

Deep reinforcement learning for optimising angle selection and dose allocation in CT reconstruction

Tianyuan Wang , Daniël M. Pelt , Felix Lucka , Tristan van Leeuwen & Kees Joost Batenburg

To cite this article: Tianyuan Wang , Daniël M. Pelt , Felix Lucka , Tristan van Leeuwen & Kees Joost Batenburg (2026) Deep reinforcement learning for optimising angle selection and dose allocation in CT reconstruction, Applied Mathematics in Science and Engineering, 34:1, 2705107, DOI: [10.1080/27690911.2026.2705107](https://doi.org/10.1080/27690911.2026.2705107)

To link to this article: <https://doi.org/10.1080/27690911.2026.2705107>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 23 Jul 2026.



Submit your article to this journal [↗](#)



Article views: 54



View related articles [↗](#)



View Crossmark data [↗](#)

RESEARCH ARTICLE



Deep reinforcement learning for optimising angle selection and dose allocation in CT reconstruction

Tianyuan Wang^{a,b}, Daniël M. Pelt^b, Felix Lucka^a, Tristan van Leeuwen^{a,c} and Kees Joost Batenburg^{a,b}

^aCentrum Wiskunde & Informatica, Amsterdam, The Netherlands; ^bLeiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands; ^cMathematics Institute, Utrecht University, Utrecht, The Netherlands

ABSTRACT

Traditional X-ray computed tomography (CT) scanning strategies typically select projection angles uniformly and allocate dose equally. In practice, however, CT scans often need to be fast, radiation-efficient, and adaptive. Sparse-view tomography addresses these requirements by reducing both the number of angles and the total dose budget. Under such constraints, angle selection and dose allocation should be information-driven, with more dose assigned to informative directions. To this end, we propose a dose-aware acquisition and reconstruction framework that combines a PWLS-PnP reconstruction backbone with an RL-based strategy for adaptive angle selection, explicitly accounting for angle-dependent photon statistics. Numerical experiments show that the proposed approach improves overall reconstruction quality and enhances defect detectability compared with conventional strategies, particularly when only a small number of projections or a constrained dose budget is available.

ARTICLE HISTORY

Received 22 April 2026
Accepted 13 July 2026

KEYWORDS

X-ray CT; dose allocation; sequential optimal experimental design; defect detection; dose-aware penalized weighted least-squares; reinforcement learning

SUBJECT CLASSIFICATION CODE

Primary: 65R32 (Inverse problems); Secondary: 68T05 (Learning and adaptive systems); 65K10 (Optimization and variational techniques); 92C55 (Biomedical imaging)

1. Introduction

X-ray Computed Tomography (CT) is a widely used non-destructive imaging modality that reconstructs the internal structure of objects from external projections. However, CT inherently faces trade-offs among image quality, scan time, and radiation exposure—particularly in settings requiring rapid inspection or involving dose-sensitive subjects.

In medical CT, *dose optimisation* is critical to minimise patient radiation exposure. As reviewed in [1], conventional approaches such as tube current modulation and automatic exposure control adjust exposure according to patient size and anatomy. Although effective, these strategies remain largely heuristic and lack task-specific adaptivity. In *industrial CT*, the objective is often to concentrate measurements on regions with high uncertainty or suspected defects while avoiding redundant projections. Similar considerations arise in computational microscopy and electron or optical tomography, where optimising illumination dose is essential to prevent specimen damage. The object's position or orientation may differ between scans due to manipulation or changes in setup. These factors motivate the development of *adaptive scanning strategies* that jointly optimise projection angle selection and dose allocation under a fixed dose budget. Depending on the imaging task, the total dose can be distributed over many low-intensity views or concentrated on fewer, more informative angles, since not all measurements contribute equally to reconstruction quality [2].

Several works investigate adaptive scanning strategies that focus solely on sequential angle selection, where projection views are chosen based on information acquired during the scan [3–6]. The problem of selecting informative projection angles has often been formulated as a *bi-level optimisation* problem [7], where the lower level performs image reconstruction, and the upper level minimises reconstruction error.

CONTACT Tianyuan Wang ✉ tianyuan.wang@cwi.nl  Centrum Wiskunde & Informatica, Science Park 123, Amsterdam, 1098 XG, The Netherlands

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

This approach is high-dimensional and non-convex, as both levels depend on the design parameters (angles). Alternatively, *Bayesian experimental design* [8–10] seeks to maximise expected information gain by comparing prior and posterior distributions or reconstruction accuracy. Extensions of this framework have incorporated prior knowledge to guide angle selection and improve acquisition efficiency [3–5, 11–13]. However, extending these methods to include *dose allocation* further increases dimensionality and computational cost. In practical applications, imaging often requires adaptive, online decision-making to accommodate variations in object geometry or motion. Traditional Bayesian approaches, which rely on repeated evaluation of expected information gain, are computationally intensive and unsuitable for real-time control. Moreover, the *joint optimisation of angle and dose* remains largely unexplored, despite its strong impact on reconstruction quality.

Recent works have explored *Reinforcement Learning (RL)* for sequential CT scanning [6, 14–16]. Policy-based RL methods learn scanning strategies directly from data, eliminating repeated optimisation during acquisition and enabling adaptive decision-making based on intermediate reconstructions. Once trained, these policies can be applied efficiently in online scanning scenarios.

From a reconstruction perspective, the *incident photon count*—that is, the local dose—directly affects measurement noise according to the Beer-Lambert law. When dose allocation varies across angles, high-dose projections yield lower noise, while low-dose projections are more uncertain. This heterogeneity calls for reconstruction algorithms that explicitly account for dose-dependent noise. *Penalised Weighted Least Squares (PWLS)* methods [17, 18] address noise in low-dose CT but typically assume uniform dose across angles. Some learning-based approaches incorporate global dose levels [19], yet they do not handle unequal per-angle dose distributions.

In this work, we address this gap by deriving a *dose-aware PWLS* model derived from a second-order Taylor approximation of the Poisson log-likelihood. While conventional PWLS formulations generally assume uniform or fixed doses across views, our adaptive setting requires handling intentionally non-uniform, per-angle dose allocations. To accommodate this, we introduce an angle-dependent weighting matrix that explicitly reflects the dose assigned to each view by the adaptive policy.

1.1. Contributions

Building on recent advances in adaptive imaging, this work presents an RL-based framework for joint projection-angle selection and dose allocation in X-ray CT. The main contributions are:

- **Joint angle–dose acquisition strategy:** We formulate adaptive CT acquisition as a sequential decision-making problem in which the policy selects projection angles under a fixed dose budget. Dose allocation is represented through repeated selections of projection angles, so that more photons are assigned to views selected multiple times. The resulting policy is trained with task-dependent objectives, including reconstruction quality and defect detectability.
- **Dose-aware reconstruction under non-uniform exposure:** We extend the PWLS reconstruction model to incorporate *angle-dependent weights* induced by the non-uniform dose distribution produced by the acquisition policy. The resulting dose-weighted PWLS data-fidelity term is combined with a Total Variation (TV) prior within a Plug-and-Play (PnP) reconstruction scheme, enabling reconstruction in settings where the statistical reliability of the measured projections varies across angles.

The remainder of this paper is organised as follows. [Section 2](#) reviews background on CT reconstruction and sequential optimal experimental design using RL. [Section 3](#) presents the proposed reconstruction and scanning framework. [Section 4](#) reports experimental results, and [Sections 5](#) and [6](#) provide discussion and conclusions.

2. Background

2.1. CT reconstruction

In X-ray CT, the objective is to reconstruct an unknown image $\bar{x} \in \mathbb{R}^n$, representing the spatial distribution of attenuation coefficients within an object, from a set of external measurements. Let Θ denote the full

set of available candidate projection angles. At step k , the agent selects one angle $\theta_k \in \Theta$. The sequence of selected angles is denoted by θ . These measurements, denoted $\mathbf{y}(\theta)$, are acquired at a sequence of design parameters $\theta = \{\theta_1, \dots, \theta_M\}$, typically projection angles. The image formation process is described by a forward operator $\mathbf{A}(\theta)$, which discretizes the Radon transform under the specified acquisition geometry [20].

To account for measurement uncertainty, we adopt a statistical observation model in which the measurements are treated as random variables:

$$\mathbf{y}(\theta) \sim \pi_{\text{data}}(\mathbf{y} \mid \bar{\mathbf{x}}; \theta), \quad (1)$$

where π_{data} denotes the data likelihood.

Beer-Lambert law. X-ray attenuation follows the Beer-Lambert law [20], which relates transmitted intensity to the cumulative attenuation along an X-ray path. Let $I_0(\theta_j)$ denote the incident photon flux (dose) at angle θ_j , and $I_i(\theta_j)$ the transmitted intensity measured at detector bin i . Then the attenuation model is

$$I_i(\theta_j) = I_0(\theta_j) \exp(-[\mathbf{A}(\theta_j)\bar{\mathbf{x}}]_i), \quad (2)$$

where $[\mathbf{A}(\theta_j)\bar{\mathbf{x}}]_i$ represents the line integral of $\bar{\mathbf{x}}$ along ray i at angle θ_j . The corresponding noise-free measurement is

$$y_i(\theta_j) = -\log \left(\frac{I_i(\theta_j)}{I_0(\theta_j)} \right) = [\mathbf{A}(\theta_j)\bar{\mathbf{x}}]_i. \quad (3)$$

In practice, the observed measurements $\mathbf{y}(\theta_j)$ deviate from this ideal due to stochastic noise governed by the dose level $I_0(\theta_j)$.

Bayesian inverse formulation. Reconstructing $\bar{\mathbf{x}}$ from noisy measurements is an ill-posed inverse problem. Within a Bayesian framework, the posterior distribution combines the measurement model with prior information:

$$\begin{aligned} \hat{\mathbf{x}}(\theta) &= \arg \max_{\mathbf{x}} \pi_{\text{post}}(\mathbf{x} \mid \mathbf{y}; \theta) \\ &= \arg \max_{\mathbf{x}} \frac{\pi_{\text{data}}(\mathbf{y} \mid \mathbf{x}; \theta) \cdot \pi_{\text{prior}}(\mathbf{x})}{\pi(\mathbf{y})}, \end{aligned} \quad (4)$$

where $\pi_{\text{prior}}(\mathbf{x})$ encodes assumptions about the image, such as smoothness or sparsity. The maximum a posteriori (MAP) estimate thus balances data fidelity and prior knowledge to yield a stable reconstruction.

2.2. Sequential optimal experimental design via reinforcement learning

Reinforcement learning (RL) provides a natural framework for solving sequential optimal experimental design (OED) problems in CT imaging by modelling the acquisition process as a *Partially Observable Markov Decision Process (POMDP)* [21]. This perspective has been effectively applied in prior work [6,15,16], where the agent interacts with the CT environment to iteratively select measurements and refine the reconstruction.

The interaction starts from an initial estimate $\hat{\mathbf{x}}_1$, typically a zero image. At each time step k , the agent selects an acquisition angle θ_k based on the current reconstruction $\hat{\mathbf{x}}_k$. A noisy measurement $\mathbf{y}_k$ is then acquired using the forward model in Equation (1), and a new reconstruction $\hat{\mathbf{x}}_{k+1}$ is computed based on all measurements up to step k . A reward r_k is subsequently calculated by comparing $\hat{\mathbf{x}}_{k+1}$ to the ground truth image.

