FROM ANATOMY TO DELIVERY: PHYSICS-ALIGNED CREDIT ASSIGNMENT FOR VMAT RADIOTHERAPY PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiation therapy planning translates a precise spatial balance between curative target coverage and healthy organ sparing into executable machine instructions. Volumetric modulated arc therapy (VMAT) realizes this delivery through a rotational trajectory of tightly coupled multileaf collimator apertures and beam intensities. Existing automated approaches struggle to connect clinical objectives directly to these machine parameters. A common approach to addressing this challenge is to use two-stage pipelines that optimize intermediate dose representations without physical constraints, causing the clinical trade-offs to degrade when eventually converted into deliverable actions. Conversely, methods that act directly on machine variables typically rely on a single plan-level clinical reward. This global scalar creates a severe credit assignment bottleneck, failing to isolate which specific leaf or intensity adjustment improved the target or harmed an organ. Differentiating the clinical objective provides local dose sensitivities for every machine variable. We find that translating these independent local gradients into a continuous, deliverable trajectory requires separating physical credit construction from policy learning. We introduce Physics-Aligned Credit for VMAT (PAC-VMAT) to realize this separation. The framework evaluates the evolving plan through a dose engine to construct explicit, fixed credit targets. Parameterized Leaf and MU policy agents then learn to coordinate these local clinical sensitivities into mechanically feasible aperture and intensity updates. Evaluated on lung and head-and-neck benchmarks, PAC-VMAT substantially improves the clinical target–organ balance over scalar-reward reinforcement learning and direct machine optimization, achieving plan quality comparable to closed-source commercial engines. This direct translation from patient anatomy to executable trajectories reduces optimization time to minutes, safely lowering collateral organ exposure while strictly satisfying mechanical delivery constraints.

1 INTRODUCTION

Radiation therapy treats cancer by delivering a curative dose of ionizing radiation to a tumor while minimizing collateral exposure to surrounding organs at risk (OARs). Volumetric modulated arc therapy (VMAT) balances this spatial trade-off by delivering radiation continuously as the treatment machine rotates around the patient (Otto, 2008). During this trajectory, multileaf collimator (MLC) apertures actively reshape the radiation field, and monitor unit (MU) increments dictate the beam intensity delivered from each angle. Protecting healthy organs while maintaining target coverage thus requires coordinating thousands of tightly coupled aperture and intensity controls throughout the entire delivery sequence. These controls must also satisfy the strict mechanical delivery limits of the treatment machine, which constrain leaf motion and intensity increments along the trajectory (Papp & Unkelbach, 2014; Unkelbach et al., 2015).

Because directly optimizing this high-dimensional machine sequence is difficult, some studies have proposed two-stage pipelines that first predict a desired dose distribution or per-angle fluence from patient anatomy and then recover machine parameters that reproduce it, either through the optimizer of a treatment planning system (McIntosh et al., 2017; Babier et al., 2021) or through a learned leaf

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sequencer (Gao et al., 2024; Heilemann et al., 2025; Arberet et al., 2026). The resulting plan is then highly dependent on the quality of the dose distribution or fluence map predicted in the first stage.

Besides, VMAT planners that learn direct machine control from a plan-level clinical cost (Hrinivich et al., 2024; Mekki et al., 2025) face a severe credit assignment bottleneck. Such a scalar cost compresses the physical consequences of the entire machine trajectory into a single plan-level score, leaving an optimization process to infer the local contribution of every specific control. Other reinforcement learning planners that act directly on leaf positions and dose rates likewise use scalar dose-based feedback (Hrinivich & Lee, 2020; Shaffer et al., 2026). The resulting policy is then highly dependent on exploration, so the clinical criteria cannot reach the individual aperture and intensity updates, leaving learning highly inefficient.

To address these challenges, we propose Physics-Aligned Credit for VMAT (PAC-VMAT), a framework that generates deliverable radiotherapy trajectories directly from patient anatomy. While differentiating the clinical objective through a dose engine yields dense local sensitivities for every machine variable (Simkó et al., 2025), we find that directly optimizing against these raw gradients struggles to coordinate the thousands of tightly coupled controls required for mechanical delivery. PAC-VMAT bridges this gap by separating physical credit construction from policy learning. Optimization begins from an anatomy-guided initialization that establishes a physically feasible, target-aligned starting plan. The framework then continually evaluates the evolving dose to construct explicit, normalized credit targets for each machine parameter. Parameterized Leaf and MU policy agents use these targets to learn coordinated aperture and intensity updates under configured mechanical constraints. This continuous alignment between local clinical credit and the physical delivery state provides the structure needed to translate independent sensitivities into a unified, executable trajectory.

We validate PAC-VMAT on lung and head-and-neck benchmarks, where it substantially improves the clinical target–organ balance over scalar-reward reinforcement learning and direct machine optimization. The resulting plans achieve quality comparable to closed-source commercial engines through optimization that takes only minutes. These results establish an efficient route from patient anatomy to executable treatment trajectories, reducing collateral organ exposure while strictly satisfying mechanical delivery constraints.

Our contributions are as follows:

- We identify the limitation of proxy-based automated VMAT planning, where clinical objectives are optimized separately from executable machine delivery. Existing approaches typically predict or optimize intermediate dose or fluence representations before generating machine parameters, causing the clinical trade-off to be fixed before delivery realization. This separation prevents physical delivery constraints from providing feedback to refine the clinical objective.
- We introduce Physics-Aligned Credit for VMAT (PAC-VMAT), a framework that separates physical credit construction from policy learning. PAC-VMAT differentiates the clinical objective through a dose engine to convert local sensitivities into explicit credit targets, enabling Leaf and MU policy agents to learn coordinated, deliverable treatment trajectories.
- We demonstrate on lung and head-and-neck benchmarks that PAC-VMAT substantially improves target–organ balance over scalar-reward reinforcement learning and direct machine-variable optimization, achieves plan quality comparable to closed-source commercial engines, and reduces optimization time to minutes while satisfying mechanical delivery constraints.

2 VMAT Planning Formulation

2.1 Sequential VMAT Planning

Given a computed tomography (CT) volume x and anatomical structure masks s that delineate the planning target volume (PTV) and the organs at risk (OARs), VMAT planning aims to determine, how to irradiate the patient across a sequence of control points so that the PTV receives sufficient dose while exposure to the OARs is kept as low as possible (Unkelbach et al., 2015) (Figure 1).

At each control point $i \in \{1, \dots, N\}$, the MLC positions ℓ_i define the aperture geometry, while the MU increment u_i determines the beam intensity delivered from that point. A plan collects these

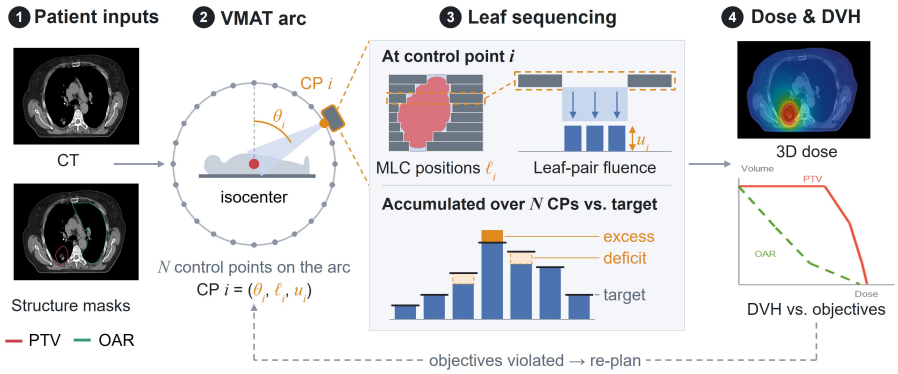

Figure 1: Conventional fluence-based VMAT planning. Clinical objectives define a target fluence, which leaf sequencing approximates through MLC positions ℓ_i and MU increments u_i across the gantry trajectory. The resulting 3D dose is checked against dosimetric criteria, with unmet criteria triggering further adjustment.

per-control-point settings over the whole delivery sequence (Papp & Unkelbach, 2014; Unkelbach et al., 2015), and the machine variables operate under strict mechanical delivery limits, which restrict the plan to the feasible plan space $\mathcal{F}$:

$$P = (\ell, u) \in \mathcal{F}, \qquad \ell = \{\ell_i\}_{i=1}^{N}, \qquad u = \{u_i\}_{i=1}^{N}, \tag{1}$$

We formulate treatment planning as an iterative refinement process over T steps. At step t, the current feasible plan P_t produces a three-dimensional dose distribution $d_t = \mathcal{D}(x; P_t)$ through a dose engine. Because aperture geometry and beam intensity jointly determine this spatial dose, their optimization is physically coupled (Unkelbach et al., 2015). Any intensity adjustment inherently acts upon the specific geometry established by the leaves. To preserve this directional physical dependency during optimization, the plan state transitions sequentially:

$$(\mathrm{P}_t, d_t) \xrightarrow{\text{Leaf}} (\bar{\mathrm{P}}_t, \bar{d}_t) \xrightarrow{\text{MU}} (\mathrm{P}_{t+1}, d_{t+1}) \tag{2}$$

The intermediate plan $\bar{P}_t$ and its recomputed dose $\bar{d}_t = \mathcal{D}(x; \bar{P}_t)$ ensure that subsequent intensity decisions respond directly to the newly realized aperture geometry. This formulation explicitly isolates the distinct physical consequences of spatial and intensity updates.

2.2 Clinical Plan Evaluation

The overall objective of the treatment planning process carried out by radiation dosimetrists or medical physicists can be expressed as a unified reward function $r(d)$ that aggregates weighted dose-objective criteria for the planning target volume and all relevant organs at risk $k \in \mathcal{O}(s)$ (Shaffer et al., 2026):

$$r(d) = \sum_{i} w_{\mathrm{PTV},i}\, r(d,\, o_{\mathrm{PTV},i}) + \sum_{k \in \mathcal{O}(s)} \sum_{j} w_{\mathrm{OAR},k,j}\, r(d,\, o_{\mathrm{OAR},k,j}). \tag{3}$$

where d is the dose distribution, $\mathcal{O}(s)$ is the set of organs at risk for treatment site s, $o_{\mathrm{PTV},i}$ and $o_{\mathrm{OAR},k,j}$ are the dosimetric objectives for the PTV and the k-th OAR respectively, $r(d,\, o)$ computes a scalar reward for each dose-objective pair, and $w_{\mathrm{PTV},i}$ and $w_{\mathrm{OAR},k,j}$ are the corresponding weights controlling the relative importance of each objective.

3 Method

3.1 PAC-VMAT Framework

In this section, we propose PAC-VMAT, a novel framework that addresses the credit assignment bottleneck of optimizing a scalar plan-level reward over executable machine variables. PAC-VMAT

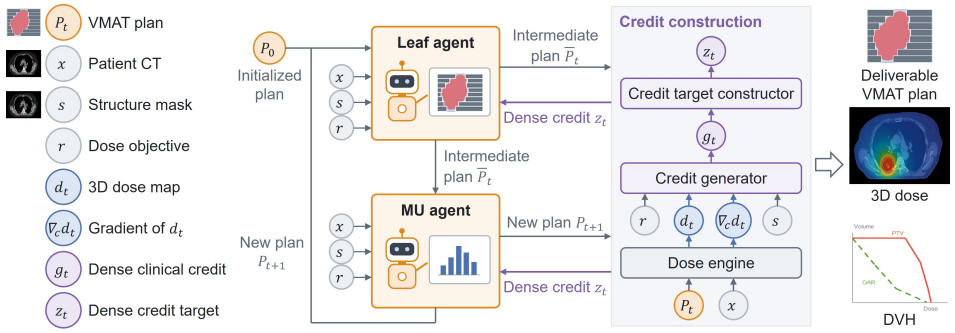

Figure 2: Overview of the PAC-VMAT framework.

separates the construction of physical credit from policy learning. A credit construction module (CCM) differentiates the clinical reward through a differentiable dose engine into a dense, signed clinical credit for every machine control, converts it into fixed targets in the action coordinates of the agents (Section 3.2). A policy learning module (PLM) then trains a pair of Leaf and MU policy agents to organize these targets into coordinated aperture and intensity updates that respect the configured mechanical limits (Section 3.3). We further introduce an anatomy-guided initialization that supplies a physically feasible starting plan (Section 3.4). Figure 2 illustrates the framework.

3.2 Credit Construction Module

Dense Clinical Credit from a Differentiable Dose Engine. Let $p \in \{\mathrm{Leaf}, \mathrm{MU}\}$ index the two policy agents, with controlled variables $c^{\mathrm{Leaf}} = \ell$ and $c^{\mathrm{MU}} = u$. Their decision states are $P_t^{\mathrm{Leaf}} = P_t$ and $P_t^{\mathrm{MU}} = \bar{P}_t$, respectively. We define the dense clinical credit at P_t^p as the local reward sensitivity:

$$g_t^p = \nabla_{c^p} r(\mathcal{D}(x; P_t^p)) = \left(\frac{\partial \mathcal{D}(x;P)}{\partial c^p}\bigg|_{P=P_t^p} \right)^\top \nabla_d r(d_t^p), \qquad d_t^p = \mathcal{D}(x; P_t^p). \tag{4}$$

This formulation connects clinical requirements to machine control through two physical stages. The dose gradient $\nabla_d r(d_t^p)$ identifies the spatial regions where local dose adjustments would improve the target–OAR balance. The dose Jacobian translates these spatial requirements into sensitivities of the specific leaf and MU controls. The policy agents thereby receive explicit directions describing the local effect of each variable on the current clinical objective.

