Multi-Agent Reinforcement Learning Meets Leaf Sequencing in Radiotherapy

Riqiang Gao [1] **Florin C. Ghesu** [2] **Simon Arberet** [1] **Shahab Basiri** [3] **Esa Kuusela** [3] **Martin Kraus** [2]
Dorin Comaniciu [1] **Ali Kamen** [1]

Abstract

In contemporary radiotherapy planning (RTP), a key module leaf sequencing is predominantly addressed by optimization-based approaches. In this paper, we propose a novel deep reinforcement learning (DRL) model termed as *Reinforced Leaf Sequencer* (RLS) in a multi-agent framework for leaf sequencing. The RLS model offers improvements to time-consuming iterative optimization steps via large-scale training and can control movement patterns through the design of reward mechanisms. We have conducted experiments on four datasets with four metrics and compared our model with a leading optimization sequencer. Our findings reveal that the proposed RLS model can achieve reduced fluence reconstruction errors, and potential faster convergence when integrated in an optimization planner. Additionally, RLS has shown promising results in a full artificial intelligence RTP pipeline. We hope this pioneer multi-agent RL leaf sequencer can foster future research on machine learning for RTP.

1. Introduction

Radiotherapy (RT) stands as an essential cornerstone in the realm of cancer treatment, recommended for approximately half of cancer patients (Huynh et al., 2020). Despite technological advances, a substantial part of RT still depends on labor-intensive and time-consuming planning pipelines from a diverse healthcare team (Elmore et al., 2019).

Radiotherapy Planning (RTP), originates from 1890s, refers to the process to plan the appropriate external beam RT treatment for patients with cancer. Modern RTP involves tailoring personalized treatment plans to effectively target tumors through Planning Target Volume (PTV) while safeguarding healthy tissues, as indicated by Organs at Risk (OARs). In practical RT, the RTP should also consider that the plan can be delivered in a reasonable time. Current RTPs are dominant by optimization pipelines in practical platforms such as Varian (Varian, 2023b) and Elekta (Elekta, 2023). Though different treatment types are existing, e.g., intensity-modulated radiation therapy (IMRT) (Mundt & Roeske, 2005) and volumetric-modulated arc therapy (VMAT) (Otto, 2008), their optimization pipelines share similar key components: *dose objectives definition*, *optimal fluence prediction*, and *leaf sequencing* to get feasible machine parameters, as shown in Figure 1[1]. *Optimal fluence prediction* estimates a 2D intensity map of the radiation beam for each beam angle allowing for customizable dose delivery to meet the objectives, however, the practicality of the fluence may be limited as it could potentially violate certain physical rules necessary for achievability. *Leaf sequencing* is to approximate optimal fluences with feasible machine movements. Depending on the implementation details of algorithms, the fluence map optimization and leaf sequencing can be performed sequentially and iteratively.

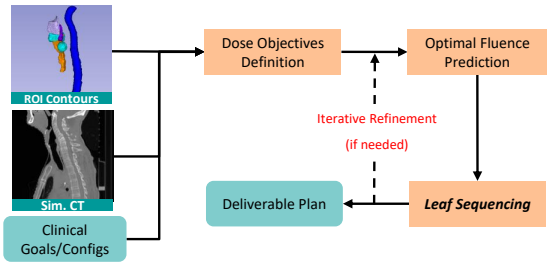


Figure 1. Illustration of a typical RTP process. Three common components are shown in the orange boxes. We focus on *leaf sequencing* in this work. The term "optimization" in this paper refers to a series of methods that are not machine learning.

To learn knowledge from large-scale existing plans, deep supervised learning has been widely applied to *dose objectives definition* and *optimal fluence prediction* modules (details in Section 2). *Leaf sequencing* is relatively challenging

[1]Digital Technology and Innovation, Siemens Healthineers, Princeton NJ, USA [2]Digital Technology and Innovation, Siemens Healthineers, Erlangen, Germany [3]Varian Medical Systems, Siemens Healthineers, Helsinki, Finland. Correspondence to: Riqiang Gao <riqiang.gao@siemens-healthineers.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]In-depth RTP introduction (background, terms, visuals, etc.) is in Appendix A. Some leaf sequencing optimizations for VMAT skip fluence prediction. Here we mainly discuss the branch with fluence prediction in the pipeline, such as RapidArc™ (Varian).

with supervised learning because its output space is large in the serial nature, especially for VMAT. For example, a typical HD-120 VMAT plan with 178 control points (CPs) (Bergman et al., 2014) has over 20K degrees of serial actions from leaf positions and monitor units.

Apart from being unable to leverage learning from existing data, the absence of a differentiable leaf sequencing module poses a challenge for training the overall end-to-end planning pipeline. In the spirit of recent success in decision-making tasks, we propose a new deep reinforcement learning (DRL) model, termed as *Reinforced Leaf Sequencer* (RLS), for a learning-based leaf sequencing. Our major contributions can be summarized as below:

- To our best knowledge, we are the first to successfully formulate *leaf sequencing* problems in practical radiotherapy scenarios with a deep multi-agent RL model.
- We propose five reward components to reasonably guide the movement of leaves and monitor units. The move pattern is controllable by tuning the weight of rewards, which enables human preferences in the loop.
- Compare to vanilla RL or optimization methods, RLS abandons iterative processing and only executes once for each CP for accelerated inference.
- Experiments conducted on four datasets across two cancer sites, RLS achieved 1) better or close performances compared to a leading optimization sequencer in a VMAT optimization platform (i.e., PORIx mentioned later); 2) effective prediction in a full-AI pipeline.

2. Related Work

Leaf Sequencing in RTP Optimization. RTP is mainly solved by optimization-based methods currently in both IMRT and VMAT (Xia & Verhey, 1998; Shepard et al., 2002; Earl et al., 2003; Ripsman et al., 2022; Carrasqueira et al., 2023; Fallahi et al., 2022; Jhanwar et al., 2023; Cedric & Tang, 2011). There are different pipelines of RTP optimization, one major branch (as in Figure 1), such as Cao et al. (2006); Shepard et al. (2007); Wang et al. (2008); Bedford (2009), is including the leaf sequencing to convert the optimized beam intensities into deliverable multi-leaf collimator (MLC) segments to form arc(s) or fields by starting from fluence maps. RapidArc™ from Varian is a representative commercial example in this branch which initially adopted the algorithm in Otto (2008) and continually upgraded in subsequent years. RapidArc has directly promoted the large-scale clinical implementation of VMAT (Cedric & Tang, 2011; Infusino, 2015). Compared to machine learning, disadvantages of optimization algorithms include time-consuming iterative execution and the lack of leveraging knowledge from large-scale training data.

AI for RTP. Recently, researchers aim to translate artificial intelligence (AI) to RTP to achieve fast and high-quality planning (Huynh et al., 2020; Luchini et al., 2022) to replace one or more components of an optimization pipeline. To replace conventional point-based objectives, RapidPlan™ introduce dose-volume histograms (DVHs) proposal as line-objectives (Fogliata et al., 2019; Varian, 2023a). Dose prediction from patient data (e.g., CT and contouring) with deep learning may serve as 3D objectives (Barragán-Montero et al., 2019; Kearney et al., 2020; Wang et al., 2022; Gao et al., 2023; Feng et al., 2023; Jiao et al., 2023; Gronberg et al., 2023). Following dose prediction, deep learning methods are used to replace fluence map optimization (Romeijn et al., 2003; 2004) to predict 2D fluence maps for each beam angle by feeding projections from predicted 3D dose (Wang et al., 2020; 2021; Ma et al., 2020; Yuan et al., 2022).

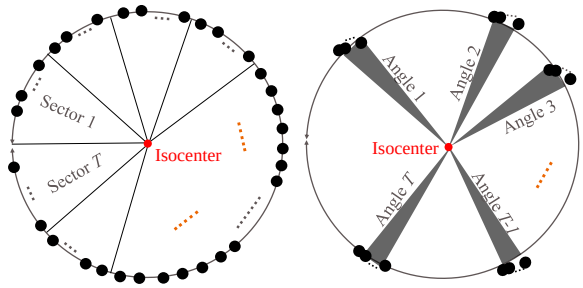
Deep RL and Multi-Agent. Deep RL has made remarkable progress in sequential decision-making tasks, including strategy games (Mnih et al., 2013; Silver et al., 2016), robotic control (Mnih et al., 2015; Schulman et al., 2017), autonomous driving (Kiran et al., 2021), and landmark detection (Ghesu et al., 2017). Many of these applications involve multiple agents, necessitating a systematic approach to modeling as multi-agent RL (MARL) (Zhang et al., 2021). Compared to value-based RLs (Mnih et al., 2015; Coulom, 2006), policy-based methods directly explore the policy space, with better convergence guarantees (Konda & Tsitsiklis, 1999; Yang et al., 2018; Wang et al., 2019; Zhang et al., 2021). Actor-critic algorithms in policy-based branch, such as PPO (Schulman et al., 2017) and SAC (Haarnoja et al., 2018), have achieved state-of-the-art performance in many applications. PPO showed surprising effectiveness in cooperative multi-agent games (Yu et al., 2022).

RL has been applied to RTP tasks. Hrinivich & Lee (2020); Ganeshan (2021) explored DQN for planning, however, these studies lie in single-leaf-pair simulations. A concurrent work (Hrinivich et al., 2024) extended Hrinivich & Lee (2020) to enable VMAT in a 3D beam context. Differently, we target modularizing RTP pipeline to be more flexible; making leaf sequencing can leverage advances of other AI models (e.g., dose/fluence predictions) and enabling integration within optimization loops. *To the best of our knowledge, the proposed RLS is the first MARL-based leaf sequencer in practical RTP scenarios (i.e., outputs include leaf positions at each leaf pair and monitor unit for each control point).*

3. Problem Formulation and Intuition

This paper focuses on leaf sequencing, the final module in RTP as in Figure 1, to achieve deliverable outcomes—a domain scarcely explored within advanced MARL.

There are two common approaches for treatment: IMRT



(a) VMAT (b) IMRT

Figure 2. Control point distribution: Each black dot signifies a control point. (a) In VMAT, control points can be approximate-evenly spaced along the designated arc during therapy, and may be divided into multiple sectors during planning and optimization. (b) The IMRT plan is comprised of several fields, each encompassing multiple control points at identical gantry angles.

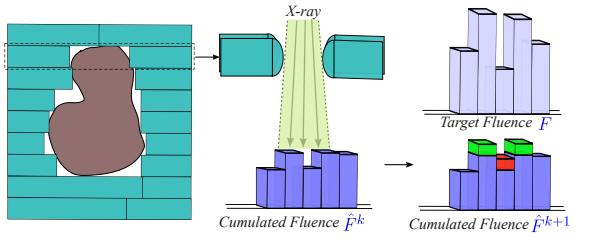
(Mundt & Roeske, 2005) and VMAT (Otto, 2008) (as in Figure 2 and more details in Appendix A.1). IMRT employs multiple beam-angles (i.e., fields) with modulated intensity using multileaf collimators (MLCs) to create precise dose profiles. VMAT offers improved delivery efficiency compared to IMRT, using numerous beam directions along arc trajectory to delivery dose dynamically (Teh et al., 1999; Teoh et al., 2011; Quan et al., 2012). It is customary for IMRT to generate a sequence of control points for a specific beam angle. In VMAT scenarios, various optimization techniques, including RapidArc, address leaf sequencing by consolidating control points within a sector during optimization (Unkelbach et al., 2015; Amendola et al., 2013). *Consequently, the leaf sequencing in both VMAT and IMRT can be viewed as the process of deriving a sequence of control points, specifying leaf positions and monitor units based on target fluences, i.e., optimal fluences.*

Note that the fluence prediction component's target fluence F cannot be directly translated into outcomes for the RT machine. The leaf sequencer is designed to predict outcomes that are feasible for the machine. Given X leaf pairs in the MLC for all K control points contributing to the fluence F and denote the leaf sequencer as $f(\cdot)$, we have:

$$(P, M) = f(F), \tag{1}$$

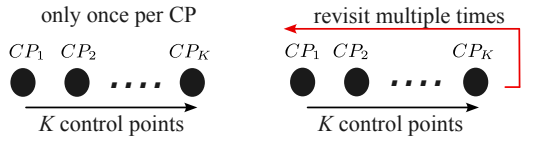
where P denotes the leaf positions with a size of $K \times X \times 2$ and M is a $K \times 1$ vector representing monitor unit (MU).

