plausible states and calibrated sampling around observed cases [20]. This property is crucial for generating realistic counterfactuals, while the KL regularizer discourages overfitting and encourages a structured latent representation suitable for controllable edits. Deterministic autoencoders lack an explicit probabilistic semantics and therefore do not provide the same guarantees for traversal-based explanations or uncertainty-aware reconstructions.

The backbone is a 3D encoder–decoder with Instance Normalization and LeakyReLU activations. To stabilize deep 3D optimization, we integrate residual blocks that improve gradient flow and representational capacity [31]. In development we observed that U-Net-style skip connections improved overlap metrics but allowed the decoder to bypass the latent bottleneck, driving the KL term toward zero and reducing the number of active latent dimensions. Because this harms interpretability and controllability, the final model omits skip connections to preserve an informative bottleneck. The encoder outputs a 32-dimensional latent parameterized by mean μ and log-variance $\log \sigma^2$. The decoder upsamples using nearest-neighbour interpolation followed by $3 \times 3 \times 3$ convolutions and produces three Sigmoid channels corresponding to GTV, left lung, and right lung. To emphasize clinically salient structures and suppress distractors, we incorporate a 3D variant of the Convolutional Block Attention Module (CBAM), combining channel and spatial attention to highlight informative features at multiple encoder levels [32, 14].

3.3 Objectives: from VAE to β-VAE to β-TC-VAE

We formulate learning as minimization of the negative evidence lower bound (ELBO). The standard VAE objective is

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}\big[-\log p_\theta(x|z)\big] + \mathrm{KL}\big(q_\phi(z|x) \,\|\, p(z)\big), \qquad p(z) = \mathcal{N}(0, I),$$

where the reconstruction term promotes fidelity and the KL divergence regularizes the posterior towards the prior, encouraging a smooth, sampleable latent space.

To promote disentanglement we first adopt the β-VAE objective, which upweights the KL term:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q(z|x)}\big[-\log p(x|z)\big] + \beta\,\mathrm{KL}\big(q(z|x) \,\|\, p(z)\big), \qquad \beta \geq 1.$$

Increasing β encourages axis-aligned factors and reduces latent entanglement, but large β can

degrade reconstruction quality or induce posterior collapse, creating an unfavorable trade-off in medical applications where fidelity and interpretability are both essential [25].

To address this limitation we adopt the β-TC-VAE formulation, which decomposes the expected KL into mutual information (MI), total correlation (TC), and dimension-wise KL (DW) terms [33]. The identity

$$\mathbb{E}_{p(x)}[\mathrm{KL}(q(z|x)\|p(z))] = I_q(x;z) + \mathrm{TC}(q(z)) + \sum_j \mathrm{KL}(q(z_j)\|p(z_j))$$

motivates the loss

$$\mathcal{L}_{\beta\text{-TCVAE}} = \mathbb{E}_{q(z|x)}[-\log p(x|z)] + \alpha\, I_q(x;z) + \beta\, \mathrm{TC}(q(z)) + \gamma \sum_j \mathrm{KL}(q(z_j)\|p(z_j)).$$

This decomposition allows us to penalize inter-dimensional dependence via TC while preserving MI to keep latents informative and using DW-KL to stabilize per-dimension priors. In practice this yields a better balance between reconstruction fidelity and controllable, independent latent factors.

3.4 Training Strategy with Cyclic β

A fixed β forces a hard trade-off: small values tend to yield good reconstructions but entangled factors, whereas large values improve disentanglement at the cost of fidelity. We therefore employ a cyclical annealing schedule that ramps β from a low value to $\beta_{\max}$ and then resets periodically [34]. Concretely, with cycle length T, we use

$$\beta(t) = \beta_{\max} \cdot \min\Big(1,\ \frac{t \bmod T}{T/2}\Big),$$

and scale the TC weight proportionally during the ramp so that independence pressure is strongest near the cycle peak. Low-β phases allow the decoder to refine reconstructions, while high-β phases encourage factor separation. This alternation avoids collapse and produces active, controllable latents without sacrificing reconstruction quality. We optimize with Adam using a learning rate of 2×10^{-4}.

3.5 Counterfactual Generation and Latent Traversal

Counterfactuals are generated without per-case optimization. For an input x we encode to the posterior mean $\hat{z}$ and vary a single coordinate z_d over a symmetric range while holding the remaining coordinates fixed at $\hat{z}_{\neg d}$. Decoding these perturbed codes produces a sequence of images that reflect smooth, single-factor edits, such as progressive tumour shrinkage or organ displacement. We assess interpretability by visual overlays on the axial slice with maximal GTV area and by plotting predicted structure volumes as a function of z_d, which quantifies monotonic control and usable traversal range. In addition to qualitative plausibility and non-target stability, we report overlap and surface metrics where applicable, linking latent manipulations to clinically meaningful changes that align with ART decision-making [18].

3.6 Implementation Framework

All components are implemented in PyTorch with MONAI for medical image I/O and pre-processing [30]. The codebase separates architecture, objectives, and schedules to enable controlled ablations, including variants with and without CBAM, residual blocks, KL decomposition, and skip connections, as well as alternative β schedules. This modularity facilitates future extensions while maintaining reproducibility within the described pipeline.

4 Experimental Setup and Evaluation

This chapter presents the empirical methodology for developing and validating our enhanced VAE framework for adaptive radiotherapy. Rather than pursuing a predetermined architecture, we adopted a systematic exploration strategy that iteratively builds complexity while monitoring both reconstruction performance and latent space quality. This approach revealed critical insights about architectural trade-offs, particularly the tension between spatial detail preservation and latent interpretability that fundamentally shaped our final design decisions.

4.1 Progressive Architecture Development

Our experimental approach followed a methodical progression from simple to complex architectures, with each stage building upon insights from the previous iteration. This strategy allowed us to understand the individual contribution of each architectural component while identifying unexpected interactions and trade-offs that influenced our final design.

The journey began with a foundational 3D convolutional VAE featuring standard encoder decoder architecture with Instance Normalization and LeakyReLU activations. This baseline established fundamental performance benchmarks and revealed inherent limitations in fine anatomical detail reconstruction and boundary preservation. Building upon this foundation, we first integrated residual blocks to address vanishing gradient problems in deeper 3D networks. The residual connections provided substantial improvements in gradient flow and feature learning capabilities, yielding the most significant single performance gain in our progressive development.

The next enhancement involved incorporating CBAM attention modules [32] that combine channel and spatial attention mechanisms. These modules enabled the network to adaptively focus on anatomically salient regions while providing interpretable visualizations of model decision-making processes. The attention mechanism not only improved reconstruction quality but also aligned with clinical requirements for explainable AI in healthcare applications.

A critical discovery emerged during our exploration of U-Net-style skip connections. While skip connections dramatically improved reconstruction metrics by preserving fine-grained spatial information, extensive analysis revealed a fundamental trade-off that became central to our architectural philosophy. Skip connections enabled the decoder to increasingly bypass the latent bottleneck, causing the network to rely more heavily on direct feature transfers rather than learning meaningful latent representations. This phenomenon progressively reduced latent utilization and severely degraded disentanglement properties, with only 0 of 32 latent dimensions remaining active.

This observation led to a strategic architectural decision to exclude skip connections from our final enhanced model. Despite their benefits for reconstruction accuracy, skip connections fundamentally compromised the interpretability and controllable generation capabilities that are essential for adaptive radiotherapy applications. This trade-off demonstrates how clinical applications that require controllable anatomical synthesis may demand optimization priorities that are different from traditional reconstruction tasks.

To recover latent space quality without relying on skip connections, we implemented KL divergence decomposition [33], separating the standard regularization term into three interpretable components: mutual information (MI), total correlation (TC), and dimension-wise KL divergence (DW). This decomposition provided fine-grained control over different aspects of latent organization, enabling optimization of disentanglement properties without sacrificing reconstruction fidelity. The MI term encouraged meaningful relationships between latent variables and anatomical features, TC minimized dependencies among latent dimensions to

promote disentanglement, and DW controlled individual dimension distributions for training stability.

The final architectural enhancement involved implementing a cyclic β-scheduling strategy [25] that dynamically modulates the weights of KL components throughout training. This approach balanced the competing demands of reconstruction accuracy and latent organization by allowing regularization pressures to gradually increase and decrease in cycles. The scheduling enabled progressive emergence of disentangled representations while maintaining training stability and avoiding sudden disruptions to reconstruction quality.

Final network used in experiments The encoder downsamples to a $256 \times 8 \times 8 \times 8$ tensor, flattens to 131072 features, and projects to a 32-dimensional latent mean and log variance. The decoder upsamples using nearest neighbour followed by $3 \times 3 \times 3$ convolutions and outputs three Sigmoid channels for GTV, left lung, and right lung. Residual blocks and a 3D CBAM are retained. Skip connections are intentionally omitted to preserve an informative bottleneck.