Credit Targets in Policy Action Coordinates. The credit in Equation 4 is expressed in the physical machine variables and is not yet a learning target for the policy agents. Let a^p denote the action coordinates of policy agent p, which parameterize a machine update through a differentiable map $c^p = f_p(a^p)$ at the current state. Their local reward sensitivity is $\tilde{g}_t^p = (\partial f_p / \partial a^p)^\top g_t^p$, where the additional Jacobian accounts for the agent's action parameterization on top of the dose Jacobian already contained in g_t^p. We normalize these action-coordinate sensitivities by their maximum norm within each control group and detach the result from the computation graph to obtain the credit target

$$z_t^p = \mathrm{sg}\left[\frac{\tilde{g}_t^p}{\max\left(\|\tilde{g}_{t,g}^p\|_\infty, \epsilon \right)} \right], \qquad \tilde{g}_t^p = \left(\frac{\partial f_p}{\partial a^p} \right)^\top g_t^p, \tag{5}$$

where $\|\tilde{g}_{t,g}^p\|_\infty$ is the maximum norm within the control group g containing each coordinate, ϵ is a numerical floor that bounds the denominator away from zero and limits the amplification of near-zero gradients. The stop-gradient operator $\mathrm{sg}[\cdot]$ keeps the target fixed, so that the policy update in Section 3.3 optimizes θ_p against a constant direction instead of differentiating through the dose engine. This target organizes the raw physical gradients into a consistent optimization scale while preserving the relative clinical importance of each local machine control.

3.3 Policy Learning Module

Policy Agents as Structured Action Parameterization. To turn the independent improvement directions of the CCM into a coordinated, mechanically feasible trajectory, the PLM trains a pair

of Leaf and MU policy agents. By sharing parameters across the sequence of control points, the networks couple adjacent leaf movements and balance intensity distributions across the arc, bridging the gap between independent physical sensitivities and a continuous, deliverable plan.

Each agent acts on an observation o_t^p that concatenates, at every control point along the trajectory, the beam's-eye-view anatomy of the current state, the realized machine variables, the current dose metrics, and the local clinical credit. The observation is rebuilt after the Leaf update, so that the MU agent conditions its intensity decision on the realized aperture geometry. Appendix C describes the network capacity and the action coordinates of each agent.

Learning against Fixed Credit Targets. To translate this local guidance into executable machine actions, policy learning aligns the network outputs with the credit targets. At each refinement step, policy agent $p \in \{\mathrm{Leaf}, \mathrm{MU}\}$ maps its observation o_t^p to a proposed machine update $u_t^p = \mu_{\theta_p}(o_t^p)$, and the regression term fits this update to the fixed clinical target z_t^p:

$$\mathcal{L}_{\mathrm{credit}}(u_t^p, z_t^p) = \lambda_{\mathrm{fit}} \frac{\sum_{j \in \mathcal{C}^p} w_j \, h\big(u_{t,j}^p - z_{t,j}^p\big)}{\sum_{j \in \mathcal{C}^p} w_j}, w_j = \mathrm{clip}\left(\frac{\big|z_{t,j}^p\big|}{\max\Big(\bar{z}_{t,g(j)}^p, \epsilon_w\Big)}, \, w_{\min}, \, w_{\max} \right), \tag{6}$$

where $h(\cdot)$ is the unit-transition Huber penalty, $\bar{z}_{t,g(j)}^p$ is the mean absolute credit within the control group $g(j)$ containing coordinate j, and ϵ_w, $w_{\min}$, and $w_{\max}$ are the mean-sensitivity floor and the baseline and cap of the coordinate weight given in Table 9. The sensitivity weighting makes the agent fit the controls with a larger local clinical response more closely, while coordinates whose clinical effect is negligible contribute little to the loss. Holding the target z_t^p fixed separates the physical evaluation of the dose from the network optimization, allowing the policy agent to reliably fit the specific clinical needs of the current state before the machine executes the update.

Combined Objective and Sequential Refinement. Fitting the credit alone is not sufficient: independent coordinate-wise responses would violate the mechanical limits and break the continuity that delivery requires. We therefore combine the regression term with a structural regularizer $\mathcal{R}_p$ that constrains how the proposed update varies across controls. The objective optimized at each refinement step is

$$\mathcal{L}_t^p(\theta_p) = \mathcal{L}_{\mathrm{credit}}\big(\mu_{\theta_p}(o_t^p), z_t^p\big) + \mathcal{R}_p\big(\mu_{\theta_p}(o_t^p), z_t^p\big) \, . \tag{7}$$

The regularizer $\mathcal{R}_p$ restricts abrupt changes across adjacent controls: it penalizes directional disagreement with the credit target at the global, leaf-pair, and arc levels, suppresses motion on coordinates whose clinical response is negligible, and bounds the magnitude, roughness, and curvature of the update (Appendix C.2, Equation 12). These terms supply the structural continuity that independent variable tuning lacks. Because aperture geometry determines the spatial distribution of radiation, applying the learned Leaf update alters the clinical impact of subsequent intensity adjustments. We therefore execute the leaf movements and recompute the intermediate dose and clinical credit before querying the MU policy agent, keeping intensity guidance aligned with the newly realized aperture shape.

3.4 Anatomy-Guided Initialization

Optimizing a random plan requires coordinating thousands of coupled leaf positions and MU increments to satisfy competing clinical criteria. We therefore initialize refinement from an anatomy-guided plan that provides a clinically meaningful and physically feasible starting point. The leaf apertures ℓ_0 follow the beam's-eye-view projection of the planning target volume at each gantry angle. We apply a uniform intensity sequence u_0 and scale it to exactly meet the prescribed target coverage $q_\star$. Letting $q(d)$ denote the coverage of the unscaled dose, the initialized plan becomes

$$P_0 = (\ell_0, \gamma_0 u_0) \in \mathcal{F}, \qquad \gamma_0 = \frac{q_\star}{q(\mathcal{D}(x; \ell_0, u_0))}. \tag{8}$$

This target-aligned baseline allows subsequent updates to focus on spatial dose redistribution between the target and surrounding organs. Appendix B details the aperture construction and coverage normalization.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate PAC-VMAT on a lung cohort ($n = 70$) and a head-and-neck cohort ($n = 67$) from the 2025 AAPM GDP-HMM Grand Challenge (Gao et al., 2025). The lung cohort is optimized with 180 control points per arc (Lung VMAT). PAC-VMAT acts as a direct trajectory optimization process that trains a separate pair of Leaf and MU policy agents for each patient. Using PyDoseRT as the differentiable dose engine enables these agents to generate executable plans directly from clinical criteria, establishing a planning route independent of commercial engines or reference trajectories.

To evaluate how effectively these generated plans resolve the spatial trade-off between curative radiation and healthy tissue preservation, we use dosimetric outcomes based on established clinical guidelines (Radiation Therapy Oncology Group, 2013; NRG Oncology, 2016). We measure target coverage (D_{95}) and localized target hotspots ($D_{0.03\,\mathrm{cc}}$), alongside standard volume- and dose-based constraints to quantify collateral organ exposure. All generated trajectories and the reproduced automated baselines (Gao et al., 2024; Shaffer et al., 2026) are evaluated at matched target coverage to isolate their spatial dose redistribution. Complete cohort eligibility, metric definitions, and baseline configurations are detailed in Appendix D. The implementation and evaluation code are available in an anonymized repository at `https://github.com/anonymous13756/PAC-VMAT`.

4.2 MAIN RESULTS

To isolate how explicit clinical credit coordinates high-dimensional machine variables, we evaluate generated VMAT trajectories against random plans, anatomy-guided initializations, and existing automated baselines across lung and head-and-neck geometries. Matched prescribed target coverage removes overall intensity scaling from the comparison, allowing us to assess how local clinical guidance redistributes dose between the target and surrounding organs.

We observe that PAC-VMAT contracts intermediate and high-dose regions around the target (Figures 3 and 4). In the lung cohort, it reduces target hotspots relative to the initial and random plans and lowers organ exposure relative to random plans (Table 1). In the head-and-neck cohort, it lowers spinal-cord, laryngeal, and brachial-plexus doses relative to the reproduced baselines (Table 2). Against the closed-source commercial reference, PAC-VMAT remains comparable, with a similar target hotspot and lower lung, spinal-cord, esophageal, and heart doses in the lung cohort and comparable target and organ endpoints in the head-and-neck cohort (gray rows, Tables 1 and 2; Appendix D.5). These results demonstrate how explicit credit targets guide coordinated machine updates to reshape the target–organ balance.

4.3 EFFECT OF CLINICAL CREDIT

To evaluate how clinical feedback guides policy learning, we compare policy agents trained with scalar plan-level rewards, directly against the clinical objective, and with explicit clinical credit. The

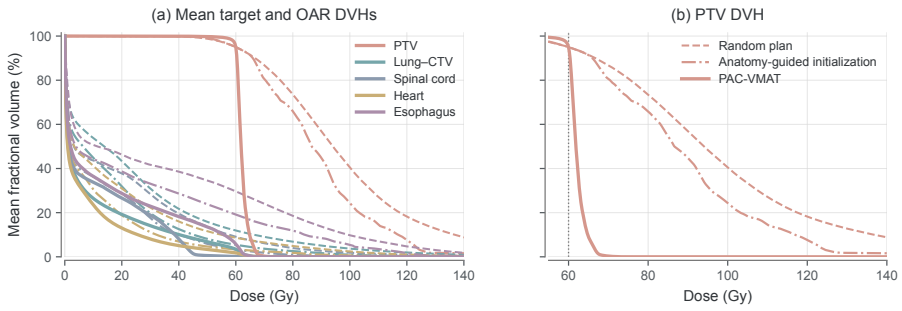

Figure 3: Mean dose–volume histograms (DVHs) at matched target coverage. Individual plans are scaled to PTV $D_{95} = 60$ Gy before pointwise averaging with equal patient weights across the evaluation cohort.

Table 1: Dosimetric evaluation of the lung cohort (Lung VMAT) at matched target coverage (PTV $D_{95} = 60$ Gy). The gray-shaded method row reports the GDP-HMM reference plan as computed in the closed-source commercial engine (Eclipse), all remaining method rows are computed with the open-source PyDoseRT engine.

Method	**PTV**		**Lung–CTV**		**Spinal cord**	**Esophagus**	**Heart**
	D_{95} (Gy)	$D_{0.03\,\mathrm{cc}}$ (Gy) ↓	$V_{20\,\mathrm{Gy}}$ (%) ↓	Mean (Gy) ↓	$D_{0.03\,\mathrm{cc}}$ (Gy) ↓	Mean (Gy) ↓	$D_{66.7\%}$ (Gy) ↓
Criterion	≥ 60	≤ 72	≤ 37	≤ 20	≤ 50.5	< 34	< 45
Eclipse reference (closed-source)	60.00 ± 0.23	63.38 ± 0.99	20.64 ± 10.64	11.79 ± 5.01	41.61 ± 26.09	19.73 ± 4.20	2.44 ± 2.75
Lung VMAT							
Proprietary plans	$\mathbf{60.00 \pm 0.00}$	112.77 ± 19.08	26.48 ± 15.24	14.79 ± 7.36	37.08 ± 20.63	22.42 ± 14.93	2.81 ± 2.74
Random	$\mathbf{60.00 \pm 0.00}$	145.20 ± 42.85	43.38 ± 23.97	24.49 ± 15.71	62.35 ± 40.06	36.03 ± 24.32	6.18 ± 9.21
Conformal arc	$\mathbf{60.00 \pm 0.00}$	98.52 ± 19.01	37.57 ± 16.31	18.93 ± 7.89	46.96 ± 20.46	32.47 ± 14.87	3.36 ± 4.40
Gao et al. (2024)	$\mathbf{60.00 \pm 0.00}$	99.41 ± 19.98	37.02 ± 15.96	18.67 ± 7.64	46.67 ± 20.20	31.96 ± 14.56	3.33 ± 4.35
Shaffer et al. (2026)	$\mathbf{60.00 \pm 0.00}$	98.52 ± 18.98	37.55 ± 16.28	18.92 ± 7.89	46.99 ± 20.48	32.47 ± 14.86	3.34 ± 4.36
PAC-VMAT	$\mathbf{60.00 \pm 0.00}$	$\mathbf{64.12 \pm 2.73}$	$\mathbf{19.20 \pm 9.05}$	$\mathbf{10.39 \pm 4.63}$	$\mathbf{36.37 \pm 14.13}$	$\mathbf{15.16 \pm 9.49}$	$\mathbf{1.79 \pm 2.81}$

Table 2: Dosimetric evaluation of the head-and-neck cohort at matched target coverage (PTV70 $D_{95} = 70$ Gy). The gray-shaded method row reports the GDP-HMM reference plan as computed in the closed-source commercial engine (Eclipse), all remaining method rows are computed with the open-source PyDoseRT engine.