Intuition. We summarize our intuition in Figure 3 and 4. Leaf sequencing can be conceptualized as a sequence of decision-making tasks. In this sequence, the goal is to approximate target fluence with a series of control points (CPs). These decisions involve agents selecting the appropriate leaf movements and MUs for each CP. Figure 3a shows the multi-leaf openings of one CP. Each leaf pair can filter X-rays to balance the radiation on PTV and OARs, as in Figure 3b. By comparing cumulated and target fluences



(a) multi-leaf pairs (b) one-leaf pair (c) one-leaf pair fluences

Figure 3. (a) shows a 2D illustration of multi-leaf pairs, with the middle depicting PTV projection. (b) provides a 3D view of a leaf pair and its connection to cumulated fluences. (c) illustrates motivations of Reward 1 (green) and Reward 2 (red) by comparing cumulated and target fluences. Details are in Appendix A.



(a) RLS: finite horizon RL (b) vanilla RL or conventional methods

Figure 4. Using finite horizon RL for accelerated inference. Conventional optimization methods start with an estimate of the leaf/MU positions and iteratively refine the estimate until converge or the stopping criteria are met. In principle, we can also apply vanilla RL in an infinite horizon context, and iteratively refine the estimates. However, to achieve greater efficiency during inference, we train RLS to execute only once for each CP.

at CPs (as well as other practical regularization), a set of rewards can guide movements of leaves and MUs, as shown in Figure 3c and Section 4.

To reduce time complexity, we formulate leaf sequencing under a finite horizon RL adapted from Pardo et al. (2018), where agents have to maximize expected return only over a fixed episode length, as in Figure 4. To cover multiple leaf pairs and MU prediction, we leverage the multi-agent framework. Consider practical function of leaf pairs and MUs, we propose a two-level RL framework as in Figure 5.

4. Methodology: Reinforced Leaf Sequencer

Our framework is shown in Figure 5. The backbone of RLS is in spirit of PPO (Schulman et al., 2017). We also conducted ablation studies of other actor-critic configurations.

4.1. Key Components of RL

Enviroment. The environment takes the output of agents (leaf actors and MU actor) and compute the cumulated fluences. Target fluence generation and fluence map computation from leaf sequencing are based on methods in a planning environment PORIx (details in Appendix B).

Actions. We define the leaf movement creating the fluence map as a sequence of discrete actions $la \sim \mathbb{N}^2, -S \leq la \leq S$. S is the maximum step size of each leaf. The MU action

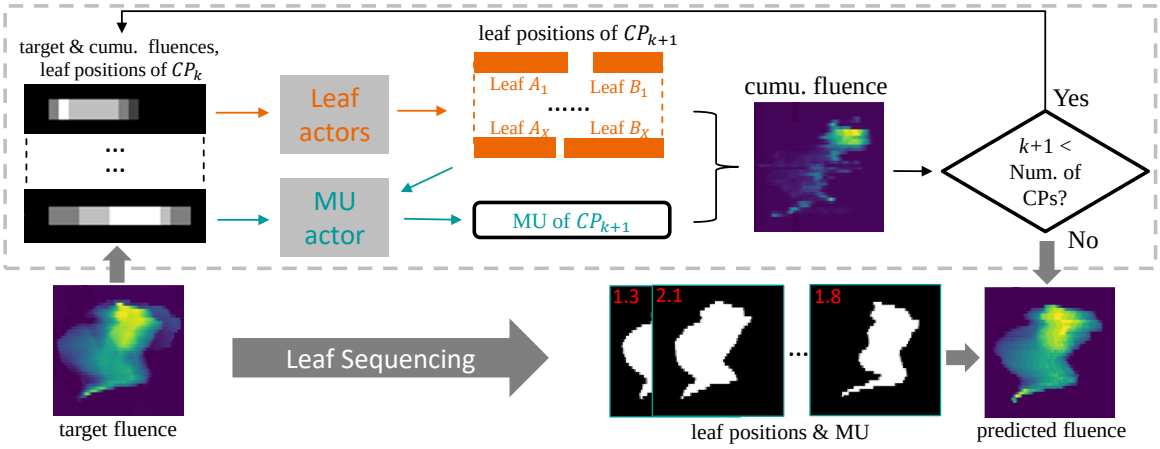


Figure 5. Illustration of the proposed RLS. The upper shows the methodology and the lower shows the input/output of RLS. The target fluence is splitted into X rows, each row is related to one leaf-pair and one leaf actor. x-th leaf actor predicts the positions of Leaf A_x and B_x. All rows in k-th control point (CP) shares the same monitor unit, which is predicted by MU actor after all leaf positions are obtained. The state of leaf actor at CP_{k+1} includes target fluence, cumulated fluence of $CP_1 \sim CP_k$, leaf positions of CP_k. The state of MU actor is similar but replace leaf positions of CP_k with that of CP_{k+1}.

ma is continuous in the range of (R_b, R_e) where R_e can be the maximum MU allowed by the radiotherapy machine. In k-th control point CP_k, x-th leaf pair has a separate leaf movement action $la_{x,k}$, while all leaf pairs of CP_k share the same MU action ma_k.

State. The state s_k of the leaf actors at CP_k is based on following information: target fluence F, cumulated fluence $\hat{F}^k$, and leaf positions (A^k, B^k), the step index k, and leaf position index x. Since MU actor is to predict the optimal MU, whose state should be updated with new leaf positions after leaf actors take actions. This is the so-call *two-level* actions motivated from clinical practice. For simplicity, we use (s_k, la_k) to represent the state for MU actor, as in the afterwards Eq. 5.

Rewards. Constructing the reward system is pivotal for framing the leaf sequencing with RL. We define the cumulated fluence $\hat{F}^k$ as the cumulation of fluences of control points CP_1 to CP_k, i.e., $\hat{F}^k = \sum_{i=1}^{k} M^k \cdot \tilde{F}^k$, where M^k and $\tilde{F}^k$ are MU and unit fluence of CP_k respectively. Considerations for reward design are as follows: ***1) Encouragement to approach target fluence (Reward 1):*** Positive rewards for actions bringing cumulative fluence closer to the target, seen in the green of Figure 3c. ***2) Punishing overdosing (Reward 2):*** Penalties for actions causing cumulative fluence to exceed target intensity, represented by the red portion in Figure 3c. ***3) Avoiding cross leaves (Reward 3):*** Encouraging avoidance of left-right leaf intersection, considering its impracticality. ***4) Leaf/MU changing regularization between CPs (Reward 4):*** Regularizing machine movement between control points, with user-adjustable parameter λ_4 in training. ***5) Aperture regularization (Reward 5)***: Regularizing the shape of the aperture formed by (A^k, B^k), with area and perimeter denoted as `area` and `peri` respectively, which aligns with Younge et al. (2012).

Rewards of k-th CP for x-th leaf pair are defined as below:

$$\begin{aligned}
R_1^k &= \sum_{y=1}^{Y} M^k \cdot \tilde{F}_{x,y}^k \cdot \mathbb{1}(F_{x,y} - \hat{F}_{x,y}^{k-1} > M^k) \\
R_2^k &= -\sum_{y=1}^{Y} \mathbb{1}(F_{x,y} - \hat{F}_{x,y}^{k} < M^k) \\
R_3^k &= \mathbb{1}(B_x^k - A_x^k) - \mathbb{1}(A_x^k - B_x^k) \\
R_4^k &= \Sigma_{I \in \{A,B,M\}}(1 - \sigma(|I_x^k - I_x^{k-1}|)) \\
R_5^k &= \texttt{area}(A^k, B^k)/\texttt{peri}(A^k, B^k)
\end{aligned} \tag{2}$$

where Y is the second dimension size of fluence F. The total reward is defined as $R^k = \lambda_1 \cdot R_1^k + \lambda_2 \cdot R_2^k + \lambda_3 \cdot R_3^k + \lambda_4 \cdot R_4^k + \lambda_5 \cdot R_5^k$. Details with pseudo-code are shown in Appendix F and math symbol summary is in Appendix A.

4.2. Two-level Multi-agent PPO

The proposed RLS adapts a multi-agent PPO framework. To accommodate RT characteristics, it incorporates two levels of actions: leaf movements and MUs prediction. RLS trains three networks: leaf policy net π_{θ_l} shared by all X leaf agents parameterized by θ_l, MU policy π_{θ_m} for MU agent parameterized by θ_m, and critic net V_ϕ parameterized by ϕ.

The policy nets are trained to maximize the objective:

$$L(\theta) = L_{\text{sur}}^{\text{CLIP}}(\theta) + c\hat{\mathbb{E}}_{x,k}\left[H(\pi_{\theta_l})\right] + c\hat{\mathbb{E}}_{k}\left[H(\pi_{\theta_m})\right], \tag{3}$$

where $H(\pi_\theta)$ is the entropy of the policy distribution, and c is the entropy coefficient hyperparameter. The clipped

Algorithm 1 RLS training

Input: X leaf agents with shared policy π_{θ_l} and an MU agent with policy π_{θ_m}
Output: trained leaf policy π_{θ_l}, MU policy π_{θ_m}, critic V_ϕ

Initialize parameters $\{\theta_l, \theta_m\}$ for policies $\{\pi_{\theta_l}, \pi_{\theta_m}\}$ and parameters ϕ for critic V_ϕ
for n = 1 to *RL iterations* **do**
  Initialize data buffer D
  for i = 1 to *batch size* **do**
    Run policy $\pi_{\theta_{l,old}}$ in the environment for X leaf agents
    Run policy $\pi_{\theta_{m,old}}$ in the environment for MU agent
    Compute the advantages using GAE
    Adding timestamp data to buffer D
  end for
  for e = 1 to *update epochs* **do**
    for i = 1 to *batch size* **do**
      Compute the policy loss based on Eq. 3
      Compute the value loss based on Eq. 6
      AdamW update $\{\theta_l, \theta_m\}$ on $\{\pi_{\theta_l}, \pi_{\theta_m}\}$ like PPO
      AdamW update ϕ on π_{θ_m} like PPO
    end for
  end for
  $\theta_{l,old} \leftarrow \theta_l$
  $\theta_{m,old} \leftarrow \theta_m$
end for

surrogate objective $L^{\text{CLIP}}_{\text{sur}}(\theta)$ is given by:

$$L^{\text{CLIP}}_{\text{sur}}(\theta) = \hat{\mathbb{E}}_{x,k}\left[\min\left(r_{x,k}(\theta)\hat{A}_{x,k}, \text{clip}\left(r_{x,k}(\theta), 1 \pm \epsilon\right)\hat{A}_{x,k}\right)\right], \quad (4)$$

where the advantages $\hat{A}_{x,k}$ are computed using GAE (Schulman et al., 2015) with rewards definition in Section 4.1. The probability ratio $r_{x,k}(\theta)$ for x-th leaf pair of CP_k is

$$r_{x,k}(\theta) = \frac{\pi_{\theta_l}(la_{x,k}|s_{x,k}) \cdot \pi_{\theta_m}(ma_k|s_k, la_{x,k})}{\pi_{\theta_{l,\text{old}}}(la_{x,k}|s_{x,k}) \cdot \pi_{\theta_{m,\text{old}}}(ma_k|s_k, la_{x,k})}. \quad (5)$$

The critic net is trained to minimize the loss function (with clipping following CleanRL (Huang et al., 2022)):

$$L(\phi) = \hat{\mathbb{E}}_{x,k}\left[\frac{1}{2}\left(V_\phi(s_{x,k}) - (r_{x,k} + \gamma V_\phi(s_{x,k+1}))\right)^2\right], \quad (6)$$

The RLS training algorithm is shown in Algo. 1.

