Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER • OPEN ACCESS

Multi-model study of fast VMAT segment dose calculation with deep learning

To cite this article: Fan Xiao *et al* 2026 *Phys. Med. Biol.* **71** 095015

IOP Publishing Phys. Med. Biol. 71 (2026) 095015 https://doi.org/10.1088/1361-6560/ae6413

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
3 December 2025

REVISED
1 April 2026

ACCEPTED FOR PUBLICATION
23 April 2026

PUBLISHED
7 May 2026

PAPER

Multi-model study of fast VMAT segment dose calculation with deep learning

Fan Xiao[1], **Niklas Wahl**[2,3], **Claus Belka**[1,4,5], **Christopher Kurz**[1], **George Dedes**[6,1,7] and **Guillaume Landry**[1,5,7,*]

[1] Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany
[2] Department of Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[3] National Center for Radiation Oncology (NCRO), Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany
[4] German Cancer Consortium (DKTK), partner site Munich, a partnership between DKFZ and LMU University Hospital, Munich, Germany
[5] Bavarian Cancer Research Center (BZKF), Munich, Germany
[6] Department of Medical Physics, LMU Munich, Munich, Germany
[7] These authors contributed equally to this work as senior authors.
[*] Author to whom any correspondence should be addressed.

E-mail: guillaume.landry@med.uni-muenchen.de

Keywords: dose calculation, deep learning, photon therapy, VMAT

Supplementary material for this article is available online

Abstract

Objective. Deep learning (DL) methods enable photon dose calculation under two main coordinate representations: Beam's Eye View (BEV) and patient coordinates. We evaluate dose calculation accuracy and speed under these coordinate paradigms and with representative DL models within a unified dataset and pipeline, and introduce two lightweight models for fast photon dose calculation. *Approach.* Planning computed tomography (CT) scans and volumetric modulated arc therapy plans from 24 prostate cancer patients were used. Monte Carlo simulation generated 5940, 540, and 3053 segment doses for training (11 patients), validation (3), and testing (10), respectively. For BEV, we used a combination of convolutional neural network (CNN) and convolutional long short-term memory network (ConvLSTM) called CNN-ConvLSTM, a CNN-Mamba combination (CNN-Mamba), a transformer-based architecture (DoTA), and a cascaded 3D UNet (C3D). These were trained on CT and segment-projection BEV cuboids. For patient coordinates, the DeepDose individual segment dose prediction framework implemented with C3D (DeepDose-C3D) was trained on cropped CT volumes with four physical inputs. Segment and plan dose accuracy were assessed using local gamma passing rates γ_{PR} (2%/3 mm and 1%/3 mm) and dose-volume histogram metrics. Dose calculation times (inference plus pre/post-processing) were measured on three different graphics processing unit (GPUs). *Results.* All five models achieved mean local γ_{PR} values $\geqslant$91.0% (2%/3 mm) for segment doses and $\geqslant$99.0% (1%/3 mm) for plan doses. Mean per-segment dose calculation times were 79, 67, 298, 490, 356 ms for CNN-ConvLSTM, CNN-Mamba, DoTA, C3D, and DeepDose-C3D, respectively. On the latest-generation GPU available, the corresponding per-plan (average 305 segments) dose calculation times were 5.5, 6.2, 33.6, 38.7, 35.4 s. *Significance.* Both BEV- and patient-coordinate DL methods achieved accurate photon plan dose calculation, with BEV-based approaches showing more robust segment performance. CNN-ConvLSTM and CNN-Mamba retain comparable accuracy at lower computational cost, enabling fast photon dose calculation.

1. Introduction

In external photon-beam radiotherapy, intensity modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) deliver highly conformal dose distributions via variable apertures formed by multileaf collimators (MLCs) (Otto 2008, Hussein *et al* 2018). As workflows advance toward online and real-time adaptation (Kontaxis *et al* 2015, Green *et al* 2018), fast and accurate dose computation is critical for rapid plan evaluation and iterative re-optimization that accounts for intrafraction motion (Lombardo *et al* 2024, Keall *et al* 2025).