Method	**PTV70**	**Spinal cord**	**Larynx**	**Brachial plexus**
	D_{95} (Gy)	$D_{0.03\,\mathrm{cc}}$ (Gy) ↓	$D_{\max}$ (Gy) ↓	$D_{\max}$ (Gy) ↓
Criterion	≥ 70	≤ 48	< 45	≤ 60
Eclipse reference (closed-source)	70.00 ± 0.00	29.91 ± 3.63	73.70 ± 2.63	68.68 ± 5.73
Proprietary plans	$\mathbf{70.00 \pm 0.00}$	81.62 ± 38.07	164.61 ± 73.26	147.39 ± 72.74
Random	$\mathbf{70.00 \pm 0.00}$	118.57 ± 51.17	143.45 ± 53.73	131.23 ± 56.74
Conformal arc	$\mathbf{70.00 \pm 0.00}$	76.30 ± 24.72	124.88 ± 22.98	90.57 ± 40.39
Gao et al. (2024)	$\mathbf{70.00 \pm 0.00}$	75.49 ± 24.65	124.85 ± 23.20	89.13 ± 39.56
Shaffer et al. (2026)	$\mathbf{70.00 \pm 0.00}$	76.17 ± 24.65	124.61 ± 22.95	90.41 ± 40.31
PAC-VMAT	$\mathbf{70.00 \pm 0.00}$	$\mathbf{30.94 \pm 4.93}$	$\mathbf{75.65 \pm 2.72}$	$\mathbf{66.11 \pm 11.07}$

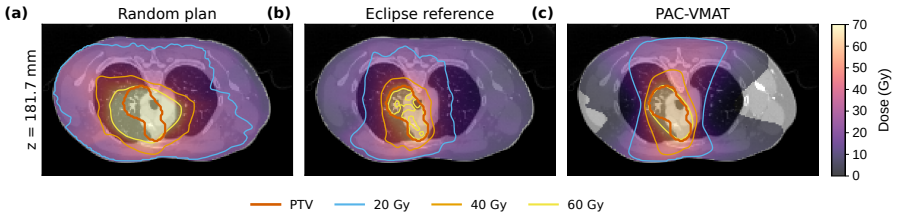

Figure 4: Axial dose distributions for the random and PAC-VMAT plans evaluated in PyDoseRT, alongside the Eclipse reference plan.

latter two variants use the same Leaf and MU policy agent architecture, allowing us to examine how local clinical sensitivities are used during learning. Comparing the resulting target–OAR balance tests whether separating physical credit construction from policy learning improves the joint refinement of aperture geometry and beam intensity (Table 3).

We find that policy learning with explicit clinical credit yields a substantially better target–OAR balance than both alternatives. Training the policy agents directly against the clinical objective reduces organ exposure compared to scalar rewards, while PAC-VMAT further lowers exposure across all evaluated organs at risk.

4.4 Effect of Policy Parameterization

To examine whether parameterized policy agents coordinate high-dimensional machine updates better than independent variable adjustments, we compare PAC-VMAT against direct constrained optimization. Both methods refine the identical anatomy-guided initial plan against the same clinical

Table 3: Clinical outcomes under different policy training signals at matched target coverage.

		PTV (Gy)		Lung–CTV		Spinal cord	Heart	Esophagus
Variant	Dose engine	D_{95}	$D_{0.03\,\mathrm{cc}}\downarrow$	$V_{20\,\mathrm{Gy}}(\%)\downarrow$	Mean (Gy)↓	$D_{0.03\,\mathrm{cc}}$(Gy)↓	$D_{66.7\%}$ (Gy)↓	Mean (Gy)↓
PPO (outcome reward)	PyDoseRT	60.00 ± 0.00	83.00 ± 8.50	36.80 ± 11.00	18.70 ± 5.80	79.22 ± 30.15	3.90 ± 4.00	17.58 ± 5.20
Policy (clinical objective)	PyDoseRT	60.00 ± 0.00	73.00 ± 4.20	28.10 ± 8.30	14.40 ± 4.60	58.15 ± 25.40	2.70 ± 3.00	16.25 ± 3.90
PAC-VMAT (ours)	PyDoseRT	$\mathbf{60.00 \pm 0.00}$	$\mathbf{64.12 \pm 2.73}$	$\mathbf{19.20 \pm 9.05}$	$\mathbf{10.39 \pm 4.63}$	$\mathbf{36.37 \pm 14.13}$	$\mathbf{1.79 \pm 2.81}$	$\mathbf{15.16 \pm 9.49}$

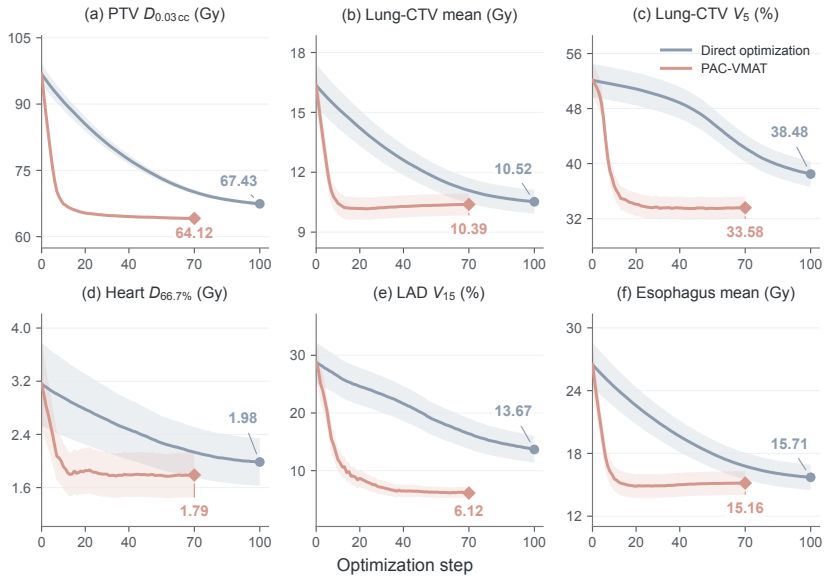

Figure 5: Clinical dose outcomes at matched target coverage (PTV $D_{95} = 60$ Gy). Curves show patient means at each optimization step, and shaded bands denote standard errors across patients.

objective and physical delivery constraints. While direct optimization updates individual machine variables using their local clinical sensitivities, PAC-VMAT trains Leaf and MU policy agents to predict these updates. Evaluating both trajectories at a matched target coverage isolates how effectively each formulation translates local clinical credit into spatial dose redistribution (Figure 5).

We find that coordinating machine updates through policy networks yields a superior target–OAR balance and accelerates convergence. PAC-VMAT reaches comparable lung, LAD, and esophageal dose outcomes to 100-step direct optimization in approximately 20 steps, reducing the estimated optimization runtime from 11.70 to 4.90 minutes on an H100 GPU (Appendix E.5). Beyond this computational efficiency, the policy agents achieve a lower final target hotspot (mean PTV $D_{0.03\,\mathrm{cc}}$ drops from 67.43 Gy to 64.12 Gy) and substantially contract intermediate dose spread, lowering the LAD $V_{15\,\mathrm{Gy}}$ from 13.67% to 6.12%. This simultaneous improvement in target homogeneity and organ sparing confirms that shared network parameters effectively link local clinical credit across the delivery sequence, enabling the agents to organize independent sensitivities into a unified, deliverable trajectory.

4.5 Effect of Anatomy-Guided Initialization

To evaluate the effect of anatomy-guided initialization, we compare refinement from a random feasible plan with refinement from an anatomy-guided initial plan, changing only the starting plan (Table 4). The advantage of anatomy-guided initialization persists after refinement. Starting from the anatomy-guided plan yields lower spinal-cord exposure than starting from the random plan. The remaining separation shows that subsequent refinement continues to depend on the initial geometry.

Table 4: Effect of initialization on plan quality at matched target coverage.

		PTV (Gy)		Lung–CTV		Spinal cord	Heart	Esophagus
Method	Dose engine	D_{95}	$D_{0.03\,\mathrm{cc}}\downarrow$	$V_{20\,\mathrm{Gy}}(\%)\downarrow$	Mean (Gy)↓	$D_{0.03\,\mathrm{cc}}$(Gy)↓	$D_{66.7\%}$ (Gy)↓	Mean (Gy)↓
Random plan	PyDoseRT	60.00 ± 0.00	145.20 ± 42.85	43.38 ± 23.97	24.49 ± 15.71	62.35 ± 40.06	6.18 ± 9.21	36.03 ± 24.32
+ PAC-VMAT	PyDoseRT	60.00 ± 0.00	89.20 ± 12.30	43.00 ± 13.80	22.30 ± 8.20	78.15 ± 32.40	4.10 ± 4.00	22.94 ± 6.80
Anatomy-guided initial plan	PyDoseRT	60.00 ± 0.00	96.80 ± 19.35	31.91 ± 18.99	16.35 ± 9.08	42.55 ± 23.56	3.16 ± 5.24	26.48 ± 17.94
+ PAC-VMAT	PyDoseRT	$\mathbf{60.00 \pm 0.00}$	$\mathbf{64.12 \pm 2.73}$	$\mathbf{19.20 \pm 9.05}$	$\mathbf{10.39 \pm 4.63}$	$\mathbf{36.37 \pm 14.13}$	$\mathbf{1.79 \pm 2.81}$	$\mathbf{15.16 \pm 9.49}$

5 Related Work

Automated VMAT planning maps patient anatomy to deliverable aperture and intensity trajectories. Existing methods differ in whether clinical criteria reach the machine variables and at what resolution. PAC-VMAT optimizes the evolving executable plan with explicit per-control clinical credit.

Planning through predicted dose or fluence. Knowledge-based planning predicts dose from anatomy (McIntosh et al., 2017; Babier et al., 2021), while inverse mapping predicts fluence from a desired dose (Ma et al., 2020; Arberet et al., 2025). Dose mimicking or learned leaf sequencing then converts these targets into machine settings (McIntosh et al., 2017; Gao et al., 2024). This commits the clinical trade-off to an intermediate representation before delivery: the realization stage follows the predicted target rather than receiving local clinical guidance on each leaf and MU update.

Direct prediction of machine parameters. Networks also predict apertures and MU directly from anatomy (Vandewinckele et al., 2023; Achlatis et al., 2025). These models learn from reference machine settings, sometimes with an additional dose-matching loss (Achlatis et al., 2025). One-pass pipelines can produce complete plans from CT and contours in seconds (Heilemann et al., 2025; Arberet et al., 2026). PAC-VMAT instead refines each patient's executable trajectory against its own clinical criteria, without reference-plan supervision.

Direct optimization of machine variables. Reinforcement-learning planners act directly on leaves and intensity using scalar dose-based feedback. A deep-Q agent controlled leaf pairs and dose rate in a simplified two-dimensional model (Hrinivich & Lee, 2020). Single- and dual-arc policies minimize plan costs (Hrinivich et al., 2024; Mekki et al., 2025), while tandem MLC and MU policies use stepwise changes in a DVH-based objective (Shaffer et al., 2026). These signals assign credit through sampled trajectories, without exposing the local clinical sensitivity of each machine control. Direct leaf-trajectory optimization incorporates deliverability into a dose objective (Papp & Unkelbach, 2014; Unkelbach et al., 2015), and differentiable dose engines expose gradients through the machine settings (Simkó et al., 2025). PAC-VMAT converts those sensitivities into fixed per-control credit targets, trains Leaf and MU policies with parameters shared across control points to coordinate the updates, and refreshes MU credit after the Leaf update. The refresh makes intensity guidance respond to the dose produced by the newly realized aperture. This jointly optimizes aperture geometry and beam intensity for each patient under mechanical constraints, connecting the clinical objective to the realized trajectory. Sections 4.3 and 4.4 compare this route with scalar-reward learning and direct gradient updates.

6 Conclusion

Directly optimizing radiotherapy trajectories requires coordinating thousands of tightly coupled machine variables to balance curative radiation and healthy tissue preservation. We find that separating physical credit construction from policy learning overcomes the severe credit assignment bottleneck inherent in this high-dimensional control space. By differentiating the clinical objective through a dose engine, global clinical goals are translated into explicit, local sensitivities for every machine parameter. Parameterized policy agents then organize these independent targets into continuous, mechanically feasible aperture and intensity updates. Results across lung and head-and-neck benchmarks demonstrate that this physics-aligned credit enables the networks to rapidly reshape the spatial dose distribution, reducing collateral organ exposure under strict mechanical constraints. Connecting local clinical sensitivities to structured policy updates establishes a direct and efficient formulation for mapping patient anatomy into executable treatment plans at the prescribed target coverage.