4.3. Dealing with Heterogeneous Fluence Patterns

As shown in Figure 6, even if target fluences are from the same site (e.g., head-and-neck), their pattern can be significantly different due to patient distinction (e.g., different size and shape of PTV). To reduce such heterogeneity, we propose a *cropping* strategy to normalize the target fluence.

As in Figure 6, we first detect positive values in target fluence with coordination $\{\hat{x_1}, \hat{x_2}, \hat{y_1}, \hat{y_2}\}$, and crop the target fluence with bounding box $\{x_1 = \hat{x_1}, x_2 = \hat{x_2}, y_1 = \hat{y_1}/2, y_2 = (\hat{y_2} + Y)/2\}$. In the training, y_1 and y_2 are sampled from $(0, \hat{y_1})$ and $(\hat{y_2}, Y)$ respectively for data augmentation. $\hat{P}$ is predicted leaf positions in the output space

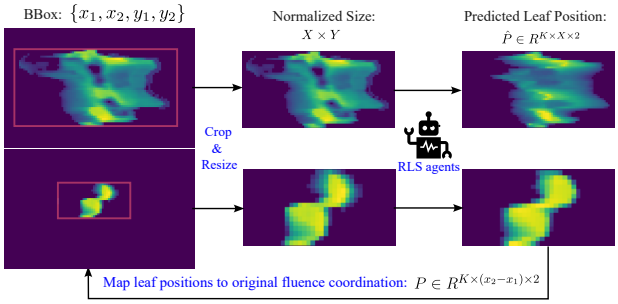


Figure 6. Illustration of dealing with heterogeneous fluences with proposed *cropping* strategy. The red box is ROI covers fluence locations with positive intensity. The RLS agents module only deal with normalized fluence maps.

of RLS, which is transformed to original fluence coordination as P obtained by linear interpolation and resize $P = \text{Resize}(\hat{P} \cdot \frac{y_2 - y_1}{Y} + y_1, (K, x_2 - x_1, 2))$.

4.4. Dealing with Various Length of Sector

Due to various of machine or optimization settings, the number of control points in one sector may be different. To make a single model that can be applied to various lengths, we provide a novel post-processing strategy: In order to merge two (for example) control points presented as coordinates $p_k = p_{k-1} + d_k$, and $p_{k+1} = p_k + d_{k+1}$ (where p_k and d_k are the position and predicted position change associated to CP_k); the merged position at CP_k is defined as $\bar{p}_k = \frac{p_k + p_{k+1}}{2} = p_{k-1} + d_k + \frac{d_{k+1}}{2}$. Thus, RLS can deal with various number of CPs in a sector with post-processing and without fundamentally change the RL module.

4.5. Innovation Highlight

In this study, we apply PPO and multi-agent learning principles to develop an innovative RL model that incorporates two-level agents (leaf agents and MU agent) in a finite horizon context, aiming to reduce time complexity. This model is tailored to address real-world challenges in RL. Our proposed reward mechanism empowers controllable movement patterns of the leaf sequence, a crucial element in RTP. Our study demonstrates an inaugural application of deep multi-agent RL methods of leaf sequencing. This approach has the potential to partially replace optimization methods in RTP and facilitate end-to-end learning with other AI modules.

5. Experiment

5.1. Dataset, Pre-processing, and Plan Evaluation

Two sites of RT cancer treatment from four datasets are included: head-and-neck (HN) and prostate (Pros). The HN site includes three datasets: 1) the HNd set contains

493 patients after quality assurance used for training and test with a train/validation/test split, 2) two external test sites from TCIA: HNe1 (Bejarano et al., 2018) with 31 patients and HNe2 (Grossberg et al., 2020) with 140 patients after filtering. The Pros site is from a public dataset with access permission requirements, including 555 patients after filtering. Each patient has up to 525 target fluences from PORIx environment. Details are in Appendix G.

We have three contexts to evaluate the proposed model. 1) Photon Optimizer Research Inference (PORIx) of a leading commercial radiotherapy company Varian; 2) end-to-end full-AI research pipeline for VMAT; and 3) simulation with IMRT. The PORIx is a leading optimization environment which provides callbacks of target fluence and leaf positions/MUs. It also allows the user to overwrite the output of the leaf sequencer, which can compare our method with the leaf sequencer in PORIx from multiple perspectives (see Appendix B). The leaf sequencing optimization in PORIx, terms as PORIx for short in some contexts, serves as a baseline to compare with our RLS.

5.2. Evaluation Metrics

We include three levels of evaluation for leaf sequencing: **(1)** Mean of Normalized Square Error (MNSE) to evaluate the fluence reconstruction performance, which is defined as $\frac{1}{n}\sum_{i=1}^{n}\frac{||F-\hat{F}||_2}{||F||_2}$. **(2)** Iterations Reducing $p\%$ Error (IRE$p\%$) to evaluate the optimization convergence speed of PORIx when its leaf sequencer is using its default model vs. our RLS. Specifically, two types of IRE are defined: The *absolute* IRE$p\%$ (AIRE$p\%$) is defined as the smallest of iteration whose planning cost is smaller than $E_1 \cdot (1-p\%)$, where E_1 is the cost after the first iteration in PORIx. The *relative* IRE$p\%$ (RIRE$p\%$) is defined as $(E_1 - \min(E)) \cdot (1-p\%) + \min(E)$. The list E is costs from the planning costs of iterations in PORIx. **(3)** To evaluate the reconstruction performance from 3D dose perspective (3D doses are computed from target vs. predicted fluences), we use the Dose score and DVH score from the OpenKBP challenge (Babier et al., 2020).

5.3. Experimental Settings

Our implementation is motivated by CleanRL (Huang et al., 2022). Our major backbone is based on PPO, which has shown success in many applications including GPT-4 (OpenAI, 2023). The actions of leaves/MUs follow the physical regularization in the planning system, and MU actions of a sector/field are refined with ridge regression. We follow hyper-parameters related to PPO settings in CleanRL unless mentioned. We apply the AdamW optimizer with an initial learning rate of 1e-4, weight decay 1e-4, and a Cosine Annealing scheduler. The reward ratios $\{\lambda_i\}$ are set to $\{1, 2, 2, 1, 1\}$ except in ablation studies. The action hyperparameters S, R_b, R_e are set as 4, 0.5, 2.5, respectively. More details can be found in Appendix C.

	HNd	HNe1	HNe2	Pros	Pros(e)
PORIx	.219	.257	.241	.079	.079
ours	**.149**	**.165**	**.146**	**.042**	**.043**

Table 1. The MNSE (↓) from target fluence and predicted fluence.

	HNd	HNe1	HNe2	Pros	Pros(e)
RIRE97% metric (↓)					
PORIx	18.8	26.3	20.6	4.13	4.13
ours	**12.6**	**10.7**	**14.4**	**3.85**	**3.75**
AIRE97% metric (↓)					
PORIx	27.2	46.8	41.0	4.21	4.21
ours	**17.6**	**27.9**	**31.6**	**3.93**	**3.78**
AIRE99% metric (↓)					
PORIx	56.7	71.0	62.2	8.65	8.65
ours	**50.1**	**66.3**	**53.1**	**7.25**	**7.05**

Table 2. RIRE97%, AIRE97% and AIRE99% of PORIx vs. ours.

5.4. Main Experimental Results

Planning with PORIx. Table 1 and 2 show the comparison between our RLS and leaf sequencing optimization in PORIx. We achieve the lower MNSE across all compared contexts, indicating our RLS can reconstruct target fluence better than PORIx with executable leaf positions and MUs. The reduced AIRE and RIRE suggest that an optimization planner, such as PORIx, incorporating our RLS, has the potential for quicker convergence. Figure 7 and 8 depict typical outcomes of our RLS and the PORIx optimizer. The RLS excels in challenging cases, especially in large and heterogeneous PTV cases of the head-and-neck (first row in both figures). In easier cases of head-and-neck and prostate (second and third rows), RLS performs close or slightly outperforms PORIx optimizer. Notably, easy cases exhibit better reconstruction and faster convergence, which may not pose a significant challenge in RTP.

	Dose score (↓)		DVH score (↓)	
	MAE	MSE	MAE	MSE
OpenKBP S1	2.4 (1st)	15.5 (1st)	1.5 (1st)	5.9 (1st)
OpenKBP S2	2.6 (2nd)	16.6 (2nd)	1.7 (12nd)	6.8 (10th)
OpenKBP S3	2.7 (4th)	18.1 (5th)	1.5 (2nd)	6.0 (2nd)
Teng et al. (2024)	2.1	-	0.98	-
ours (VMAT, 8 cps)	0.19	0.18	0.88	1.2
ours (IMRT, 32 degree)	0.23	0.28	0.97	1.6

Table 3. Dose and DVH scores when comparing a 3D dose from prediction and a 3D dose from target. The upper panel is best performances in the OpenKBP leaderboard (Babier et al., 2020) and the latest breakthrough; the lower panel is for evaluating leaf sequencing when target fluence from fluence prediction module.

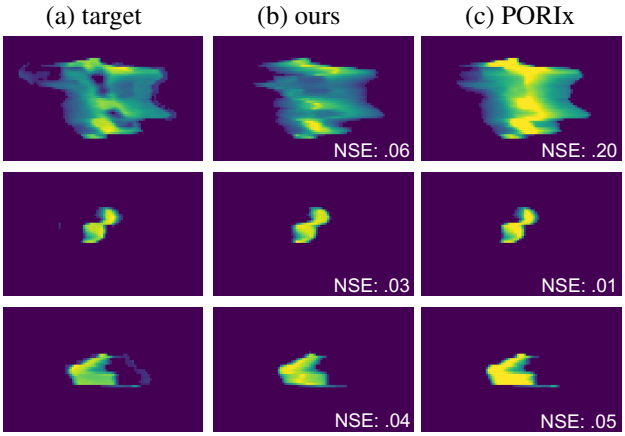


Figure 7. Predicted fluence vs. target fluence. The reconstruction error is shown right bottom of the predicted fluences. First row: hard case from HN; second row: easy case from HN; third row: prostate case. This figure is matched with cases in Figure 8.

Full-AI End-to-end VMAT Planning (Prostate). As in Figure 1, a typical RTP pipeline can consist of three major components. Here, we evaluate our RLS in the context of a full-AI end-to-end framework without optimization iteration. The objective definition is replaced by dose prediction adapted from Gao et al. (2023), and the fluence prediction is motivated by Wang et al. (2020) (which was originally for IMRT field dose, here has been adapted for VMAT). More details of dose/fluence predictions are in Appendix E.

As in Figure 9 and Table 3, the computed 3D doses from target fluence (not executable for machine) and predicted fluence from RLS (executable from machine) are close, both visually and in terms of Dose/DVH scores. Compared to dose prediction (a module needed in E2E pipeline) from OpenKBP challenge (Babier et al., 2020), RLS achieves better Dose score and DVH score compared to state-of-the-art models in OpenKBP [2].

Simulation for IMRT. Although RLS was designed for VMAT contexts, it can be applied to leaf sequencing of IMRT with a simulation testbed. We group 16 control points (∼32 degree between two adjacent beam angles) as a single field from fluence prediction module of the end-to-end pipeline in a IMRT-simulation context. As in Table 3 (last row), our RLS also achieves promising performance.