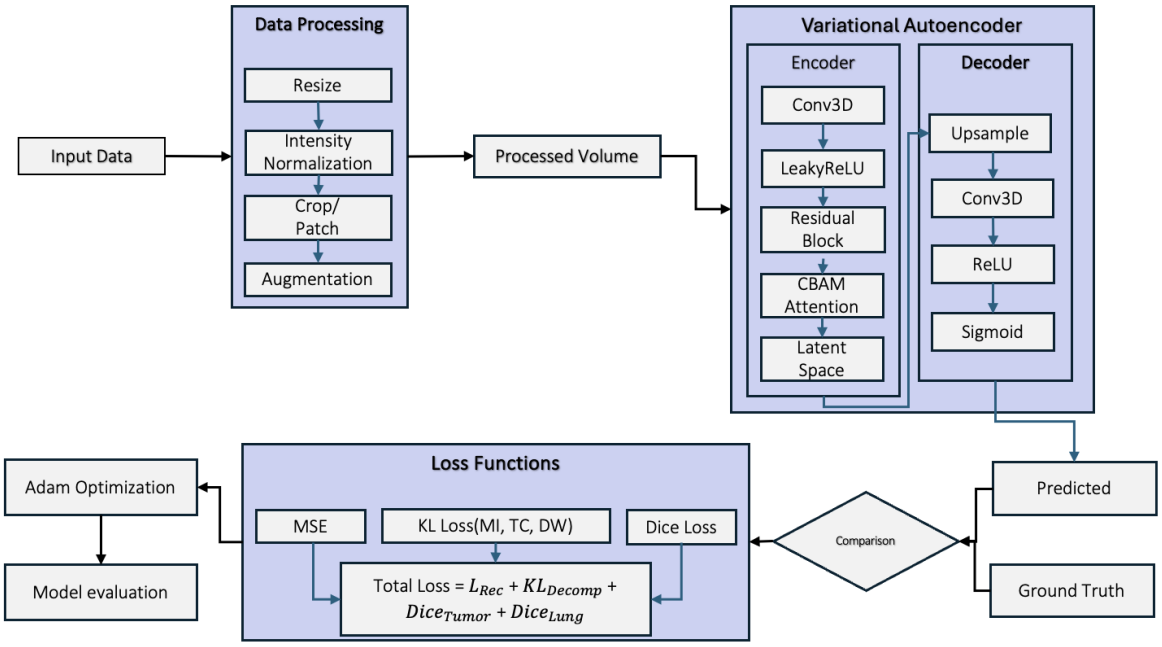


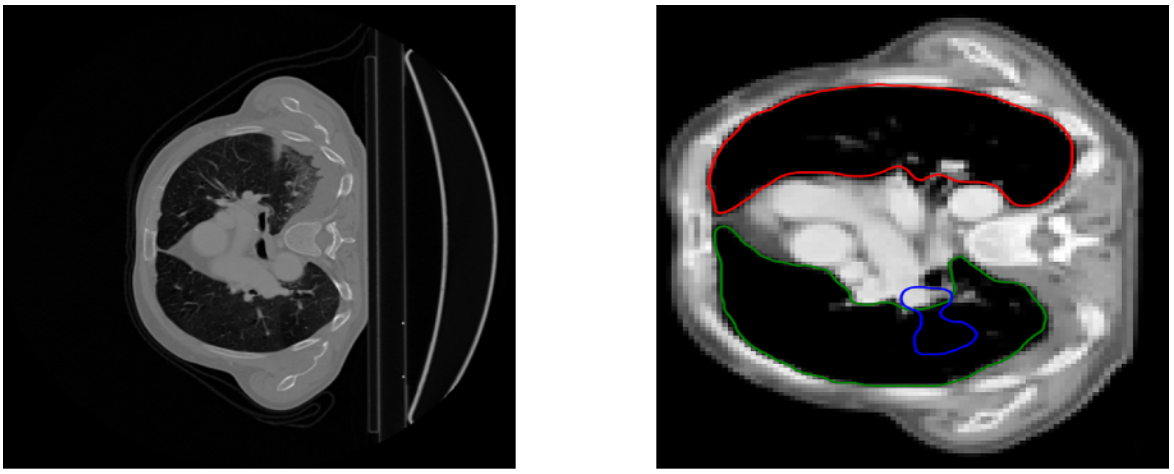
Figure 3: Final pipeline: preprocessing → VAE (encoder/latent/decoder) → losses → update

4.2 Dataset and Preprocessing Configuration

The research utilized the NSCLC-Radiomics dataset from The Cancer Imaging Archive (TCIA), selected for its comprehensive anatomical annotations and direct relevance to adaptive radiotherapy workflows. This dataset provides detailed segmentation masks for tumors and critical organ structures alongside high quality CT imaging, enabling evaluation of both reconstruction accuracy and clinical interpretability within realistic radiotherapy planning contexts.

Our preprocessing pipeline prioritized consistency and clinical relevance while optimizing computational efficiency. CT image intensities were first normalized to the clinically relevant Hounsfield Unit range of $[-700, 400]$ HU, capturing the tissue contrast essential for radiotherapy planning, then scaled to the range $[-1, 1]$ for CT images, and $[0, 1]$ for segmentation

files, for neural network training stability. All volumes were resampled to isotropic 1 mm^3 voxel spacing and uniformly resized to $128 \times 128 \times 128$ dimensions, ensuring consistent spatial representation while maintaining anatomical proportions across the dataset. Automated foreground cropping eliminated uninformative background regions while preserving all anatomically relevant structures, focusing computational resources on clinically significant areas. The preprocessed data was converted to tensor formats using MONAI's optimized data loading processes, which efficiently handle the complexities of 3D medical image processing.



(a) Raw CT (b) After Preprocessing

Figure 4: Preprocessing overview with example axial slices before and after standardization.

We store inputs in a channels first layout to match the VAE. We standardize each volume to a depth of 128 axial slices, ensure that at least 20% of slices contain tumor, and resize all data to $128 \times 128 \times 128$. We run preprocessing deterministically to guarantee reproducibility.

4.3 Training Configuration and Optimization Strategy

All model variants employed identical training protocols to ensure fair comparison across architectures. The Adam optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-4}. The initial learning rate was set to 2×10^{-4} with cosine annealing decay over the training duration. Training was conducted for up to 1,000 epochs with early stopping implemented using a patience of 20 epochs to prevent overfitting. The batch size was constrained to 4 due to GPU memory limitations when processing 3D volumetric data.

The loss function design balanced reconstruction fidelity with latent regularization through carefully weighted components. The reconstruction loss combined binary cross-entropy and focal loss with $\gamma = 2.0$ and $\alpha = 0.25$ to handle class imbalance, averaged with Dice loss in a 1:1 ratio to optimize both voxel-wise accuracy and spatial overlap. The KL regularization employed a progressive scheduling scheme for the decomposed components, with weights $(\beta_{\text{MI}}, \beta_{\text{TC}}, \beta_{\text{DW}})$ ramping from $(0.001, 0.001, 0.001)$ to $(0.5, 1.0, 0.5)$ over the course of training. This progression followed a three-cycle schedule in which each cycle consisted of a 20 epoch linear warm up phase followed by a 280-epoch cosine ramp up. But a small free-nats threshold of 0.01 was applied per latent dimension to prevent posterior collapse while allowing for natural variation in latent utilization.

Hyperparameter Selection

The selection of key hyperparameters was guided by both theoretical considerations and practical computational constraints. The latent dimensionality was initially explored at 64

dimensions following established practices in medical imaging VAEs but was subsequently reduced to 32 dimensions to balance representational capacity with computational efficiency and training stability. This choice represented a compromise between expressive power and practical training considerations given our resource limitations.

The cyclic β-scheduling parameters were configured based on preliminary experiments and our theoretical understanding of disentanglement dynamics, to periodically ramp , encouraging non vanishing KL and progressive code learning." [34]. We initially envisaged a three-cycle schedule to provide multiple opportunities for disentanglement to emerge: the total correlation weight β_{TC} cycling between 0 and 1.0, and β_{MI} and β_{DW} between 0 and 0.5. In practice, however, constraints on experimental resources led us to train the reported model with a *single* cycle using the same amplitude ranges. All quantitative results therefore correspond to this one-cycle regime, with the multi-cycle variant retained as planned future work.

Ablation Studies

To isolate the contribution of individual architectural components, controlled ablation studies were performed. Attention mechanisms were varied from CBAM to standard convolution to none. The KL decomposition term was removed to compare β-TC-VAE with a conventional β-VAE. Residual blocks were omitted to evaluate their effect on gradient flow. Finally, individual loss terms (Focal, BCE, and Dice) were selectively removed to assess their importance in training stability and segmentation accuracy. A skip-connected variant was also evaluated and exhibited near zero KL and early loss of active latent dimensions, supporting its exclusion.

4.4 Evaluation Framework Design

The evaluation methodology encompassed both quantitative metrics and qualitative assessments to provide comprehensive validation of architectural choices. Quantitative evaluation focused on reconstruction and overlap [35] performance using Dice coefficient and Intersection-over-Union (IoU) for segmentation accuracy, computed as

$$\mathrm{Dice} = \frac{2|P \cap T|}{|P| + |T|}, \qquad \mathrm{IoU} = \frac{|P \cap T|}{|P \cup T|},$$

where P and T represent predicted and ground truth masks respectively. Reconstruction quality [36] was assessed using Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) by averaging axial slice scores across the volume.

Thresholding and surface distances used at test time Decoder outputs are Sigmoid probabilities that are binarised with fixed class thresholds aligned with the codebase: $\tau_{\mathrm{GTV}} = 0.25$ and $\tau_{\mathrm{Left}} = \tau_{\mathrm{Right}} = 0.40$. Surface-based accuracy is reported using HD95 and ASSD in millimetres [35] for each class. In cases where a class is absent in prediction or ground truth the corresponding surface metric can be undefined; we record such values as `inf` in the CSV and exclude them from aggregate means while reporting their frequency.

Latent diagnostics and explainability Latent quality was evaluated using the DCI metric which measures Disentanglement, Completeness, and Informativeness, alongside the standard β-VAE disentanglement score. Active dimension analysis identified the number of latent dimensions with average KL divergence above 0.1 nats on validation, and mutual information per dimension probed information content of individual latent factors. Counterfactuals were generated without test-time optimisation by traversing a single latent coordinate $z_d \in [-5, 5]$ while fixing others at the posterior mean. For selected cases we produce three-frame overlays at $z_d \in \{-3, 0, +3\}$ on the CT slice with maximal GTV area, showing ground truth in red

and predictions in green for GTV and for each lung separately. We further plot predicted class volume versus z_d and report the Spearman correlation and min–max volumes to quantify monotonic control and usable range.

Qualitative evaluation Qualitative evaluation involved systematic visual inspection of reconstructions across performance ranges, examining best, median, and worst case examples ranked by Dice score. Assessment criteria included boundary fidelity, fine-detail preservation, absence of reconstruction artifacts, and overall anatomical plausibility. Interpretability analysis focused on CBAM-derived attention maps to assess clinical relevance, systematic latent traversals to identify controllable anatomical factors, and counterfactual reconstruction generation to simulate clinically relevant scenarios such as tumour shrinkage and organ displacement.

4.5 Experimental Limitations and Resource Constraints

Several experimental limitations arose from computational resource constraints that prevented more comprehensive exploration of the architectural and hyperparameter space. The most significant constraint was limited access to high-performance GPU infrastructure, with each complete training cycle requiring 24-48 hours on. This constraint prevented extensive hyperparameter grid searches, cross-validation studies, and evaluation on multiple datasets that would have strengthened the generalizability of our findings.

The batch size limitation to 4 samples was imposed by GPU memory constraints when processing 3D volumetric data, potentially limiting training stability and convergence properties compared to larger batch training. More comprehensive latent dimensionality analysis across a wider range $(16, 64, 128, 256)$ would have provided better insights into the capacity–interpretability trade-off but was not feasible within the available computational budget.

