ISSN: 3156-3953 eISSN: 3156-4437

AZMAN JOURNAL OF HEALTH AND SCIENCES (AJOHAS)

Volume 1, Issue 1, June, 2026 www.ajohas.azmanuniversity.edu.ng

Adaptive Voxel Importance Prediction for Rapid Radiotherapy Treatment Planning

Khadijah M. Hanga[1] and Maryam Alka [2]

[1]Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham B4 7XG, United Kingdom

[2]School of Mathematics, University of Birmingham, Birmingham B15 2TT, United Kingdom

Correspondence: dijahanga@gmail.com

Abstract

Large-scale inverse optimisation remains a practical bottleneck in online adaptive radiotherapy, where treatment replanning must be completed within minutes. Existing voxel-reduction methods can accelerate optimisation, but many rely on patient-specific probing or hand-crafted heuristics before each plan. This work presents AdaVIP, an extension of the voxel-importance framework of Mair et al. that replaces iterative probing with one-pass voxel-importance prediction. AdaVIP combines frozen nnU-Net anatomical features with compact physics-informed descriptors to rank voxels for reduced optimisation. Stochastic inverse-probability weighting preserves the target objective in expectation for the sampling stage, while post-sampling safeguards improve coverage of clinically critical regions. Across prostate, liver, and head-and-neck planning cohorts, AdaVIP reduces planning time by about sevenfold relative to full optimisation while keeping major dose–volume histogram endpoints within predefined clinical equivalence margins. The method adds less than one second of inference time and preserves compatibility with the downstream convex optimisation workflow.

Keywords: Radiotherapy, AdaVIP, nnU-Net, voxel-reduction

INTRODUCTION

Modern inverse treatment planning for intensity-modulated radiotherapy (IMRT) and volumetric-modulated arc therapy (VMAT) is formulated as a large-scale convex optimisation problem. Beamlet intensities are adjusted to deliver conformal dose to the planning target volume (PTV) while limiting exposure to surrounding organs at risk (OARs) (Bortfeld, 2006; Shepard et al., 1999). Because voxel-based discretisations routinely contain millions of dose points, each solver iteration remains expensive even with modern optimisation methods (Romeijn et al., 2003; Craft et al., 2014).

This cost is especially restrictive in online adaptive radiotherapy (ART), where imaging, contour review, replanning, and quality assurance must often be completed within a single treatment session (Archambault et al., 2023; Byrne et al., 2022; Sibolt et al., 2024). Several acceleration strategies have been proposed, including clustering, stochastic sampling, and gradient-guided voxel selection (Ungun et al., 2019; Breedveld et al., 2017; Mair et al., 2024). Among these, the voxel-importance framework of Mair et al. (2024) is attractive because it focuses computation on the voxels that contribute most strongly to the objective.

Its main limitation is the need for patient-specific probing before sampling can begin. That probing stage requires repeated gradient evaluations on the full voxel set and therefore consumes part of the same time budget that adaptive workflows are trying to save.

Meanwhile, pretrained medical-imaging encoders such as nnU-Net capture organ boundaries and broader anatomical context with strong transferability across segmentation tasks (Isensee et al., 2021). We hypothesise that these anatomical representations, combined with compact physics-informed descriptors derived from the dose-influence matrix, can predict voxel importance directly and shift the probing burden into offline training.

We therefore propose AdaVIP, an extension of the voxel-importance framework of Mair et al. that replaces per-patient probing with learned voxel-importance prediction. We evaluate the method on prostate, liver, and head-and-neck cohorts and report planning efficiency, dosimetric fidelity, plan quality, and statistical equivalence under a common optimisation pipeline.

Materials and Methods

Patient cohorts

Experiments are conducted on three radiotherapy cohorts: prostate (n = 80), liver (n = 45), and head-and-neck (n = 55) cases. For each site, a held-out test split is reserved for final evaluation (prostate: 20 cases; liver: 15 cases; head-and-neck: 15 cases). The remaining cases form the development split used for model fitting, with a validation subset drawn from that development split for hyperparameter tuning and early stopping. The same held-out test cases are then used for voxel-importance prediction and downstream fluence map optimisation comparisons under consistent evaluation criteria including wall-clock time, DVH endpoints, plan quality score, and statistical equivalence testing.