ETHICS STATEMENT

Automated radiotherapy planning directly intersects with patient safety and treatment efficacy. We conduct our evaluation using lung and head-and-neck benchmark data from the 2025 AAPM GDP-HMM Grand Challenge, derived from the NSCLC-Cetuximab, NSCLC-Radiomics, and Head-Neck Cetuximab collections in The Cancer Imaging Archive (TCIA), adhering strictly to the applicable privacy and data-use guidelines. While our results demonstrate explicit improvements in computational plan quality and mechanical deliverability, these generated trajectories reflect algorithmic capabilities rather than deployable medical protocols. Translating such automated plans into physical execution requires rigorous, patient-specific validation by qualified radiation oncology professionals to guarantee clinical safety.

REPRODUCIBILITY STATEMENT

We provide the complete formulation of physical credit construction and policy optimization within the main text to establish the theoretical foundation of our method. To support independent replication of the empirical outcomes, the appendices supply the comprehensive experimental configuration. We detail the clinical objectives, policy network architectures, and anatomy-guided initialization procedures necessary to construct the optimization environment. The exact evaluation cohort, baseline adaptations, and hyperparameter settings are explicitly documented alongside the strict mechanical feasibility constraints required to validate the generated delivery sequences.

AI USE STATEMENT

We utilized generative AI models to refine the language, improve readability, and assist with the LaTeX formatting of this manuscript. These tools additionally supported the structural presentation of our experimental discussions. All algorithmic designs, empirical evaluations, and data analyses were conducted exclusively by the human authors. We maintain full responsibility for the scientific claims, the accuracy of the reported results, and the final intellectual content of this work.

REFERENCES

Stefanos Achlatis, Efstratios Gavves, and Jan-Jakob Sonke. Physics-guided radiotherapy treatment planning with deep learning. *arXiv preprint arXiv:2506.19880*, 2025. URL https://arxiv.org/abs/2506.19880.

Hugo J. W. L. Aerts, Leonard Wee, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, André Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Data from NSCLC-Radiomics. The Cancer Imaging Archive, 2014. URL https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.

Simon Arberet, Florin C. Ghesu, Riqiang Gao, Martin Kraus, Jonathan Sackett, Esa Kuusela, and Ali Kamen. Beam's eye view to fluence maps 3D network for ultra fast VMAT radiotherapy planning. *Medical Physics*, 52(5):3183–3190, 2025. doi: 10.1002/mp.17673.

Simon Arberet, Riqiang Gao, Martin Kraus, Florin C. Ghesu, Wilko Verbakel, Mamadou Diallo, Anthony Magliari, Venkatesan Karuppusamy, Sushil Beriwal, REQUITE Consortium, Ali Kamen, and Dorin Comaniciu. AI end-to-end radiation treatment planning under one second. *arXiv preprint arXiv:2603.06338*, 2026. URL https://arxiv.org/abs/2603.06338.

Aaron Babier, Binghao Zhang, Rafid Mahmood, Kevin L. Moore, Thomas G. Purdie, Andrea L. McNiven, and Timothy C. Y. Chan. OpenKBP: The open-access knowledge-based planning grand challenge and dataset. *Medical Physics*, 48(9):5549–5561, 2021. doi: 10.1002/mp.14845.

Walter R. Bosch, William L. Straube, John W. Matthews, and James A. Purdy. Head-Neck Cetuximab [data set]. The Cancer Imaging Archive, 2015. URL https://doi.org/10.7937/K9/TCIA.2015.7AKGJUPZ.

Jeffrey Bradley and Ken Forster. Data from NSCLC-Cetuximab. The Cancer Imaging Archive, 2018. URL https://doi.org/10.7937/TCIA.2018.jze75u7v.

K. Bush, I. M. Gagne, S. Zavgorodni, W. Ansbacher, and W. Beckham. Dosimetric validation of Acuros XB with Monte Carlo methods for photon dose calculations. *Medical Physics*, 38(4): 2208–2221, 2011. doi: 10.1118/1.3567146.

Karl Bzdusek, Henrik Friberger, Kjell Eriksson, Björn Hårdemark, David Robinson, and Michael Kaus. Development and evaluation of an efficient approach to volumetric arc therapy planning. *Medical Physics*, 36(6):2328–2339, 2009. doi: 10.1118/1.3132234.

Antonella Fogliata, Giorgia Nicolini, Eugenio Vanetti, Alessandro Clivio, and Luca Cozzi. Dosimetric validation of the anisotropic analytical algorithm for photon dose calculation: fundamental characterization in water. *Physics in Medicine & Biology*, 51(6):1421–1438, 2006. doi: 10.1088/0031-9155/51/6/004.

Antonella Fogliata, Giorgia Nicolini, Alessandro Clivio, Eugenio Vanetti, Pietro Mancosu, and Luca Cozzi. Dosimetric validation of the Acuros XB advanced dose calculation algorithm: fundamental characterization in water. *Physics in Medicine & Biology*, 56(6):1879–1904, 2011. doi: 10.1088/0031-9155/56/6/022.

Riqiang Gao, Florin-Cristian Ghesu, Simon Arberet, Shahab Basiri, Esa Kuusela, Martin Kraus, Dorin Comaniciu, and Ali Kamen. Multi-agent reinforcement learning meets leaf sequencing in radiotherapy. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14723–14746. PMLR, 2024. URL https://proceedings.mlr.press/v235/gao24g.html.

Riqiang Gao, Mamadou Diallo, Han Liu, Anthony Magliari, Jonathan Sackett, Wilko Verbakel, Sandra Meyers, Rafe McBeth, Masoud Zarepisheh, Simon Arberet, Martin Kraus, Florin C. Ghesu, and Ali Kamen. Automating RT planning at scale: High quality data for AI training. *arXiv preprint arXiv:2501.11803*, 2025. URL https://arxiv.org/abs/2501.11803.

Gerd Heilemann, Lukas Zimmermann, Tufve Nyholm, Attila Simkó, Joachim Widder, Gregor Goldner, Dietmar Georg, and Peter Kuess. Ultra-fast, one-click radiotherapy treatment planning outside a treatment planning system. *Physics and Imaging in Radiation Oncology*, 33:100724, 2025. doi: 10.1016/j.phro.2025.100724.

William T. Hrinivich, Mahasweta Bhattacharya, Lina Mekki, Todd McNutt, Xun Jia, Heng Li, Daniel Y. Song, and Junghoon Lee. Clinical VMAT machine parameter optimization for localized prostate cancer using deep reinforcement learning. *Medical Physics*, 51(6):3972–3984, 2024. doi: 10.1002/mp.17100.

William Thomas Hrinivich and Junghoon Lee. Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning. *Medical Physics*, 47(12):6140–6150, 2020. doi: 10.1002/mp.14544.

Lin Ma, Mingli Chen, Xuejun Gu, and Weiguo Lu. Deep learning-based inverse mapping for fluence map prediction. *Physics in Medicine & Biology*, 65(23):235035, 2020. doi: 10.1088/1361-6560/abc12c.

Chris McIntosh, Mattea Welch, Andrea McNiven, David A. Jaffray, and Thomas G Purdie. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Physics in Medicine & Biology*, 62(15):5926–5944, 2017. doi: 10.1088/1361-6560/aa71f8.

Lina Mekki, William T. Hrinivich, and Junghoon Lee. Dual-arc VMAT machine parameter optimization for localized prostate cancer using deep reinforcement learning. *Physics in Medicine & Biology*, 70(22):225007, 2025. doi: 10.1088/1361-6560/ae1803.

NRG Oncology. RTOG 0522: A randomized phase III trial of concurrent accelerated radiation and cisplatin versus concurrent accelerated radiation, cisplatin, and cetuximab (C225) [followed by surgery for selected patients] for stage III and IV head and neck carcinomas, 2016. URL https://www.cancerimagingarchive.net/wp-content/uploads/

RTOG0522_Am10_Protocol_pv012616.pdf. Amendment 10, January 26, 2016, Sections 6.1.2.4 and 6.5.

Karl Otto. Volumetric modulated arc therapy: IMRT in a single gantry arc. *Medical Physics*, 35(1): 310–317, 2008. doi: 10.1118/1.2818738.

Dávid Papp and Jan Unkelbach. Direct leaf trajectory optimization for volumetric modulated arc therapy planning with sliding window delivery. *Medical Physics*, 41(1):011701, 2014. doi: 10.1118/1.4835435.

Radiation Therapy Oncology Group. RTOG 0617/NCCTG N0628/CALGB 30609/ECOG R0617: A randomized phase III comparison of standard-dose (60 Gy) versus high-dose (74 Gy) conformal radiotherapy with concurrent and consolidation carboplatin/paclitaxel with or without cetuximab in patients with stage IIIA/IIIB non-small cell lung cancer, 2013. URL https://www.nrgoncology.org/Clinical-Trials/Protocol/rtog-0617/. Protocol version date April 4, 2013, Sections 6.1.5 and 6.5.

Nathan Shaffer, Avinash Reddy Mudireddy, and Joel St-Aubin. A tandem reinforcement learning framework for localized prostate cancer treatment planning and machine parameter optimization. *Medical Physics*, 53(2):e70306, 2026. doi: 10.1002/mp.70306.

Eric Simiele, Ignacio O. Romero, Jen-Yeu Wang, Yizheng Chen, Yuliia Lozko, Yuliia Severyn, Lawrie Skinner, Yong Yang, Lei Xing, Iris Gibbs, Susan M. Hiniker, and Nataliya Kovalchuk. Automated contouring, treatment planning, and quality assurance for VMAT craniospinal irradiation (VMAT-CSI). *Frontiers in Oncology*, 14:1378449, 2024. doi: 10.3389/fonc.2024.1378449.

Attila Simkó, Matthias Kronsteiner, Simon Glatzer, Minh Vu, Josef A. Lundman, Joakim Jonsson, Jörgen Olofsson, Kristina Sandgren, Wolfgang Lechner, Dietmar Georg, Tommy Löfstedt, Tufve Nyholm, Anders Garpebring, and Gerd Heilemann. A physics-informed, plug-and-play dose engine for gradient-based radiotherapy treatment planning. *arXiv preprint arXiv:2512.18863*, 2025. URL https://arxiv.org/abs/2512.18863.

W. Ulmer, J. Pyyry, and W. Kaissl. A 3D photon superposition/convolution algorithm and its foundation on results of Monte Carlo calculations. *Physics in Medicine & Biology*, 50(8):1767–1790, 2005. doi: 10.1088/0031-9155/50/8/010.

Jan Unkelbach, Thomas Bortfeld, David Craft, Markus Alber, Mark Bangert, Rasmus Bokrantz, Danny Chen, Ruijiang Li, Lei Xing, Chunhua Men, Simeon Nill, Dávid Papp, Edwin Romeijn, and Ehsan Salari. Optimization approaches to volumetric modulated arc therapy planning. *Medical Physics*, 42(3):1367–1377, 2015. doi: 10.1118/1.4908224.

L. Vandewinckele, T. Reynders, C. Weltens, F. Maes, and W. Crijns. Deep learning based MLC aperture and monitor unit prediction as a warm start for breast VMAT optimisation. *Physics in Medicine & Biology*, 68(22):225013, 2023. doi: 10.1088/1361-6560/ad07f6.

Qiong Zhou, Liwen Qian, Chong Shen, Xinyan Bei, Gaojie Liu, and Xiaonan Sun. Automated VMAT planning for short-course radiotherapy in locally advanced rectal cancer. *PLOS ONE*, 20 (6):e0325567, 2025. doi: 10.1371/journal.pone.0325567.

A Clinical Objective

A.1 Formulation of the Clinical Reward

The clinical cost combines target-dose uniformity with continuous penalties for OAR dose violations. The same cost supplies the local sensitivities used by PAC-VMAT and the objective used by direct machine-variable optimization.

For an OAR containing N voxels with doses v, let $v^{\downarrow}$ denote the doses sorted in descending order and $[x]_+ = \max(x, 0)$. The upper-tail and mean-dose penalties are

$$\begin{aligned} \mathcal{U}(v; k, L) &= \frac{1}{N-k} \sum_{i=k+1}^{N} [v_i^{\downarrow} - L]_+, \\ \mathcal{M}(v; L) &= [N^{-1} \textstyle\sum_i v_i - L]_+. \end{aligned} \tag{9}$$

The parameter k determines the number of voxels permitted above the dose threshold L, with $0 \le k < N$. The first k entries account for this permitted volume. Every threshold-violating voxel among the remaining entries contributes to the cost and its local dose gradient. For a volume-at-dose limit of $p\%$, $k = \lfloor pN/100 \rfloor$. For a small-volume dose limit at 0.03 cc, $k = \lceil 0.03/v_{\text{voxel}} \rceil - 1$, where v_{voxel} is the voxel volume in cc.