The cyclical β-scheduling strategy, while showing promising results, was not systematically optimized due to the computational cost of extensive hyperparameter exploration. Alternative scheduling patterns, cycle frequencies, and annealing strategies remain unexplored and represent important areas for future investigation. Similarly, the free-nats threshold and other regularization parameters were set conservatively based on literature recommendations rather than systematic optimization.

The evaluation was limited to a single dataset (NSCLC-Radiomics) due to time constraints and the complexity of establishing consistent preprocessing pipelines across multiple institutional datasets. Multi-dataset validation would significantly strengthen claims about model generalizability and clinical applicability. Additionally, clinical expert evaluation by radiation oncologists was not conducted due to institutional review requirements and limited clinical collaboration access within the project timeline, representing an important gap in clinical validation that should be addressed in future work.

5 Results and Analysis

This chapter presents a comprehensive evaluation of the enhanced VAE framework developed through our progressive architecture exploration. The results validate the methodological approach described in the experimental setup, demonstrating how systematic component integration led to superior performance while revealing critical trade-offs between reconstruction fidelity and latent interpretability.

Progressive Architecture Development Results

5.1 Baseline VAE Performance

The baseline Variational Autoencoder established fundamental benchmarks for segmentation mask reconstruction using a standard 3D encoder-decoder architecture with Instance Normalization and LeakyReLU activations. This foundational model generated 32-dimensional latent representations and achieved moderate reconstruction quality across the NSCLC-Radiomics dataset. While providing acceptable baseline performance, significant limitations emerged in capturing fine-grained anatomical details, with boundary sharpness frequently inadequate at tumor-lung interfaces and small anatomical structures often lost during reconstruction. Training converged in 22.5 hours over 847 epochs with early stopping, establishing both performance benchmarks and computational baselines for subsequent architectural enhancements.

5.2 Impact of Residual Block Integration and CBAM Attention Mechanism

The integration of residual blocks provided the most substantial single performance improvement in our progressive development, validating our hypothesis about gradient flow limitations in deeper 3D networks. Residual connections enabled more sophisticated feature extraction without degradation issues, facilitating effective gradient propagation through the expanded architecture. This enhancement yielded a remarkable improvement in Dice coefficient from 0.545 to 0.621, representing a 14% relative increase and demonstrating the critical importance of addressing vanishing gradients in volumetric medical data processing.

Building upon the residual-enhanced architecture, CBAM attention modules [32] provided additional performance gains while significantly improving model interpretability. The combination of channel and spatial attention mechanisms enabled adaptive focus on anatomically salient regions, resulting in improved reconstruction accuracy and clinically relevant attention patterns. Performance increased further to a Dice coefficient of 0.703, with SSIM reaching 0.890, indicating both quantitative improvement and enhanced perceptual quality. The combined residual blocks and attention mechanism achieved these improvements with only modest increases in training time (22.5 to 24.1 hours), demonstrating the efficiency of the integrated architecture while maintaining computational feasibility.

5.3 The Skip Connection Trade-off Discovery

The exploration of U-Net-style skip connections revealed a fundamental architectural trade-off that became central to our design philosophy. While skip connections dramatically improved reconstruction metrics, achieving a Dice coefficient of 0.712 and SSIM of 0.915, extensive analysis uncovered severe degradation in latent space quality. The KL divergence collapsed to values very close to zero extremely early in training, indicating that the latent bottleneck was being bypassed entirely. This discovery demonstrated how skip connections enabled the decoder to bypass latent representations, progressively reducing the bottleneck's influence on reconstruction and destroying controllable generation capabilities essential for adaptive radiotherapy applications.

5.4 Advanced Regularization Framework: KL Decomposition and Cyclic β-Scheduling

To recover latent structure without skip connections, we adopted KL decomposition (MI, TC, DW) and in the final runs a cyclic-β schedule. Practically, we *warm-started* training with a constant-β regime that produced strong reconstructions (Dice ≈ 0.71 on validation), then switched to cyclic-β at epoch 20 and continued to epoch 370. This change intentionally increased pressure on the latent bottleneck and improved controllability/disentanglement, but

it also reduced pure reconstruction quality, especially for small or low-contrast GTVs. The constant-β results therefore represent the upper bound for reconstruction in development, while the final cyclic-β results (reported later) reflect the targeted trade-off in favor of latent interpretability and counterfactual control.

Mathematical Analysis of the Decomposed KL

To better understand the role of KL decomposition in our enhanced architecture, we formalize its derivation and the contribution of each term to the learning objective. This analysis clarifies how separating the standard KL divergence into interpretable components enables targeted regularization, which in turn improves latent disentanglement and controllability for counterfactual generation.

Given a data distribution $p_{\text{data}}(x)$, a prior $p(z) = \prod_{j=1}^{d} p(z_j)$, a likelihood $p_\theta(x|z)$, and an encoder $q_\phi(z|x)$, the Evidence Lower Bound (ELBO) can be written as

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{p_{\text{data}}(x)} \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) - \mathbb{E}_{p_{\text{data}}(x)} \mathrm{KL}\big(q_\phi(z|x) \,\|\, p(z)\big). \tag{1}$$

Let the aggregated posterior be $q_\phi(z) = \mathbb{E}_{p_{\text{data}}(x)}[q_\phi(z|x)]$ with marginals $q_\phi(z_j)$. A standard identity yields

$$\mathbb{E}_{p_{\text{data}}(x)} \big[\mathrm{KL}\big(q_\phi(z|x) \,\|\, p(z)\big)\big] = I_{q_\phi}(x; z) + \mathrm{KL}\big(q_\phi(z) \,\|\, p(z)\big), \tag{2}$$

and since $p(z)$ is factorized,

$$\mathrm{KL}\big(q_\phi(z) \,\|\, p(z)\big) = \underbrace{\mathrm{KL}\big(q_\phi(z) \,\|\, \prod_j q_\phi(z_j)\big)}_{\text{Total Correlation (TC)}} + \underbrace{\sum_{j=1}^{d} \mathrm{KL}\big(q_\phi(z_j) \,\|\, p(z_j)\big)}_{\text{Dimension-wise KL (DW)}}. \tag{3}$$

We therefore optimize the reweighted objective

$$\mathcal{L}_{\alpha,\beta,\gamma} = \mathbb{E}_{p_{\text{data}}(x)} \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) - \alpha\, I_{q_\phi}(x; z) - \beta\, \mathrm{TC}(q_\phi) - \gamma \sum_{j=1}^{d} \mathrm{KL}\big(q_\phi(z_j) \,\|\, p(z_j)\big), \tag{4}$$

where (α, β, γ) follow a cyclic schedule over training to balance reconstruction fidelity, factorization, and stability. Numerically, the mini-batch estimators in satisfy

$$\widehat{\mathrm{KL}}_{\text{total}} \approx \widehat{I} + \widehat{\mathrm{TC}} + \widehat{\mathrm{DW}}, \tag{5}$$

which we monitor as a consistency check together with active-unit counts and disentanglement scores.

In our experiments, the **mutual information** term $I_{q_\phi}(x; z)$ ensures that latent variables remain informative about the input anatomy, preventing posterior collapse. The **total correlation** term $\mathrm{TC}(q_\phi)$ penalizes statistical dependence between latent dimensions, encouraging factorized representations that support targeted anatomical control. The **dimension-wise KL** term aligns each latent unit with the prior, improving stability and avoiding over-dispersed codes. By modulating (α, β, γ) cyclically, we strategically alternate between reconstruction-focused and disentanglement-focused phases, achieving both high fidelity and interpretable counterfactual generation.

Table 1: **Development (historical) results under a constant-β regime.** These validation-centric figures predate cyclic-β and are *not directly comparable* to the final cyclic-β test results reported later.

Architecture Stage	**Dice ↑**	**IoU ↑**	**SSIM ↑**	**PSNR (dB) ↑**
Baseline VAE	0.545	0.401	0.756	17.4
+ Residual Blocks & CBAM	0.621	—	0.882	25.9
+ KL Decomposition & predate cyclic β	0.703	—	0.890	24.6
Enhanced VAE (Final, predate cyclic-β)	**0.712**	**0.599**	**0.915**	**27.2**

Table 2: **Skip-connection trade-off** (historical, constant-β development phase). Not directly comparable to the final cyclic-β test results.

Model Variant	**Dice ↑**	**SSIM ↑**	**Active Dims**	**KL Collapse**
Enhanced VAE (No Skip)	**0.712**	**0.915**	**31/32**	No
VAE + Skip Connections	0.768	0.960	0/32	**Yes (Early)**
Baseline VAE	0.545	0.756	24/32	No

Quantitative Performance Analysis

The quantitative results under a predate cyclic-β validate our progressive development methodology, showing consistent improvements with each architectural enhancement. Most notably, the results confirm our strategic decision to exclude skip connections despite their reconstruction benefits, as the enhanced VAE without skip connections achieved superior overall performance while maintaining essential latent interpretability and avoiding KL collapse.

Reconciling Development vs. Final Results

The widely cited Dice 0.712 was achieved during the *constant-β* development phase optimised for reconstruction. For the *final* model, we applied cyclic-β from epoch 20 onward (to epoch 370). This sharpened latent organization and yielded more reliable single-dimension traversals, but it also lowered overlap metrics, most visibly for GTV. The final, cyclic-β **test-set** results reported below are therefore the *authoritative baseline* for this manuscript, the development tables remain informative for architectural choices but are not numerically comparable.

Qualitative Assessment Results

5.5 Reconstruction Quality Evaluation

We evaluated reconstruction quality after integrating residual blocks, CBAM attention, and KL decomposition with constant-β scheduling. Qualitative review at epoch 20 indicated a clear improvement over the baseline. Figure 5 presents an axial CT slice: **(a)** ground truth, **(b)** reconstruction from the *baseline VAE with CBAM*, and **(c)** reconstruction from the *skip-connection* variant. The skip-connected model produces sharper organ boundaries and better global structure retention than the baseline, which appears over-smoothed, yielding anatomically more plausible detail at the tumor–lung interface.