5.5. Ablation Studies

Backbone. Our major experiments follow the settings of PPO in CleanRL (Huang et al., 2022), i.e., the networks of actor and critic are separated. We include the ablation

[2]It is a compromise to use records of OpenKBP to compare with our model since our tasks are not the same as in OpenKBP. See detailed considerations in Appendix E.

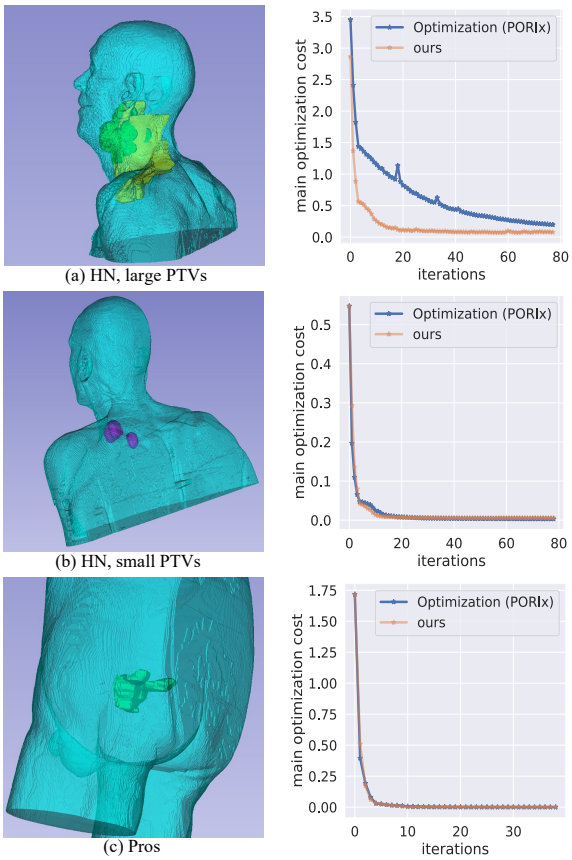


Figure 8. Typical cases in different scenarios. The left panel shows the PTV contours within body mask; right panel shows the main optimization cost with the number of iterations. Upper: large PTVs (PTV 54 and PTV 60) in HN, middle: small PTVs in HN (PTV 66), lower: PTV in prostate. The RLS brings clear improvements for hard cases (e.g., those with large PTVs).

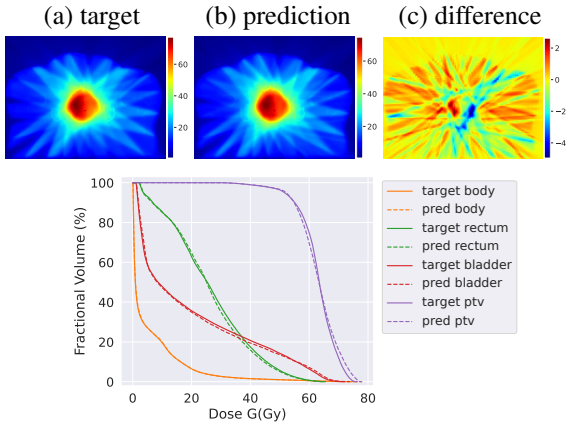


Figure 9. RT plan from an AI end-to-end preliminary solution. (a) *target* is the 3D dose computed from fluences (not machine executable) of *Fluence Prediction* module, while (b) *prediction* is computed from executable leaf sequencing. The lower part shows DVHs from *target* (solid lines) and *prediction* (dash lines).

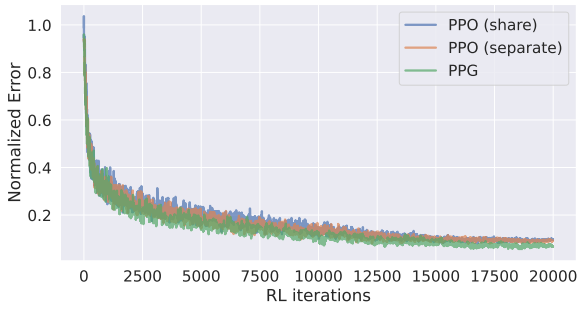
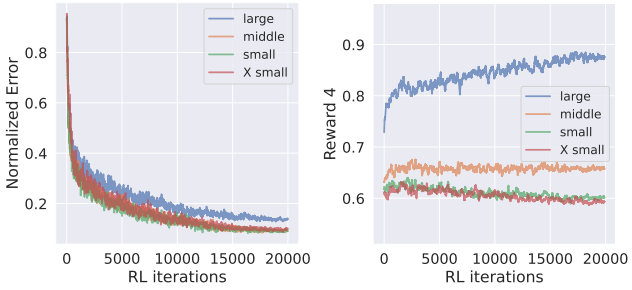


Figure 10. The ablation of different actor-critic structures with normalized reconstruction errors. *share* and *separate* represent the policy-net and critic-net are shared and separated, respectively.



(a) Normalized Errors (b) Reward 4

Figure 11. Different weights on leaf speed regularization (i.e., Reward 4 in Eq. 2). Four scales of λ_4 are compared; Large: 5, middle: 1, small: 0.1, X small: 0.01.

studies of sharing backbone between actor and critic, and PPG (Cobbe et al., 2021) which separates actor and critic training into distinct phases. As shown in Figure 10, in our leaf sequencing context, the shared-network setting is a little worse than the separate-network under the PPO setting. Overall, the PPG framework has minor superiority over PPO backbones. Considering differences are not significant in these three backbones and for simplicity purposes, we apply the most widely used PPO (separated) for the majority of our experiments. We leave careful comparisons of different backbones in future work.

The Ratio of Reward Components. To better understand the sensitivity of different reward components, we conduct the ablation studies with five different ratios of each reward, as shown in Table 4.

Reward 1 and 2 are the major rewards driving the feasibility of RLS, as depicted in Figure 13c, which are more sensitive to weights. Reward 3-5 are auxiliary with clinical or physical considerations, which are less sensitive to the weight. The default weights achieved a reasonably good performance, exposing such ablation studies guides users to choose weights in different scenarios.

Weight	0	0.01	0.1	1	10
λ_1	0.300	0.289	0.229	0.149	0.181
λ_2	0.574	0.528	0.271	0.151	0.226
λ_3	0.155	0.152	0.150	0.149	0.151
λ_4	0.148	0.147	0.147	0.149	0.167
λ_5	0.150	0.148	0.149	0.149	0.165

Table 4. The reconstruction errors (MNSE) of five weights in different scale for each reward ratio λ_i (other $\lambda_{j\neq i}$) are set to default ratio. The default ratios are $\{\lambda_1 : 1, \lambda_2 : 2, \lambda_3 : 2, \lambda_4 : 1, \lambda_5 : 1\}$, which achieve 0.149 of MNSE.

	HNe1		Pros(e)	
	w/o Crop	w Crop	w/o Crop	w Crop
RIRE97% (↓)	13.2	**10.7**$_{-19\%}$	6.13	**3.85**$_{-38\%}$
AIRE97% (↓)	**27.3**	27.9$_{+2\%}$	6.25	**3.93**$_{-37\%}$
AIRE99% (↓)	69.4	**66.3**$_{-4\%}$	9.44	**7.25**$_{-23\%}$
MNSE (↓)	.181	**.165**$_{-9\%}$	.060	**.043**$_{-28\%}$

Table 5. With vs. without cropping comparison. RLS is trained on HNd, and externally tested on HNe1 and prostate sites.

Figure 11 shows a detailed ablation study of the leaf change in adjacent control points (i.e., R_4^k in Eq. 2). High leaf speed may increase the instability and violate the max speed of the RT machine, while low speed may reduce the flexibility to reconstruct the target fluence. We demonstrate that by tuning the weight of the reward (i.e., λ_4) in training, our model allows user preference in the balance between leaf speed and reconstruction performance.

Cropping Strategy. The utilization of the *cropping* strategy facilitates the normalization of heterogeneous fluence patterns. A comprehensive comparison, as in Table 5, highlights the impact of *cropping* vs. *no cropping* and the optimization solution. Significantly, the *cropping* strategy has enhanced the performance of leaf sequencing, particularly evident in external testing across diverse sites, as indicated by the columns for Pros(e) in Table 5. This scenario involves training on head-and-neck cancer site and subsequently external testing on prostate cancer site. More experimental results can be found in Appendix H.

6. Discussion

Conclusion. This paper introduces the *Reinforced Leaf Sequencer* (RLS), a pioneer multi-agent reinforcement learning approach in practical leaf sequencing for radiotherapy. RLS can potentially supplant commonly used optimization-based methods, producing executable plans. Our technical contribution involves extending multi-agent and Proximal Policy Optimization (PPO) into a two-level practical leaf sequencing framework, incorporating novel rewards and actions. We assess the performance of the proposed RLS

using four datasets (three from head-and-neck site and one from prostate site) in three distinct contexts: 1) a practical radiotherapy environment, 2) a full-AI end-to-end research pipeline, and 3) IMRT simulation. Notably, our AI model, even without an iteration loop, exhibits improvements in terms of reconstruction error and early iteration coverage rate when compared to optimization methods.

Limitation and Future Work. As an initial endeavor to address practical leaf sequencing with MARL, our work has several limitations. ***First***, unlike well-established research domains, our comparative analyses are constrained by the scarcity of directly relevant literature. Moreover, the absence of open-source models, often attributable to commercial or confidentiality constraints, further hinders direct comparisons. Fortunately, the PORIx provides a contemporary baseline for evaluation and facilitates meaningful comparisons. ***Second***, the training of RLS is focused on leaf sequencing module so that main-loop costs in PORIx may not be always well-optimized. This limitation is similarly present in the end-to-end AI planning pipeline. One future work is to train different modules end-to-end for global optimal plan. ***Third***, we currently concentrate solely on the initial ≤ 100 iterations of PORIx as a preliminary plan. In clinical practice, the number of iterations can theoretically be infinite, and planners can use multi-resolution planning (as in Eclipse) to fine-tune solutions. We simplified this process for large-scale automation. One potential future work could involve initiating the planning process with deep learning plans, which offer rapid and automated solutions, followed by a limited number of human/optimization intervention steps. More contexts can be found in Appendix B. ***Fourth***, we opt for one of the most representative RL frameworks, specifically PPO, as our backbone. Its effectiveness has been validated across various RL tasks, including renowned LLMs (OpenAI, 2023). Although we perform some ablations on its actor-critic configuration, the exploration of experimental comparisons among various RL backbones, such as with model-based RLs (Moerland et al., 2023), is left for future work.

Outlook. Excitement surrounds the potential of deep learning to fully or partially replace conventional optimization in practical RT, despite acknowledged limitations. The proposed RLS stands out for three key reasons: 1) it enables potential end-to-end learning for the entire planning pipeline to seek global optimal solutions; 2) unlike the leaf sequencer method used in iterative optimizers (such as typical VMAT), there is no need to give as input a "seed sequence" (i.e., initial state of leaf positions and MUs) from previous iteration, allowing a more flexible algorithm design; and 3) with proper learned knowledge, RLS can potentially deliver superior performance than optimization with user-preference. We are excited about future work when more advanced algorithms and scenarios are applied.

Disclaimer

The information in this paper is based on research results that are not commercially available. Future commercial availability cannot be guaranteed.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning for Radiotherapy. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

We acknowledge the support of Varian for the research interface of practical planning environment, and its feasibility to evaluate leaf sequencing performances.

We thank Vivek Singh and Yue Zhang from Digital Technology and Innovation, Siemens Healthineers for the helpful discussions on this study.

We thank all the data contributors to the REQUITE project, including patients, clinicians and nurses. The core REQUITE consortium consists of David Azria, Erik Briers, Jenny Chang-Claude, Alison M. Dunning, Rebecca M. Elliott, Corinne Faivre-Finn, Sara Gutiérrez-Enríquez, Kerstie Johnson, Zoe Lingard, Tiziana Rancati, Tim Rattay, Barry S. Rosenstein, Dirk De Ruysscher, Petra Seibold, Elena Sperk, R. Paul Symonds, Hilary Stobart, Christopher Talbot, Ana Vega, Liv Veldeman, Tim Ward, Adam Webb and Catharine M.L. West.