The target term is the mean squared relative deviation from the prescription dose D_{rx} over all PTV voxels, with $D_{\text{rx}} = 60$ Gy for the lung cohorts and 70 Gy for the head-and-neck cohort. With $\mathcal{E}_{\text{OAR}}(s)$ denoting the OAR penalty terms, the implemented cost is

$$\mathcal{C}(d) = \frac{1}{N_{\text{PTV}}} \sum_{i \in \text{PTV}} \left(\frac{d_i}{D_{\text{rx}}} - 1 \right)^2 + \sum_{j \in \mathcal{E}_{\text{OAR}}(s)} w_j^{\text{OAR}} \frac{P_j(v_{S_j})}{s_j}. \tag{10}$$

Each P_j is an upper-tail or mean-dose penalty from Equation 9, weighted by its clinical priority w_j^{OAR}. Dose-at-volume and mean-dose penalties use $s_j = L$. Volume-at-dose penalties scale the residual by the prescription dose $s_j = D_{\text{rx}}$. The reward is $r(d) = -\mathcal{C}(d)$, with the target and organ terms corresponding to Equation 3. The target residual separates into cold and hot components, providing dose-increase and dose-decrease sensitivities around the prescription. The objective is evaluated at matched PTV $D_{95} = 60$ Gy for the comparison in Figure 5.

A.2 Optimization Settings and Reporting Metrics

The lung optimization uses the released Lung–GTV structure with $V_{20\,\text{Gy}} \le 37\%$, $V_{5\,\text{Gy}} \le 65\%$, and mean dose ≤ 20 Gy. Mean-dose penalties also use 20 Gy for the heart and 34 Gy for the esophagus. Small-volume penalties use $D_{0.03\,\text{cc}}$ thresholds of 45 Gy for the spinal cord and 60 Gy for the heart and esophagus. LAD $V_{15\,\text{Gy}}$ uses a 10% threshold. These nine OAR penalties enter the cost with unit weight after the normalization above.

Lung $V_{5\,\text{Gy}}$, LAD $V_{15\,\text{Gy}}$, heart mean dose, and the heart and esophageal small-volume thresholds are study-specific optimization settings. The 45 Gy spinal-cord optimization threshold provides a margin below the protocol's 50.5 Gy maximum-dose limit (Appendix D.2). Target uniformity is optimized around 60 Gy, and target maximum dose is evaluated against the 72 Gy protocol limit. The reported lung outcomes use Lung–CTV, and the reported heart dose endpoint is $D_{66.7\%}$.

The head-and-neck evaluation uses the same objective structure, with PTV70 as the target and $D_{\text{rx}} = 70$ Gy. PAC-VMAT and the reproduced baselines optimize the identical head-and-neck clinical cost with organ penalties that include the endpoints reported in Table 2. The spinal-cord penalty receives a weight of 10, while all remaining target and organ terms enter the cost with unit weight.

B Anatomy-Guided Initialization Details

Anatomy-guided initialization supplies target-aligned aperture geometry before clinical refinement begins. PAC-VMAT constructs an initial deliverable plan P_0 that maps the patient's anatomical

geometry into the configured machine constraints. The isocenter is placed at the volumetric centroid of the PTV, with the delivery geometry and machine constraints specified in Appendix F. Establishing a common coordinate frame between the patient's computed tomography (CT) volume and the machine trajectory allows the initial aperture shapes to be derived analytically from the target anatomy.

At each control point, the multileaf collimator (MLC) positions follow the beam's-eye-view (BEV) projection of the PTV. For a PTV voxel with patient-space displacement $\Delta = (\Delta_x, \Delta_y, \Delta_z)$ from the isocenter, its BEV projection at gantry angle θ and collimator angle ϕ is given by

$$\begin{aligned} h_\theta &= \Delta_x \cos\theta + \Delta_y \sin\theta, \\ x_{\mathrm{BEV}} = h_\theta \cos\phi + \Delta_z \sin\phi, &\qquad y_{\mathrm{BEV}} = -h_\theta \sin\phi + \Delta_z \cos\phi. \end{aligned} \tag{11}$$

The physical leaf pairs align with the y_{BEV} axis. For each active pair, the initializer bounds the projected target points and applies a 5-mm margin along the x_{BEV} axis to account for physical penumbra effects.

The aperture sequence defines the spatial field, and the initial intensity allocation establishes a baseline distribution of radiation along the trajectory. A uniform sequence of monitor unit (MU) increments yields equal delivery across successive angular intervals. We use the prescribed target volume (PTV for lung and PTV70 for head-and-neck) to define the coverage measure $q(d) = D_{95}(d)$ and set $q_\star = D_{\mathrm{rx}}$ in Equation 8. The dose engine evaluates the unscaled target coverage, after which all MU increments receive the common scaling coefficient γ_0 defined there. This scaling anchors the initial plan P_0 to the prescribed target coverage (60 Gy for lung or 70 Gy for head-and-neck) without altering the conformal aperture geometry or relative angular intensity profile. The subsequent optimization trajectory refines OAR sparing and target dose homogeneity from this clinically aligned state.

C Policy Agent Architecture and Clinical Credit

C.1 Policy Agent Networks

In the default PAC-VMAT configuration, both the Leaf and MU policy networks are initialized from scratch for each patient, framing treatment planning as patient-specific optimization. The policy agents act on the control-point-aligned observations described in Section 3.3. The MU policy agent receives the post-Leaf state, including the realized leaf-position changes and corresponding intermediate dose changes, so that its intensity decisions respond to the updated aperture geometry.

The Leaf policy agent uses a one-dimensional convolutional backbone along the control-point axis to aggregate local angular context within each arc. A residual depthwise mixing block combines features across neighboring leaf pairs and control points. Bounded continuous outputs specify leaf-position adjustments.

The MU policy agent modulates beam intensity over the geometry established by the Leaf policy agent using a separate one-dimensional convolutional backbone on the updated state sequence. Its architecture separates the local intensity profile across control points from the total radiation allocation of each arc. A convolutional head predicts the relative intensity profile within each arc, and arc-specific multilayer perceptrons predict the total monitor unit allocation from sequence-level features. This decoupled parameterization maintains explicit control over both within-arc intensity redistribution and arc-level dose allocation when producing a deliverable intensity sequence.

C.2 Clinical Credit Targets and Policy Learning

The physics-aligned clinical credit in Equation 4 provides local clinical sensitivities with respect to the physical machine controls, and Equation 5 converts them into the credit target z_t^p in the action coordinates of each policy agent. The dose-space gradient has already been translated into machine-variable feedback through the dose Jacobian; the second transformation accounts for the policy agent's action parameterization. The machine constraints in Appendix F govern the realized perturbation used in the subsequent dose calculation.

For aperture geometry, the credit follows the leaf-position controls along the delivery trajectory. For beam intensity, the action coordinates distinguish the relative intensity profile within each arc from

its total MU allocation. The profile component removes the common-intensity direction to guide redistribution among delivered angular intervals, while arc-total credit guides changes in the overall radiation allocated to each arc. Both express the local sensitivity of the same clinical objective to the corresponding machine controls.

The normalization in Equation 5 applies a common positive scale within each control group, preserving the relative magnitudes and directions within that group. The numerical floor ϵ bounds the scaling denominator away from zero, limiting the amplification of near-zero gradients. The target remains fixed during each policy update, tying the learned response to the physical state at which the clinical sensitivity was evaluated.

The objective in Equation 7 combines the regression term in Equation 6 with the structural regularizer in Equation 12. The regression term is implemented as a sensitivity-weighted Huber loss over normalized credit, with the fitting weight, coordinate-weight baseline, and cap listed in Table 9; the mean sensitivity in the coordinate-weight denominator is floored at the value given there.

Fitting the credit alone is not sufficient: independent coordinate-wise responses would violate the mechanical limits and break the continuity that delivery requires. The regularizer therefore constrains how the proposed update varies across controls:

$$\begin{aligned}\mathcal{R}_p(u_t^p, z_t^p) = \lambda_{\text{dir}}\Big[1 - \cos(u_t^p, z_t^p) + \frac{1}{|\mathcal{G}^{\text{pair}}|}\sum_P \Big(1 - \cos\Big(u_{t,P}^p, z_{t,P}^p\Big)\Big) \\ + \frac{1}{|\mathcal{G}^{\text{arc}}|}\sum_A \Big(1 - \cos\Big(u_{t,A}^p, z_{t,A}^p\Big)\Big)\Big] \\ + \lambda_{\text{inact}}\frac{1}{|\mathcal{I}_t^p|}\sum_{j\in\mathcal{I}_t^p}\big(u_{t,j}^p\big)^2 + \lambda_{\text{mag}}\left\|u_t^p\right\|_2^2 + \lambda_{\text{sm}}\left\|\Delta u_t^p\right\|_2^2 \\ + \lambda_{\text{curv}}\left\|\Delta^2 u_t^p\right\|_2^2 + \lambda_{\text{pair}}\mathbb{1}_{\{p=\text{Leaf}\}}\left\|\Delta_{\text{pair}} u_t^p\right\|_2^2,\end{aligned} \tag{12}$$

The regularizer $\mathcal{R}_p$ uses the set $\mathcal{I}_t^p = \{j : |z_{t,j}^p| < \tau_{\text{inact}}\}$ to identify inactive coordinates where the absolute normalized credit falls below the activity threshold. By penalizing mean squared output on these coordinates, it suppresses motion with negligible clinical guidance while aligning updates with clinical credit at the global, leaf-pair, and arc levels. The restricted vectors $u_{t,P}^p$ and $u_{t,A}^p$ describe updates within leaf pair P and arc A. Directional agreement is averaged over target vectors whose norms exceed the cosine target threshold. The operators Δ and Δ^2 measure first and second differences along the control-point axis, and Δ_{pair} measures first differences between adjacent leaf pairs. Their penalties limit roughness and curvature across the delivery trajectory, while the magnitude penalty bounds the overall update. Table 9 specifies all weights and thresholds.

D Experimental Setup and Evaluation

D.1 Datasets and Evaluation Cohorts

The evaluated cases are drawn from the 2025 AAPM GDP-HMM Grand Challenge dataset (Gao et al., 2025). The patient imaging and available clinical target data originate from three collections in The Cancer Imaging Archive (TCIA): NSCLC-Cetuximab (RTOG 0617) (Bradley & Forster, 2018), NSCLC-Radiomics (LUNG1) (Aerts et al., 2014), and Head-Neck Cetuximab (RTOG 0522) (Bosch et al., 2015). The GDP-HMM pipeline uses these data to generate the released reference RTPLAN and RTDOSE files with automated optimization in Eclipse (Gao et al., 2025). Thus, the source collections identify the patients and anatomy, whereas the GDP-HMM release supplies the reference treatment plans used here. The RTOG 0617 and RTOG 0522 protocols define the dose criteria used to evaluate these plans.

Lung VMAT. Lung VMAT evaluates all 70 patients in the benchmark's lung validation cohort with 180 control points per arc. This configuration optimizes dual-arc trajectories and is evaluated at the prescribed PTV $D_{95} = 60$ Gy. Table 5 summarizes anatomical variation in the Lung VMAT cohort.

Head-and-neck. The official GDP-HMM head-and-neck validation split contains 77 patients. We retain 67 RTOG 0522 cases from the Head-Neck Cetuximab collection that have a VMAT plan and

the target and organ structures needed for evaluation under the RTOG 0522 criteria; the remaining ten are excluded by these eligibility requirements. These are distinct patients from the lung cohort. Their plans use four arcs with 180 control points per arc and a PTV70 prescription of 70 Gy in 35 fractions. Table 2 evaluates target and organ endpoints at PTV70 $D_{95} = 70$ Gy using criteria adapted from the RTOG 0522 protocol (NRG Oncology, 2016). Available plan geometry and target and organ structures define eligibility, while the RTOG 0522 protocol defines the dose criteria.

PAC-VMAT initializes and optimizes separate Leaf and MU policy agents for each patient from the CT volume, anatomical masks, and prescription criteria.

Table 5: Anatomical variation in the lung VMAT benchmark. Targets vary widely in volume and frequently contact the lung, heart, or esophagus.

Structure	Volume, median [IQR] (cm^3)	PTV distance, median [IQR] (mm)
PTV	275.1 [93.0, 426.9]	–
Total lung–CTV	3325.6 [2850.8, 4207.7]	0.0 [0.0, 0.0]
Spinal cord	10.6 [7.9, 12.4]	23.0 [13.1, 37.2]
Heart	628.6 [491.2, 790.1]	0.0 [0.0, 10.0]
LAD	4.3 [3.7, 4.8]	29.7 [22.2, 42.9]
Esophagus	21.1 [16.9, 26.1]	2.0 [0.0, 15.5]

D.2 RTOG 0617 Reference Criteria

The evaluation uses dose criteria from the April 4, 2013 version of RTOG 0617 (Radiation Therapy Oncology Group, 2013), with a common prescription of 60 Gy. Section 6.1.5 requires prescription-dose coverage of 95% of the PTV and limits maximum target dose to 120% of the prescription. These give $D_{95} \geq 60$ Gy and $D_{0.03\,\mathrm{cc}} \leq 72$ Gy for the target endpoints in Table 1. The protocol defines target maximum and minimum point doses over 0.03 cc and also requires a minimum target dose of at least 95% of the prescription, or 57 Gy. Table 1 reports target coverage and maximum dose as the selected target endpoints.