Treatment planning problem formulation

We follow the voxel-wise penalty formulation used in voxel-importance-based treatment planning (Mair et al., 2024). Let $x \in \mathbb{R}^{n_b}_{\geq 0}$ denote non-negative beamlet intensities, $y \in \mathbb{R}^{n}_{\geq 0}$ denote voxel doses, and let $A \in \mathbb{R}^{n \times n_b}_{\geq 0}$ be the dose-influence matrix such that

$$y = Ax. \tag{1}$$

Voxels are partitioned into disjoint structures V_s for $s \in S = \{PTV, OAR, BDY\}$. We solve a convex optimisation problem of the form

$$\min_{x \geq 0} f(Ax) \quad \text{s.t.} \quad Ax \in C, \tag{2}$$

where C encodes affine dose constraints and $f(\cdot)$ is a separable convex penalty,

$$f(y) = \sum_{s \in \mathcal{S}} \frac{1}{|V_s|} \sum_{v \in V_s} f_s(y_v). \tag{3}$$

This formulation is consistent across full optimisation and voxel-reduced variants, enabling fair runtime and plan-quality comparison.

Reference voxel importance framework (probing / GradNorm)

We build on the voxel-importance sampling framework of Mair et al. [10]. A probing optimisation is run for K iterations of PGD to obtain a per-voxel gradient signal $g_K(v)$, from which voxel importance is defined as

$$\chi(v) = \|g_K(v)\|_2. \tag{4}$$

A sampling distribution is formed by normalisation $q(v) \propto \chi(v)$, and a reduced voxel set is sampled to approximate the full objective with unbiased importance weights [10].

AdaVIP overview

AdaVIP replaces the iterative probing stage with a single forward pass of a neural regressor that predicts voxel-wise importance scores using (i) frozen pretrained anatomical features and (ii) physicsinformed features derived from the dose-influence matrix and geometric context. The pipeline

is: (i) feature construction, (ii) voxel-wise importance prediction, (iii) importance-weighted subsampling with safeguards, and (iv) optimisation on the reduced voxel set.

Voxel-wise feature construction

For each voxel v, we form a fixed feature vector $z_v \in \mathbb{R}^{524}$ by concatenating:

- **A 512-dimensional anatomical descriptor:** selected multiscale nnU-Net encoder channels, interpolated to the voxel grid and kept frozen during AdaVIP training.
- **A 12-dimensional physics-informed descriptor:** eight dose-influence summary statistics (including max, mean, standard deviation, ℓ_2 norm, sparsity, angular-span, prescription-ratio, and gradient-proxy terms) together with four geometric or structural summaries (signed distance to the PTV, distance to the nearest OAR, beam-depth proxy, and structure-class code).

This fixed 512 + 12 = 524 decomposition is used across all sites. All features are standardised prior to training.

Neural regressor and training objective

We use a lightweight 4-layer multilayer perceptron (MLP) with ∼180k parameters that maps $z_v \mapsto s_v \in [0,1]$ via a sigmoid output. Training uses oracle labels derived from gradient-based sensitivity analysis of full-resolution plans (teacher signal) and a composite loss combining voxel-wise regression with distribution matching:

$$L = \mathrm{MSE}(s, s_{\mathrm{oracle}}) + \beta\, \mathrm{KL}(p_{\mathrm{oracle}} \| p), \qquad (5)$$

where p and p^{oracle} denote the score vectors normalised into sampling distributions over voxels. The weighting coefficient β controls the trade-off between local score accuracy and calibration of the global sampling distribution; we set $\beta = 0.1$ based on validation performance. The MLP is trained with AdamW (initial learning rate 5×10^{-4}, weight decay 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$), cosine decay to 10^{-6} over at most 100 epochs, batch size 4 patients, and early stopping with patience 15 epochs.