The agent learns a parameterised policy $\pi_a(\theta \mid \hat{\mathbf{x}}_k; \mathbf{w}_a)$, which maps the current reconstruction to a probability distribution over candidate angles. After M steps, the acquisition process produces a trajectory τ defined as:

$$\tau = \{\hat{\mathbf{x}}_1, (\theta_1, \hat{\mathbf{x}}_2, r_1), \dots, (\theta_M, \hat{\mathbf{x}}_{M+1}, r_M)\}.$$

This trajectory follows a distribution governed by:

$$\pi_{\text{chain}}(\tau; \mathbf{w}_a) = \prod_{k=1}^M \pi_a(\theta_k \mid \hat{\mathbf{x}}_k; \mathbf{w}_a) \cdot \pi_t(\hat{\mathbf{x}}_{k+1} \mid \hat{\mathbf{x}}_k, \theta_k), \quad (5)$$

where $\pi_t(\hat{\mathbf{x}}_{k+1} \mid \hat{\mathbf{x}}_k, \theta_k)$ represents the transition model capturing the effect of measurement and reconstruction.

The learning objective is to maximise the expected cumulative reward, where future rewards may be discounted using a factor $\gamma \in [0, 1]$ to prioritise immediate outcomes. The state-value function under policy π_a is defined as:

$$V^{\pi_a}(\hat{\mathbf{x}}_k) = \mathbb{E}_{\tau \sim \pi_{\text{chain}}} \left[\sum_{t=k}^M \gamma^{t-k} r_t \mid \hat{\mathbf{x}}_k \right], \quad (6)$$

where γ^{t-k} discounts the reward obtained at time step k relative to the current step k . This formulation captures the preference for receiving rewards earlier rather than later.

The overall objective in RL is to maximise the expected return starting from the fixed initial zero image $\hat{\mathbf{x}}_1$:

$$J(\mathbf{w}_a) = V^{\pi_a}(\hat{\mathbf{x}}_1), \quad (7)$$

where optimisation is performed over the policy parameters $\mathbf{w}_a$.

2.3. Policy gradient methods

Policy gradient methods aim to optimise the expected return $J(\mathbf{w}_a)$ by computing gradients with respect to the policy parameters $\mathbf{w}_a$. A popular approach is the *actor-critic* framework, where the *actor* represents the policy, and the *critic* estimates the state-value function to guide policy improvement.

To reduce variance and improve learning efficiency, policy updates are often scaled by an *advantage estimate* $\hat{A}_k$, which measures how much better (or worse) an action is compared to the critic's baseline value estimate $V^{\pi_a}(\hat{\mathbf{x}}_k)$. A simple and widely used choice is the one-step Temporal-Difference (TD(0)) estimator [22]:

$$\hat{A}_k = r_k + \gamma V^{\pi_a}(\hat{\mathbf{x}}_{k+1}) - V^{\pi_a}(\hat{\mathbf{x}}_k), \quad (8)$$

where γ is the discount factor. More general estimators, such as Generalised Advantage Estimation (GAE) [23], can further reduce variance by incorporating multi-step returns.

The policy gradient can then be approximated from sampled trajectories as:

$$\nabla_{\mathbf{w}_a} J(\mathbf{w}_a) \approx \frac{1}{N} \sum_{n=1}^N \left(\sum_{k=1}^M \nabla_{\mathbf{w}_a} \log \pi_a(\theta_k \mid \hat{\mathbf{x}}_k; \mathbf{w}_a) \cdot \hat{A}_k \right), \quad (9)$$

where N is the number of sampled trajectories and M is the number of timesteps per trajectory.

3. Methods

This section introduces the proposed framework, which combines a dose-aware reconstruction algorithm with prior information to handle unequal dose allocation, and a RL approach to address the adaptive CT scanning problem.

3.1. Reconstruction with unequal dose allocation

Building on the Beer-Lambert formulation in Section 2, we now incorporate measurement noise arising from photon statistics. Since photon detection follows a Poisson distribution, the measured photon count at detector bin i for angle θ_j can be modelled as (according to Equation (2))

$$I_i(\theta_j) \sim \text{Poisson}(I_0(\theta_j) \exp(-[\mathbf{A}(\theta_j)\bar{\mathbf{x}}]_i)). \quad (10)$$

where $\mathbf{I}_0(\theta_j) = [I_{0,1}(\theta_j), \dots, I_{0,n_{\text{det}}}(\theta_j)]^T$ denotes the incident photon flux at all detector bins. (For the experiments in this paper, we consider a spatially uniform incident dose so that $\mathbf{I}_0(\theta_j) = I_0(\theta_j)\mathbf{1}$.)

Normalising by $\mathbf{I}_0(\theta_j)$ and applying a negative logarithm yields the noisy log-transformed measurements:

$$\mathbf{y}(\theta_j) = -\log \left(\frac{\mathbf{I}(\theta_j)}{\mathbf{I}_0(\theta_j)} \right). \quad (11)$$

From the Bayesian inverse formulation in Equation (4), the negative log-likelihood (up to constants) gives the dose-aware reconstruction objective:

$$\begin{aligned} \hat{\mathbf{x}}(\boldsymbol{\theta}) \propto \arg \min_{\mathbf{x}} \{ & \sum_j \mathbf{I}_0(\theta_j)^T \exp(-\mathbf{A}(\theta_j)\mathbf{x}) \\ & + \sum_j [\mathbf{I}_0(\theta_j) \odot \exp(-\mathbf{y}(\theta_j))]^T \mathbf{A}(\theta_j)\mathbf{x} - \log \pi_{\text{prior}}(\mathbf{x}) \}. \end{aligned} \quad (12)$$

This formulation explicitly incorporates unequal per-angle dose levels $\mathbf{I}_0(\theta_j)$, which directly affect both the noise characteristics of the measurements and their contribution to the reconstruction objective.

PWLS formulation. Direct minimisation of Equation (12) is computationally demanding. Using a second-order Taylor expansion around the measured log-data [20] yields a Penalised Weighted Least-Squares (PWLS) problem [17,20,24]:

$$\hat{\mathbf{x}}(\boldsymbol{\theta}) \propto \arg \min_{\mathbf{x}} \left\{ \frac{1}{2} (\mathbf{y}(\boldsymbol{\theta}) - \mathbf{A}(\boldsymbol{\theta})\mathbf{x})^T \mathbf{W}(\boldsymbol{\theta}) (\mathbf{y}(\boldsymbol{\theta}) - \mathbf{A}(\boldsymbol{\theta})\mathbf{x}) - \log \pi_{\text{prior}}(\mathbf{x}) \right\}, \quad (13)$$

with diagonal weight matrix

$$\mathbf{W}(\boldsymbol{\theta}) = \text{diag}(\mathbf{I}_0(\boldsymbol{\theta}) \odot \exp(-\mathbf{y}(\boldsymbol{\theta}))). \quad (14)$$

Here, $\mathbf{I}_0(\boldsymbol{\theta})$ expands the per-angle dose allocation $\mathbf{I}_0(\theta_j)$ to all bins. In contrast to conventional PWLS formulations [17] that assume a uniform incident dose across views, this model explicitly accounts for angle-dependent dose allocation. This *dose-aware* weighting provides a principled mechanism to assign higher statistical confidence to high-dose projections and lower confidence to noisy low-dose measurements, effectively capturing the signal-dependent variance characteristic of Poisson noise in X-ray CT.

Plug-and-Play priors In particular when using low radiation dosages distributed over few projection angles, minimising the PWLS objective directly is ill-posed and leads to very noisy reconstructions ridden with sparse-angle artifacts. To regularise the inversion in the most flexible way, we describe here how to employ the Plug-and-Play (PnP) framework [25]. PnP decouples the reconstruction into two alternating steps: a gradient step for the PWLS objective followed by the application of an image denoiser. To illustrate this, we implemented denoising with a Total Variation (TV) energy [26,27], denoted by $\text{prox}_{\tau\text{TV}}(\cdot)$. TV penalises the ℓ_1 -norm of image gradients and is used to suppresses noise in homogeneous regions while preserving sharp edges. While this choice means that the PnP framework reduces to a proximal gradient descent with an explicit image prior $\pi_{\text{prior}}(\mathbf{x})$, using, e.g. a trained DNN would correspond to using an implicit prior.

Algorithm 1. Dose-aware PnP–PWLS reconstruction (Gradient + TV prior).

Require: Forward operator $\mathbf{A}(\boldsymbol{\theta})$; measured log-projections $\mathbf{y}(\boldsymbol{\theta})$; dose allocation $\mathbf{d}(\boldsymbol{\theta})$; number of iterations T ; initial estimate $\mathbf{x}_0$; PnP denoiser strength $\tau > 0$ (TV parameter); positivity flag.

Ensure: Reconstructed image $\mathbf{x}_T$.

```

1:  $\mathbf{x} \leftarrow \mathbf{x}_0$ 
2: Normalize dose:  $\mathbf{d}(\boldsymbol{\theta}) \leftarrow \mathbf{d}(\boldsymbol{\theta}) / \text{mean}(\mathbf{d}(\boldsymbol{\theta}))$ 
3: Dose–Poisson weights:  $\mathbf{R}(\boldsymbol{\theta}) \leftarrow \mathbf{d}(\boldsymbol{\theta}) \odot \exp(-\mathbf{y}(\boldsymbol{\theta}))$ 
4: Define weighting operator:  $\mathcal{W}(\cdot) : \mathbf{u} \mapsto \mathbf{R}(\boldsymbol{\theta}) \odot \mathbf{u}$ 
5: Estimate Lipschitz constant  $L$  via Appendix A; set step size  $\alpha \leftarrow h/L$  with  $h \in [0, 2]$ 
6: for  $k = 1$  to  $T$  do                                 $\triangleright$  Data fidelity + PnP prior
7:    $\mathbf{y}_{\text{pred}} \leftarrow \mathbf{A}(\boldsymbol{\theta}) \mathbf{x}$                                  $\triangleright$  Forward projection
8:    $\mathbf{r} \leftarrow \mathcal{W}(\mathbf{y}_{\text{pred}} - \mathbf{y}(\boldsymbol{\theta}))$                                  $\triangleright$  Dose-aware residual
9:    $\mathbf{g} \leftarrow \mathbf{A}(\boldsymbol{\theta})^\top \mathbf{r}$                                  $\triangleright$  Backprojection (gradient of data term)
10:   $\mathbf{x} \leftarrow \mathbf{x} - \alpha \mathbf{g}$                                  $\triangleright$  Gradient step (data fidelity)
11:   $\mathbf{x} \leftarrow \text{prox}_{\tau \text{TV}}(\mathbf{x})$                                  $\triangleright$  PnP denoiser: TV proximal operator (e.g.
    Chambolle’s algorithm)
12:  if positivity flag then
13:     $\mathbf{x} \leftarrow \max(\mathbf{x}, 0)$ 
14:  end if
15: end for
16: return  $\mathbf{x}$ 

```

Explanation. Algorithm 1 performs gradient descent on the PWLS objective with explicit modelling of angle-dependent dose allocation. At each iteration, the current reconstruction is forward-projected and compared with the measured log-projections to form a residual. This residual is reweighted by a Poisson-dose factor, which increases the influence of high-dose views and suppresses noisy low-dose views. The weighted residual is then backprojected to compute the gradient, and the step size is normalised by the estimated Lipschitz constant (Appendix A) to ensure stable convergence. After the gradient update, a TV denoiser $\text{prox}_{\tau \text{TV}}(\cdot)$ is applied in the PnP framework, providing noise suppression while preserving edges. Finally, a positivity constraint can be optionally enforced to maintain physically valid attenuation values.

Compared with standard iterative reconstruction, this dose-aware PnP scheme explicitly accounts for non-uniform photon allocation across projection angles, yielding statistically consistent updates under unequal-dose settings while benefiting from edge-preserving regularisation.