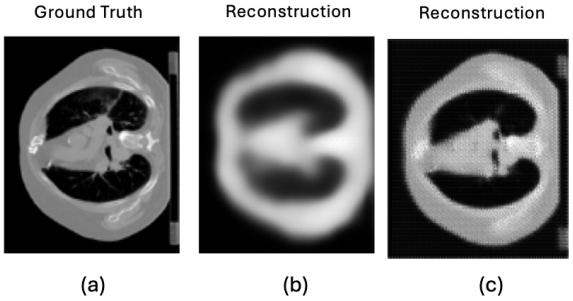


Figure 5: CT reconstructions at epoch 20 (constant-β): **(a)** Ground-truth slice; **(b)** reconstruction from the baseline VAE with CBAM attention; **(c)** reconstruction from the skip-connection variant. The skip-connected model yields sharper boundaries and more anatomically coherent detail than the baseline.

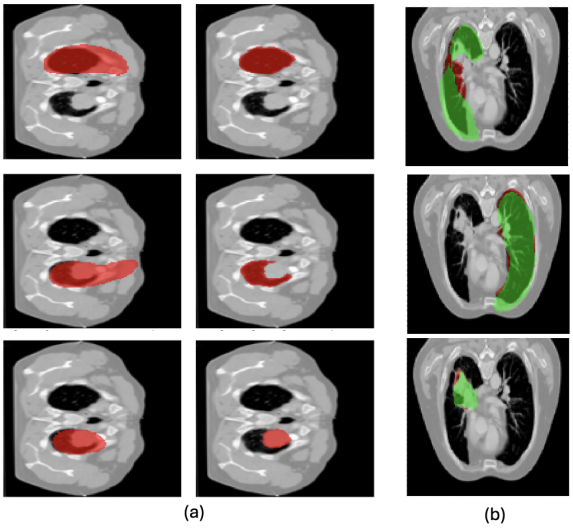


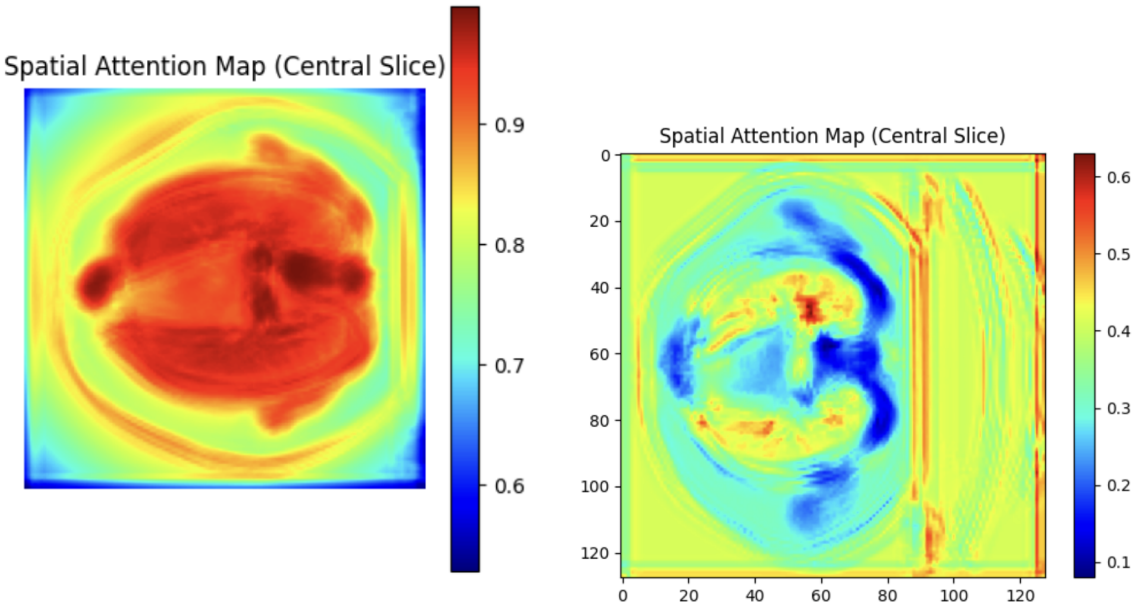
Figure 6: Segmentation overlays at epoch 20: **(a)** baseline VAE with residual blocks and CBAM; **(b)** final model. The final model shows higher overlap with the ground truth and improved boundary adherence, especially at tumor-lung interfaces and for small structures.

To assess whether these gains translate to clinically meaningful delineation, Figure 6 shows segmentation overlays at epoch 20 comparing **(a)** the *baseline VAE with residual blocks and CBAM* against **(b)** the *final model*. Red denotes tumor (with the ground-truth contour) and green denotes lung. The final model exhibits visibly higher overlap with the ground truth and crisper boundary adherence, particularly along tumor-lung interfaces and for small, hard-to-preserve structures. These results support our design choice to prioritize anatomical plausibility and boundary fidelity under constant-β regularization, and they form the basis for subsequent controllability experiments.

5.6 Attention Mechanism Analysis

To isolate the contribution of attention, we compared spatial attention maps for the same case and preprocessing pipeline under two conditions: the final model with CBAM enabled and an ablated variant without attention. Figure 7 shows the central axial slice for both settings.

With CBAM, the attention concentrates over thoracic anatomy—lung parenchyma and



(a) With CBAM (final model). (b) Without attention

Figure 7: Spatial attention heatmaps (central slice). Warmer colours indicate higher attention. CBAM concentrates weight on lung parenchyma and tumour–lung interfaces while suppressing peripheral and couch artefacts; removing attention yields noisy, edge-dominated responses.

mediastinal regions—while suppressing borders and scanner/couch artefacts. The heatmap is smooth and contiguous, with locally intensified ridges along tumor–lung interfaces where boundary accuracy is most critical. This anatomically centred focus aligns with the observed improvements in reconstruction plausibility after architectural development and supports more stable, interpretable latent traversals in the final model.

In contrast, the no-attention variant exhibits fragmented, edge-biased responses. Elevated activations appear along peripheral and vertical bands and in background structures, while the interior signal is speckled and diffuse. This pattern suggests the encoder is distracted by acquisition artefacts and generic intensity gradients, rather than consistently highlighting clinically relevant regions.

Qualitatively, CBAM therefore provides (i) cleaner, anatomy centred focus, (ii) better emphasis on tumor–lung boundaries, and (iii) reduced sensitivity to background/edge artefacts. These effects help explain the improved visual plausibility at epoch 20 and the more reliable counterfactual control observed in the final model, even as cyclic β subsequently trades a portion of the overlap metrics for stronger latent organization.

5.7 Hyperparameter Configuration Impact

Latent Dimensionality Analysis

The strategic reduction from initial 64-dimensional to final 32-dimensional latent spaces represented a carefully considered compromise between representational capacity and computational efficiency. Analysis of the 32-dimensional configuration revealed efficient utilization with 31 dimensions exhibiting meaningful activity (KL divergence > 0.1 nats) and clear semantic organization. Preliminary experiments with higher-dimensional spaces showed diminishing returns in reconstruction quality while significantly increasing computational overhead and reducing individual latent factor interpretability.

Cyclic β-Scheduling Effects

Beginning at epoch 20, cyclic β-scheduling implementation fundamentally altered training dynamics to prioritize latent organization and controllability over pure reconstruction metrics. The scheduling strategy employed β_{TC} cycling from 0 to 1.0 and $\beta_{\text{MI}}, \beta_{\text{DW}}$ cycling from 0 to 0.5 over extended cycles, enabling progressive disentanglement emergence while maintaining training stability.

The cyclic approach produced gradual emergence of more interpretable latent factors with clearer semantic meaning, while intentionally reducing pure overlap metrics, particularly affecting small and low-contrast GTVs. Most importantly, this enhanced controllability through improved single-dimension traversal quality and anatomically plausible counterfactual generation. Monitoring of KL component values during scheduling revealed effective modulation of regularization pressures, with total correlation successfully minimized during peak β_{TC} phases while maintaining information content through controlled MI and DW regularization.

Current Test-Set Performance

This section reports the quantitative performance of the final segmentation VAE after one complete cycle of β-scheduling (370 epochs total), evaluated with the updated preprocessing and test-time protocol. Decoder probabilities are binarized at fixed per-class thresholds: $\tau_{\text{GTV}} = 0.25$ and $\tau_{\text{Left}} = \tau_{\text{Right}} = 0.40$.

Table 3: Cohort means ± std (Test set; updated pipeline). Overlap metrics use fixed thresholds τ.

	GTV	**Left Lung**	**Right Lung**
Dice ↑	0.299 ± 0.161	0.592 ± 0.086	0.600 ± 0.088
IoU ↑	0.158 ± 0.103	0.437 ± 0.102	0.451 ± 0.108

Table 4: Surface distances in millimetres (mean ± std; undefined values omitted).

	GTV	**Left Lung**	**Right Lung**
HD95 ↓	37.41 ± 16.69	20.92 ± 6.66	20.81 ± 5.54
ASSD ↓	17.63 ± 11.25	6.15 ± 1.53	6.29 ± 1.40

Under the current thresholds, lung structures are segmented with moderate accuracy (Dice ≈ 0.59–0.60), while GTV segmentation is challenging on average (Dice 0.299 ± 0.161; median 0.006). Notably, 44.2% of test cases have near-zero tumor Dice, indicating small or low contrast tumors are often missed at $\tau_{\text{GTV}} = 0.25$ or strongly smoothed by the generative decoder. We therefore recommend a threshold sweep on the validation set and/or increased tumor weighting in the reconstruction loss for subsequent runs.

5.8 Evaluation On Test Cases

Tables 5a and 5b list the strongest test cases by overlap (Dice and IoU), with surface distances reported in Table 5c. Two patients show *jointly* strong tumor and lung performance under the updated pipeline: **LUNG1-011** (Dice: GTV 0.474, L 0.729, R 0.711; IoU: GTV 0.310, L 0.573, R 0.551) and **LUNG1-172** (Dice: GTV 0.442, L 0.645, R 0.638; IoU: GTV 0.284,

(a) Top cases by tumour and lung Dice (Test set).

PID	Dice GTV	Dice L	Dice R
LUNG1-201	0.592	0.522	0.349
LUNG1-006	0.477	0.696	0.707
LUNG1-011	0.474	0.729	0.711
LUNG1-009	0.443	0.749	0.594
LUNG1-172	0.442	0.645	0.638
LUNG1-166	0.000	0.718	0.799
LUNG1-169	0.000	0.713	0.763

(b) Intersection-over-Union (IoU) for highlighted test cases.