Inference, sampling and unbiased reweighting

At inference, AdaVIP predicts importance scores for all voxels in a single pass (GPU inference < 1 second per patient). Scores are normalised into a sampling distribution

$$q_v = \frac{s_v}{\sum_{u \in V} s_u}. \qquad (6)$$

For the stochastic importance-sampling component, we sample $m = \lfloor \alpha n \rfloor$ voxels with replacement, with $\alpha \approx 0.15$, and solve the reduced optimisation problem using inverse-probability weights. Let $s(v)$ denote the structure containing voxel v. Each sampled voxel contributes

$$\frac{1}{|V_{s(v)}|\, m q_v} f_{s(v)}(y_v) \qquad (7)$$

to the reduced objective, with repeated samples aggregated by summing their weights. This estimator is unbiased for the separable objective in Eq. (3) provided that all voxels with non-zero contribution under the full objective have strictly positive sampling probability and that multiplicities are retained in the reweighting step. After stochastic sampling, deterministic safeguard voxels are added to improve coverage of clinically critical regions; these additions prioritize robustness and constraint coverage over strict unbiasedness of the final augmented subset.

Clinical safeguards

To enhance robustness, AdaVIP includes safeguard strata in sampling, ensuring adequate representation of PTV-critical regions and OAR hotspot regions. Safeguards are applied after the stochastic importance sampling stage through enforced PTV strata and OAR hotspot oversampling.

Baselines and evaluation metrics

We compare against full optimisation (100% voxels), uniform subsampling (15%), the probing-based gradient-norm baseline of Mair et al. (2024), and clustering-based reduction (Ungun et al., 2019). Metrics include wallclock planning time, DVH endpoints, plan quality score (PQS),

and statistical tests based on Wilcoxon signed-rank and TOST equivalence procedures.

Evaluation Metrics

Plan quality is assessed using standard clinical endpoints including PTV $V_{95\%}$, $V_{107\%}$, $D_{1\%}$, and $D_{99\%}$, together with OAR D_{mean}, $D_{0.1\text{cc}}$, and $D_{1\text{cc}}$ where applicable. Overall plan quality is quantified using a clinical plan quality score (PQS) (Nelissen et al., 2023). Wall-clock optimization time is recorded on identical hardware (dual AMD EPYC 7763, 512 GB RAM, NVIDIA A100 80 GB). Statistical equivalence is tested using two one-sided TOST procedures with margins Δ = 2 Gy or 2% as appropriate, together with Wilcoxon signed-rank tests at $\alpha = 0.05$ with Bonferroni correction.

Implementation Details

All experiments are implemented in Python with PyTorch and nnU-Net v2. Data from prostate, liver, and head-and-neck cohorts are organised according to the nnU-Net raw data specification, with images and labels stored in NIfTI format. Anatomical representations are obtained from nnU-Net v2 models trained for segmentation using the standard three-dimensional full-resolution configuration, stochastic gradient descent, and a combined Dice plus cross-entropy loss. These nnU-Net models are used only to produce pretrained anatomical features; during AdaVIP training and inference, the encoder is frozen.

The downstream voxel-importance predictor is a separate 4-layer MLP trained on voxel-wise feature vectors using the AdamW, cosine-schedule, and MSE+KL setup described above. Preprocessing and experiment planning are performed automatically by nnU-Net and include dataset integrity verification, intensity normalisation, resampling to target spacing, and automatic configuration of network architecture, patch size, and batch size for the segmentation models. For reproducibility, we preserve preprocessed inputs, trained checkpoints, predicted importance maps, sampled voxel indices, and solver outputs for every experiment, and all final comparisons are reported only on the held-out test sets described above.

Results

Computational efficiency

AdaVIP substantially reduces end-to-end planning time across all sites. Table 1 reports wall-clock planning times (minutes) for full optimisation and baselines at 15% voxel subsampling in the present study. Table 2 provides a compact site-level summary combining runtime reduction and clinical-equivalence status. Across the three held-out test cohorts, AdaVIP achieves an average time reduction of 85.1% relative to full optimisation and keeps the reported key DVH endpoints within the prespecified equivalence margins.

Table 1: Wall-clock planning time (minutes) across sites.