References

Amendola, B. E., Amendola, M., Perez, N., Iglesias, A., and Wu, X. Volumetric-modulated arc therapy with rapidarc: An evaluation of treatment delivery efficiency. *Reports of practical oncology and radiotherapy*, 2013.

Babier, A., Zhang, B., Mahmood, R., Moore, K. L., Purdie, T. G., McNiven, A. L., and Chan, T. C. Y. OpenKBP: The open-access knowledge-based planning grand challenge. *Medical Physics*, 2020.

Barragán-Montero, A. M., Nguyen, D., Lu, W., Lin, M. H., Norouzi-Kandalan, R., Geets, X., Sterpin, E., and Jiang, S. Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations. *Medical Physics*, 2019.

Bedford, J. L. Treatment planning for volumetric modulated arc therapy. *Medical physics*, 2009.

Bejarano, T., Ornelas-Couto, M., and Mihaylov, I. Head-

and-neck squamous cell carcinoma (hnscc) patients with 3d ct taken during pre-treatment, mid-treatment, and post-treatment dataset. *The Cancer Imaging Archive*, 2018.

Bergman, A. M., Gete, E., Duzenli, C., and Teke, T. Monte carlo modeling of hd120 multileaf collimator on varian truebeam linear accelerator for verification of 6x and 6x fff vmat sabr treatment plans. *Journal of applied clinical medical physics*, 2014.

Cao, D., Earl, M. A., Luan, S., and Shepard, D. M. Continuous intensity map optimization (cimo): a novel approach to leaf sequencing in step and shoot imrt. *Medical physics*, 2006.

Carrasqueira, P., Rocha, H., Dias, J., Ventura, T., Ferreira, B., and Lopes, M. An automated treatment planning strategy for highly noncoplanar radiotherapy arc trajectories. *International Transactions in Operational Research*, 2023.

Cedric, X. Y. and Tang, G. Intensity-modulated arc therapy: principles, technologies and clinical implementation. *Physics in Medicine & Biology*, 2011.

Cobbe, K. W., Hilton, J., Klimov, O., and Schulman, J. Phasic policy gradient. In *International Conference on Machine Learning*, 2021.

Coulom, R. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, 2006.

Earl, M. A., Shepard, D. M., Naqvi, S., Li, X. A., and Yu, C. X. Inverse planning for intensity-modulated arc therapy using direct aperture optimization. *Physics in Medicine and Biology*, 2003.

Elekta. High-precision treatment planning for radiation therapy, 2023. URL https://www.elekta.com/products/oncology-informatics/elekta-one/treatment-applications/monaco/. Accessed: 2023-10-12.

Elmore, S. N. C., Prajogi, G. B., Rubio, J. A. P., and Zubizarreta, E. The global radiation oncology workforce in 2030: estimating physician training needs and proposing solutions to scale up capacity in low-and middle-income countries. *Appl Rad Oncol*, 2019.

Fallahi, A., Mahnam, M., and Niaki, S. T. A. A discrete differential evolution with local search particle swarm optimization to direct angle and aperture optimization in imrt treatment planning problem. *Applied Soft Computing*, 2022.

Feng, Z., Wen, L., Wang, P., Yan, B., Wu, X., Zhou, J., and Wang, Y. Diffdp: Radiotherapy dose prediction via a diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023.

Fogliata, A., Cozzi, L., Reggiori, G., Stravato, A., Lobefalo, F., Franzese, C., Franceschini, D., Tomatis, S., and Scorsetti, M. RapidPlan knowledge based planning: Iterative learning process and model ability to steer planning strategies. *Radiation Oncology*, 2019.

Ganeshan, A. R. Reinforcement learning applied to mlc tracking, 2021.

Gao, R., Lou, B., Xu, Z., Comaniciu, D., and Kamen, A. Flexible-cm gan: Towards precise 3d dose prediction in radiotherapy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Ghesu, F.-C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., and Comaniciu, D. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

Gronberg, M. P., Jhingran, A., Netherton, T. J., Gay, S. S., Cardenas, C. E., Chung, C., Fuentes, D., Fuller, C. D., Howell, R. M., Khan, M., et al. Deep learning–based dose prediction to improve the plan quality of volumetric modulated arc therapy for gynecologic cancers. *Medical physics*, 2023.

Grossberg, A., Elhalawani, H., Mohamed, A., Mulder, S., Williams, B., White, A., Zafereo, J., Wong, A., Berends, J., AboHashem, S., Aymard, J., Kanwar, A., Perni, S., Rock, C., Chamchod, S., Kantor, M., Browne, T., Hutcheson, K., Gunn, G., Frank, S., Rosenthal, D., Garden, A., and Fuller, C. M.d. anderson cancer center head and neck quantitative imaging working group. (2020) hnscc [dataset]. *The Cancer Imaging Archive*, 2020.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning*, 2018.

Hrinivich, W. T. and Lee, J. Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning. *Medical physics*, 2020.

Hrinivich, W. T., Bhattacharya, M., Mekki, L., McNutt, T., Jia, X., Li, H., Song, D. Y., and Lee, J. Clinical vmat machine parameter optimization for localized prostate cancer using deep reinforcement learning. *Medical physics*, 2024.

Huang, S., Dossa, R. F. J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., and Araújo, J. G. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *The Journal of Machine Learning Research*, 2022.

Huynh, E., Hosny, A., Guthier, C., Bitterman, D. S., Petit, S. F., Haas-Kogan, D. A., Kann, B., Aerts, H. J., and Mak, R. H. Artificial intelligence in radiation oncology. *Nature Reviews Clinical Oncology*, 2020.

Infusino, E. Clinical utility of rapidarc™ radiotherapy technology. *Cancer Management and Research*, 2015.

Jhanwar, G., Tefagh, M., Taasti, V. T., Alam, S. R., Tuomaala, S., Nadeem, S., and Zarepisheh, M. Portpy: An open-source python package for planning and optimization in radiation therapy including benchmark data and algorithms. *AAPM 65th Annual Meeting and Exhibition*, 2023.

Jiao, Z., Peng, X., Wang, Y., Xiao, J., Nie, D., Wu, X., Wang, X., Zhou, J., and Shen, D. Transdose: Transformer-based radiotherapy dose prediction from ct images guided by super-pixel-level gcn classification. *Medical Image Analysis*, 2023.

Kearney, V., Chan, J. W., Wang, T., Perry, A., Descovich, M., Morin, O., Yom, S. S., and Solberg, T. D. DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. *Scientific Reports*, 2020.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in neural information processing systems*, 1999.

Luchini, C., Pea, A., and Scarpa, A. Artificial intelligence in oncology: current applications and future perspectives. *British Journal of Cancer*, 2022.

Ma, L., Chen, M., Gu, X., and Lu, W. Deep learning-based inverse mapping for fluence map prediction. *Physics in Medicine & Biology*, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 2015.

Moerland, T. M., Broekens, J., Plaat, A., Jonker, C. M., et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 2023.

Mundt, A. J. and Roeske, J. C. *Intensity modulated radiation therapy: a clinical perspective*. PMPH-USA, 2005.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Otto, K. Volumetric modulated arc therapy: Imrt in a single gantry arc. *Medical physics*, 2008.

Pardo, F., Tavakoli, A., Levdik, V., and Kormushev, P. Time limits in reinforcement learning. In *International Conference on Machine Learning*, 2018.

Quan, E. M., Li, X., Li, Y., Wang, X., Kudchadker, R. J., Johnson, J. L., Kuban, D. A., Lee, A. K., and Zhang, X. A comprehensive comparison of imrt and vmat plan quality for prostate cancer treatment. *International Journal of Radiation Oncology Biology Physics*, 2012.

Ripsman, D. A., Purdie, T. G., Chan, T. C., and Mahmoudzadeh, H. Robust direct aperture optimization for radiation therapy treatment planning. *INFORMS Journal on Computing*, 2022.

Romeijn, H. E., Ahuja, R. K., Dempsey, J. F., Kumar, A., and Li, J. G. A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. *Physics in Medicine & Biology*, 2003.

Romeijn, H. E., Dempsey, J. F., and Li, J. G. A unifying framework for multi-criteria fluence map optimization models. *Physics in Medicine & Biology*, 2004.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Seibold, P., Webb, A., Aguado-Barrera, M. E., Azria, D., Bourgier, C., Brengues, M., Briers, E., Bultijnck, R., Calvo-Crespo, P., Carballo, A., Choudhury, A., Cicchetti, A., Claßen, J., Delmastro, E., Dunning, A. M., Elliott, R. M., Farcy-Jacquet, M. P., Gabriele, P., Garibaldi, E., Gómez-Caamaño, A., Gutiérrez-Enríquez, S., Higginson, D. S., Johnson, K., Lobato-Busto, R., Mollà, M., Müller, A., Payne, D., Peleteiro, P., Post, G., Rancati, T., Rattay, T., Reyes, V., Rosenstein, B. S., De Ruysscher, D., De Santis, M. C., Schäfer, J., Schnabel, T., Sperk, E.,

Symonds, R. P., Stobart, H., Taboada-Valladares, B., Talbot, C. J., Valdagni, R., Vega, A., Veldeman, L., Ward, T., Weißenberger, C., West, C. M., and Chang-Claude, J. REQUITE: A prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer. *Radiotherapy and Oncology*, 2019.

Shepard, D., Cao, D., Afghan, M., and Earl, M. An arc-sequencing algorithm for intensity modulated arc therapy. *Medical physics*, 2007.

Shepard, D. M., Earl, M. A., Li, X. A., Naqvi, S., and Yu, C. Direct aperture optimization: a turnkey solution for step-and-shoot imrt. *Medical physics*, 2002.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 2016.

Society, C. C. external radiation therapy, 2024. URL https://cancer.ca/en/treatments/treatment-types/radiation-therapy/external-radiation-therapy. Accessed: 2024-01-10.

Teh, B. S., Woo, S. Y., and Butler, E. B. Intensity Modulated Radiation Therapy (IMRT): A New Promising Technology in Radiation Oncology. *The Oncologist*, 1999.

Teng, L., Wang, B., Xu, X., Zhang, J., Mei, L., Feng, Q., and Shen, D. Beam-wise dose composition learning for head and neck cancer dose prediction in radiotherapy. *Medical Image Analysis*, 2024.

Teoh, M., Clark, C. H., Wood, K., Whitaker, S., and Nisbet, A. Volumetric modulated arc therapy: A review of current literature and clinical use in practice. *British Journal of Radiology*, 2011.

Unkelbach, J., Bortfeld, T., Craft, D., Alber, M., Bangert, M., Bokrantz, R., Chen, D., Li, R., Xing, L., Men, C., et al. Optimization approaches to volumetric modulated arc therapy planning. *Medical physics*, 2015.

Varian. RapidPlan Knowledge-Based Planning — Varian, 2023a. URL https://www.varian.com/products/radiotherapy/treatment-planning/rapidplan-knowledge-based-planning. Accessed: 2023-10-12.

Varian. Treatment planning, 2023b. URL https://www.varian.com/products/radiosurgery/treatment-planning. Accessed: 2023-10-14.

Wang, B., Teng, L., Mei, L., Cui, Z., Xu, X., Feng, Q., and Shen, D. Deep Learning-Based Head and Neck Radiotherapy Planning Dose Prediction via Beam-Wise Dose Decomposition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022.