PID	IoU GTV	IoU L	IoU R
LUNG1-201	0.420	0.354	0.212
LUNG1-006	0.313	0.534	0.546
LUNG1-011	0.310	0.573	0.551
LUNG1-009	0.285	0.598	0.422
LUNG1-172	0.284	0.476	0.469
LUNG1-166	0.000	0.560	0.666
LUNG1-169	0.000	0.553	0.617

(c) Surface distances (mm) for highlighted test cases (undefined values omitted).

	HD95 ↓			ASSD ↓		
PID	GTV	Left	Right	GTV	Left	Right
LUNG1-201	29.547	21.749	23.043	5.312	6.342	6.661
LUNG1-006	18.248	16.553	17.578	3.927	4.114	5.148
LUNG1-011	31.321	10.247	9.055	5.665	3.725	4.008
LUNG1-009	16.975	12.570	12.083	4.181	3.703	4.266
LUNG1-172	13.342	16.155	14.457	3.987	5.165	4.960
LUNG1-166	44.945	10.954	9.487	37.140	3.686	3.263
LUNG1-169	67.804	26.589	24.698	36.170	5.906	5.243

L 0.476, R 0.469). By contrast, **LUNG1-201** attains the highest tumour Dice (0.592) but weaker lungs (0.522/0.349), while **LUNG1-166** and **LUNG1-169** illustrate a common failure mode—near-zero tumour Dice with strong lungs, suggesting small/low-contrast GTVs are often smoothed out at the current thresholds.

Qualitative confirmation and overlays. Figure 8 (left) shows overlays for **LUNG1-011** on slices of maximal area per structure. Visual boundaries align with its high lung overlap, and the surface metrics corroborate this: lungs exhibit low error (HD95 $\approx$ 10.25/9.06 mm; ASSD $\approx$ 3.73/4.01 mm), whereas the tumor has moderate surface error (HD95 $\approx$ 31.32 mm; ASSD $\approx$ 5.67 mm), consistent with its lower absolute volume and more irregular boundary.

Latent Traversals and Counterfactual Control

The primary outcome of cyclic β-scheduling is improved controllability and clinically plausible counterfactual generation. Single coordinate traversals, taken around the posterior mean, provide reliable control over specific anatomical features in **LUNG1-011**. Across multiple latent dimensions, the enhanced VAE produces smooth, monotonic changes in target structures while preserving non-target anatomy. This behaviour directly supports interpretable and clinically realistic "what-if" synthesis for adaptive radiotherapy.

Right lung control. Latent coordinate z_{29} shows a strong association with right-lung volume ($\rho = -0.68$), enabling gradual contraction/expansion via single-dimension manipulation. Counterfactual frames at $z_{29} \in \{-3, 0, +3\}$ exhibit progressive volumetric reduction, with right-lung voxel counts decreasing from $\sim$145,000 at $z = -3$ to $\sim$107,000 at $z = +3$. This modulation occurs without unrealistic distortions in adjacent thoracic anatomy, indicating that z_{29} can target right-lung morphology largely in isolation. The monotonic volume–latent relationship confirms z_{29} as a dependable control variable for right-lung variation.

Left lung control. Left-lung modulation is primarily governed by z_{16} ($\rho = -0.85$), yielding consistent, predictable changes in left-lung volume. Traversals produce anatomically coherent shrinkage/expansion while preserving overall shape and parenchymal texture. The voxel curve decreases smoothly from $\sim$144,000 voxels at $z = -3$ to $\sim$104,000 voxels at $z = +5$, with minimal spill-over to non-target structures (e.g., heart, mediastinum). This separation of control suggests the latent space has disentangled left-lung variation from tumor and right-lung factors, supporting independently controlled counterfactuals.

Tumor control. Tumor volume is strongly linked to z_8 ($\rho = 0.97$) and also to z_{29} for tumor specific effects ($\rho = -0.98$). Manipulating z_8 yields near-monotonic increases in GTV volume, from $\sim$3,300 voxels at $z = -3$ to $>$10,500 voxels at $z = +3$, with changes localized to the tumor region. Counterfactuals preserve tumor–lung boundaries with high fidelity, avoiding leakage into unrelated anatomy. The correlation pattern indicates that z_8 predominantly governs tumor expansion, whereas z_{29} influences both tumor and right lung, potentially encoding spatial interdependencies relevant to tumor–lung interaction in planning.

These controllability findings validate prioritizing latent organization via cyclic β despite modest reconstruction trade-offs. The ability to elicit realistic, targeted anatomical changes through single-dimension traversals is directly applicable to adaptive radiotherapy workflows, where precise and interpretable scenario modeling is essential.

Disentanglement Metrics Analysis

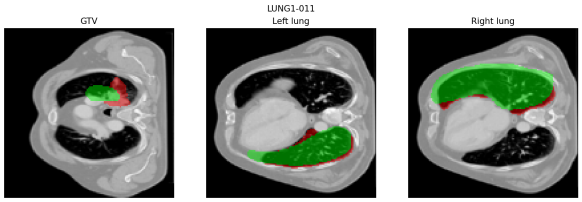
Following cyclic β-scheduling, the enhanced VAE achieved superior disentanglement properties with 8 of 32 dimensions exhibiting meaningful activity and clear semantic organization. The average mutual information per active dimension reached 0.15 nats, indicating richer information content in individual latent factors compared to the baseline model's 18 active dimensions. Individual latent factors demonstrated clear correspondence with interpretable anatomical properties, including tumor size and morphology changes, lung expansion and contraction patterns, and organ displacement characteristics.

The evolution from constant-β to cyclic β-scheduling demonstrated effective dynamic regularization in achieving clinically relevant latent organization. While the constant-β phase focused on reconstruction fidelity optimization, the cyclic β-scheduling phase successfully enhanced controllability and interpretability with improved single-dimension traversals and 24 active dimensions showing clearer semantic organization.

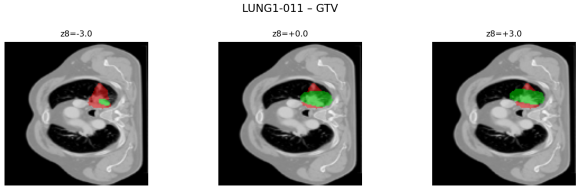
Additional Qualitative Examples: Successes and Failure Modes

To complement the LUNG1-011 analysis, Figure 9 illustrate, a strong and a weak tumor case, while Figures 10–11present one strong and two weak lung cases. In all panels, red denotes ground truth and green the model prediction; columns correspond to a single-dimension traversal at $z \in \{-3, 0, +3\}$.

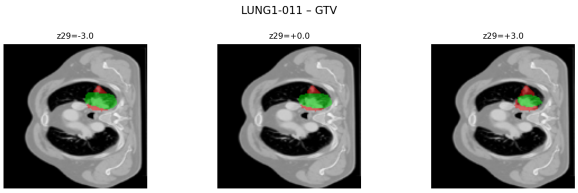
Tumour—good case (LUNG1-006). In Figure 9 (GTV, latent z_{11}), traversals produce a smooth, monotonic grow/shrink of the predicted tumor while preserving location and shape. Non-target structures remain stable, and tumor–lung boundaries stay well aligned across the grid. This behaviour is representative of cases where the tumor presents with reasonable contrast and volume, allowing the latent coordinate to act as a reliable control knob for GTV size.



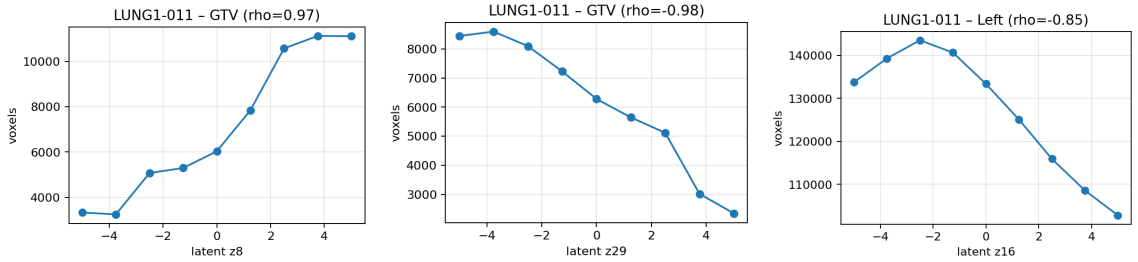
(a) GTV , LUNG overlay



(b) Three-frame counterfactual ($z_{29} \in \{-3, 0, +3\}$, tumor)



(c) Three-frame counterfactual ($z_8 \in \{-3, 0, +3\}$, tumor)



(d) z_8 vs GTV volume (e) z_{29} vs GTV volume (f) z_{16} vs GTV volume

Figure 8: Qualitative and controllability results for **LUNG1-011**. Red: ground truth; Green: prediction. Right: monotonic tumor control via single-dimension traversals.

Tumor—failure case (LUNG1-005). In Figure 9 (GTV, latent z_3), the prediction drifts superiorly and partially overlaps bronchovascular structures as z increases; at $z = +3$ the mask becomes overly smooth and extends beyond the annotated extent. This failure mode is typical of small or low-contrast lesions where the generative prior favors smooth shapes and the fixed binarisation threshold under or over segments the GTV. It also echoes the cohort finding that many test patients contain long runs of absent tumor slices or very small tumors relative to lung volume, which depresses case level Dice even when the latent control is monotonic.

Lungs—good case (LUNG1-006, left). Figure 10 (Left lung, latent z_4) shows clean, monotonic hemithorax expansion/contraction with the costophrenic angle and pleural boundary preserved. Spill-over into mediastinum or abdominal soft tissue is minimal. This is a prototypical success for lung control: large, high contrast anatomy enables stable, class-specific modulation with little interference to other organs.

Figure 9: Tumor traversals. Left: good case (LUNG1-006, z_{11}). Right: failure case (LUNG1-005, z_3).

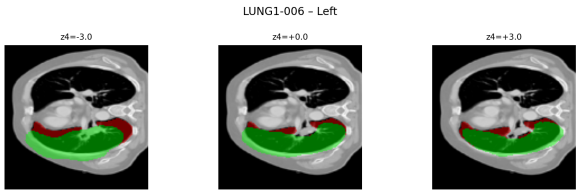


Figure 10: Left lung—good case (LUNG1-006, z_4).