Site	Full	Uniform	GradNorm	Cluster	AdaVIP
Prostate	42.1	8.2	7.9	9.1	6.1
Liver	68.4	12.5	11.8	13.4	9.3
H&N	98.2	18.1	16.7	19.8	13.8

Table 2: Site-level summary of planning-time reduction and equivalence outcome.

Site	Full → AdaVIP (min)	Reduction (%)	Key endpoints within margin
Prostate	42.1 → 6.1	85.5	Yes
Liver	68.4 → 9.3	86.4	Yes
H&N	98.2 → 13.8	86.0	Yes

Dosimetric fidelity and equivalence (Head-and-Neck example)

Across disease sites, AdaVIP maintains dosimetric fidelity with < 1.5% DVH deviation under the predefined equivalence criteria. The site-level summary in Table 2 indicates that the reported primary endpoints remained inside the prespecified 2%/2 Gy equivalence margins for all three cohorts. Table 3 then shows representative head-and-neck DVH endpoints in more detail, where AdaVIP remains close to full optimisation while competing baselines more frequently violate serial-organ endpoints.

Table 3: Representative head-and-neck DVH endpoints from the present study.

Metric	Full	Uniform	GradNorm	AdaVIP
PTV70 V95% (%)	95.0	91.2	92.1	94.1
PTV70 D1% (Gy)	73.2	75.8	74.9	73.6
Cord D0.1cc (Gy)	44.1	48.3	47.2	44.8
Parotid (L) Dmean (Gy)	25.3	28.0	27.3	25.9

Ablation study (Prostate)

Ablations confirm that both pretrained anatomical representations and physics-informed features are important contributors to performance. Table 4 reports planning time and selected clinical deviations for prostate variants in the present study.

Table 4: Ablations (Prostate): time and selected deviations relative to full.

Variant	Time (min)	ΔV95% (%)	ΔDmean, bladder (Gy)
No pretraining	15.2	3.1	2.8
No physics	11.8	2.4	2.1
Random regressor	16.4	4.0	3.5
AdaVIP	**6.1**	**1.1**	**0.7**

Discussion

The results show a consistent pattern across the three evaluated disease sites. AdaVIP reduces optimisation time substantially relative to the full problem, remains faster than the voxel-reduction baselines considered here, and keeps the reported key DVH endpoints within the prespecified clinical margins. The ablation study further indicates that both frozen anatomical representations and physics-informed features contribute materially to the final performance.

From a translational perspective, the observed planning times fall into a range that is more compatible with online adaptive workflows than full optimisation alone. That does not by itself establish clinical readiness, but it does suggest that learned voxel prioritisation can remove a meaningful portion of the computational bottleneck without changing the downstream optimisation engine.

The present evaluation still has important scope limits. Although AdaVIP performs consistently across prostate, liver, and head-and-neck cohorts, those sites do not span the full range of treatment geometries, motion patterns, re-irradiation settings, or institutional planning conventions encountered in practice. Additional external validation is therefore needed to determine how often retraining or recalibration would be required when moving to new tumour sites, beam arrangements, or optimisation objectives.

The regulatory and validation implications should also be stated carefully. AdaVIP changes voxel selection upstream of the solver rather than replacing the final convex optimisation step, which is advantageous for integration because the dose engine, constraints, and solver remain unchanged. Even so, deployment would still require software verification, failure-mode analysis, prospective

workflow testing, and site-specific quality-assurance procedures before routine clinical use could be justified.

Limitations

Reliance on Benchmark Optimisation as Reference

Plan quality is assessed relative to Full optimisation, which serves as a benchmark rather than an absolute clinical optimum. Although Full optimisation reflects standard high-quality planning practice, it does not necessarily represent the globally optimal solution. Consequently, deviations reported relative to this benchmark should be interpreted as comparative rather than absolute measures of plan quality.

Evaluation Metrics

The analysis focuses on selected dose–volume histogram (DVH) endpoints that are commonly used in clinical practice. While these metrics capture key aspects of target coverage and organ-at-risk sparing, they do not fully characterise overall plan quality. Other considerations, such as spatial dose distribution, robustness to anatomical variation, and correlations with long-term clinical outcomes, are not explicitly evaluated in this study.