3.2. Proximal policy optimisation for CT environment

To stably optimise the objective $J(\mathbf{w}_a)$ in Equation (7), Proximal Policy Optimisation (PPO) [28] is adopted as the RL framework. The vanilla policy-gradient method in Equation (9) updates the policy parameters by directly ascending an unbiased stochastic estimate of $\nabla_{\mathbf{w}_a} J(\mathbf{w}_a)$, computed from sampled trajectories, without explicitly constraining the size of policy updates. PPO replaces this unconstrained update with a clipped surrogate objective that approximates the same gradient while preventing excessively large policy changes, thereby improving training stability:

$$\text{Loss}^{\text{actor}}(\mathbf{w}_a) = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^M \min(e_k(\mathbf{w}_a) \hat{A}_k, c_{1-\epsilon}^{1+\epsilon}(e_k(\mathbf{w}_a)) \hat{A}_k), \quad (15)$$

where $e_k(\mathbf{w}_a) = \frac{\pi_a(\theta_k | \hat{\mathbf{x}}_k; \mathbf{w}_a)}{\pi_a^{\text{old}}(\theta_k | \hat{\mathbf{x}}_k)}$ is the probability ratio between new and old policies, $\hat{A}_k$ is the estimated advantage, and $\epsilon > 0$ is a clipping parameter. Here, $c_{1-\epsilon}^{1+\epsilon}(\cdot)$ clips the probability ratio $e_k(\mathbf{w}_a)$ to the interval $[1 - \epsilon, 1 + \epsilon]$, ensuring stable and conservative policy updates. The objective is maximised during training to update the actor parameters $\mathbf{w}_a$.

The critic is trained using a mean squared error (MSE) loss with respect to the Monte Carlo return G_k (Formally, $G_k = \sum_{t=k}^M \gamma^{t-k} r_t$):

$$Loss^{\text{critic}}(\mathbf{w}_v) = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^M (G_k - V(\hat{\mathbf{x}}_k; \mathbf{w}_v))^2, \quad (16)$$

Reward Definition. The reward function depends on the objective of the reconstruction task. For general image quality assessment, we use the Peak Signal-to-Noise Ratio (PSNR), computed over a region of interest (RoI) Ω :

$$\text{PSNR}_{\Omega}(\bar{\mathbf{x}}, \hat{\mathbf{x}}) = 10 \log_{10} \left(\frac{H^2}{\frac{1}{|\Omega|} \sum_{i \in \Omega} (\bar{x}_i - \hat{x}_i)^2} \right), \quad (17)$$

where H denotes the dynamic range of the image.

For defect-focused tasks, the reward can instead be based on the Contrast-to-Noise Ratio (CNR):

$$\text{CNR}_{\mathcal{D}, \mathcal{B}}(\bar{\mathbf{x}}, \hat{\mathbf{x}}) = \frac{\mu_{\mathcal{D}} - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}}}, \quad (18)$$

where $\mu_{\mathcal{D}}$ is the mean intensity within the defect mask $\mathcal{D}$, and $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ denote the mean and standard deviation in a background region $\mathcal{B}$.

Training Procedure. The RL agent interacts with the CT environment under a fixed dose budget B , which is spent in base quanta Δb per projection. Dose allocation is modelled implicitly by repeating angle selections: if an angle is chosen z times, the allocated dose is $z\Delta b$. The episode terminates after $L = \lfloor B/\Delta b \rfloor$ steps, corresponding to exhaustion of the total dose budget B . The policy maps the current reconstruction $\hat{\mathbf{x}}_k$ to a probability distribution over candidate projection angles, thereby jointly determining angle selection and dose allocation. The overall training loop is summarised in Algorithm 2. A trajectory buffer stores observations, actions, rewards, and log-probabilities of the old policy. After each rollout, advantages and returns are computed, and the policy and value networks are updated by Equations (15) and (16). Implementation details including advantage estimation, entropy regularisation, and minibatch updates are provided in Appendix B.

Algorithm 2. PPO training in a CT environment with a dose budget.

Require: Environment $\mathcal{E}$, policy $\pi_a(\cdot; \mathbf{w}_a)$, value function $V(\cdot; \mathbf{w}_v)$, total dose budget B , fixed dose per step Δb , number of trajectories per iteration n

- 1: Set trajectory length $L \leftarrow \lfloor \frac{B}{\Delta b} \rfloor$
- 2: Set horizon $T \leftarrow n \cdot L$ ▷ total steps per PPO iteration
- 3: **for** each iteration **do**
- 4: Initialize empty trajectory buffer $\mathcal{T}$
- 5: **for** $i = 1$ **to** n **do** ▷ collect n trajectories
- 6: Reset environment: $\hat{\mathbf{x}}_1$ (zero matrix) $\leftarrow \mathcal{E}.\text{reset}()$
- 7: **for** $k = 1$ **to** L **do**
- 8: Sample angle $\theta_k \sim \pi_a(\cdot | \hat{\mathbf{x}}_k; \mathbf{w}_a)$
- 9: $\log \pi_k \leftarrow \log \pi_a(\theta_k | \hat{\mathbf{x}}_k; \mathbf{w}_a)$, $v_k \leftarrow V(\hat{\mathbf{x}}_k; \mathbf{w}_v)$
- 10: Apply action: $(\hat{\mathbf{x}}_{k+1}, r_k) \leftarrow \mathcal{E}.\text{step}(\theta_k)$
- 11: $d_k \leftarrow \mathbb{I}[k = L]$
- 12: Store $(\hat{\mathbf{x}}_k, \theta_k, r_k, \hat{\mathbf{x}}_{k+1}, \log \pi_k, v_k, d_k)$ in $\mathcal{T}$
- 13: **end for**
- 14: **end for**
- 15: Compute advantages $\hat{A}_k$ and returns G_k from $\mathcal{T}$ (Appendix B)
- 16: Normalize advantages
- 17: Update actor by maximizing the clipped PPO objective (Equation (15)),
where π^{old} uses stored $\log \pi_k$ from $\mathcal{T}$
- 18: Update critic by minimizing the MSE loss (Eq. (16))
- 19: **end for**

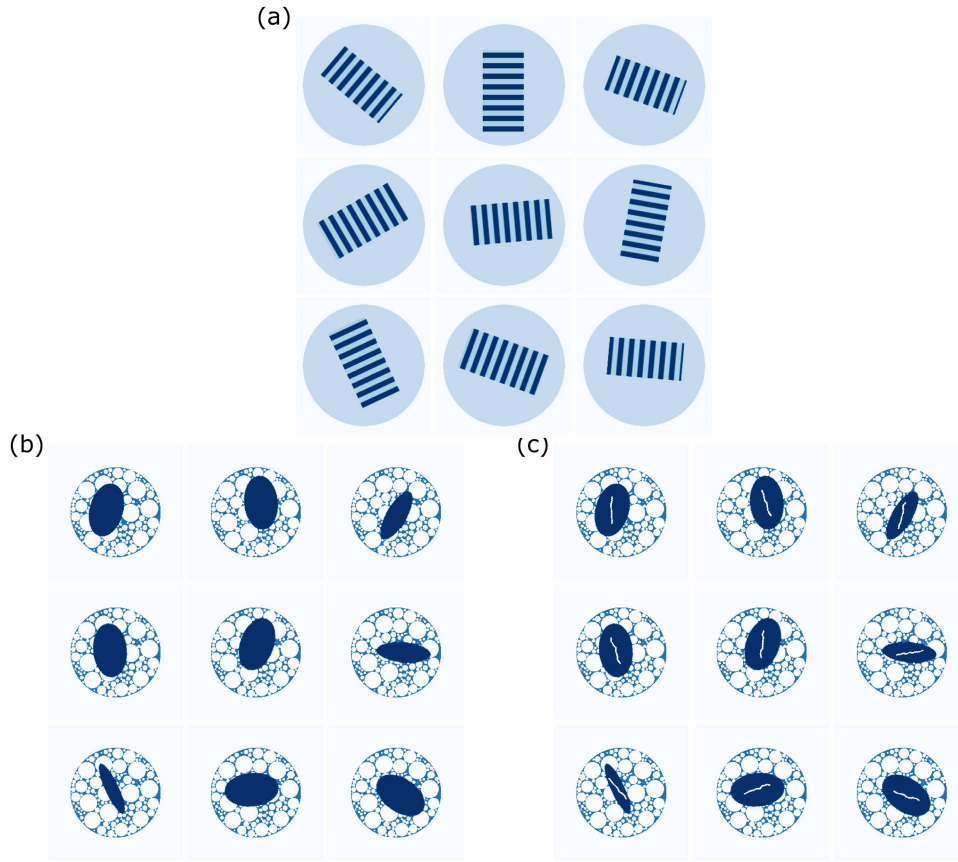


Figure 2. Example training dataset consisting of three types of phantoms. (a) Wedge-shaped phantoms composed of two different materials, with variations in rotation, position, and size. (b) Ellipse- and foam-based phantoms with different material properties. (c) Ellipse-based phantoms containing defects. The ellipses vary in rotation, position, size, and roundness, providing a diverse set of training samples.

Dataset (a) contains wedge-shaped phantoms composed of different materials, with variations in rotation, position, and size. These phantoms are inspired by industrial CT inspection scenarios involving fibre-reinforced or layered composite materials, where internal structures exhibit strong directional characteristics. Datasets (b) and (c) consist of ellipse- and foam-based phantoms [29] with different material properties. These phantoms are inspired by polymer or metal foam-like materials that act as containers for embedded foreign objects or defects with preferential orientations. Such configurations are representative of industrial CT inspection tasks, where the goal is to detect and characterise internal objects or anisotropic defects within a heterogeneous, porous background. In particular, dataset (c) includes ellipse-based phantoms containing defects. All datasets include phantoms at different rotations, represented by 36 equally spaced angles between 0° and 179° . In addition, the phantoms exhibit variations in scaling and spatial shifts.

4.2. Implementation

Tomosipo [30] was used to generate simulated projections and perform reconstructions. The image size was set to 256×256 , and a projection size of 384 was chosen to ensure that each view fully covered the image domain.

The photon budgets in our synthetic experiments are normalised photon-count parameters in the Poisson noise model and are not directly calibrated to physical dose units such as mGy. Such a conversion would require scanner-specific information, including the X-ray spectrum, filtration, exposure time, detector efficiency, acquisition geometry, and object attenuation. We therefore use these budgets to define

controlled low-count regimes for comparing angle and dose selection policies under matched relative dose constraints. They should be interpreted as representative of dose-scarce imaging scenarios, such as low-dose CT, repeated or interventional imaging, and industrial CT inspection with limited exposure or acquisition time, rather than as a one-to-one mapping to a specific clinical or industrial dose value.

For the reconstruction algorithm, the parameter h in Algorithm 1 was set to 1.8, and the PnP-TV denoiser employed a fixed proximal step size of 0.0005. In the RL framework (Algorithm 2), the discount factor was set to $\gamma = 0.99$, and the replay buffer capacity was fixed at 500 transitions. The actor and critic losses were weighted by 1.0 and 0.5, respectively, and an entropy regularisation term with weight 0.01 was included to promote adequate exploration during training (see Appendix B for details). These hyperparameters were selected empirically to provide a stable balance between policy optimisation, accurate value-function estimation, and robust exploratory behaviour. A convolutional encoder was employed to extract low-dimensional feature representations from the input images, following the design in [6]. All network parameters were optimised using the Adam algorithm [31] with a learning rate of 10^{-4} and a weight decay of 10^{-5} . Further architectural and algorithmic details are provided in Table B1 in Appendix B.

4.3. Experiment 1: reconstruction using dose-aware PnP-PWLS with a TV prior

This experiment evaluates the proposed reconstruction algorithm against standard iterative methods. We focus on datasets (a) and (b) described in Section 4.1. Reconstructions were performed with three equal angular sampling settings: 36, 48, and 60 angles. To simulate unequal dose allocation across views, odd-indexed angles were assigned a low dose (100 photons), and even-indexed angles a high dose (1000 photons). The number of gradient descent iterations was varied from 5 to 100 in increments of 5.

Four methods were compared: Simultaneous Iterative Reconstruction Technique (SIRT), conventional PWLS, dose-aware PWLS, and dose-aware PnP-PWLS with a TV prior. All reconstructions were performed under a non-negativity constraint.