Wang, C., Luan, S., Tang, G., Chen, D. Z., Earl, M. A., and Cedric, X. Y. Arc-modulated radiation therapy (amrt): a single-arc form of intensity-modulated arc therapy. *Physics in Medicine & Biology*, 2008.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Wang, W., Sheng, Y., Wang, C., Zhang, J., Li, X., Palta, M., Czito, B., Willett, C. G., Wu, Q., Ge, Y., Yin, F. F., and Wu, Q. J. Fluence Map Prediction Using Deep Learning Models – Direct Plan Generation for Pancreas Stereotactic Body Radiation Therapy. *Frontiers in Artificial Intelligence*, 2020.

Wang, W., Sheng, Y., Palta, M., Czito, B., Willett, C., Hito, M., Yin, F.-F., Wu, Q., Ge, Y., and Wu, Q. J. Deep Learning-Based Fluence Map Prediction for Pancreas Stereotactic Body Radiation Therapy With Simultaneous Integrated Boost. *Advances in Radiation Oncology*, 2021.

Xia, P. and Verhey, L. J. Multileaf collimator leaf sequencing algorithm for intensity modulated beams with multiple static segments. *Medical Physics*, 1998.

Yang, Z., Zhang, K., Hong, M., and Başar, T. A finite sample analysis of the actor-critic algorithm. In *IEEE conference on decision and control*, 2018.

Younge, K. C., Matuszak, M. M., Moran, J. M., McShan, D. L., Fraass, B. A., and Roberts, D. A. Penalization of aperture complexity in inversely planned volumetric modulated arc therapy. *Medical physics*, 2012.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 2022.

Yuan, Z., Wang, Y., Hu, P., Zhang, D., Yan, B., Lu, H.-M., Zhang, H., and Yang, Y. Accelerate treatment planning process using deep learning generated fluence maps for cervical cancer radiation therapy. *Medical Physics*, 2022.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, 2021.

Appendix

Appendix A introduces the background of radiotherapy to readers who do not have much background in radiotherapy. We include cartoon illustrations, explanations of terminologies in Appendix A.1, and a summary of mathematical symbols in Appendix A.2.

Appendix B provides a more detailed introduction to the planning environment PORIx.

Appendix C includes more hyperparameters compared to the main text.

Appendix D details the network structures of our model.

Appendix E introduces the other modules (except for leaf sequencing) in the full-AI end-to-end planning pipeline, and the considerations of evaluating leaf sequencing in 3D dose space and comparison with dose prediction metrics (Table 3).

Appendix F includes more details of reward components and their PyTorch style pseudo code.

Appendix G details the datasets.

Appendix H shows multiple examples of the optimization cost curve of PORIx, comparing the proposed RLS, RLS without *cropping*, and the optimization sequencer in PORIx.

Appendix I includes additional examples illustrating reconstruction errors.

A. Introduction to Radiotherapy

In this section, we offer a comprehensive overview of radiotherapy concepts relevant to our research. Our aim is to assist readers in understanding our work from a machine learning standpoint, even if they lack a background in radiotherapy. Explanations of related terminology and mathematical symbols are summarized in Appendix A.1 and A.2. Cartoon illustrations of the radiotherapy machine and MLC/fluences are shown in Figure 12 and 13, respectively.

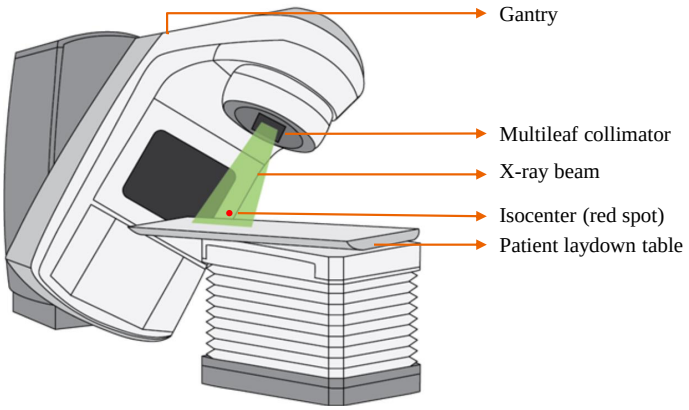


Figure 12. Cartoon illustration of a typical radiotherapy machine. The machine figure (except the annotations) is borrowed from (Society, 2024). The multileaf collimator is connected with concept in Figure 13.

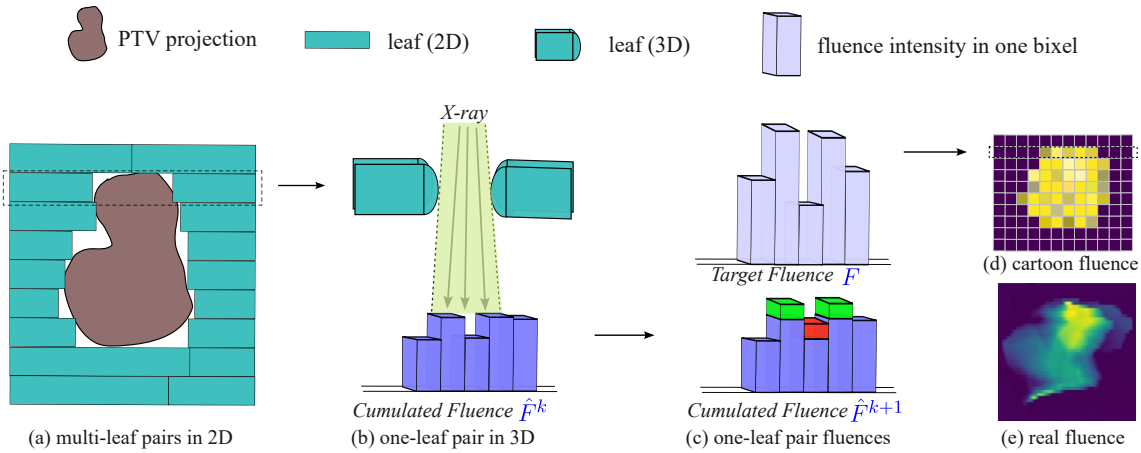


Figure 13. Cartoon illustration of MLC and fluences, an extension figure of Figure 3 in the main text. (a) shows multi-leaf pairs in 2D representing Multi-leaf Collimator and PTV projections when observed from the direction of radiation source. (b) provides a 3D view of a leaf pair and its connection to cumulated fluences. (c) illustrates motivations of Reward 1 (green) and Reward 2 (red) by comparing cumulated and target fluences. (d) shows a cartoon illustration of fluences from multi-leaf pairs, and (e) represents a real fluence map.

A.1. Aberrations, Explanations of Terminologies

Radiotherapy (RT): also known as radiation therapy, it is one of the primary medical treatments that uses high-energy radiation (such as photons, electrons, or protons) to target and kill cancer cells or shrink tumors. The goal of radiation therapy is to damage the DNA inside the cancer cells, preventing them from growing and dividing, while minimizing damage to nearby healthy cells. There are two types of RT: External Beam Radiation Therapy and Internal Radiation Therapy. In this paper, we mainly focus on photon External Beam Radiation Therapy.

Radiotherapy Planning (RTP): In radiation oncology, radiotherapy planning is a precise process that strategically administers radiation to treat conditions such as cancer, with the goal of maximizing tumor destruction while minimizing

damage to adjacent healthy tissues. In photon External Beam Radiation Therapy, the planning task involves optimizing incoming fields/sectors from different directions.

Volumetric Modulated Arc Therapy (VMAT) and Intensity-Modulated Radiation Therapy (IMRT): VMAT and IMRT are two major modern radiotherapy treatments, both aiming for precise dose delivery. In some literature, VMAT has been broadly categorized under IMRT. The illustrations of VMAT and IMRT are summarized in Figure 2. The gantry is static for IMRT and dynamic for VMAT during X-ray delivery.

Planning Target Volume (PTV): A three-dimensional (3D) volume delineated on specific medical planning images (e.g., CT images), outlining the region where the prescribed dose level is targeted to achieve the desired tumor control. The PTV generally encompasses the visually identified tumor, adjacent healthy tissue at a high risk of harboring proliferative cancer cells, and geometric margins to accommodate positioning inaccuracies between the planning images and the patient's anatomy during treatment.

Organ at Risk (OAR): This term refers to any healthy organ or tissue (such as the heart, lung, eye, bladder, etc.) located in proximity to the region undergoing radiation treatment, and which could potentially experience adverse effects due to exposure to radiation.

RT Dose: Also known as 3D Dose or Dose within the context of this paper, it denotes the precise quantity of radiation administered to a Planning Target Volume (PTV) within a patient's body throughout a course of radiation treatment. This measurement represents the absorbed amount of radiation energy by the tissues undergoing treatment. RT Dose is a 3D matrix that shares the same shape with CT and ROI contours.

Fluence Map: This term refers to a depiction of the intensity of a radiation beam across the treatment field. It is a two-dimensional (2D) map that visualizes how the radiation dose is distributed across a specific cross-sectional area of the patient's body. Fluence maps play a crucial role in contemporary radiation therapy treatment planning and delivery.

Isocenter: The isocenter is determined during the treatment planning process, where medical professionals use imaging techniques such as CT scans to identify the tumor's location and define the treatment volume. Once the treatment plan is established, the linear accelerator or other radiation delivery equipment is adjusted to ensure that the beams intersect precisely at the isocenter.

Control Point: This term denotes a specific time point in the administration of radiation therapy. The details of a control point encompass the positions of all leaf pairs and the monitor unit at that specific time point.

Multileaf Collimator (MLC): Positioned in close proximity to the patient's body and situated in the path of the radiation beam, the MLC plays a key role in shaping and controlling the radiation beam. Its primary function is to enable precise and conformal radiation treatment, facilitating highly targeted delivery.

Leaf: The MLC consists of a series of individual "leaves" or "slats" arranged in pairs. These leaves are typically made of high-density material, such as tungsten, to effectively block the radiation.

Photon Optimizer Research Interface from Varian (PORIx): An optimizer for photon VMAT planning that provides an interface for the necessary components for RL agent training: the implemented callbacks of the interface allows user to replace the prediction in the PORIx so the AI model can be evaluated in this environment. More details are in Section B and Figure 14.

Main Optimization Cost: This term originates from the cost callbacks provided by PORIx, where it represents the dosimetric cost derived from the intermediate dose and dose objectives for each iteration in the optimization process. In Figure 8 and 15, the "main optimization cost" serves as the y-axis.

A.2. Summary of Mathematical Symbols

Different Indices:

k: index of control points, the number of control points of a fluence map is K.

x: index of leaf agents, the number of leaf agents of a fluence map is X.

y: index of the second dimension of fluence maps, the second dimension of fluence map is Y, indicating the range of leaf positions is $[0, Y]$.

Fluence Map. The target fluence is denoted as F has the size of $X \times Y$. The part of fluence associated to the x-th leaf pair is denoted as F_x. The predicted fluence is termed $\hat{F}$. The k-th cumulated fluence is termed as $\hat{F}^k$, which is cumulated from CP_1 to CP_k, the unit fluence of CP_k is termed as $\tilde{F}^k$, so $\hat{F}^k = \sum M^k \cdot \tilde{F}^k$. Given K control points for the target fluence, we have $\hat{F} = \hat{F}^K$.

3D Dose. Tn the end-to-end AI pipeline. the reference dose is termed as D, and predicted dose is $\hat{D}$.

Leaf Pair. Tn a leaf pair, the leaf with smaller index is termed as A and the other is B. The position of x-th leaf pair at k-th control point is termed as (A_x^k, B_x^k).

Control Point: denoted as CP. A fluence is associated with K control points, and k-th control point is termed as CP_k.

Monitor Unit: termed as M. The monitor unit at k-th control point is M^k.

Leaf Policy: termed as π_{θ_l}. All leaf pairs share the same leaf policy.

MU Policy: termed as π_{θ_m}.

Critic Net: termed as V_ϕ.

Reward: the total reward is termed as R, and Reward i is termed as R_i. As the shown in the Eq. 2, the rewards related to x-th leaf pair is R_i^k.