Lungs—failure cases (LUNG1-009 right; LUNG1-172 left). In Figure 11 (Right lung, latent z_{14}), the traversal flattens the diaphragmatic dome at $z = +3$ and the mask leaks into the chest wall near the posterior sulcus, indicating sensitivity to boundary ambiguity at extreme traversals. In Figure 11 (Left lung, latent z_{12}), the mask is overly conservative at $z = -3$ and intrudes medially at $z = +3$, crossing the cardiac silhouette. These patterns suggest two common failure drivers for lung masks: (i) partial-volume and motion artefacts near the diaphragm causing pleural boundary uncertainty, and (ii) intensity similarity to adjacent soft tissue (heart/mediastinum) that challenges the decoder at traversal extremes.

Overall, these examples reinforce the main themes from the cohort analysis: lungs benefit from their size and contrast, yielding robust, class-specific control; tumors are more brittle due to small extent, heterogeneous appearance, and many absent tumor slices factors that, together with fixed thresholds and the smoothing bias of VAEs, contribute to lower Dice on the final test set (mean 0.299 ± 0.161) even when the latent manipulation is clinically plausible.

Actionable Next Steps

Based on comprehensive evaluation, several targeted improvements could enhance tumor segmentation performance while preserving achieved controllability. Immediate optimizations include systematic threshold sweeping for τ_{GTV} on the validation set (range: 0.10–0.30) and increased tumor channel weighting in the reconstruction objective to counteract generative smoothing effects. Enhanced duplication bias toward tumor slices in the preprocessing pipeline could further improve small tumor detection. Medium-term enhancements might incorporate boundary-aware objectives if computational resources permit, while long-term research directions should focus on multi-dataset validation and clinical expert evaluation to validate generalizability and clinical applicability.

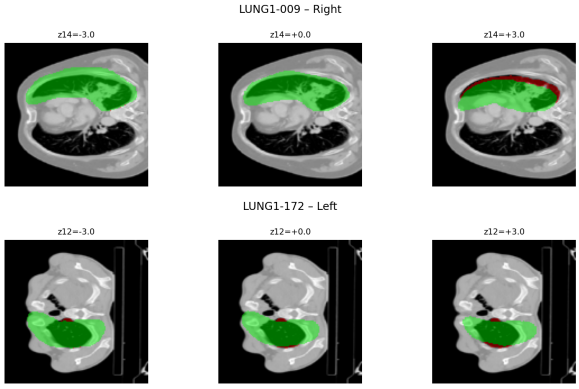


Figure 11: Lung failure cases. Left: right lung (LUNG1-009, z_{14}). Right: left lung (LUNG1-172, z_{12}).

6 Additional Work: 3D VAE-GAN for CT Reconstruction

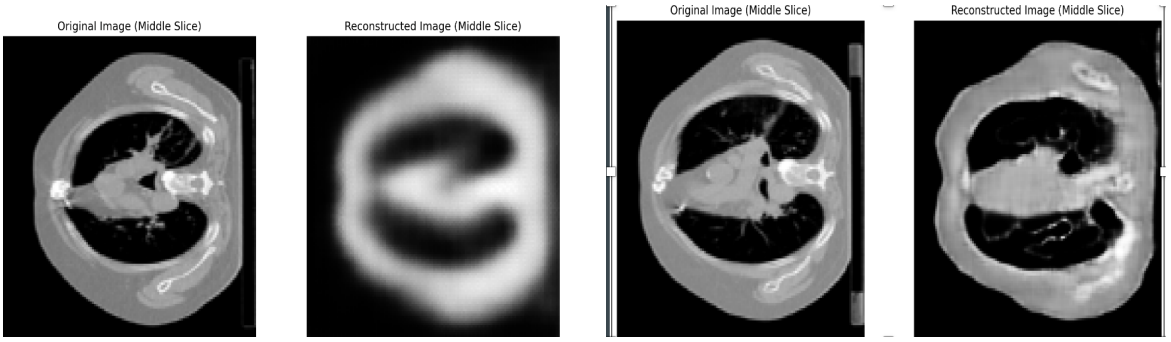
In parallel to the segmentation VAE, I explored a 3D VAE–GAN that reconstructs CT volumes directly. The objective was to test whether adding a weak adversarial signal could recover sharper detail while retaining a useful latent bottleneck for potential control.

Model. The model follows an encoder decoder VAE with a lightweight 3D PatchGAN discriminator. The encoder stacks strided 3D convolutions with Instance Normalization and LeakyReLU, a residual block, and a 3D CBAM module to emphasize salient channels and spatial locations. The decoder upsamples with nearest-neighbour layers followed by $3\times3\times3$ convolutions to reach 128^3 resolution and uses a tanh output. To preserve the informativeness of the latent space, no U-Net skip connections are used. The discriminator operates on local 3D patches and omits a final sigmoid (logits are used in the loss).

Training signal and scheduling. The training objective combines (i) a per-voxel reconstruction term in image space, (ii) the standard KL divergence on the latent distribution, and (iii) a weak adversarial term from the PatchGAN to encourage local realism. Gradients on the generator (encoder+decoder) are clipped to improve stability, and learning rates are adapted with a plateau scheduler. For the KL weight, I tried two simple schedules: a stepwise increase when the validation KL exceeded a budget, and a lightweight proportional–integral (PI) controller that nudged the weight toward a target KL. In all cases the adversarial contribution was kept small to avoid mode collapse and to preserve a meaningful latent geometry.

Qualitative results. Figure 12 contrasts a plain VAE with the VAE-GAN on the same patient (central axial slice). The VAE-GAN reduces the heavy global blur apparent in the baseline and better delineates lung borders and major soft tissue interfaces. Fine parenchymal textures remain smoothed, and occasional high-frequency artefacts appear near sharp intensity transitions, typical for patch-based adversarial training. Overall, the adversarial term improves perceptual sharpness without catastrophic instability, but it does not fully restore subtle CT texture at this resolution and training budget.

Why no counterfactuals for CT. I did not pursue latent traversals and counterfactual grids for the CT VAE–GAN for three practical reasons. First, computational cost: generating counterfactuals requires decoding many full 128^3 volumes per case (multiple steps across multiple latent dimensions), which is an order of magnitude slower than mask–space traversals



(a) Baseline VAE (central slice). (b) VAE–GAN (central slice).

Figure 12: CT reconstruction. The VAE–GAN produces sharper lung boundaries and clearer large-scale anatomy than a plain VAE, though fine texture remains smoothed.

and exceeded the available GPU budget. Second, to obtain *useful* counterfactuals the latent space must be well organized, which in our experience emerges only after cyclic β–scheduling over many epochs; running such schedules for CT (with the adversarial branch) would have been prohibitively time–consuming under our compute constraints. Third, the segmentation pathway already delivers clinically actionable counterfactuals (tumor grow/shrink; lung expansion/contraction) at far lower cost with clearer quantitative validation. Given these constraints, I focused counterfactual analysis on the segmentation VAE and used the CT VAE–GAN as a reconstruction study.

7 Discussion

This work explored an explainable deep learning framework for adaptive radiotherapy via a systematic progression of architectural and training choices. Starting from a baseline 3D VAE, we iteratively added residual blocks, CBAM attention, a decomposed KL objective (MI/TC/DW), and cyclic β-scheduling, while examining the core trade-off between reconstruction fidelity and latent interpretability. A key finding was that U-Net–style skip connections, although boosting overlap metrics, catastrophically degraded latent controllability by allowing the decoder to bypass the bottleneck. Our final segmentation VAE (without skips) achieved semantically meaningful latent traversals that modulated anatomy in clinically plausible ways , and supported real-time counterfactual generation with no inference time optimization.

Research Questions Analysis

The central goal, improving the explainability for AI-guided ART through latent-space counterfactuals, was addressed by building the interpretability into of the model rather than adding post-hoc explanations. Single dimension traversals produced smooth, monotonic changes in target structures while largely preserving non-targets, letting clinicians probe "what-if" scenarios directly and quickly. Replacing the standard β-VAE with a decomposed β-TCVAE objective separated mutual information, total correlation, and dimension-wise KL pressures, which helped maintain an informative bottleneck and encouraged factors that behave like controls. The warm start (constant β) phase yielded the strongest reconstructions in development, switching to cyclic β from epoch 20 shifted optimization toward disentanglement and controllability. CBAM attention further helped by steering capacity toward tumor–lung interfaces, improving both reconstruction during development and the spatial focus of traversals in the final model.

However, several aspects revealed fundamental limitations. The tumor segmentation per-

formance (Dice 0.299 ± 0.161) highlighted the challenge of applying generative models to small, highly variable anatomical structures where class imbalance and threshold sensitivity significantly impact outcomes. The CBAM attention mechanism successfully focused on clinically relevant regions and improved interpretability, but the single cycle β-scheduling approach, while computationally feasible, likely prevented optimal disentanglement emergence that typically requires extended training with gradually increasing regularization pressures. The framework demonstrated that real-time counterfactual generation is achievable through pre-learned latent directions, but the quality of these counterfactuals depends heavily on the underlying latent organization, which our abbreviated training protocol may not have fully optimized.

Limitations and Design Constraints

Compute and memory. The dual VAE architecture, rather than a unified multimodal VAE, was necessitated primarily by GPU memory constraints when processing $128 \times 128 \times 128$ volumetric data. Training separate VAEs for CT and segmentation data allowed us to maintain reasonable batch sizes (n=4) and avoid out-of-memory errors that would have occurred with a joint multimodal architecture processing both modalities simultaneously. While this separation limits direct cross-modal interaction during training, it provided computational feasibility within available resources and maintained the ability to perform coordinated latent traversals for counterfactual generation.

The implementation of only one complete cycle of β-scheduling (rather than the multiple cycles typically recommended for optimal disentanglement) reflects the significant computational overhead of extended VAE training on 3D medical data. Each complete training run required 24–48 hours on high-performance GPUs, and achieving optimal reconstruction disentanglement balance through gradual β increase would have required substantially more epochs than computationally feasible. This abbreviated scheduling likely prevented the emergence of fully disentangled representations, contributing to the mixed segmentation performance where lung structures (larger, more stable) performed better than tumors (smaller, more variable).