Figure 3 compares reconstruction performance on datasets (a) and (b). Both dose-aware variants (PWLS and PnP-PWLS with a TV prior) consistently achieve higher PSNR and exhibit greater robustness to the number of iterations across different angular sampling settings. In contrast, SIRT and conventional PWLS degrade noticeably at higher iteration counts, making reconstruction quality more sensitive to the iteration choice and less reliable for subsequent experiments. These results demonstrate that dose-aware weighting improves reconstruction stability, with PnP-PWLS with a TV prior consistently providing the best overall performance.

Remark. The performance gap between the proposed dose-aware methods and the baselines becomes more pronounced at higher iteration counts. While SIRT and conventional PWLS tend to overfit the noise and degrade in quality, dose-aware weighting preserves relative stability by scaling the contribution of each view according to its photon statistics, although some degradation remains visible on dataset (a). This highlights that dose-awareness is particularly beneficial in long-iteration regimes, where conventional methods become unreliable. In addition, incorporating a TV prior provides effective denoising while preserving edges, further improving reconstruction quality.

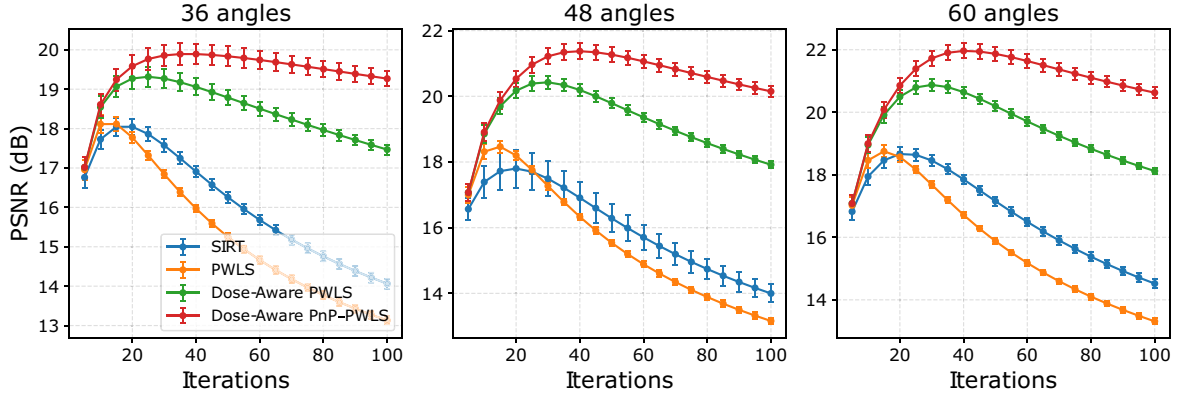
These visual results are consistent with the quantitative PSNR values reported in Figure 4, where the proposed dose-aware methods achieve improvements of approximately 1.5–3 dB over SIRT and conventional PWLS when considering their best reconstructions within the iteration range of 0–100. This demonstrates that explicitly modelling angle-dependent dose allocation not only enhances numerical performance but also leads to visibly cleaner and higher-contrast reconstructions.

4.3.1. Experiment 2: angle selection and dose allocation for reconstruction quality improvement

This experiment evaluates whether the RL agent can jointly select projection angles and allocate dose to improve reconstruction quality when the reward function is based solely on PSNR. We used dataset (a) for this study and adopted the dose-aware PnP-PWLS reconstruction method with a TV prior, which showed the best performance in the previous experiments.

We compared the learned RL policy with two baselines. The first baseline is the *Golden Ratio (GR) policy* [32,33], which is commonly used in sparse-view CT. The GR policy generates projection angles by

(a) PSNR vs. Iterations for different angle counts on dataset



(b) PSNR vs. Iterations for different angle counts on dataset

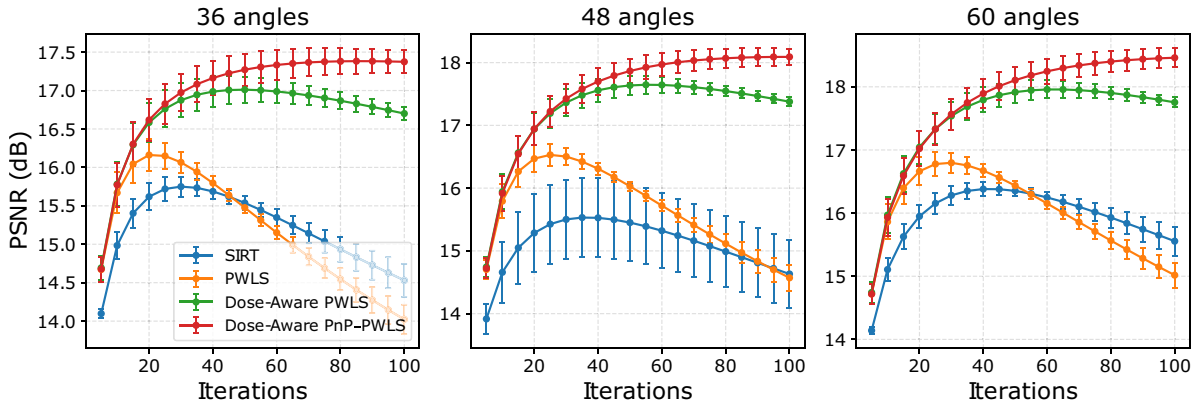


Figure 3. PSNR versus number of iterations for different angular sampling settings (36, 48, and 60 angles, left to right) on datasets (a) and (b). The four methods compared are SIRT, conventional PWLS, dose-aware PWLS, and dose-aware PnP-PWLS. Curves show the mean PSNR across all non-defective phantoms in the dataset, and error bars indicate the standard deviation.

applying an irrational angular increment, resulting in a non-repeating sequence that approximately uniformly fills the angular range over time. The second baseline is a *Projected-Uncertainty (PU) adaptive policy*, which provides a simple adaptive non-RL comparison.

For the PU baseline, an uncertainty map is computed at acquisition step k from the change between consecutive reconstructions,

$$u_{k-1} = |\hat{x}_{k-1} - \hat{x}_{k-2}|. \quad (19)$$

Each candidate angle $\theta \in \Theta$ is scored by forward-projecting this uncertainty map,

$$S_k(\theta) = \frac{\|A(\theta)u_{k-1}\|_1}{1 + n_k(\theta)}, \quad (20)$$

where $A(\theta)$ is the single-view forward operator at angle θ , and $n_k(\theta)$ is the number of previous selections of angle θ . The denominator discourages repeated selection of the same angle while still allowing additional dose to be assigned through repeated measurements. The next angle is selected greedily as

$$\theta_k = \arg \max_{\theta \in \Theta} S_k(\theta). \quad (21)$$

For the first acquisition step, the first GR angle is used to avoid the degenerate case of uniform initial uncertainty. This baseline uses the same candidate angle set, dose budget, repeated-angle dose-allocation

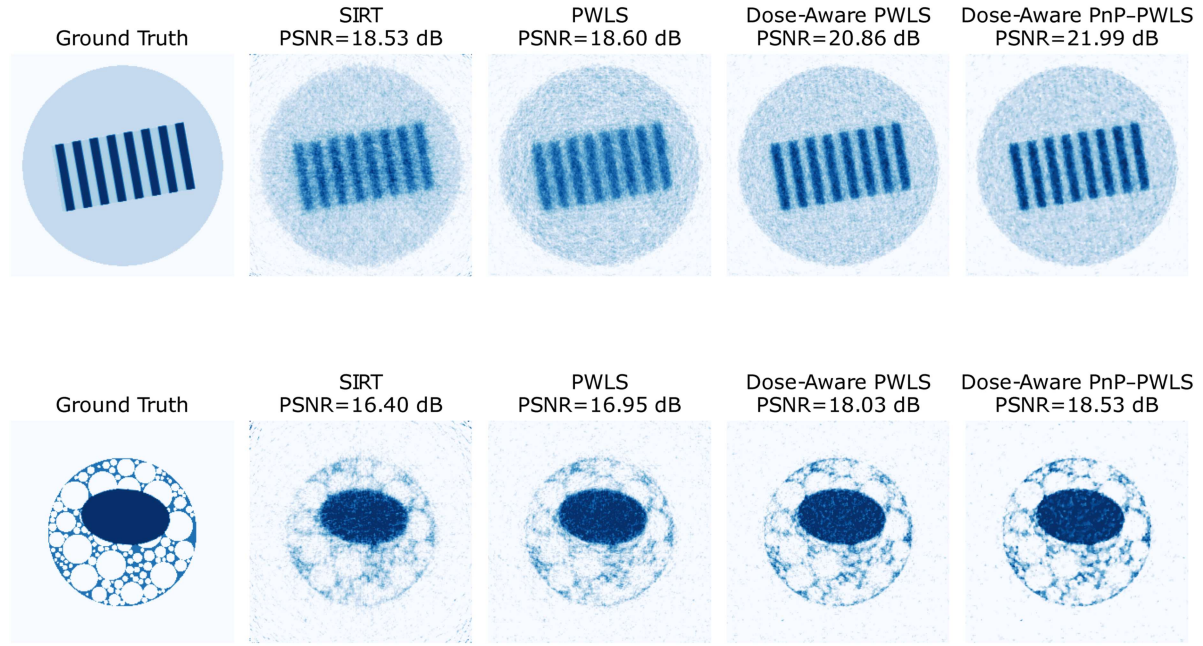


Figure 4. Reconstruction results for two phantoms with 60 projection angles using SIRT, conventional PWLS, and the proposed dose-aware methods. For each method, the iteration count was selected from 0 to 100 to yield its best reconstruction. SIRT reconstructions suffer from low contrast and artifacts, while PWLS provides modest improvements but remains noisy. The proposed dose-aware approaches achieve the highest PSNR and visibly improve contrast.

mechanism, Poisson noise model, and dose-aware PnP-PWLS reconstruction method as the RL policy, but it is myopic and does not learn a task-specific acquisition strategy.

Training procedure. The agent was trained on 150 phantoms from dataset (a). For the GR policy, 20 angles were used with an equal dose assigned to each angle. Three separate RL policies were trained for three different total photon budgets of 6000 photons (300 photons per angle), 7000 photons (350 photons per angle) and 8000 photons (400 photons per angle). For the RL policy, the agent could allocate more dose to informative angles by repeating them. For example, with a base dose of 300 photons per angle, if angle a is repeated n times, the total dose assigned to that angle becomes $n \times 300$. Training was performed for 1500 episodes using the reward function in Equation (22), assuming that the ground truth $\bar{x}$ was non-defective. The weight parameter α in Equation (22) was set to 0.2, and the RoI Ω corresponded to the full image.

Validation procedure. Validation was performed on a separate test set of 100 phantoms with unseen rotation angles. All results reported in this section are obtained from this validation set using the trained policy.

The reward functions used in this work require ground-truth images and are therefore used only during offline training in simulation. During inference, the trained policy does not require access to the ground truth; it selects angles and allocates dose based only on the current reconstruction state and acquisition history.

Quantitative comparison. Table 1 summarises the performance of the GR, PU, and RL policies on the unseen-rotation test set across total photon budgets of 6000, 7000, and 8000. The RL policy consistently achieves higher PSNR than both the non-adaptive GR policy and the adaptive PU baseline. Compared with GR, the mean PSNR improvements of the RL policy are 2.80, 2.46, and 2.68 dB for total photon budgets of 6000, 7000, and 8000, respectively. The corresponding 95% paired bootstrap confidence intervals are strictly positive, and the improvements remain significant after Benjamini–Hochberg FDR correction at $\alpha = 0.05$. In contrast, the PU baseline does not outperform GR, indicating that a simple myopic projected-uncertainty strategy is insufficient for this wedge reconstruction task. As the total photon budget increases, the RL policy selects a larger number of distinct angles, suggesting that repeated angle selection is reduced when more dose is available.