Environment: is termed as Env, which follows radiotherapy-related computations in PORIx.

Expection: is termed as $\hat{\mathbb{E}}[\cdot]$.

Sigmod Function: is termed as $\sigma(\cdot)$.

Indicator Function. In this paper, $\mathbb{1}(\cdot)$ is defined as

$$\mathbb{1}_{(0,\infty)}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases}$$

B. Introduction of PORIx

Photon Optimizer Research Interface (PORIx) originates from a prominent commercial company, Varian Medical Systems, a Siemens Healthineers company. The framework of PORIx[3] is depicted in Figure 14. This research interface is adept at generating radiotherapy (RT) plans through scripting. Callbacks within the interface allow users to substitute predictions, facilitating the evaluation of AI models within the PORIx environment. It is worth noting that clinical planning is a complex and often subjective process, involving experts with diverse backgrounds. Various metrics and considerations may be included for planning optimization, beyond the metrics used in this paper.

In this study, our focus centers on assessing the leaf sequencing functions within the initial 100 iterations. During this phase, the number of control points in one sector is fixed at either 16 or 10, similar to the VMAT photon optimization in Varian's Eclipse. The main focus in this paper is on the first-level resolution of the multi-resolution optimization and no manual intervention during optimization.

One potential future work could involve initiating the planning process with deep learning plans, which offer rapid and automated solutions, followed by a limited number of human/optimization intervention steps. This approach may promise a faster and more efficient path compared to conventional methods that start from scratch, while also leading to safe plans approved by experts.

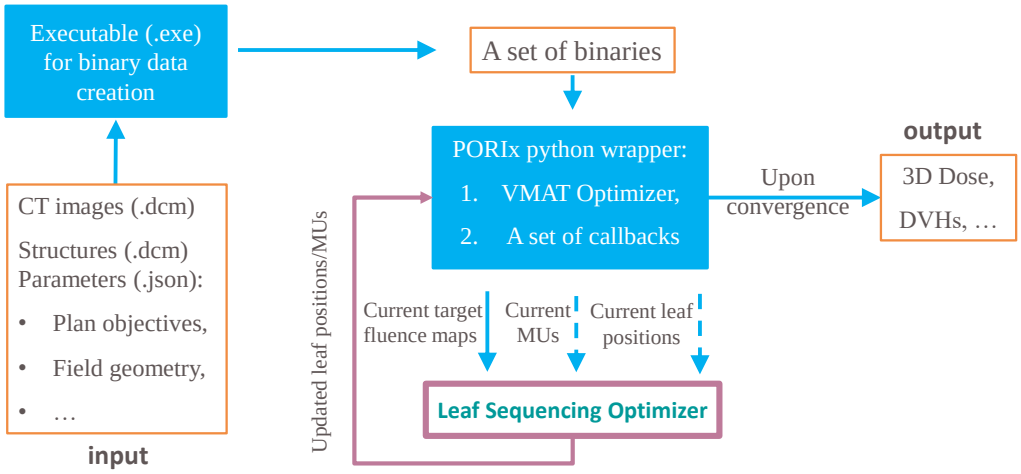


Figure 14. Illustration of RT planning with PORIx. The input of the planning includes CT, RT structures, and a set of planning parameters. The PORIx is a python wrapper of VMAT optimizer with a set of callbacks that can provide target fluences, monitor units, and leaf positions of the iterations. Leaf Sequencing Optimizer is another optimization layer for leaf sequencing only. Our RLS can replace the Leaf Sequencing Optimizer module to get an AI-integrated solution. Note our RLS do not take MUs/leaf positions from previous iteration which can work beyond a optimization environment.

We estimate the computational requirements (GPU memory occupation) of RLS and whole planning pipelines, as in Table 6. The RLS itself is computationally efficient due to the lightweight network structure described in Appendix D, and it does not pose a bottleneck within end-to-end pipelines.

Model	Estimated GPU memory occupation
RLS	∼0.35 GB
AI E2E pipeline (dose/fluence prediction + RLS)	∼2.8 GB
PORIx pipeline	∼3.0 GB

Table 6. Estimated GPU memory occupation when running in different contexts. *Disclaimer: The comparison is based on settings in this study, the number can be varied when the settings are different.*

[3]Our experiments are based on a version of PORIx in 2023. Some features may vary across different versions.

C. Additional Settings and Hyperparameters

```
#--------------------------------- General Parameters ---------------------------------#
deep learning platform: PyTorch 1.13
GPU type: NVIDIA RTX A4500
RL iterations: 20000 # similar concept as args.num_iterations in CartPolePPO of CleanRL
Update epochs: 2 # similar concept as args.update_epochs in CartPolePPO of CleanRL
batch_size: 96
#---------------------------------- Parameters for Environment -----------------------#
number of leaf agents: 50
max movement of leaf: 4
number of steps in an episode of training: 16
monitor unit range: 0.5 to 2.5
Reward 1 ratio: 1
Reward 2 ratio: 2
Reward 3 ratio: 2
Reward 4 ratio: 1
Reward 5 ratio: 1
#------------------------------- Parameters for Loss Functions -----------------------#
discount factor gamma: 0.99
GAE lambda: 0.95
toggles advantages normalization: True
surrogate clipping coefficient: 0.2
clip value loss: True
entropy coefficient: 0.01
maximum norm for the gradient clipping: 0.5

#------------------------------- Parameters for Optimizer ---------------------------#
learning rate: 1e-4
weigth decay: 1e-4
optimizer: AdamW
optimizer scheduler: CosineAnnealingLR
```

D. Details of Network Structures

Leaf Actor:

```
----------------------------------------------------------------
Layer type           Output Shape      Param #
================================================================
Input                  [-1, 205]               0
Linear-1               [-1, 512]          105472
ELU-1                  [-1, 512]               0
Linear-2               [-1, 512]          262656
ELU-2                  [-1, 512]               0
Linear-3               [-1, 512]          262656
ELU-3                  [-1, 512]               0
Linear-4               [-1, 512]          262656
ELU-4                  [-1, 512]               0
Linear-5               [-1, 18]             9234
ELU-5                  [-1, 18]                0
================================================================
Total params: 902674
Trainable params: 902674
----------------------------------------------------------------
```

MU Actor:

```
----------------------------------------------------------------
Layer type           Output Shape      Param #
================================================================
Input                  [-1, 205]               0
Linear-1               [-1, 512]          105472
ELU-1                  [-1, 512]               0
Linear-2               [-1, 512]          262656
ELU-2                  [-1, 512]               0
Linear-3               [-1, 512]          262656
ELU-3                  [-1, 512]               0
Linear-4               [-1, 2]              1026
ELU-4                  [-1, 2]                 0
================================================================
Total params: 631810
Trainable params: 631810
----------------------------------------------------------------
```

Critic:

```
----------------------------------------------------------------
Layer type           Output Shape      Param #
================================================================
Input                  [-1, 205]               0
Linear-1               [-1, 512]          105472
ELU-1                  [-1, 512]               0
Linear-2               [-1, 512]          262656
ELU-2                  [-1, 512]               0
Linear-3               [-1, 512]          262656
ELU-3                  [-1, 512]               0
Linear-4               [-1, 512]          262656
ELU-4                  [-1, 512]               0
Linear-5               [-1, 1]               513
ELU-5                  [-1, 1]                 0
================================================================
Total params: 893953
Trainable params: 893953
----------------------------------------------------------------
```

E. The Full-AI End-to-end (E2E) Pipeline

E.1. Introduction of Other Modules

Dose Prediction. Dose prediction with deep learning is a relatively well-studied topic. It can be considered an extension of planning objectives prediction compared to the dot objectives in conventional optimization or line objectives in RapidPlan. In most literature, the input of this module includes the CT and masks of PTV/OARs (along with other auxiliary conditions), and the output is the estimated 3D dose required for cancer treatment.

In this work, the dose prediction module is trained in a conditional GAN framework, inspired by Gao et al. (2023); Kearney et al. (2020). The adversarial training is to make the predicted 3D dose realistic. In the end-to-end AI pipeline, this dose prediction module serves as the *dose objectives definition* as in Figure 1.

Fluence Prediction. After obtaining the 3D dose, we predict this fluence map of each beam angle as a 2D intensity map. This module is motivated by Wang et al. (2020) and we have made significant changes to extend the fluence map prediction for VMAT. The fluence prediction module predicts all fluence maps jointly, whose input is projections of the 3D dose map to the beam's eye view of gantry angles. In the end-to-end AI pipeline, this dose prediction module serves as the *optimal fluence prediction* as in Figure 1. Following a similar setting in PORIx that multiple control points contribute to each fluence, the resolution of target fluences in the VMAT context is eight control points per fluence except the beginning and ending fluences is nine control points (as in bottom of Table 3). A full arc has 178 control points in total, which has been divided into 22 sectors for leaf sequencing.

The primary difference between this full-AI pipeline and frequently used optimization methods lies in its ability to be executed in a single (or few) iteration(s), thus the former can potentially accelerate the planning process significantly. It is important to note that this full-AI pipeline is currently in the preliminary stage and may yield lower quality plans than commercial optimization platforms. Also, some detailed clinical settings may be simplified in this AI pipeline. Despite these considerations, we see the full-AI pipeline as a valuable testbed, particularly since this study mainly focuses on evaluating the leaf sequencing module.

E.2. Consideration of Table 3

The rationale behind utilizing the dose prediction record presented in Table 3 is to assess the error scale of our model by comparing it to an extensively studied task (i.e., dose prediction) in a full-AI end-to-end pipeline. A good match between the target and prediction indicates promising performance of RLS in this end-to-end pipeline. Notably, there is currently no established reference in the literature for a direct comparison with our model in such a pipeline. The decision to refer to the OpenKBP record is based on its status as the most widely recognized open challenge in radiotherapy, with researchers consistently updating and improving the top-performing models through open-sourced evaluations. We acknowledge those two (i.e., dose prediction (upper part of Table 3) and evaluation fluence prediction in 3D dose perspective (lower part of Table 3) [4] are not the same task and show their similarity and differences below:

Similarity: both those two tasks are evaluated by the difference between two 3D dose maps based modules in radiotherapy plan workflow.

Differences: as mentioned above, the outcome in dose prediction is a 3D estimation of dose distribution, and there is no guarantee that the machine can achieve the predicted dose. Contrastively, the two doses from the second task are obtained by the same dose calculation method (i.e., Acuros XB of Eclipse[5]) and the target/prediction fluences. Fundamentally, this is for evaluating the performance of leaf sequencing but using an auxiliary way. The predicted dose of this module is machine executable.

[4] The number of test patients for end-to-end pipeline is 64 since treatment configurations of some patients are not applicable to dose/fluence prediction modules.

[5] https://www.varian.com/products/radiotherapy/treatment-planning/eclipse

F. Details/Pseudo-Code of Rewards

This section provides supplementary information for rewards in Section 4.1. The trade-off between reconstruction accuracy and other clinically relevant factors is carefully considered in the design of rewards. Specifically, Reward 1 and 2 are to guide agents to reconstruct the target fluence, and Reward 3, 4, and 5 take into account various clinical considerations.

Reward 1. As depicted in Figure 13c, the green parts indicate the Reward 1, which pushes the cumulative fluence closer to target fluence.

Reward 2. As depicted in Figure 13c, the red parts indicate Reward 2, which punishes agents if the intensity of cumulative fluence exceeds that of target fluence.

Reward 3: avoiding cross leaves. This reward component is to encourage the agent to avoid predictions where left- and right-leaf positions intersect, which violates the physical laws. In addition to this, the environment enforces that the final leaf positions do not intersect and do not extend beyond the radiation fields.