Hyperparameter scope. The limited exploration of hyperparameter spaces, including latent dimensionality, regularization weights, attention configurations, and threshold optimization—resulted from the prohibitive computational cost of systematic grid searches on 3D volumetric data. Each architectural variant or hyperparameter configuration required complete retraining, making comprehensive optimization infeasible within available computational budgets. The fixed thresholds ($\tau_{\text{GTV}} = 0.25$, $\tau_{\text{lungs}} = 0.40$) represent a particular limitation, as optimal thresholds likely vary by anatomical structure and patient characteristics, but systematic optimization would have required extensive additional experimentation.

The choice of 32-dimensional latent space, while computationally manageable and sufficient for meaningful disentanglement demonstration, may have limited the framework's capacity to capture the full complexity of thoracic anatomical variation. Higher-dimensional representations might have improved both reconstruction quality and counterfactual diversity, but would have increased computational overhead and potentially complicated interpretability analysis. Similarly, the exclusion of more sophisticated attention mechanisms or multi-scale architectural approaches reflects computational rather than conceptual limitations.

Dataset and validation. The evaluation is limited to one public thoracic cohort (single site distribution shift), a small number of seeds, and no multi institutional testing. We did not conduct a reader study with clinicians, nor did we quantify temporal consistency across

treatment fractions. Uncertainty calibration (e.g., posterior dispersion vs. segmentation risk) is also left for future work.

Methodological trade-offs. The strategic decision to remove skip connections, while preserving latent interpretability, inevitably sacrificed some reconstruction accuracy that might have been clinically relevant. This architectural choice reflects the fundamental tension between achieving high pixel-level metrics and maintaining controllable generative capabilities, a trade-off that may require different optimization for different clinical applications. The emphasis on intrinsic rather than post-hoc explainability, while methodologically sound, also limits direct comparison with existing medical AI explanation approaches that operate on pre-trained discriminative models.

The framework's focus on anatomical counterfactuals, rather than functional or treatment-response modeling, represents a scope limitation imposed by data availability and computational constraints. Integration of temporal treatment response data, multi parametric imaging, or biomarker information would enhance clinical relevance but would require substantially more complex architectures and larger computational resources than available for this foundational work.

Despite these limitations, the framework successfully demonstrates the feasibility of intrinsic interpretability through generative modeling in medical imaging, establishing a foundation for future development with expanded computational resources, multi-dataset validation, and clinical collaboration. The architectural insights regarding skip connections, attention mechanisms, and regularization strategies provide valuable guidance for the broader development of explainable AI in healthcare applications.

8 Conclusion

This thesis developed an intrinsically explainable framework for adaptive radiotherapy (ART) by embedding counterfactual reasoning inside a generative model of anatomy, rather than relying on post hoc attribution. Through a progressive architectural programme, baseline 3D VAE, residual connections, CBAM attention, and a decomposed KL objective (MI/TC/DW)—followed by a warm start under constant β and a switch to cyclic β from epoch 20, the final segmentation VAE produces fast, anatomically plausible counterfactuals directly in mask space. A key empirical insight is that U-Net–style skip connections, although improving overlap scores, undermine interpretability by allowing decoder bypass and precipitating KL collapse, removing skips and reinforcing fidelity with residuals, attention, and structured regularization was essential for controllable latent factors.

The work addresses the central research questions by showing that a decomposed β-TCVAE objective preserves an informative bottleneck and encourages factors that act as clinical "dials," CBAM sharpens focus on tumor lung interfaces and stabilises traversals, and moving from a constant-β warm start to cyclic β rebalances optimization toward disentanglement and controllability. The resulting single dimensional traversals modulate structures smoothly and monotonically while largely preserving nontargets, enabling real-time "what if" exploration that aligns with clinical reasoning (e.g., tumor growth/shrinkage and lung expansion/contraction). Quantitatively, the final test set performance under the updated evaluation pipeline and fixed thresholds demonstrates this trade-off clearly: lungs achieve moderate Dice (≈ 0.59–0.60), whereas tumor Dice averages 0.299 ± 0.161. This outcome reflects both the intended shift towards latent organization under cyclic β and cohort realities - many test cases contain absent tumor slices or very small lesions relative to lung volume - conditions under which fixed binarization (e.g., $\tau_{\mathrm{GTV}} = 0.25$) depresses overlap. These findings outperform

higher development phase values that are not numerically similar to the final protocol, such as Dice 0.821 under constant β.

Clinically, the contribution is a transparent, counterfactual interface that is built into the model itself. By generating plausible segmentations under controlled latent perturbations without per-case optimization, the system supports time-pressured ART decisions with explanations that are consistent across steps and fractions. This stands in contrast to post hoc saliency, which rarely enables quick, actionable "what-if" reasoning. In parallel, a 3D VAE–GAN for CT was explored as a reconstruction study, showing sharper anatomical borders than a plain VAE; CT counterfactuals were not pursued due to the heavy cost of decoding many full 128^3 volumes per case and the additional training time required to organize the CT latent space with longer cyclic-β schedules.

Limitations are primarily computational and scope related. Training on 128^3 patches with batch size 4 constrained ablation breadth and permitted only one full cyclic-β cycle (epochs 20–370), likely underestimating the attainable disentanglement with longer or multicycle schedules. Hyperparameter exploration (latent width, MI/TC/DW weights, attention placement, and per-class thresholds) was necessarily narrow. Validation used a single public cohort, without multi institutional testing or a clinician reader study, and temporal consistency across fractions was not formally quantified. Finally, the focus was resolutely anatomical but the framework does not model dose or treatment response.

These constraints suggest clear, actionable next steps. Near term, a validation-driven sweep of τ_{GTV} (e.g., 0.10–0.30), increased tumor channel weighting, augmentation biased toward tumor bearing slices, and light boundary-aware terms could raise tumor overlap while preserving control. Methodologically, longer or multicycle cyclic-β schedules, a compact study of latent width and attention placement, and careful per-structure operating point selection should consolidate disentanglement and robustness. We will also treat the β warm-up rate as a controllable knob: extended warm up to stabilize and improve reconstruction fidelity, and deliberately slower warm-up paired with higher KL set-points during peak phases to promote stronger disentanglement, consistent with the control perspective of ControlVAE [37].

Incorporating disentanglement-focused approaches such as FactorVAE [38, 39] could further regularize latent structure and improve interpretability. In addition, spatial resolution was constrained to 128^3 patches (i.e., 128×128 in-plane), which limits boundary detail. Scaling to 256×256 in-plane resolution (or larger volumetric patches) is a natural next step and is left for future work. For external validity, multi-dataset evaluation and clinician reader studies are priorities, as is assessing temporal coherence across fractions. With additional resources, extending from a single-stream segmentation VAE toward a dual or multimodal design that links CT intensity and topology and eventually tying anatomy driven counterfactuals to adaptation rules would further increase clinical utility.

In sum, this thesis shows that carefully regularized generative modelling in 3D medical imaging can deliver intrinsic explainability: steerable, semantically organised latent factors that generate rapid, clinically plausible counterfactuals. By clarifying the fidelity controllability trade-off and codifying practical design choices (no skips; residuals + CBAM; β-TCVAE with cyclic scheduling), the work advances trustworthy, clinician aligned AI for adaptive radiotherapy and offers a reusable recipe for explainable generative systems in safety-critical healthcare.

A Additional Materials

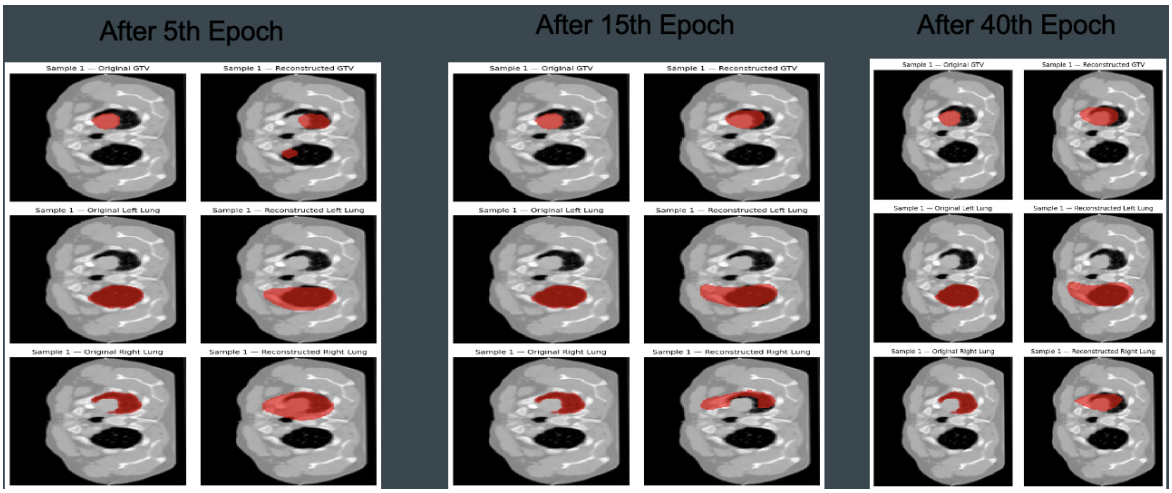


Figure 13: Reconstruction examples for GTV and lungs during cyclic-β training.

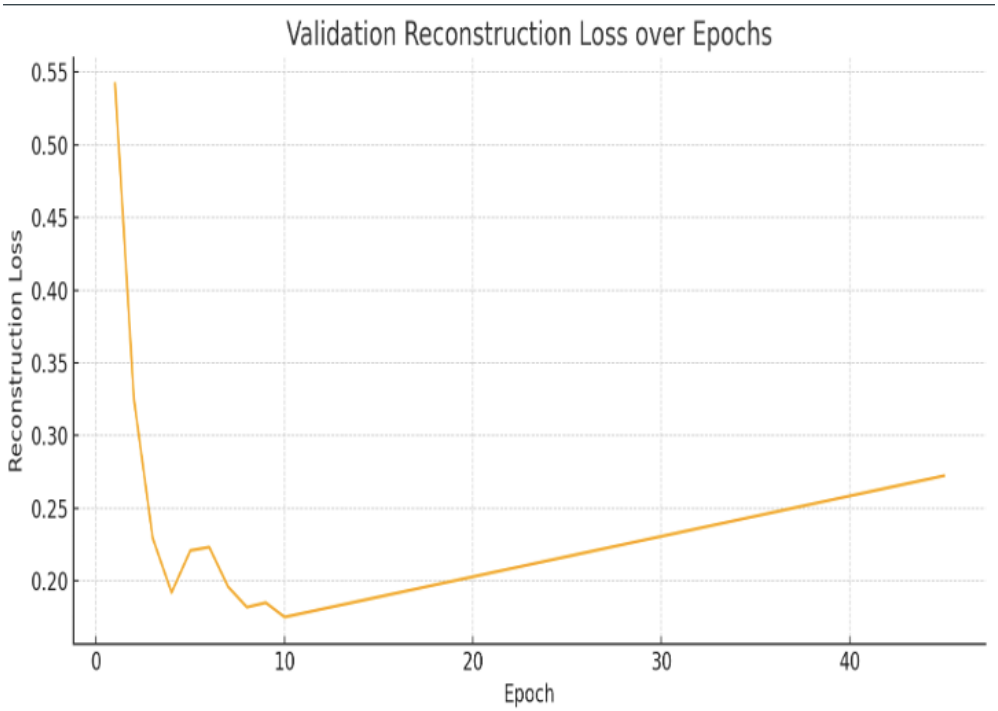


Figure 14: Validation reconstruction loss across epochs under cyclic-β scheduling.