Table 1. Performance comparison of GR, projected-uncertainty (PU), and RL policies on unseen-rotation tests for the wedge phantom experiment. Results are reported as mean $\pm$ standard deviation. The RL–GR difference is reported with 95% paired bootstrap confidence intervals. One-sided p -values for RL outperforming GR were corrected using Benjamini–Hochberg FDR control at $\alpha = 0.05$.

Total photons	Policy	PSNR	Average number of angles	PSNR diff. RL–GR [95% CI]
6000	GR	16.82 $\pm$ 0.40	20	2.80 [2.76, 2.83]*
	PU	15.64 $\pm$ 1.15	20	
	RL	19.61 $\pm$ 0.67	7.26 $\pm$ 1.13	
7000	GR	17.25 $\pm$ 0.43	20	2.46 [2.42, 2.50]*
	PU	15.92 $\pm$ 1.23	20	
	RL	19.71 $\pm$ 0.89	7.92 $\pm$ 1.44	
8000	GR	17.60 $\pm$ 0.47	20	2.68 [2.65, 2.71]*
	PU	16.15 $\pm$ 1.30	20	
	RL	20.28 $\pm$ 0.63	8.81 $\pm$ 1.32	

*Significant after Benjamini–Hochberg FDR correction at $\alpha = 0.05$.

Angle selection and reconstruction analysis. Figures 5 and 6 (a1-a2) illustrate the angle selections produced by the GR and RL policies for wedge-shaped phantoms at total doses of 6000 and 8000 photons, respectively. Each coloured spoke denotes the beam direction of a parallel-beam view at angle θ (degrees), measured counter-clockwise from the x -axis, with θ and $\theta + 180^\circ$ being equivalent. The colour coding indicates the number of repeats, corresponding to the relative photon allocation at that angle. The GR policy distributes views uniformly across the angular range, whereas the RL policy adapts its strategy by concentrating angles and dose around preferential directions corresponding to the wedge orientations. This adaptive allocation provides more informative measurements along high-contrast directions while reducing sampling in less informative regions.

Figures 5 and 6 (b1–b4) present reconstruction results for wedge-shaped phantoms using the GR and RL policies at total doses of 6000 and 8000 photons. For both dose levels, reconstructions obtained with the GR policy appear noisy and exhibit poor contrast between materials. In contrast, the RL policy produces reconstructions with clearer separation of material regions and reduced noise, demonstrating the advantage of adaptive angle selection and dose allocation.

4.3.2. Experiment 3: angle selection and dose allocation for defect detectability improvement

This experiment addresses the task of defect detection. Both dataset (b) (non-defective) and dataset (c) (defective) phantoms were used to evaluate whether the agent can select angles and allocate dose in a way that maintains reconstruction quality for non-defective phantoms while improving defect detectability. We still used the dose-aware PnP-PWLS with a TV prior as the reconstruction algorithm. The elliptical region is assumed to be defect-prone, and the ground-truth defect mask is used during training. Accordingly, we define an ROI covering the ellipse and compute rewards within this region. Figure 7 illustrates the ROI, the ground-truth defect mask, and the surrounding background region used for computing the CNR in Equation (18).

Training procedure. The agent was trained on a dataset of 300 phantoms, comprising 150 non-defective and 150 defective cases. Three separate RL policies were trained for three different total photon budgets of 1000, 1500, and 2000. Training was performed for 2000 episodes using the reward function defined in Equation (22). The weight parameters α and β in Equation (22) were set to 0.2 and 3.5, respectively, to balance the two tasks.

$$R(\bar{\mathbf{x}}, \hat{\mathbf{x}}) = \begin{cases} \alpha \text{PSNR}_{\Omega}(\bar{\mathbf{x}}, \hat{\mathbf{x}}), & \text{if } \bar{\mathbf{x}} \text{ is non - defective,} \\ \beta \text{CNR}_{\mathcal{D}, \mathcal{B}}(\bar{\mathbf{x}}, \hat{\mathbf{x}}), & \text{if } \bar{\mathbf{x}} \text{ is defective,} \end{cases} \quad (22)$$

where α and β control the relative importance of reconstruction accuracy and defect detectability.

Validation procedure. Evaluation was carried out on a separate test set of 200 phantoms with unseen rotation angles, including 100 non-defective and 100 defective cases. All results reported in this section are obtained from this validation set using the trained policy.

Quantitative comparison. Table 2 summarises the performance of the GR, PU, and RL policies on the unseen-rotation test set across total photon budgets of 1000, 1500, and 2000. For non-defective phantoms, the RL policy achieves PSNR comparable to or higher than GR at 1000 and 1500 photons, with significant

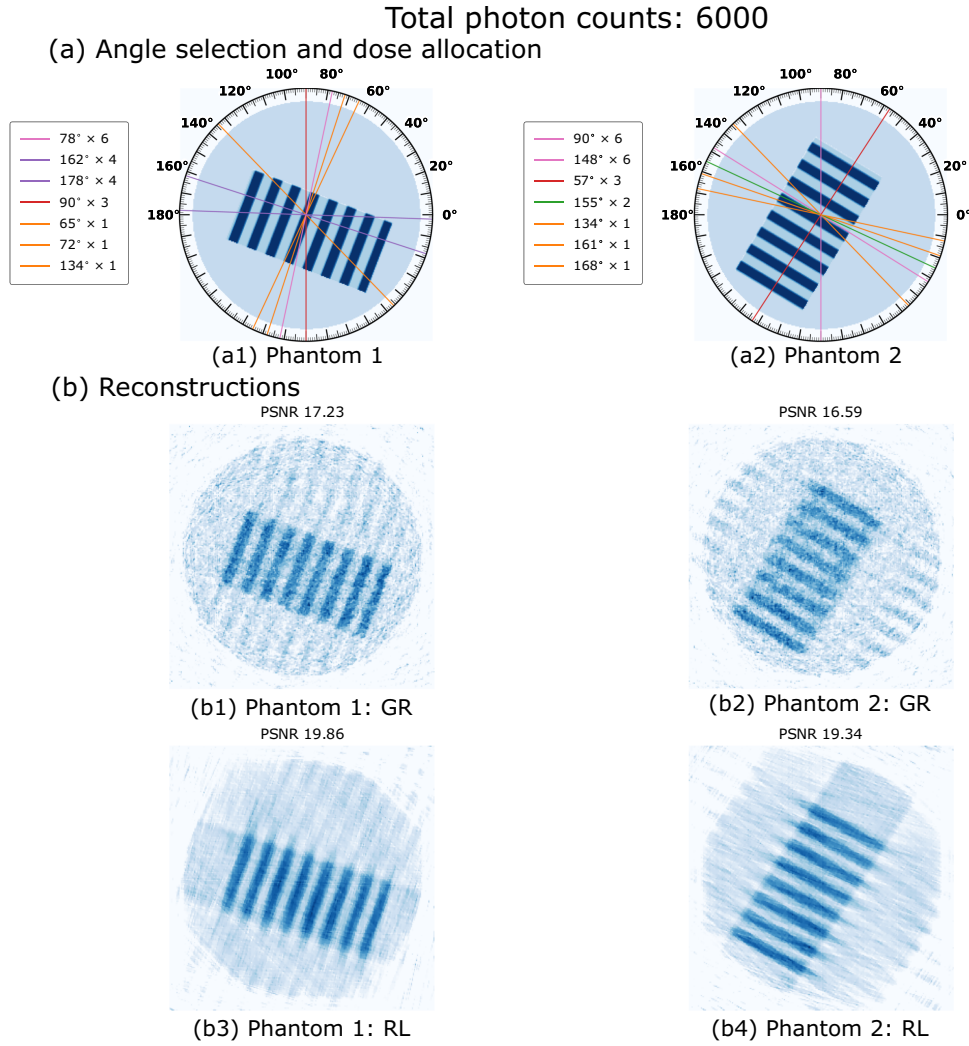


Figure 5. (a) Angle selection and dose allocation at a total dose of 6000 photons for wedge-shaped phantoms (a1–a2). Each coloured spoke represents a parallel-beam view at angle θ , measured counter-clockwise from the x -axis; angles θ and $\theta + 180^\circ$ are equivalent. Colours indicate the number of repeated selections and thus the allocated photons at that angle. With 6000 photons in total, each selection corresponds to 300 photons; for example, an orange spoke denotes one selection (300 photons), and a green spoke denotes two selections (600 photons). (b) Reconstructions obtained under the GR (b1–b2) and RL (b3–b4) policies at the same total dose. The RL policy adaptively concentrates measurements in more informative directions, yielding improved reconstruction quality.

mean improvements of 0.56 and 0.06 dB, respectively, based on paired bootstrap confidence intervals and Benjamini–Hochberg FDR correction. At 2000 photons, the RL policy obtains a slightly lower PSNR than GR, with a mean difference of -0.26 dB. For defective phantoms, however, the RL policy consistently achieves substantially higher CNR than GR at all dose levels. The mean CNR improvements are 0.47, 0.61, and 0.59 for total photon budgets of 1000, 1500, and 2000, respectively, and all corresponding 95% paired bootstrap confidence intervals are strictly positive after FDR correction. The PU baseline does not outperform GR in either PSNR or CNR, indicating that simple myopic uncertainty-based adaptivity is insufficient for this defect-detection task. Overall, these results show that the learned RL policy prioritises task-specific defect detectability while largely preserving reconstruction quality.

Angle selection and reconstruction analysis. Figures 8 and 9 (a1–a4) illustrate the angle selections produced by the GR and RL policies for non-defective and defective phantoms at total doses of 1000 and 2000 photons, respectively. While the GR policy distributes views uniformly regardless of phantom type, the RL policy adapts its strategy: for defective phantoms, angles and dose are concentrated around the

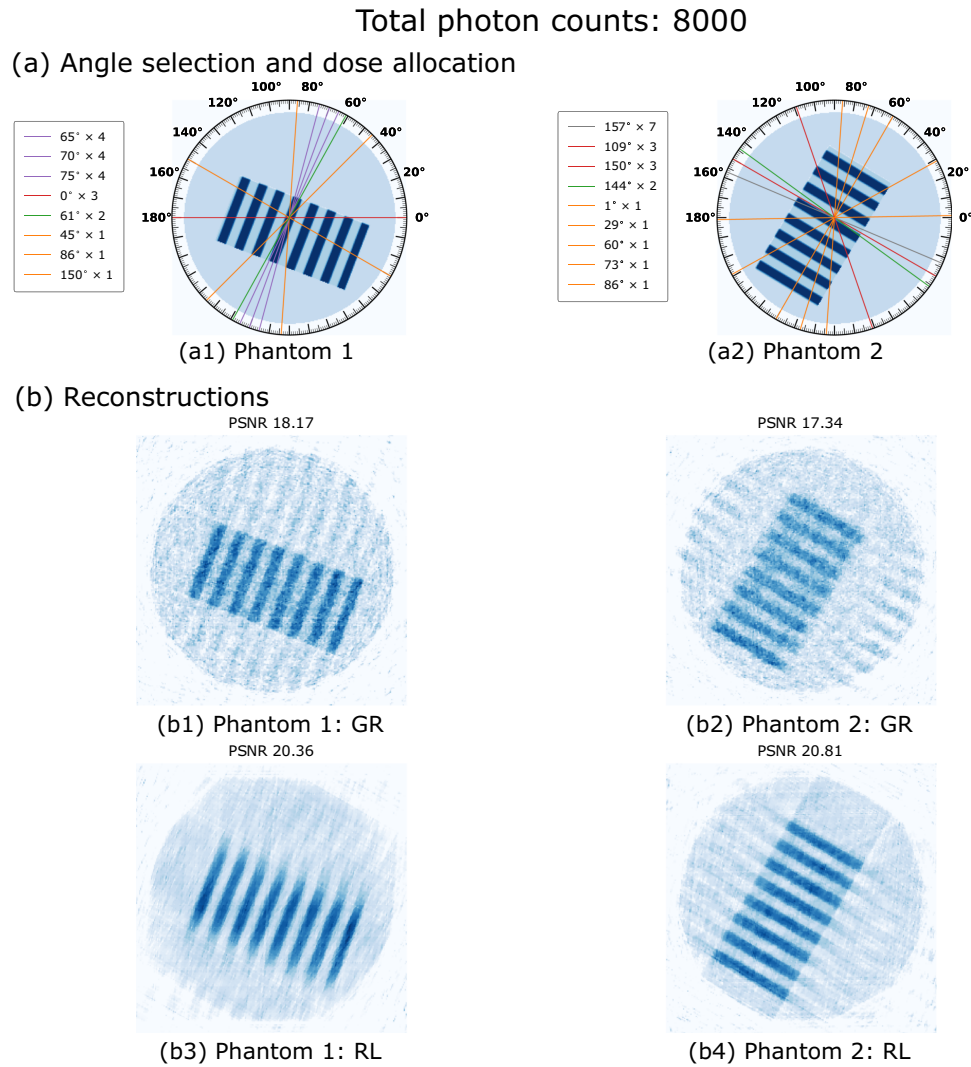


Figure 6. (a) Angle selection and dose allocation at a total dose of 8000 photons for wedge-shaped phantoms (a1–a2). Each coloured spoke represents a parallel-beam view at angle θ , measured counter-clockwise from the x -axis; angles θ and $\theta + 180^\circ$ are equivalent. Colours indicate the number of repeated selections and thus the allocated photons at that angle. With 8000 photons in total, each selection corresponds to 400 photons; for example, an orange spoke denotes one selection (400 photons), and a green spoke denotes two selections (800 photons). (b) Reconstructions obtained under the GR (b1–b2) and RL (b3–b4) policies at the same total dose. The RL policy adaptively concentrates measurements in more informative directions, yielding improved reconstruction quality.