Computational Environment

All experiments are conducted in a controlled computational environment with consistent hardware and software configurations. Although relative comparisons between optimisation strategies remain valid, absolute planning times may vary across systems with different hardware capabilities, parallelisation strategies, or optimisation engines. Additional evaluation on clinical treatment planning systems would be required to assess real-world computational performance.

Generalisation of Learned Prioritisation

The proposed AdaVIP framework learns adaptive objective prioritisation from available training data. While the ablation study demonstrates robustness within the evaluated scenarios, the generalisability of learned prioritisation strategies across substantially different treatment planning problems or optimisation settings remains an open question. Further investigation is required to assess transferability and the potential need for retraining in new contexts.

Conclusion

This work presents AdaVIP, a learned replacement for patient-specific voxel probing in importance sampled radiotherapy optimisation. By combining frozen nnU-Net features with compact physics-informed descriptors, the method reduces planning time across prostate, liver, and head-and-neck cohorts while keeping the reported key DVH endpoints within predefined clinical equivalence margins.

Because AdaVIP changes voxel selection rather than the downstream convex solver, it offers a practical path for further validation in adaptive workflows. The next step is not broader rhetoric about speed, but broader evidence: external multi-centre testing, prospective workflow studies, and uncertainty-aware extensions.

Abbreviations

IMRT	Intensity-Modulated Radiotherapy
VMAT	Volumetric Modulated Arc Therapy
ART	Adaptive Radiotherapy
PTV	Planning Target Volume
OAR	Organ at Risk
DVH	Dose–Volume Histogram
PQS	Plan Quality Score
AdaVIP	Adaptive Voxel Importance Prediction
PGD	Projected Gradient Descent
MLP	Multilayer Perceptron
TOST	Two One-Sided Tests
Gy	Gray

References

Archambault, Y., Boylan, C., van Beek, S., Bissonnette, J.-P., & et al. (2023). Clinical implementation of an online adaptive radiotherapy system. *Journal of Applied Clinical Medical Physics*, *4*(1), e13982.

Bortfeld, T. (2006). IMRT: A review and preview. *Physics in Medicine and Biology*, *51*(13), R363–R379.

Breedveld, S., Storchi, P. R. M., & Heijmen, B. J. M. (2017). Stochastic constraint sampling for non-convex optimization in radiation therapy. *Physics in Medicine and Biology*, *62*(23), N407–N423.

Byrne, M., Archibald-Heeren, B., Hu, Y., & et al. (2022). Varian ethos online adaptive radiotherapy for prostate cancer: Early clinical experience. *Physics in Medicine and Biology*, *67*(1), 015016.

Craft, D., Bangert, M., Long, T., Papp, D., & Unkelbach, J. (2014). Shared data for intensity modulated radiation therapy (IMRT) optimization research: The CORT dataset. *Medical Physics*, *41*(8), 081703.

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*, 203–211.

Mair, S., Fu, A., & Sjölund, J. (2024). Efficient radiation treatment planning based on voxel importance. *Physics in Medicine and Biology*, *69*(16), 165014.

Nelissen, K. J., Versteijne, E., Senan, S., & et al. (2023). Same-day adaptive palliative radiotherapy without prior CT simulation. *Radiotherapy and Oncology*, *182*, 109538.

Romeijn, H. E., Ahuja, R. K., Dempsey, J. F., & Kumar, A. (2003). A new linear programming approach to radiation therapy treatment planning problems. *Operations Research*, *54*(2), 201–216.

Shepard, D. M., Olivera, G. H., Reckwerdt, P. J., & Mackie, T. R. (1999). Iterative approaches to dose optimization in tomographic radiation therapy. *Physics in Medicine and Biology*, *44*(10), 2541–2557.

Sibolt, P., Gustafsson, C., Cronholm, A., & et al. (2024). Clinical experience with daily online adaptive radiotherapy using ethos. *Technical Innovations and Patient Support in Radiation Oncology*, *29*, 100241.

Ungun, B., Xing, L., & Boyd, S. (2019). Real-time radiation treatment planning with optimality guarantees via cluster and bound methods. *INFORMS Journal on Computing*, *31*(3), 544–558.