Section 6.5.1 gives the spinal cord highest priority, limiting its total direct-plus-scatter maximum dose to 50.5 Gy. We retain that threshold and report $D_{0.03\,\mathrm{cc}}$ as a near-maximum dose measure; this small-volume reporting convention is an adaptation of the protocol's spinal-cord maximum-dose criterion. Section 6.5.2 defines lung volume as both lungs with CTV excluded. It gives $V_{20\,\mathrm{Gy}} \leq 37\%$ as the lung constraint, subject to spinal-cord priority, and mean lung dose ≤ 20 Gy as an alternative recommended target. Section 6.5.3 strongly recommends an esophageal mean dose below 34 Gy and requests calculation of esophageal $V_{60\,\mathrm{Gy}}$ without assigning it a numerical limit. The esophageal endpoint in the result tables is mean dose.

Section 6.5.4 recommends heart dose-volume limits of 60 Gy to less than one-third, 45 Gy to less than two-thirds, and 40 Gy to less than the entire heart volume. We select the two-thirds criterion and report it as $D_{66.7\%} < 45$ Gy, with 66.7% representing the rounded volume fraction. The lung mean-dose, esophageal mean-dose, and heart criteria are recommendation-based reference values. Lung $V_{5\,\mathrm{Gy}}$ and LAD $V_{15\,\mathrm{Gy}}$ in Figure 5 describe additional dose outcomes associated with the optimization settings in Appendix A; their thresholds originate from those study settings.

D.3 Dose Normalization and Evaluation

The GDP-HMM Eclipse plans serve as the reference. Evaluating the spatial coordination of generated plans requires separating their structural dose pattern from absolute intensity scaling. Each plan is therefore scaled to its cohort's prescribed target coverage before its clinical metrics are computed with PyDoseRT, the dose engine also used for optimization: $\widehat{d} = 60\,d/D_{95}(d)$ for lung and $\widehat{d} = 70\,d/D_{95}^{\mathrm{PTV70}}(d)$ for head-and-neck, with dose measured in Gy. Metrics are computed from $\widehat{d}$ before aggregation. The lung evaluation mask is the total-lung mask with the CTV removed, following Section 6.5.2 of RTOG 0617. Lung PTV $D_{95} = 60$ Gy and head-and-neck PTV70 $D_{95} = 70$ Gy are the respective reporting conditions; improvements are measured by target hotspot and OAR exposure.

The lung VMAT Eclipse results in Table 8 and the Proprietary plans row in Table 1 describe the same benchmark delivery under the two dose calculations. PyDoseRT recalculation metrics are

available for every lung validation patient, with no missing cases. The recalculation uses each patient's CT, anatomical structure masks, and released delivery parameters without reoptimizing the plan. The Proprietary plans, PAC-VMAT, and Random plan rows use the same plan-wise prescribed-coverage reporting dose defined above. For each reference plan, all metrics, including lung $V_{20\,\mathrm{Gy}}$, are recomputed from the scaled PyDoseRT voxel dose before aggregation. The lung VMAT Eclipse reference row in Table 8 retains approximate group-level scaling of dose means and standard deviations to mean PTV $D_{95} = 60$ Gy; its Lung–CTV $V_{20\,\mathrm{Gy}}$ is computed from dose scaled by the same cohort factor before thresholding. Its PTV $D_{0.03\,\mathrm{cc}}$, Heart $D_{66.7\%}$, and Lung–CTV metrics are recomputed from the released two-arc reference doses and PTV, heart, and lung masks using the stored dose scaling and voxel spacing. For dose-valued metrics, the cohort mean and sample standard deviation are multiplied by $60/\overline{D_{95}}$, with $\overline{D_{95}} = 60.1023$ Gy. The remaining Eclipse entries retain their existing reference summaries.

Dose metrics are computed from the structure's voxel doses before patient-level aggregation. PTV D_{95} is the $\lceil 0.95 N_{\mathrm{PTV}} \rceil$th hottest target voxel dose. The small-volume metric $D_{0.03\,\mathrm{cc}}$ is the $\lceil 0.03/v_{\mathrm{voxel}} \rceil$th hottest voxel dose in the relevant structure. Heart $D_{66.7\%}$ is the $\lceil 0.667 N_{\mathrm{heart}} \rceil$th hottest heart voxel dose. Lung–CTV is constructed as the total-lung mask intersected with the complement of the CTV mask. Mean dose is the arithmetic mean within the structure; $V_{x\,\mathrm{Gy}}$ is the percentage of its voxels receiving at least x Gy. Dose-valued endpoints are reported in Gy and volume-at-dose endpoints in percent.

Patient-level dose metrics are aggregated across the evaluated cohorts to assess overall plan quality. To isolate the spatial dose redistribution achieved during refinement, intermediate lung plans and final trajectories are evaluated at matched target coverage (PTV $D_{95} = 60$ Gy), with PTV70 $D_{95} = 70$ Gy used for head-and-neck evaluation. This normalization ensures that optimization dynamics and cross-method comparisons reflect changes in dose distribution at fixed target coverage rather than arbitrary intensity scaling. The PTV, lung, and heart trajectories in Figure 5 are evaluated from the machine state at each optimization step under this normalization.

Table 6: Aggregation of patient-level dose metrics. Values are reported as mean ± standard deviation.

Summary	Standard deviation
PAC-VMAT, direct optimization, initialized and random plans, Eclipse, and reproduced baselines	Sample
Proprietary plans (Lung VMAT only)	Population

D.4 Baseline Implementations

We re-implement the Gao and Shaffer baselines based on their published formulations to enable patient-specific optimization from identical Conformal arc initializations.

Inputs and shared setup. The three baselines use each patient's CT, structure contours, and prescription. Released RTPLANs define the gantry angles, collimator settings, and leaf boundaries, while anatomy determines the initial apertures and MU. PyDoseRT evaluates dose on a uniform 3-mm grid. The computational volume follows the transverse body boundary and includes both the target extent and an isocenter-centered axial interval, ensuring that initialization, optimization, and reporting share the same grid and masks. Table 7 specifies the spatial and optimization settings. Lung plans use two arcs with 180 control points per arc and 40 active leaf pairs, with 60 Gy prescribed in 30 fractions. Head-and-neck plans use four arcs with 180 control points per arc and 60 active leaf pairs, with 70 Gy prescribed in 35 fractions. Mechanical constraints follow Appendix F.

Both learned baselines are optimized separately for each patient. Their lung objective follows Appendix A. The head-and-neck objective combines mean squared relative PTV70 dose error with relative threshold-violation penalties for organ endpoints, including those reported in Table 2. The spinal-cord penalty has weight 10 and all remaining terms have unit weight. Both objectives are evaluated at prescribed target coverage.

Conformal arc. The common initial plan follows the anatomy-guided construction in Appendix B. Its isocenter is the PTV centroid, and apertures cover the PTV's BEV projection with a 5-mm leaf-

direction margin. Each arc initially distributes 100 MU uniformly over delivery intervals. A common MU scale establishes PTV $D_{95} = 60$ Gy for lung or PTV70 $D_{95} = 70$ Gy for head-and-neck. This plan is evaluated directly as Conformal arc and supplies the unchanged starting point for both learned baselines.

Gao adaptation. Following the fluence-to-delivery formulation of Gao et al. (2024), patient-specific fluence optimization provides the target for a multi-agent leaf sequencer. Each leaf agent observes the target and accumulated fluence rows, leaf positions, previous MU, and sequencing position. Discrete leaf displacements reshape the aperture, followed by MU prediction. Shared multilayer perceptrons and the five-component sequencing reward coordinate these decisions. Table 7 gives the fluence discretization, action range, network width, and training parameters.

Shaffer adaptation. The tandem structure of Shaffer et al. (2026) alternates a leaf-update pass across all control points with global MU refinement. The Leaf policy combines BEV-aligned dose, PTV, and OAR channels with local machine parameters, while the MU policy combines global channels with the full plan parameters. Convolutional and attention encoders use GroupNorm for patient-specific small-batch training. Bounded leaf increments and logarithmic intensity updates define the action space, and the decrease in clinical cost supplies the scalar reward. Both networks start from random weights and refine the common Conformal arc plan without reference-plan behavioral cloning.

Table 7: Spatial discretization and training settings for the reproduced baselines.

Method	Setting	Value
Shared	Dose grid / transverse extent	3 mm / body bounding box
	Axial extent	Union of 256-mm isocenter interval and PTV extent with 8-voxel margin
	Random seed / discount / GAE parameter	0 / 0.99 / 0.95
	PPO learning rate / epochs	10^{-4} / 2
	Improvement threshold over initialization	0.1%
Gao	Fluence bins per leaf row / sector size	100 / at most 16 intervals
	Fluence optimizer / updates / learning rate	Adam / 12 / 0.05
	Leaf displacement actions	$-4, \ldots, 4$ bins (9 choices)
	Displacement bin width	X-jaw opening / 100
	Shared MLP width	512
	Sequencing reward weights	$(1, 2, 2, 1, 1)$
	PPO clipping / minibatch size	0.2 / 96
	Complete sequencing rollouts (min–max)	4–32
	Non-improving episode patience	12
Shaffer	Global MU updates per leaf-update pass	8
	Leaf input channel dimensions	$48 \times 48 \times 128$
	MU input channel dimensions	$48 \times 128 \times 128$
	Leaf increment bound before projection	1 mm
	MU-profile / arc-total log-step sizes	0.1 / 0.05
	PPO clipping / minibatch size / KL threshold	0.15 / 2 / 0.03
	Complete Leaf–MU episodes (min–max)	2–8
	Non-improving episode patience	3

Plan selection and evaluation. Both learned baselines are optimized using Proximal Policy Optimization (PPO), with termination governed by improvement and patience criteria after a minimum training budget, or by exhaustion of the maximum budget (Table 7). Each baseline retains the plan yielding the lowest clinical cost among the common initialization and complete episode endpoints. Within each reproduced baseline comparison, the learned methods refine the same anatomy-guided starting plan in a unified dose evaluation environment. Patient-level outcomes are evaluated at prescribed target coverage.

D.5 Original Benchmark Reference Dose Results

The GDP-HMM benchmark combines clinical guidelines with automated Eclipse optimization to generate reference plans of quality comparable to manual planning (Gao et al., 2025). The original lung doses achieve cohort-mean target and organ metrics within the listed protocol thresholds, establishing the reference balance between target coverage and organ sparing (Table 8).

Recalculating the same deliveries in PyDoseRT changes their reported dose distributions. The lung VMAT mean PTV $D_{0.03\,\mathrm{cc}}$ is 63.38 Gy in the original Eclipse summary and 112.77 Gy after recalculation under the reporting conventions in Appendix D.3. Tissue heterogeneity modeling and beam commissioning are plausible contributors to this difference (Bush et al., 2011; Simkó et al., 2025).

Table 8: Original benchmark reference dose results. Panels group cohorts by their reported target and organ metrics. Values are mean ± standard deviation; the lung VMAT Eclipse normalization convention is described in Appendix D.3.

(a) Lung cohort

Cohort	Dose source	PTV		Lung–CTV		Spinal cord	Esophagus	Heart
		D_{95} (Gy)	$D_{0.03\,\mathrm{cc}}$ (Gy) ↓	$V_{20\,\mathrm{Gy}}$ (%) ↓	Mean (Gy) ↓	$D_{0.03\,\mathrm{cc}}$ (Gy) ↓	Mean (Gy) ↓	$D_{66.7\%}$ (Gy) ↓
Criterion	–	≥ 60	≤ 72	≤ 37	≤ 20	≤ 50.5	< 34	< 45
Lung VMAT	Eclipse	60.00 ± 0.23	63.38 ± 0.99	20.64 ± 10.64	11.79 ± 5.01	41.61 ± 26.09	19.73 ± 4.20	2.44 ± 2.75

(b) Head-and-neck cohort

Cohort	Dose source	PTV70	Spinal cord	Larynx	Brachial plexus
		D_{95} (Gy)	$D_{0.03\,\mathrm{cc}}$ (Gy) ↓	$D_{\max}$ (Gy) ↓	$D_{\max}$ (Gy) ↓
Criterion	–	≥ 70	≤ 48	< 45	≤ 60
Head-and-neck	RTDOSE	70.00 ± 0.00	29.91 ± 3.63	73.70 ± 2.63	68.68 ± 5.73

D.6 Dose Calculation across Engines

Low-density lung tissue increases the range of secondary electrons and disrupts local charged-particle equilibrium near tumor–lung interfaces. Dose algorithms that represent this transport differently can therefore predict distinct spatial gradients for identical machine apertures. Eclipse supports convolution/superposition through AAA and transport-equation solutions through Acuros XB (Ulmer et al., 2005; Fogliata et al., 2006; 2011). The GDP-HMM publication identifies Eclipse but does not report the algorithm selected for its reference doses (Gao et al., 2025). In heterogeneous-phantom tests, AAA discrepancies from Monte Carlo increased in low-density lung, while Acuros XB remained closer to the Monte Carlo calculation (Bush et al., 2011).