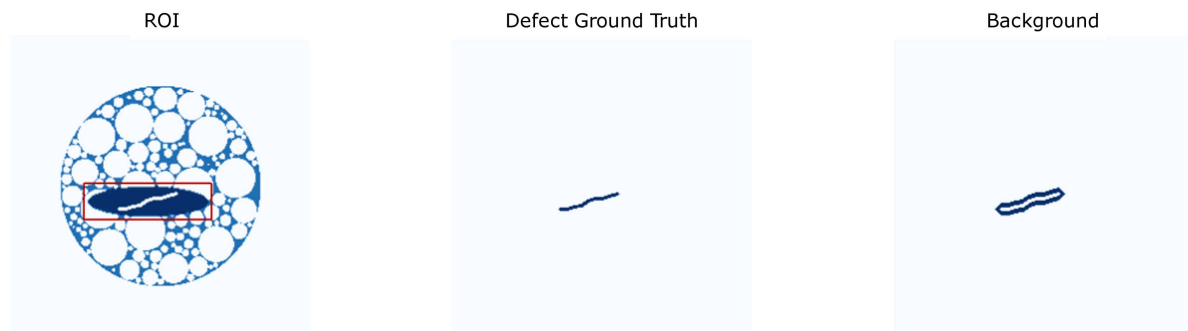


Figure 7. Illustration of the regions used for defect detection. (Left) ROI highlighted by a red rectangle. (Middle) Ground-truth defect mask. (Right) Background region used for CNR calculation.

Table 2. Performance comparison of GR, projected-uncertainty (PU), and RL policies on unseen-rotation tests for the foam-defect experiment. Results are reported as mean $\pm$ standard deviation. The RL–GR difference is reported with 95% paired bootstrap confidence intervals. One-sided p -values for RL outperforming GR were corrected using Benjamini–Hochberg FDR control at $\alpha = 0.05$.

Total photons	Policy	PSNR (Non-def.)	CNR (Def.)	Average number of angles	PSNR diff. RL–GR [95% CI]	CNR diff. RL–GR [95% CI]
1000	GR	19.50 $\pm$ 1.28	0.51 $\pm$ 0.25	10	0.56 [0.52, 0.60] *	0.47 [0.45, 0.50] *
	PU	19.19 $\pm$ 1.30	0.49 $\pm$ 0.39	10		
	RL	20.06 $\pm$ 1.52	0.98 $\pm$ 0.42	6.92 $\pm$ 1.41		
1500	GR	20.68 $\pm$ 1.23	0.62 $\pm$ 0.30	10	0.06 [0.02, 0.10] *	0.61 [0.59, 0.63] *
	PU	19.79 $\pm$ 1.35	0.59 $\pm$ 0.46	10		
	RL	20.74 $\pm$ 1.49	1.23 $\pm$ 0.42	7.15 $\pm$ 1.27		
2000	GR	21.38 $\pm$ 1.20	0.70 $\pm$ 0.32	10	−0.26 [−0.30, −0.22]	0.59 [0.57, 0.62] *
	PU	20.10 $\pm$ 1.44	0.62 $\pm$ 0.49	10		
	RL	21.12 $\pm$ 1.50	1.29 $\pm$ 0.47	7.56 $\pm$ 1.25		

*Significant after Benjamini–Hochberg FDR correction at $\alpha = 0.05$.

defect region. Comparing across dose levels shows that higher photon budgets lead to broader angular coverage, while the presence of defects drives the RL policy to focus sampling more locally.

The ultimate goal of this experiment is to evaluate whether adaptive angle and dose selection can improve defect detectability. Figures 8 and 9 (b1–b4) show reconstruction results for defective phantoms at total doses of 1000 and 2000 photons, respectively. Reconstructions using the GR policy fail to make the defects clearly visible. In contrast, reconstructions obtained with the RL policy consistently achieve higher CNR values and reveal defects more distinctly, validating the benefit of dose-aware adaptive view selection for defect detection.

Computational cost. To quantify the training cost without repeating the full training procedure, we benchmarked the runtime of representative training runs and measured the average wall-clock time per training episode. The benchmark was performed on a single NVIDIA L40 GPU with 44.39GB memory, using PyTorch 2.11.0+cu130 and CUDA 13.0. Since each training episode contains the same acquisition length and reconstruction procedure as in the full training, the total training cost was computed by multiplying the measured average per-episode runtime by the number of training episodes. For the wedge phantom experiment, 50 representative training episodes required 4.28 hours, corresponding to an average of 0.0856 hours per episode. Multiplying this by the full training schedule of 1500 episodes gives a training cost of approximately 128.33 GPU-hours. For the foam-defect experiment, 50 representative training episodes required 4.40 hours, corresponding to an average of 0.0880 hours per episode. Multiplying this by the full training schedule of 2000 episodes gives a training cost of approximately 176.06 GPU-hours. Because each run used a single GPU, the reported GPU-hours are equal to the wall-clock hours.

Summary. These experiments demonstrate that the proposed RL policy adaptively selects relative informative angles and allocates dose across projection angles according to phantom content. It preserves reconstruction quality for non-defective phantoms while substantially improving defect detectability in defective cases, consistently outperforming the GR policy in both CNR and visual clarity.

5. Discussion

These results demonstrate that the proposed dose-aware reconstruction algorithms are particularly effective when dose is distributed unequally across projection angles. Moreover, the RL policy achieves superior reconstruction quality and defect detectability by leveraging fewer angles and adaptively distributing dose within a fixed budget, compared to the GR policy, which relies on more angles with equal dose allocation.

The dose-aware PWLS formulation is based on a standard second-order approximation of the Poisson negative log-likelihood; the main novelty of this work lies in integrating this reconstruction model into an adaptive RL framework for joint angle selection and dose allocation. The angle-dependent weights provide a statistically motivated way to account for dose-dependent noise by assigning higher confidence to high-dose projections and lower confidence to noisier low-dose projections. For a fixed acquisition sequence, the PnP–PWLS update behaves as a regularised weighted least-squares iteration, and the estimated Lipschitz constant is used to stabilise the gradient step. However, because the RL policy changes the selected angles and dose distribution sequentially, the complete adaptive acquisition–reconstruction loop

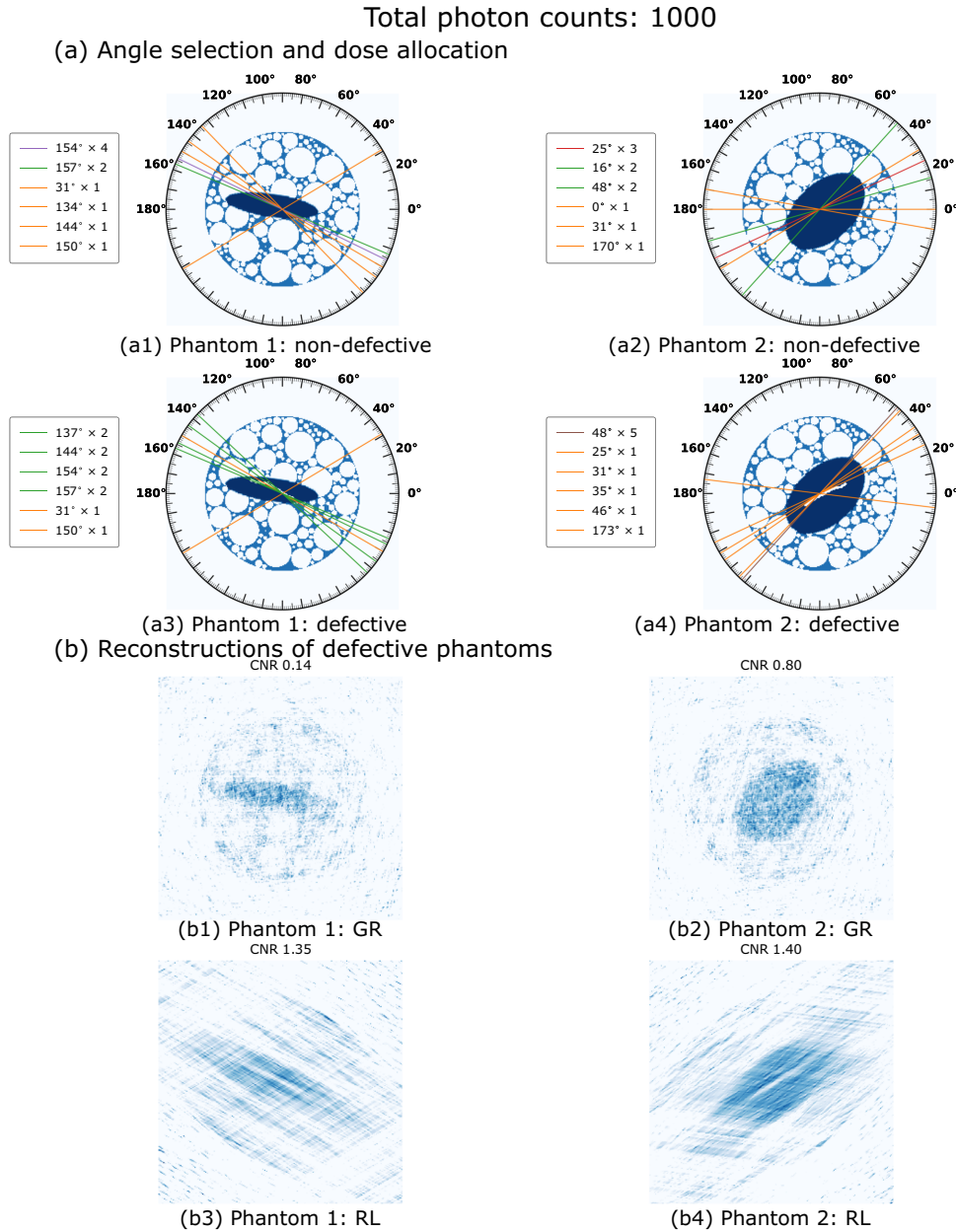


Figure 8. (a) Angle selection and dose allocation at a total dose of 1000 photons for non-defective (a1–a2) and defective (a3–a4) phantoms. Coloured spokes denote projection angles, with colour indicating the number of repeats (i.e. relative photon allocation). An orange spoke corresponds to one selection (base level: 100 photons), while colours such as green indicate multiple repeats and therefore higher allocated photons. (b) Reconstructions of defective phantoms using the GR (b1–b2) and RL (b3–b4) policies. The RL policy adaptively concentrates measurements near the defect, yielding higher CNR and improved defect visibility compared with the fixed GR policy.