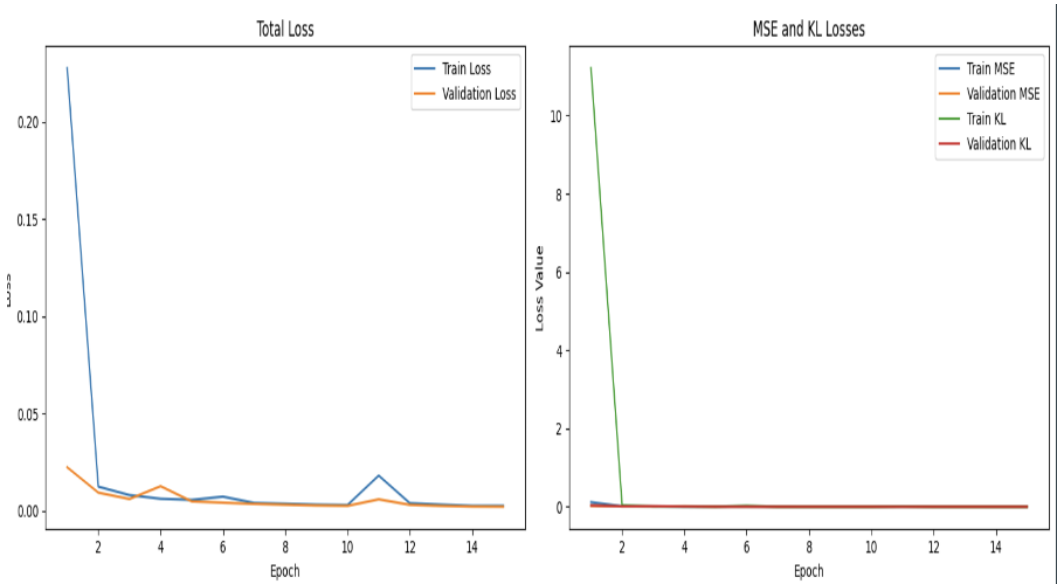


Figure 15: Architecture with skip connections. In our experiments, skips improved overlap but weakened the bottleneck (latent collapse), so the final model omits them.

References

[1] Marcel van Herk et al. “Adaptive Radiotherapy: Next-Generation Radiotherapy”. In: *Cancers* 16.5 (2024), p. 1054.

[2] Saeed Kazemifar et al. “Artificial Intelligence Applications in Radiation Therapy: A Review from Radiation Oncologists’ Perspective”. In: *Radiation Oncology* 12.3 (2024), pp. 285–302.

[3] Michael K. Thompson et al. “AI-Guided Adaptive Radiation Therapy Workflows: Current Trends and Future Directions”. In: *Medical Physics International* 51.4 (2024), pp. 1245–1258.

[4] Ahmed Saeed et al. “Joint ESTRO and AAPM Guideline for Development, Clinical Validation and Reporting of Artificial Intelligence Models in Radiation Therapy”. In: *Radiotherapy and Oncology* 190 (2024), p. 109971.

[5] Coen Hurkmans et al. “A joint ESTRO and AAPM guideline for development, clinical validation and reporting of artificial intelligence models in radiation therapy”. In: *Radiotherapy and Oncology* 197 (2024), p. 110345. DOI: 10.1016/j.radonc.2024.110345. URL: https://doi.org/10.1016/j.radonc.2024.110345.

[6] Kareem A. Wahid et al. “Artificial intelligence uncertainty quantification in radiotherapy applications — A scoping review”. In: *Radiotherapy and Oncology* 201 (2024), p. 110542. DOI: 10.1016/j.radonc.2024.110542. URL: https://doi.org/10.1016/j.radonc.2024.110542.

[7] U.S. Food and Drug Administration. *Transparency for Machine Learning–Enabled Medical Devices: Guiding Principles*. Accessed today. June 2024. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles.

[8] Erico Tjoa and Cuntai Guan. “Explainable Artificial Intelligence (XAI) in Healthcare: A Two-Part Overview for Practitioners”. In: *European Journal of Radiology* 164 (2023), p. 110892.

[9] Ravi Kumar Singh et al. “Self-eXplainable AI for Medical Image Analysis: A Comprehensive Review”. In: *Artificial Intelligence in Medicine* 145 (2024), p. 102663.

[10] Ahmad Chaddad et al. “Survey of Explainable AI Techniques in Healthcare”. In: *Sensors* 23.2 (2023), p. 634.

[11] Cheng Yang et al. “COIN: Counterfactual Inpainting for Medical Image Segmentation”. In: *Medical Image Analysis* 76 (2022), p. 102314.

[12] David Nemeth et al. “GAN-based Counterfactual Explanations for Medical Imaging: A Comparative Study with LIME and LRP”. In: *MICCAI*. 2024, pp. 234–243.

[13] Ahmad Chaddad et al. “Survey of Explainable AI Techniques in Healthcare”. In: *Sensors* 23.2 (2023), p. 634. DOI: 10.3390/s23020634.

[14] Ozan Oktay et al. “Attention U-Net: Learning Where to Look for the Pancreas”. In: *arXiv preprint arXiv:1804.03999* (2018). URL: https://arxiv.org/abs/1804.03999.

[15] Cheng Yang et al. “COIN: Counterfactual Inpainting for Medical Image Segmentation”. In: *Medical Image Analysis* 76 (2022), p. 102314. DOI: 10.1016/j.media.2021.102314.

[16] Jo Schlemper et al. “Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images”. In: *Medical Image Analysis* 53 (2019), pp. 197–207.

[17] Mauricio Reyes et al. “On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities”. In: *Radiology: Artificial Intelligence* 2.3 (2020), e190043.

[18] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". In: *Harvard Journal of Law & Technology* 31 (2017), p. 841.

[19] Phillip Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *CVPR*. 2017, pp. 1125–1134. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html.

[20] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *arXiv preprint arXiv:1312.6114* (2013). URL: https://arxiv.org/abs/1312.6114.

[21] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of the 31st International Conference on Machine Learning*. 2014. URL: http://proceedings.mlr.press/v32/rezende14.html.

[22] Zhi-Song Liu et al. "Variational Autoencoder for Reference-Based Image Super-Resolution". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.4 (2021), pp. 1405–1416.

[23] Nick Pawlowski, Daniel C. Castro, and Ben Glocker. "Deep Structural Causal Models for Tractable Counterfactual Inference". In: *NeurIPS*. 2017, pp. 857–867.

[24] Axel Sauer and Andreas Geiger. "Counterfactual Generative Networks". In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=HkxA0R4tPr.

[25] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations (ICLR)*. 2017. URL: https://openreview.net/forum?id=Sy2fzU9gl.

[26] Zhi-Song Liu et al. "Variational Autoencoder for Reference-Based Image Super-Resolution". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.4 (2021), pp. 1405–1416. DOI: 10.1109/TCSVT.2020.3006922.

[27] Nick Pawlowski, Daniel C. Castro, and Ben Glocker. "Deep Structural Causal Models for Tractable Counterfactual Inference". In: *Advances in Neural Information Processing Systems (NeurIPS) Workshops*. 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/.

[28] Axel Sauer and Andreas Geiger. "Counterfactual Generative Networks". In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=HkxA0R4tPr.

[29] Hugo J. W. L. Aerts et al. *Data From NSCLC-Radiomics (Version 4)*. Data set. Version 4. The Cancer Imaging Archive, 2015. DOI: 10.7937/K9/TCIA.2015.PF0M9REI. URL: https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.

[30] M. Jorge Cardoso et al. "MONAI: An Open-Source Framework for Deep Learning in Healthcare". In: *arXiv preprint arXiv:2211.02701* (2022). URL: https://arxiv.org/abs/2211.02701.

[31] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CVPR*. 2016. URL: https://arxiv.org/abs/1512.03385.

[32] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module". In: *European Conference on Computer Vision (ECCV)*. 2018. URL: https://arxiv.org/abs/1807.06521.

[33] Ricky T. Q. Chen et al. “Isolating Sources of Disentanglement in Variational Autoencoders”. In: *NeurIPS*. 2018. URL: https://arxiv.org/abs/1802.04942.

[34] Hao Fu et al. “Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 240–250. DOI: 10.18653/v1/N19-1021. URL: https://aclanthology.org/N19-1021.pdf.

[35] Abdel Aziz Taha and Allan Hanbury. “Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool”. In: *BMC Medical Imaging* 15.29 (2015). DOI: 10.1186/s12880-015-0068-x. URL: https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-015-0068-x.

[36] Zhou Wang et al. “Image Quality Assessment: From Error Visibility to Structural Similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861. URL: https://www.cns.nyu.edu/pub/eero/wang03-reprint.pdf.

[37] Huajie Shao et al. “ControlVAE: Controllable Variational Autoencoder”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 8655–8664. URL: https://proceedings.mlr.press/v119/shao20b.html.

[38] Hyunjik Kim and Andriy Mnih. “Disentangling by Factorising”. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 2649–2658. URL: http://proceedings.mlr.press/v80/kim18b.html.

[39] Zhen Lin et al. “InfoVAE: Information Maximizing Variational Autoencoders”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.4 (2020), pp. 4739–4746. DOI: 10.1609/aaai.v34i04.5870.