PyDoseRT constructs aperture fluence, projects it into the patient, convolves it with depth-dependent scatter kernels, and accumulates MU-weighted contributions (Simkó et al., 2025). Its pencil-beam formulation estimates water-equivalent depth along a central ray and approximates lateral scatter with two-dimensional kernels. This structure enables repeated dose calculation and differentiation, with oblique tissue boundaries and lateral density variation placing demands on the approximation. The published validation focuses on prostate plans. CT-to-density conversion, output factors, source penumbra, and MLC transmission also enter the independently commissioned beam model. Because tissue heterogeneity modeling produces distinct spatial dose gradients for identical machine apertures, comparisons across dose engines combine optimization quality with calculation physics. We therefore recalculate reference deliveries in PyDoseRT. This establishes a shared physical environment for evaluating spatial dose redistribution under a common dose model.

On the lung VMAT benchmark, PAC-VMAT achieves reported target–OAR metrics comparable to the original Eclipse reference, with lower mean spinal-cord and esophageal exposures (Tables 1 and 8). The policy agents use explicit clinical credit to coordinate aperture and intensity updates. The resulting dose metrics support the potential of this open-source planning approach to achieve competitive target coverage and organ sparing.

D.7 Hyperparameters and Compute Configuration

PAC-VMAT trains the Leaf and MU policy agents to align their proposed machine updates with clinical credit targets recomputed from the evolving plan.

For the plan-quality comparison in Figure 5, PAC-VMAT uses 70 complete Leaf–MU refinement steps and direct constrained optimization uses 100 joint updates. The policy networks use distinct coordinate-group step sizes to accommodate the different clinical sensitivities of aperture reshaping, MU-profile redistribution, and common-MU scaling (Table 9). Repeated three-dimensional dose evaluations account for most of the runtime, which varies with anatomical volume and geometry.

Table 9: Optimization settings. PAC-VMAT uses the clinical credit loss in Equation 6 and the structural regularization in Equation 12. CP denotes control point.

Group	Setting	Value
Networks	Leaf / MU hidden channels	128 / 96
	Leaf / MU trainable parameters	481,624 / 213,875
	Leaf action coordinates (Lung VMAT)	$40 \times 360 \times 2 = 28{,}800$
	MU action coordinates (Lung VMAT)	$360 + 2 = 362$
Policy update	Coordinate-group RMS step	0.25
	Backbone RMS step	0.003
	Policy-output head RMS step	0.03
Clinical credit loss	Huber transition h	unit transition
	Regression weight λ_{fit}	1.0
	Coordinate-weight baseline $w_{\min}$ / cap $w_{\max}$	0.1 / 20
	Mean-sensitivity floor ϵ_w	10^{-6}
	Inactive-coordinate weight λ_{inact}	0.1
	Inactive threshold τ_{inact}	10^{-4}
Structural regularization	Direction weights λ_{dir}: global / pair / arc	0.1 for all three levels
	Leaf $\lambda_{\mathrm{mag}}/\lambda_{\mathrm{sm}}/\lambda_{\mathrm{pair}}/\lambda_{\mathrm{curv}}$	$0.001/0.002/0.001/0.001$
	MU $\lambda_{\mathrm{mag}}/\lambda_{\mathrm{sm}}/\lambda_{\mathrm{curv}}$	$0.001/0.002/0.001$
Numerical settings	Credit-target norm floor ϵ	10^{-12}
	Cosine target threshold / norm floor	10^{-8} / 10^{-8}
Compute	Policy–direct comparison steps	70 / 100
	Accelerator for runtime comparison	1 NVIDIA H100 GPU
	PAC-VMAT estimated runtime (20 steps)	4.90 minutes (mean)
	Direct optimization runtime at 100 steps	11.70 minutes (mean)

Runtime is reported on an NVIDIA H100 GPU and covers the optimizer call, including dose calculation and optimization updates.

D.8 Optimization Comparison Protocols

Direct optimization without policy agents. Direct optimization updates Leaf and MU jointly from the incoming machine state. Starting from identical plans and anatomy under the same target–OAR objective and machine constraints, PAC-VMAT separates the two updates through sequential clinical credit evaluation. PAC-VMAT fits the two policy agents to clinical credit targets, accepts the Leaf update, and recomputes the MU gradient at the resulting state before independently accepting the MU update. Both methods use the full target–OAR objective for gradient construction, action acceptance, and final plan selection. The direct baseline runs 100 joint updates; PAC-VMAT runs 70 complete Leaf–MU steps. Evaluating both procedures at PTV $D_{95} = 60$ Gy isolates their ability to redistribute dose at matched target coverage (Section 4.4).

Direct clinical-objective training. The clinical-objective baseline evaluates how explicit credit construction changes policy learning. It retains the dual-policy architecture, observations, initialization, Leaf–MU execution order, and target-independent regularization, while training the policy agents end to end against the clinical reward differentiated through the dose engine. This comparison tests the complete formulation of policy learning with clinical credit, in which normalized local targets, sensitivity weighting, and target-dependent regularization organize the local clinical sensitivities into learning targets for the policy agents.

E Additional Results and Mechanistic Analysis

E.1 Policy Parameterization and Update Sequence

To examine how the organization of machine updates affects policy learning with clinical credit, Table 10 compares a joint policy agent with sequential updates, separate policy agents acting in parallel, and separate policy agents following the Leaf–MU sequence. Each configuration retains physics-aligned clinical credit. Parallel updates evaluate Leaf and MU from the same incoming plan, whereas the Leaf–MU sequence evaluates MU after the aperture change has been realized.

We observe similar OAR dose outcomes across the three configurations. PAC-VMAT attains the lowest spinal-cord and esophageal dose values among these configurations. These results show that comparable plan quality can be obtained with clinical credit under different policy structures and update orders. In PAC-VMAT, the Leaf–MU sequence keeps intensity credit aligned with the realized aperture geometry. The similar outcomes support this choice as a consistent execution design, without identifying the sequence or policy separation as the dominant source of improvement.

Table 10: Policy parameterization and update sequence under the PAC-VMAT optimization protocol. The controlled variants yield similar OAR dose outcomes.

Method	PTV $D_{0.03\,\mathrm{cc}}\downarrow$	Lung–CTV $V_{20\,\mathrm{Gy}}(\%)\downarrow$	Lung mean↓	Spinal cord $D_{0.03\,\mathrm{cc}}\downarrow$	Heart $D_{66.7\%}$ (Gy)↓	Esophagus mean↓
Joint Leaf–MU policy agent, sequential	65.00 ± 2.60	20.00 ± 9.20	10.70 ± 4.80	37.50 ± 14.65	2.10 ± 3.00	16.45 ± 9.80
Separate policy agents, parallel	64.40 ± 2.50	20.10 ± 9.20	10.90 ± 5.00	37.65 ± 14.80	1.70 ± 2.60	15.45 ± 9.65
PAC-VMAT	64.12 ± 2.73	19.20 ± 9.05	10.39 ± 4.63	36.37 ± 14.13	1.79 ± 2.81	15.16 ± 9.49

E.2 Joint Refinement of Leaf and MU

Aperture geometry and beam intensity jointly determine the dose distribution, with distinct contributions that evolve during refinement. Tracking aperture and intensity changes across optimization steps reveals how the two policy agents distribute their adjustments over time as MU credit is refreshed after each realized Leaf update. The accompanying dose measurements use plans scaled to the same target coverage, linking the machine-variable changes to target and OAR outcomes.

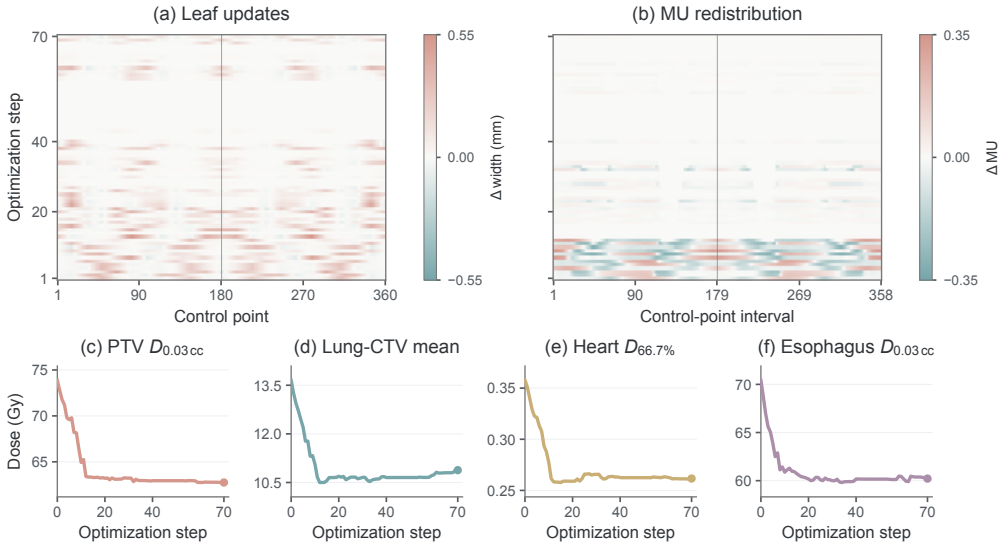

Figure 6: Optimization dynamics. Larger aperture changes concentrate in early steps, while MU redistribution remains active during refinement. The dose panels show PTV $D_{0.03\,\mathrm{cc}}$, Lung–CTV mean dose, Heart $D_{66.7\%}$, and esophageal $D_{0.03\,\mathrm{cc}}$ across optimization steps at matched target coverage.

We observe that larger aperture-width changes concentrate in earlier steps. PTV $D_{0.03\,\mathrm{cc}}$ decreases from 73.87 Gy at initialization to 62.75 Gy at completion. MU redistribution remains active after the largest early reductions in target and OAR dose. Both variables continue to update during optimization, but with different temporal emphasis. This pattern is consistent with aperture reshaping establishing the plan's spatial structure early, while sustained intensity redistribution continues to refine the dose delivered through that geometry. The complementary dynamics support coordinating Leaf and MU within the same feasible plan.

E.3 Spatial Dose Redistribution

PAC-VMAT concentrates radiation around the target and contracts the intermediate- and high-dose regions compared with random initialization (Figure 7). This spatial concentration accompanies the improved target–organ balance obtained after refinement from both random and anatomy-guided initial plans (Table 4).

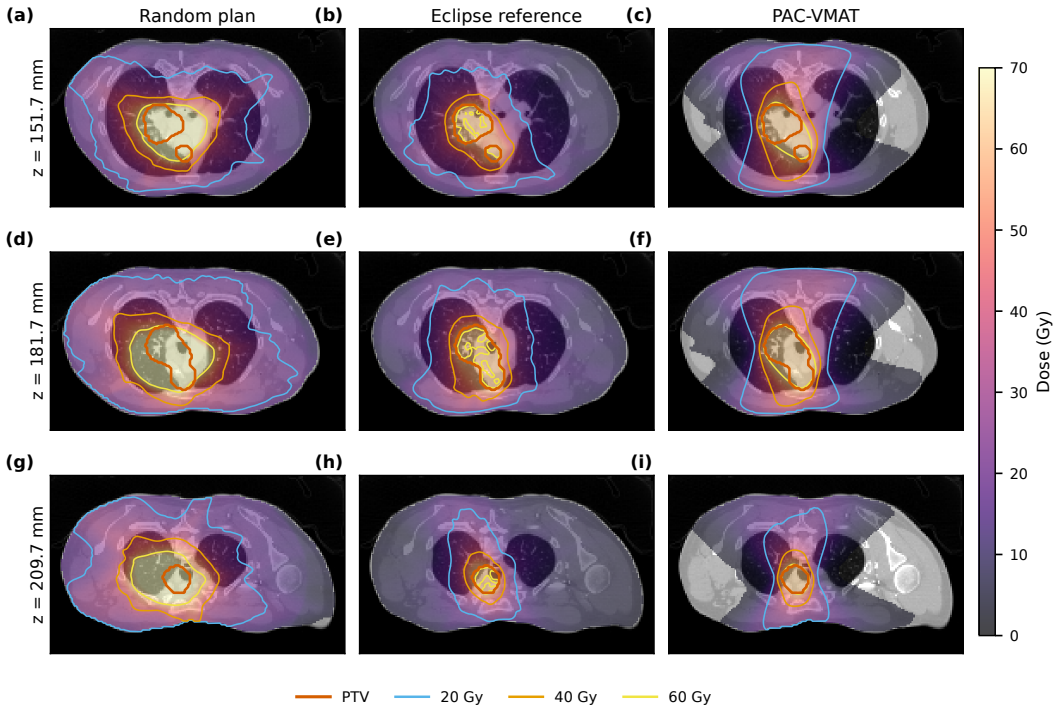

Figure 7: Spatial dose redistribution across three axial slices. PAC-VMAT contracts the intermediate and high dose regions around the PTV relative to the random plan and recovers the spatial concentration exhibited by the clinical reference.