does not optimise a single fixed objective, and we therefore do not claim a formal global convergence guarantee. The reward design also influences the learned acquisition strategy. PSNR- and CNR-based rewards are used only during offline training, where ground-truth phantoms are available; the trained policy does not require ground truth during inference. The weights α and β control the trade-off between reconstruction quality and defect detectability, making the learned policy task-dependent. Consequently, the policy is expected to generalise within the object class and imaging conditions represented in the training distribution, but different tasks, datasets, or imaging conditions may require retraining, fine-tuning, or an expanded training distribution. Alternative reward formulations based on data consistency,

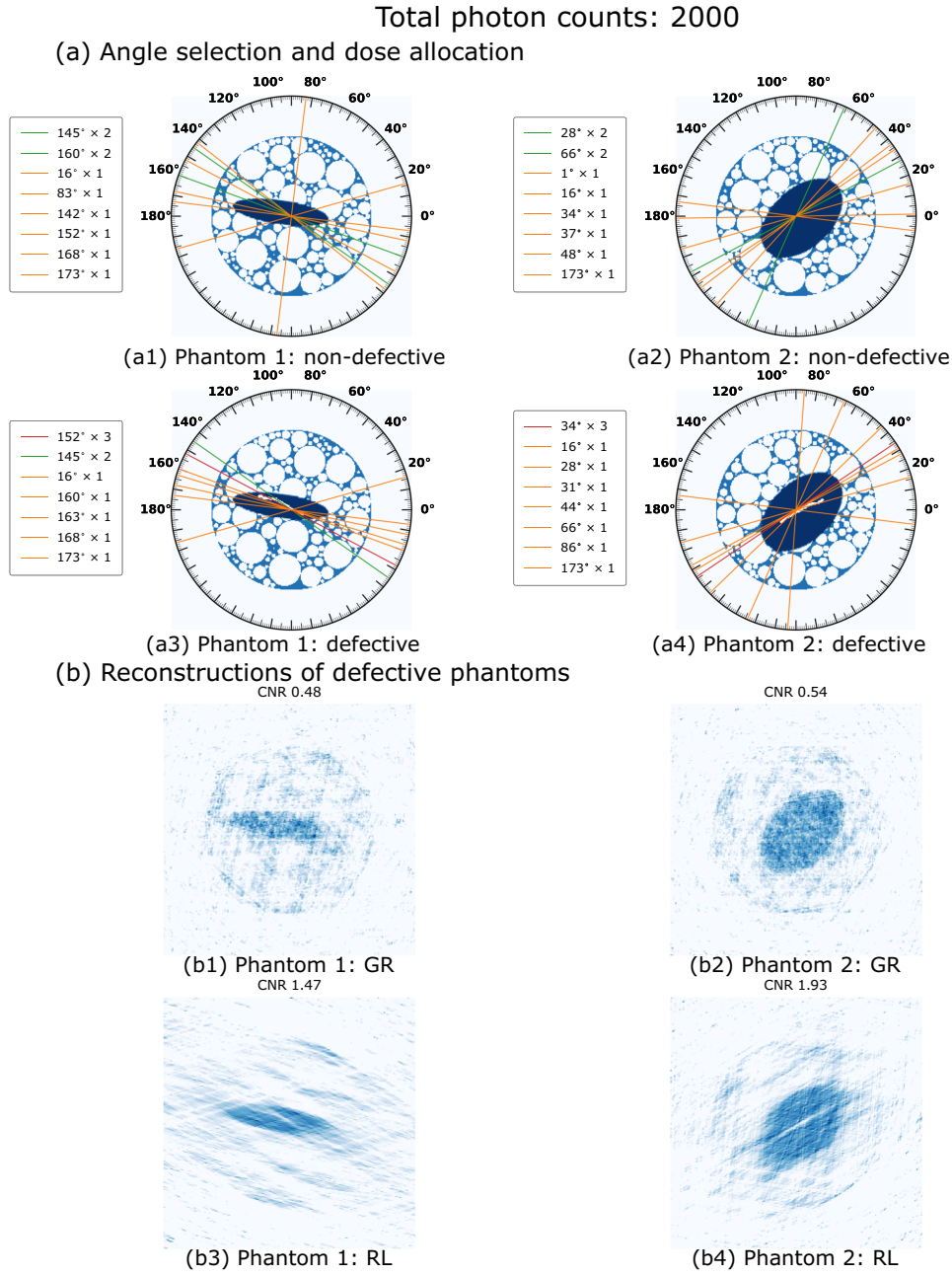


Figure 9. (a) Angle selection and dose allocation at a total dose of 2000 photons for non-defective (a1–a2) and defective (a3–a4) phantoms. Coloured spokes denote projection angles, with colour indicating the number of repeats (i.e. relative photon allocation). An orange spoke corresponds to one selection (base level: 200 photons), while colours such as green indicate multiple repeats and therefore higher allocated photons. (b) Reconstructions of defective phantoms using the GR (b1–b2) and RL (b3–b4) policies. The RL policy adaptively concentrates measurements near the defect, yielding higher CNR and improved defect visibility compared with the fixed GR policy.

uncertainty reduction, expected information gain, or downstream task confidence may further reduce reliance on ground-truth-based rewards and are important directions for future work.

Despite these encouraging findings, several limitations remain. First, only a Poisson noise model was considered, whereas real CT measurements are often better described by a mixture of Gaussian and Poisson noise [34,35]. Accurate modelling of the noise characteristics of a specific CT system will be important for translating this approach to real data. Second, dose allocation is represented through repeated selection of the same projection angle. This provides a convenient abstraction for concentrating

photons under a fixed dose budget, but practical implementation may require scanner-specific mechanisms such as exposure-time control, tube-current modulation, or repeated measurements. Third, extending the framework to three-dimensional datasets introduces additional challenges: iterative reconstruction becomes more computationally demanding, and the state and action spaces for RL grow substantially. Fourth, the validation experiments use unseen rotations, shifts, and shape variations, indicating generalisation within the training distribution. However, we do not claim generalisation to arbitrary out-of-distribution geometries or unrelated object classes. In practical industrial CT applications, this assumption is often reasonable because prior information about the inspected object class is available through CAD models, blueprints, design specifications, or previous scans. For substantially different geometries or imaging conditions, retraining, fine-tuning, or domain-randomised training would be required. Finally, future work may explore the integration of downstream tasks such as segmentation and classification, which could further highlight the practical benefits of adaptive angle and dose allocation.

6. Conclusion

In this paper, we proposed dose-aware reconstruction methods and developed a reinforcement learning strategy for joint angle selection and dose allocation. Our results demonstrate that explicitly modelling angle-dependent dose variation improves reconstruction quality and defect detectability under unequal dose distributions. Furthermore, the proposed RL-based framework shows the potential to reduce the number of projection angles while maintaining or improving image quality, thereby enabling more efficient and task-adaptive CT scanning.

Acknowledgements

The authors gratefully acknowledge the support of the xCTing project. We also acknowledge the use of OpenAI's large language model, ChatGPT based on GPT-4o, to assist in refining the manuscript text. The tool was used at the sentence level for tasks such as grammar correction and sentence rephrasing.

Author contributions

CRediT: **Tianyuan Wang:** Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; **Daniël M. Pelt:** Conceptualization, Visualization, Writing – review & editing; **Felix Lucka:** Investigation, Validation, Writing – review & editing; **Tristan van Leeuwen:** Conceptualization, Supervision, Validation; **Kees Joost Batenburg:** Conceptualization, Project administration, Resources.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was co-financed by the European Union H2020-MSCA-ITN-2020 under grant agreement no. 956172 (xCTing).

Data availability statement

The data that support the findings of this study are available from the corresponding author, T.W., upon reasonable request. The implementation and synthetic datasets used in this study are publicly available on GitHub: <https://github.com/tianyuan1wang/ct-angle-dose-rl> and Zenodo: <https://doi.org/10.5281/zenodo.20486336>.

References

- [1] McCollough CH, Bruesewitz MR, Kofler Jr JM. CT dose reduction and dose management tools: overview of available options. *Radiographics*. 2006;26(2):503–512. doi: [10.1148/rg.262055138](https://doi.org/10.1148/rg.262055138)
- [2] Kazantsev I. Information content of projections. *InvPr*. 1991;7(6):887–898. doi: [10.1088/0266-5611/7/6/010](https://doi.org/10.1088/0266-5611/7/6/010)