Reward 4: smoothness of actions (Leaf/MU change between control-points). Ensuring smooth change in leaf / MU configuration is necessary as significant or sudden changes in leaf positions may cause machine instability. We demonstrated that by tuning the weight of the reward (Figure 11), our model allows users to balance between the maximum leaf speed (machine-dependent parameter) and the reconstruction performance.

Reward 5: aperture shape regularization motivated from Younge et al. (2012). According to Younge et al., volumetric modulated arc therapy (VMAT) planning often yields small, irregular aperture shapes, leading to dosimetric inaccuracies during delivery. We included this regularization as R5 = area of aperture / perimeter of aperture following Younge et al. (2012).

Below shows the Pytorch-style pseudo code of the rewards:

```
# MU: monitor unit,  mask: unit fluence
# y_tail: leaf position of leaf A_x (i.e., the left leaf or the tail leaf)
# y_front: leaf position of leaf B_x (i.e., the right leaf or the front leaf)
rw1 = MU * torch.sum((tar_fluence - cumu_fluence >= MU) * mask, axis = 1)
cumu_fluence += mask * MU
rw2 = - torch.sum((tar_fluence - cumu_fluence < 0) * mask, axis=1)
rw3 = (y_front - y_tail > 0) - (y_front - y_tail < 0)
rw4 = 3 - sigmoid(delta_front) - sigmoid(delta_tail) - sigmoid(delta_MU)
rw5 = Aperture(y_front, y_tail)
```

G. Dataset Description

The dataset distribution is described in Table 7. Unlike tasks in other AI fields (e.g., computer vision or natural language processing tasks are usually easy to get thousands, millions, or even billions of data samples), RT planning data for research use is relatively limited. Most RTP research related to deep learning typically involves fewer than 500 patients. For example, the well-studied OpenKBP challenge includes only 340 patients. Our study is conducted on relatively larger-scale data resources with more patients.

Dataset Name	Cancer Site	# Patient after filtering	train/val/test	Availability
HNd	head-and-neck	493	370/10/113	Data will not be public available due to data regularization
HNe1	head-and-neck	31	0/0/31	Public: www.cancerimagingarchive .net/collection/hnscc-3dct-rt
HNe2	head-and-neck	140	0/0/140	Public: www.cancerimagingarchive .net/collection/hnscc
Pros	Prostate	555	471/10/84	Public with permission required (Seibold et al., 2019)
Pros(e)	the same set as Pros, but test results are reported when model trained by head-and-neck site.			

Table 7. The description of datasets used in our experiments. Note that our experiment is conducted at fluence level. Each patient has up to 525 target fluences (within 100 iterations optimization in PORIx). Compared to original dataset, some patients are filtered due to missing data, quality issue, and incompatibility with environment/software.

The experiments and evaluations on different datasets/sites/scenarios (e.g., Table 1 and 2) indicate the generalizability of our model. Furthermore, leaf sequencing is executed at the fluence-level, encompassing a sample size that is hundreds of times larger than the number of patients. Our model can have a reasonably good performance even when the training size is smaller, as in Table 8.

Model	MNSE (↓)
PORIx	0.219
RLS (trained with 50 patients)	0.158
RLS (trained with 100 patients)	0.152
RLS (trained with 370 patients)	0.149

Table 8. The reconstruction error MNSE of different number of training patients.

H. Supplementary Examples of *Cropping* Ablations

As shown in Figure 15, the *cropping* strategy helped the optimization in terms of faster convergence. Overall, our RLS can achieve smaller error and faster convergence especially in early iterations during the optimization.

I. Additional Examples Illustrating Reconstruction Errors

Continued with Figure 7, we supplement more and detailed illustrations of fluence reconstruction in Figure 16. Figure 16a to Figure 16h represents examples from eight different patients, where (a) - (f) are head-and-neck cancer cases, and (g) - (h) are from prostate cancer site. For each case, the five figures represent **target fluence**, **predicted fluence from RLS**, **difference between RLS prediction and target**, **predicted fluence from PORIx**, **difference between PORIx prediction and target**, from up to down. The reconstruction error is shown at right bottom. RLS achieves smaller error in the majority of cases.

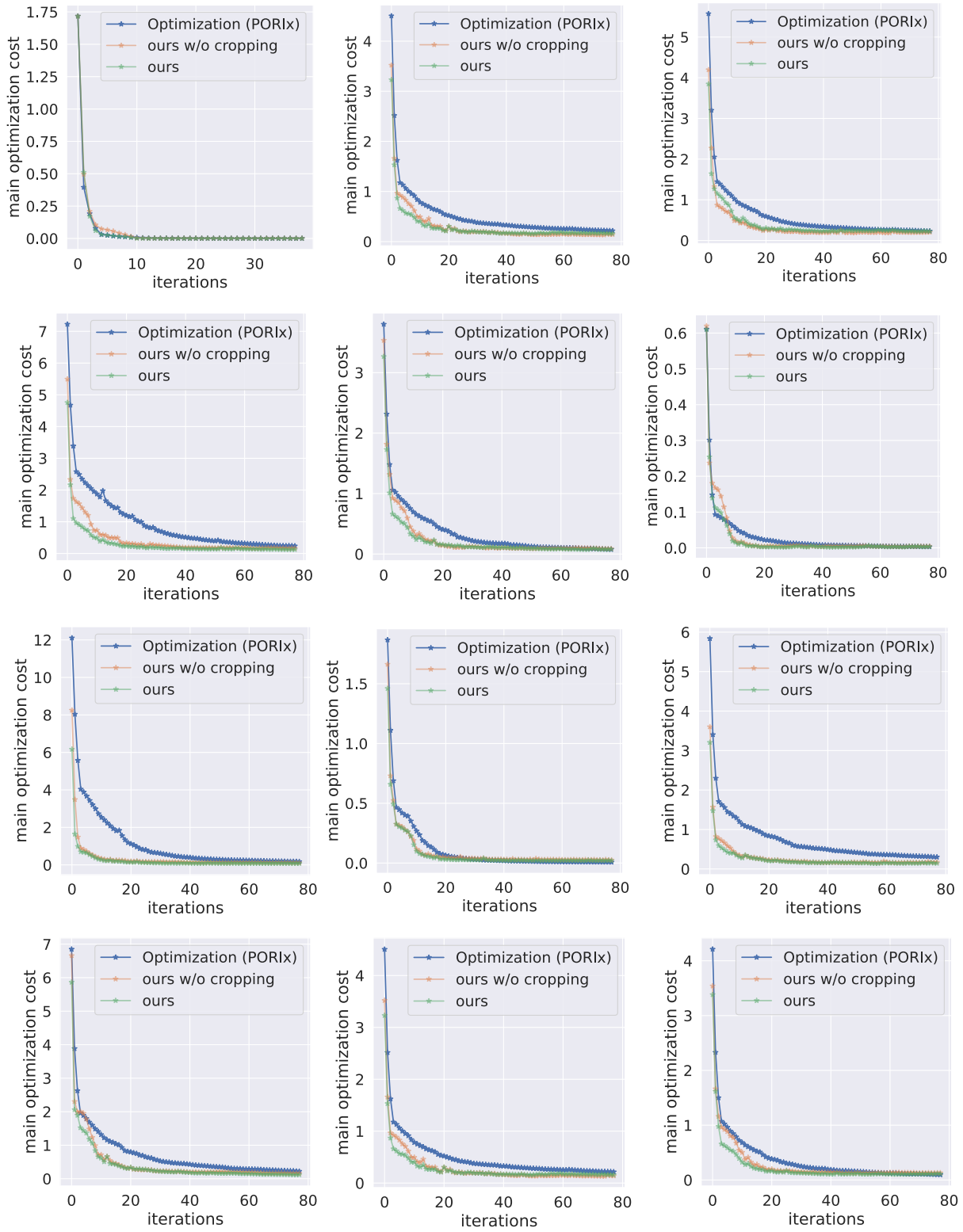


Figure 15. Ablation study with vs. without *cropping*. The main optimization cost is from callback of PORIx planning environment. **ours** denotes that our RLS has replaced the leaf sequencing optimizer of PORIx.

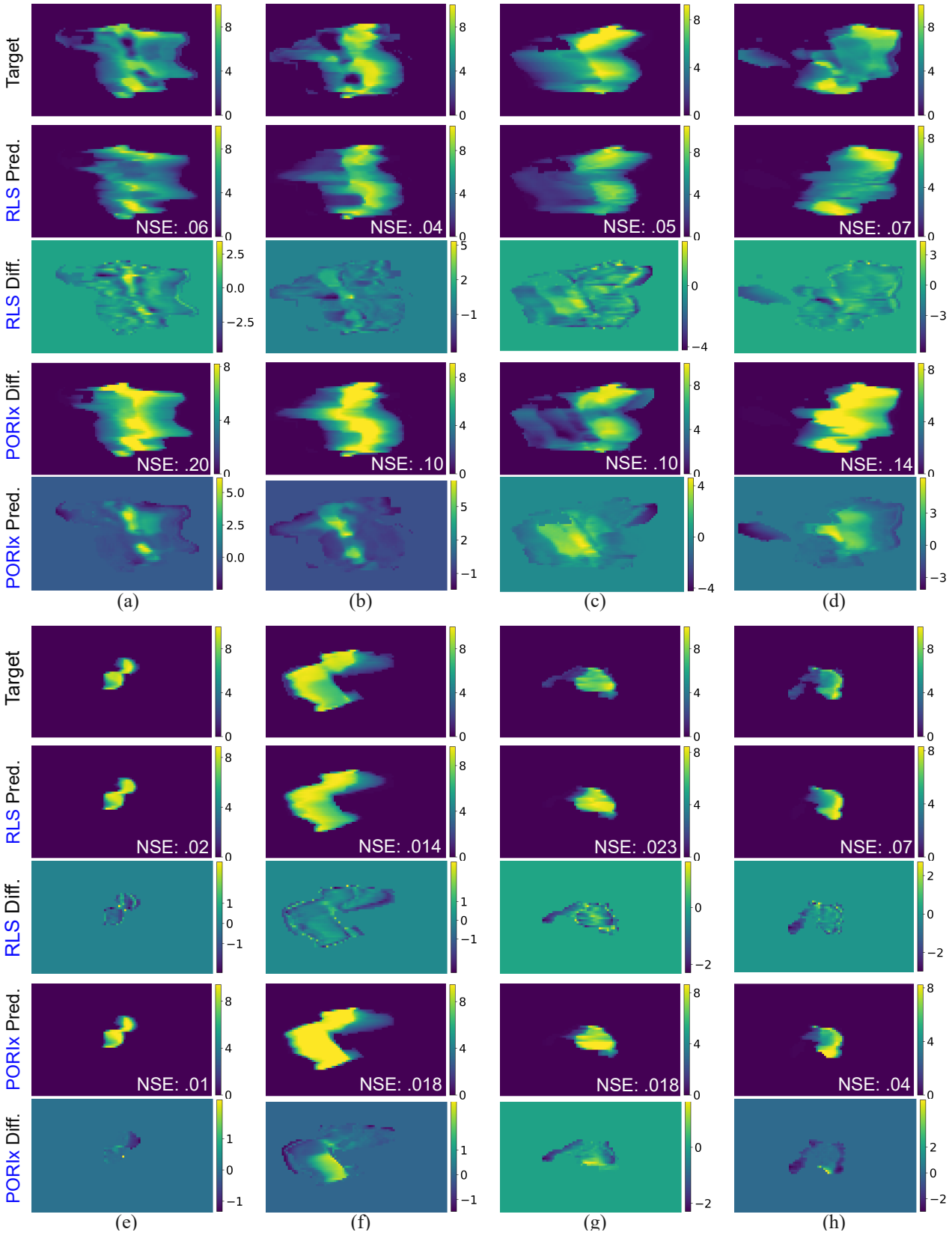


Figure 16. Supplement examples of Figure 7 in the main text. We provide difference maps between predicted and target fluences here.