This spatial contraction reduces dose exposure across the surrounding thorax while maintaining prescribed target coverage (Figure 8). The localized reduction of collateral exposure visually corroborates the dosimetric improvements, linking surrounding-organ sparing to the policy agents' spatial redistribution of dose. Both plans use the coverage normalization in Appendix D.3.

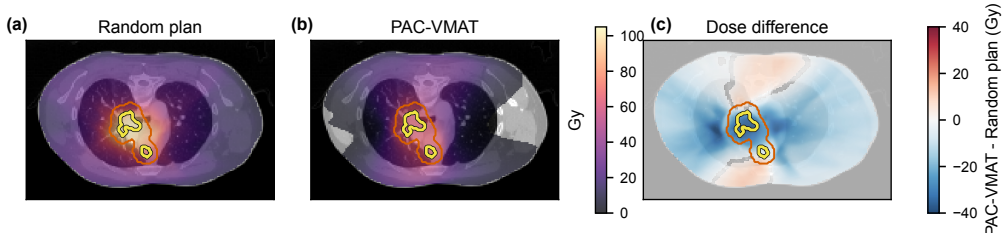

Figure 8: Spatial dose change relative to the random plan. PAC-VMAT preserves the prescription-dose region around the PTV while reducing dose across much of the surrounding thorax.

E.4 Supplementary Ablations

Network context. While clinical credit supplies independent local targets, the convolutional backbone aggregates spatial and angular context across adjacent machine controls. Restricting this feature mixing mildly increases OAR exposure, indicating that learned context provides additional coordination beyond isolated physical guidance (Table 11). The reduced-context variant retains the same observations, clinical credit targets, regularization, and output controls.

Table 11: Supplementary ablations of network context, active controls, and MU credit timing at matched target coverage.

		PTV (Gy)		Lung–CTV		Spinal cord	Heart	Esophagus
Variant	Dose engine	D_{95}	$D_{0.03\,\mathrm{cc}}\downarrow$	$V_{20\,\mathrm{Gy}}(\%)\downarrow$	Mean (Gy)↓	$D_{0.03\,\mathrm{cc}}$(Gy)↓	$D_{66.7\%}$ (Gy)↓	Mean (Gy)↓
Full-context policy agent	PyDoseRT	60.00	64.12 ± 2.73	19.20 ± 9.05	10.39 ± 4.63	36.37 ± 14.13	1.79 ± 2.81	15.16 ± 9.49
Reduced-context policy agent	PyDoseRT	60.00	64.80 ± 2.70	19.80 ± 9.10	10.70 ± 4.80	37.10 ± 14.80	1.90 ± 2.90	15.35 ± 9.60
Leaf-only	PyDoseRT	60.00	79.00 ± 2.20	32.50 ± 7.90	16.10 ± 3.90	68.95 ± 28.50	3.20 ± 2.00	15.05 ± 3.80
MU-only	PyDoseRT	60.00	72.10 ± 1.90	25.30 ± 7.40	13.00 ± 3.50	52.30 ± 30.20	2.50 ± 1.50	15.25 ± 3.65
Joint Leaf–MU	PyDoseRT	60.00	64.12 ± 2.73	19.20 ± 9.05	10.39 ± 4.63	36.37 ± 14.13	1.79 ± 2.81	15.16 ± 9.49
Fresh MU credit	PyDoseRT	60.00	64.12 ± 2.73	19.20 ± 9.05	10.39 ± 4.63	36.37 ± 14.13	1.79 ± 2.81	15.16 ± 9.49
Stale MU credit	PyDoseRT	60.00	64.60 ± 2.50	19.50 ± 9.10	10.50 ± 4.70	36.75 ± 14.30	1.80 ± 2.80	15.22 ± 9.55

Leaf and MU controls. Aperture shaping and intensity modulation play complementary roles in redistributing spatial dose. Jointly optimizing both control families substantially reduces spinal-cord exposure compared with optimizing either leaves or MU alone (Table 11), supporting the role of clinical credit in coordinating their combined degrees of freedom toward a unified objective. Esophageal dose remains similar across these variants. All three settings refine the same anatomy-guided plan under the same clinical objective and matched target coverage.

MU credit refresh. In the sequential update design, MU credit is evaluated on the intermediate dose produced by the newly updated aperture. Reverting to stale MU credit computed before the Leaf update yields slightly higher OAR exposure (Table 11). This result supports a modest advantage from aligning intensity feedback with the realized physical geometry during clinical refinement.

Policy weight transfer. We investigate whether policy weights learned from other anatomies improve how the networks express machine-variable updates for a new patient. Optimization from randomly initialized networks is compared with transferred weights while both variants refine the same anatomy-guided plan using patient-specific clinical credit. This design separates the contribution of prior policy experience from the physical guidance recomputed for the current anatomy.

Transferred and from-scratch policy agents yield similar spinal-cord and esophageal dose outcomes. The similar overall balance reached from random initialization shows that patient-specific credit already provides effective local guidance. Transferred weights provide a learned starting point for coordinating aperture and intensity updates.

Table 12: Policy weight transfer versus training from scratch at matched PTV $D_{95} = 60$ Gy. Both variants refine the same anatomy-guided initial plan.

		PTV (Gy)		Lung–CTV		Spinal cord	Heart	Esophagus
Variant	Dose engine	D_{95}	$D_{0.03\,\mathrm{cc}}\downarrow$	$V_{20\,\mathrm{Gy}}(\%)\downarrow$	Mean (Gy)↓	$D_{0.03\,\mathrm{cc}}$(Gy)↓	$D_{66.7\%}$ (Gy)↓	Mean (Gy)↓
From scratch	PyDoseRT	60.00	64.12 ± 2.73	19.20 ± 9.05	10.39 ± 4.63	36.37 ± 14.13	1.79 ± 2.81	15.16 ± 9.49
Transferred weights	PyDoseRT	60.00	63.40 ± 2.30	18.30 ± 8.60	9.90 ± 4.40	$\mathbf{36.30 \pm 14.05}$	1.60 ± 2.50	$\mathbf{15.08 \pm 9.40}$

E.5 Optimization Efficiency

As established in Section 4.4, learning coordinated machine updates through policy agents accelerates convergence. PAC-VMAT reaches comparable mean lung dose, LAD exposure, and esophageal dose to complete direct optimization in approximately 20 refinement steps. Scaling the measured optimization runtime to this step budget gives an estimated mean of 4.90 minutes on an NVIDIA H100 GPU. This early progress is consistent with the policy networks rapidly establishing a favorable spatial dose distribution, while subsequent iterations continue to refine aperture geometry and intensity redistribution.

This computational efficiency addresses the time-intensive nature of standard VMAT planning. Conventional clinical workflows require repeated manual adjustment and numerical optimization to resolve competing dose objectives, with reported durations ranging from tens of minutes for planning to several hours for broader workflows that include contouring (Bzdusek et al., 2009; Simiele et al., 2024; Zhou et al., 2025). By automating the translation of local clinical sensitivities into rapid, successive plan improvements, PAC-VMAT offers a route to generate deliverable trajectories with reduced reliance on prolonged iterative manual tuning.

F Machine Constraints and Mechanical Feasibility

F.1 Machine Constraints

The clinical improvements produced by PAC-VMAT must be realized through a sequence of apertures and intensities that satisfies the machine constraints. The feasible plan space $\mathcal{F}$ in Section 2.1 therefore restricts both leaf geometry and MU delivery. The feasible plan space accommodates the treatment geometry of each evaluated cohort. Lung VMAT uses two coplanar arcs with 180 control points per arc and optimizes 40 central leaf pairs. Head-and-neck plans use four arcs with 180 control points per arc and optimize 60 leaf pairs. The MLC has 60 leaf pairs. In the Lung VMAT configuration, a counterclockwise arc from 181° to 179° is followed by a complementary clockwise return arc. Fixed jaws enclose the union of target projections across each arc. Gantry geometry and jaw settings remain fixed during refinement, while the Leaf and MU policy agents update the aperture positions and intensity sequence within this geometry.

Leaf positions are bounded by the physical travel range and the fixed X-jaw interval of each arc. An open leaf pair must maintain a minimum gap of 0.5 mm. Pairs outside the target projection are initially parked with a 0.5-mm overlap, and mechanically closed pairs remain permitted during refinement. To preserve continuity along each arc, the displacement of either leaf bank between adjacent control points is limited to 9.30 mm whenever the pair is open at either endpoint. This displacement bound is the configured control-point motion limit used for optimization and evaluation.

Intensity delivery is represented by a positive total MU for each arc and a cumulative meterset sequence over its control points. The cumulative weights start at zero, end at one, and increase strictly between adjacent control points. Multiplying their differences by the arc-total MU yields positive MU increments, so the sequence delivers radiation in the prescribed angular order. Table 13 summarizes the geometric and meterset conditions used to define mechanical feasibility.

Table 13: Machine constraints used to evaluate the realized PAC-VMAT plans. Leaf checks apply to the active pairs within each arc.

Constraint	Requirement
Physical leaf travel	Each active leaf position lies within $[-200, 200]$ mm.
Jaw containment	Active leaf positions remain within the fixed X-jaw interval.
Open-pair separation	An open pair maintains a gap of at least 0.5 mm.
Adjacent leaf motion	Displacement is at most 9.30 mm per control-point interval when the pair is open at either endpoint.
Cumulative meterset	Finite, strictly increasing weights with endpoints zero and one.
Arc-total intensity	Finite, strictly positive total MU for each arc.
Numerical evaluation	Displacement tolerance 10^{-4} mm; cumulative meterset endpoints checked within numerical precision.

F.2 Mechanical Feasibility Evaluation

All generated PAC-VMAT trajectories are evaluated directly against the physical limits of the delivery system. The realized aperture sequences respect physical leaf travel ranges, jaw containment boundaries, open-pair separation, and the 9.30-mm adjacent-control-point motion limit under the conditions in Table 13. Furthermore, positive arc-total MU and strictly increasing cumulative metersets from zero to one across all configured arcs ensure that intensity is delivered in the prescribed angular order.

Coverage normalization preserves this structure: a positive common MU scale leaves aperture geometry and cumulative meterset weights unchanged and maintains positive intensity increments. The reported dose normalization therefore preserves the mechanical feasibility of the underlying trajectory.

F.3 Mechanical Feasibility Results

The opposing leaf banks form continuous trajectories across both arcs, with every adjacent-control-point displacement in the displayed plan remaining within the configured 9.30-mm limit (Figure 9).

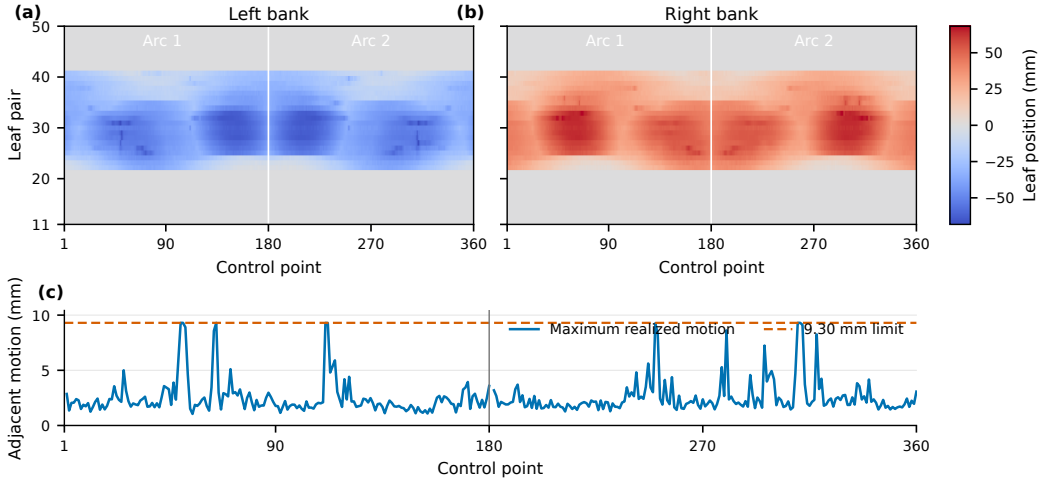

Figure 9: Kinematic structure of PAC-VMAT delivery. The opposing leaf banks form continuous trajectories across both arcs, and every adjacent-control-point motion remains within the configured 9.30-mm limit.

Across the evaluation cohort in Appendix D, all PAC-VMAT plans satisfy the complete set of mechanical constraints. The improvements in target–OAR balance are therefore realized in complete aperture and intensity trajectories that satisfy the configured machine constraints. PAC-VMAT therefore provides a direct route from patient anatomy to executable treatment plans with improved dose quality and mechanically feasible delivery trajectories.