- [3] Batenburg KJ, Palenstijn WJ, Balázs P, et al. Dynamic angle selection in binary tomography. *Comput Vis Image Underst.* 2013;117(4):306–318. doi: [10.1016/j.cviu.2012.07.005](https://doi.org/10.1016/j.cviu.2012.07.005)
- [4] Elata N, Michaeli T, Elad M. Adaptive compressed sensing with diffusion-based posterior sampling In: *European Conference on Computer Vision*. Springer; 2025. p. 290–308. doi: [10.1007/978-3-031-73229-4_17](https://doi.org/10.1007/978-3-031-73229-4_17)
- [5] Barbano R, Leuschner J, Antorán J, et al. Bayesian experimental design for computed tomography with the linearised deep image prior, *arXiv preprint arXiv:2207.05714*. 2022.
- [6] Wang T, Lucka F, van Leeuwen T. Sequential experimental design for X-ray CT using deep reinforcement learning. *IEEE Trans Comput Imaging.* 2024;10:953–968. doi: [10.1109/TCI.2024.3414273](https://doi.org/10.1109/TCI.2024.3414273)
- [7] Ruthotto L, Chung J, Chung M. Optimal experimental design for inverse problems with state constraints. *SIAM J Sci Comput.* 2018;40(4):B1080–B1100. doi: [10.1137/17M1143733](https://doi.org/10.1137/17M1143733)
- [8] Lindley DV. Bayesian statistics: A review. *SIAM*; 1972.
- [9] Ryan E, Drovandi C, McGree J, et al. Fully Bayesian optimal experimental design: a review. *Int Stat Rev.* 2016;84:128–154. doi: [10.1111/insr.12107](https://doi.org/10.1111/insr.12107)
- [10] Rainforth T, Foster A, Ivanova DR, et al. Modern Bayesian experimental design. *Stat Sci.* 2024;39(1):100–114. doi: [10.1214/23-STS915](https://doi.org/10.1214/23-STS915)
- [11] Batenburg KJ, Fortes W, Hajdu L, et al. Bounds on the difference between reconstructions in binary tomography. In: *International Conference on Discrete Geometry for Computer Imagery*. Springer; 2011. p. 369–380.
- [12] Dabrovolski A, Batenburg KJ, Sijbers J. Dynamic angle selection in X-ray computed tomography. *Nucl Instrum Methods Phys Res, Sect B.* 2014;324:17–24. doi: [10.1016/j.nimb.2013.08.077](https://doi.org/10.1016/j.nimb.2013.08.077)
- [13] Helin T, Hyvönen N, Puska J-P. Edge-promoting adaptive Bayesian experimental design for X-ray imaging. *SIAM J Sci Comput.* 2022;44(3):B506–B530. doi: [10.1137/21M1409330](https://doi.org/10.1137/21M1409330)
- [14] Shen Z, Wang Y, Wu D, et al. Learning to scan: a deep reinforcement learning approach for personalized scanning in CT imaging. *Inverse Probl Imaging.* 2022;16(1):179. doi: [10.3934/ipi.2021045](https://doi.org/10.3934/ipi.2021045)
- [15] Wang T, Florian V, Schielein R, et al. Task-adaptive angle selection for computed tomography-based defect detection. *J Imaging.* 2024;10(9):208. doi: [10.3390/jimaging10090208](https://doi.org/10.3390/jimaging10090208)
- [16] Wang T, Lucka F, Pelt DM, et al. Dynamic angle selection in X-ray CT: A reinforcement learning approach to optimal stopping, *arXiv preprint arXiv:2503.12688*. 2025.
- [17] Wang J, Li T, Lu H, et al. Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography. *IEEE Trans Med Imaging.* 2006;25(10):1272–1283. doi: [10.1109/TMI.2006.882141](https://doi.org/10.1109/TMI.2006.882141)
- [18] Kazantsev D, Bleichrodt F, van Leeuwen T, et al. A novel tomographic reconstruction method based on the robust student's t function for suppressing data outliers. *IEEE Transactions on Computational Imaging.* 2017;3(4):682–693. doi: [10.1109/TCI.2017.2694607](https://doi.org/10.1109/TCI.2017.2694607)
- [19] Xia W, Lu Z, Huang Y, et al. Ct reconstruction with pdf: parameter-dependent framework for data from multiple geometries and dose levels. *IEEE Trans Med Imaging.* 2021;40(11):3065–3076. doi: [10.1109/TMI.2021.3085839](https://doi.org/10.1109/TMI.2021.3085839)
- [20] Hansen PC, Jørgensen J, Lionheart WR. Computed tomography: algorithms, insight, and just enough theory. *SIAM*; 2021.
- [21] Shen W, Huan X. Bayesian sequential optimal experimental design for nonlinear models using policy gradient reinforcement learning. *Comput Methods Appl Mech Eng.* 2023;416:116304. doi: [10.1016/j.cma.2023.116304](https://doi.org/10.1016/j.cma.2023.116304)
- [22] Sutton RS, Barto AG. Reinforcement learning: An introduction. *MIT Press*; 2018.
- [23] Schulman J, Moritz P, Levine S, et al. Highdimensional continuous control using generalized advantage estimation, *arXiv preprint arXiv:1506.02438*. 2015.
- [24] Fessler, JA. Penalized weighted least-squares image reconstruction for positron emission tomography. *IEEE Trans Med Imaging.* 1994;13(2):290–300. doi: [10.1109/42.293921](https://doi.org/10.1109/42.293921)
- [25] Venkatakrishnan SV, Bouman CA, Wohlberg B. Plug-and-play priors for model based reconstruction. In: 2013 IEEE global conference on signal and information processing. *IEEE*; 2013. p. 945–948. doi: [10.1109/GlobalSIP.2013.6737048](https://doi.org/10.1109/GlobalSIP.2013.6737048)
- [26] Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. *Physica D.* 1992;60(1-4):259–268. doi: [10.1016/0167-2789\(92\)90242-F](https://doi.org/10.1016/0167-2789(92)90242-F)
- [27] Chambolle A. An algorithm for total variation minimization and applications. *J Math Imaging Vis.* 2004;20(1):89–97. doi: [10.1023/B:JMIV.0000011325.36760.1e](https://doi.org/10.1023/B:JMIV.0000011325.36760.1e)
- [28] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347*. 2017.
- [29] Pelt DM, Hendriksen AA, Batenburg KJ. Foam-like phantoms for comparing tomography algorithms. *Synchrotron Radiat.* 2022;29(1):254–265. doi: [10.1107/S1600577521011322](https://doi.org/10.1107/S1600577521011322)
- [30] Hendriksen AA, Schut D, Palenstijn WJ, et al. Tomosipo: fast, flexible, and convenient 3D tomography for complex scanning geometries in python. *Opt Express.* 2021;29(24):40 494–40 513. doi: [10.1364/OE.439909](https://doi.org/10.1364/OE.439909)
- [31] Kingma DP, & Ba J. Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*. 2014.
- [32] Kohler T. A projection access scheme for iterative reconstruction based on the golden section. In: *IEEE Symposium Conference Record Nuclear Science 2004 (Vol. 6) IEEE*; 2004. p. 3961–3965. doi: [10.1109/NSSMIC.2004.1466745](https://doi.org/10.1109/NSSMIC.2004.1466745)

- [33] Craig TM, Kadu AA, Batenburg KJ, et al. Real-time tilt undersampling optimization during electron tomography of beam sensitive samples using golden ratio scanning and RECAST3D. *Nanos*. 2023;15(11):5391–5402. doi: [10.1039/D2NR07198C](https://doi.org/10.1039/D2NR07198C)
- [34] Andriashen V, van Lier R, van Leeuwen T, et al. X-ray image generation as a method of performance prediction for real-time inspection: a case study. *JNE*. 2024;43(3):79. doi: [10.1007/s10921-024-01091-8](https://doi.org/10.1007/s10921-024-01091-8)
- [35] Graas A, Lucka F. Scintillator decorrelation for self-supervised X-ray radiograph denoising. *MeScT*. 2025;36(6):065415. doi: [10.1088/1361-6501/addc06](https://doi.org/10.1088/1361-6501/addc06)

Appendices

Appendix A. Power iteration for Lipschitz constant estimation

To ensure stable gradient descent updates in Algorithm 1, the step size is normalised by the Lipschitz constant

$$L = \|\mathbf{A}^\top \mathbf{W} \mathbf{A}\|,$$

which corresponds to the spectral norm of the weighted normal operator. Since exact computation of L is infeasible for large-scale CT systems, we approximate it using the standard *power iteration* method. The procedure is summarised in Algorithm 3.

Algorithm 3. Power Iteration to Estimate Spectral Norm $L = \|\mathbf{A}^\top \mathbf{W} \mathbf{A}\|$.

Require: Forward operator $\mathbf{A}(\boldsymbol{\theta})$; transpose operator $\mathbf{A}^\top(\boldsymbol{\theta})$; weighting operator $\mathcal{W}(\cdot)$; domain size n ; maximum iterations T .

Ensure: Spectral norm estimate L .

- 1: Initialize: $\mathbf{z}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$; normalize $\mathbf{z}^{(0)} \leftarrow \mathbf{z}^{(0)} / \|\mathbf{z}^{(0)}\|$
 - 2: **for** $t = 0$ **to** $T - 1$ **do**
 - 3: $\mathbf{u} \leftarrow \mathbf{A}(\boldsymbol{\theta}) \mathbf{z}^{(t)}$ ▷ Forward projection
 - 4: $\mathbf{v} \leftarrow \mathcal{W}(\mathbf{u})$ ▷ Dose-aware weighting
 - 5: $\mathbf{w} \leftarrow \mathbf{A}^\top(\boldsymbol{\theta}) \mathbf{v}$ ▷ Backprojection
 - 6: Normalize: $\mathbf{z}^{(t+1)} \leftarrow \mathbf{w} / \|\mathbf{w}\|$
 - 7: Update estimate: $L^{(t+1)} \leftarrow \|\mathbf{w}\|$
 - 8: **end for**
 - 9: **return** $L = L^{(T)}$
-

In practice, only a small number of iterations (e.g. $T \leq 20$) is sufficient to obtain an accurate estimate of the largest eigenvalue, which is then used to set the step size $\alpha = h/L$ in Algorithm 1.

Appendix B. Additional PPO training details

The training loop in Algorithm 2 (main text) summarises the high-level workflow of PPO in the CT environment. For clarity and readability, minibatch updates, entropy regularisation, and advantage estimation are omitted in the main algorithm but are detailed here for reproducibility. We denote the parametric approximation of the value function by $V(\cdot; \mathbf{w}_v)$ to distinguish it from the (unknown) true value function $V^\pi(\cdot)$. This convention follows our previous work [6].

Notation.

- $\hat{\mathbf{x}}_k$: reconstructed state at step k ,
- Θ : full set of available candidate projection angles,
- $\theta_k \in \Theta$: projection angle selected at step k ,
- $\boldsymbol{\theta}$: sequence of selected projection angles used for reconstruction,
- r_k : immediate reward (see Equation (22)),
- $\pi_a(\cdot | \hat{\mathbf{x}}_k; \mathbf{w}_a)$: stochastic policy over Θ with parameters $\mathbf{w}_a$,
- $V(\hat{\mathbf{x}}_k; \mathbf{w}_v)$: approximate value function (critic) with parameters $\mathbf{w}_v$,
- G_k : Monte Carlo return from step k ,
- $\hat{A}_k$: estimated advantage at step k ,
- $e_k(\mathbf{w}_a) = \frac{\pi_a(\theta_k | \hat{\mathbf{x}}_k; \mathbf{w}_a)}{\pi_a^{\text{old}}(\theta_k | \hat{\mathbf{x}}_k)}$: policy probability ratio.

Dose allocation is not treated as a separate action variable; instead, it is induced by repeated selections of the same projection angle.

Table B1. Key architectural and algorithmic parameters of the RL pipeline.

Parameter	Value
Input image size	256 × 256
Action-space dimension	180 candidate projection angles
Training seed	1029
Evaluation seeds	20260526 + r , where r is the replication index
Discount factor γ	0.99
GAE parameter λ	0.95
PPO clipping parameter ϵ	0.2
PPO update epochs	4
Mini-batch size	64
Timesteps per batch	540
Optimiser	Adam
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Training episodes, wedge	1500
Training episodes, foam-defect	2000
Convolutional encoder	
Conv block 1	Conv2D (1, 12, 3, stride = 2, padding = 1) + GroupNorm (4, 12) + LeakyReLU (0.2) + MaxPool2D (2)
Conv block 2	Conv2D (12, 24, 3, padding = 1) + GroupNorm (4, 24) + LeakyReLU (0.2) + MaxPool2D (2)
Conv block 3	Conv2D (24, 48, 3, padding = 1) + GroupNorm (4, 48) + LeakyReLU (0.2) + MaxPool2D (4)
Actor head	Linear(flattened features, 180) + Softmax
Critic head	Linear(flattened features, flattened features) + ReLU + Linear(flattened features, 1)

Minibatch Updates. As in standard PPO, parameter updates are performed over multiple epochs using minibatch sampling from the collected trajectory set. Although each episode produces only a small number of transitions (M steps), we construct an effective training batch by aggregating transitions from multiple episodes until a predefined batch size (e.g. 64 samples) is reached. This ‘batch-size accumulation’ strategy stabilises optimisation, ensures sufficient gradient diversity, and enables PPO updates even in environments where each trajectory is short (as in sparse-view CT).

Advantage Estimation. We use generalised advantage estimation (GAE) with discount factor γ and parameter $\lambda \in [0, 1]$. Following the PPO update scheme, the temporal-difference residuals are computed using the value function parameters $\mathbf{w}_v^{\text{old}}$, which are frozen during the update epoch:

$$\delta_k = r_k + \gamma V(\hat{\mathbf{x}}_{k+1}; \mathbf{w}_v^{\text{old}}) - V(\hat{\mathbf{x}}_k; \mathbf{w}_v^{\text{old}}).$$

The advantages are then estimated as

$$\hat{A}_k = \sum_{t=0}^{T-k-1} (\gamma\lambda)^t \delta_{k+t}, \quad G_k = \hat{A}_k + V(\hat{\mathbf{x}}_k; \mathbf{w}_v^{\text{old}}).$$

Entropy Regularisation. To encourage exploration, we add the policy entropy:

$$\mathcal{H}(\pi_a(\cdot | \hat{\mathbf{x}}_k)) = - \sum_{\theta} \pi_a(\theta | \hat{\mathbf{x}}_k) \log \pi_a(\theta | \hat{\mathbf{x}}_k).$$

Total PPO Objective. Combining the actor and critic losses from the main text with entropy regularisation, the overall objective is

$$J(\mathbf{w}_a, \mathbf{w}_v) = \mathbb{E} [L^{\text{actor}}(\mathbf{w}_a) - c_v L^{\text{critic}}(\mathbf{w}_v) + c_e \mathcal{H}(\pi_a(\cdot | \hat{\mathbf{x}}; \mathbf{w}_a))],$$

where c_v and c_e are positive coefficients weighting the critic and entropy terms, respectively.