Monte Carlo (MC) simulation is widely regarded as the gold standard for dose calculation (Panettieri *et al* 2009). With general-purpose MC codes (Onizuka *et al* 2018, Paschal *et al* 2022), dose calculations typically take hours to tens of hours because they track millions to billions of particle histories. Advances in graphics processing unit (GPU) implementations (Hissoiny *et al* 2011, Cheng *et al* 2023) with acceleration algorithms (Renaud *et al* 2015) have substantially reduced runtimes, with plan-level dose calculations typically reported in the tens-of-seconds range (Li *et al* 2021, Chen *et al* 2025). More recently, GARDEN reported gamma passing rates $\gamma(2\%/3\text{ mm})$ of 99.73% for VMAT plans in 15.4 s with 0.2% uncertainty and $\gamma(3\%/3\text{mm})$ above 99.23% for IMRT and VMAT plans in 3 s at $\sim 1\%$ uncertainty (Liu *et al* 2025). In parallel, rapid advances in artificial intelligence-based image processing (Li *et al* 2024) have yielded near real-time, high quality outputs across diverse radiotherapy applications (Rabe *et al* 2025), with deep learning (DL)-based dose calculation becoming increasingly prevalent as an open research topic in recent years.

For the photon dose calculation task, approaches operate in two coordinate representations: patient coordinates or beam's eye view (BEV) coordinates. For patient coordinates, existing methods typically rely on multiple 3D physics-based inputs, such as segment binary mask, distance from source, central beamline distance, radiological depth (Kontaxis *et al* 2020), noisy MC dose (Bai *et al* 2021), fluence projection volume (Xiao *et al* 2022), or pencil beam dose (Song *et al* 2023). These models learn a mapping between the physics-derived inputs, density information from computed tomography (CT), and the reference MC dose. In contrast, methods in BEV coordinates aim to reduce or bypass multiple physical inputs, by performing dose sequence modeling directly from beam geometry and aperture information extracted from the treatment plan: Pastor-Serrano *et al* proposed a 3D convolutional-transformer model using the tissue-density-corrected segment projection as model input and reported an average VMAT plan dose calculation time of 8 s (Pastor-Serrano *et al* 2023). Witte *et al* showed a gated recurrent unit model with 2D convolutional layers using radiotherapy plan parameters as model input, achieving an 82 times speed-up over GPUMCD (Hissoiny *et al* 2011) for VMAT plans (Witte and Sonke 2024). Fan *et al* developed a 2D Unet-like model using distance-corrected conical fluence maps as model input, with a mean runtime of 1 s per IMRT single beam (Fan *et al* 2025).

Both patient and BEV coordinate-based DL approaches have shown competitive accuracy-runtime performance; yet systematic comparisons under identical dataset, processing, and metrics remain limited. Here, we conducted a comparative study: we present a DL framework for fast photon segment dose calculation (including BEV resampling, segment projection, and inverse resampling of dose to patient coordinates, with support for large fields (40 cm $\times$ 40 cm)). Under the same protocol, we compared two factors: model architectures and coordinate representations (BEV versus patient coordinates). In addition, two fast and lightweight DL models were introduced for BEV coordinate-based methods.

2. Materials and methods

2.1. Proposed workflow

For the BEV coordinate method (figure 1(a)), the patient CT was first resampled into BEV sequences (see section 2.3.1). The DL model then took the BEV CT and segment projection cuboids as inputs to predict dose, which was subsequently inverse-resampled back onto the patient coordinate for comparison with the MC dose (ground truth MC doses were always in the patient coordinate system). Two open-source dose prediction models (transformer DoTA (Pastor-Serrano *et al* 2023) and Unet cascaded 3D (C3D) (Liu *et al* 2021)) were used as baselines. Inspired by a recent DL work on proton dose calculation (Neishabouri *et al* 2025), a convolutional neural network (CNN)-convolutional long short-term memory network (ConvLSTM) (CNN + convolutional long short-term memory) model for photon segment dose prediction was developed (see section 2.5). Besides, following the recent advances in selective state-space models, such as Mamba (Gu and Dao 2024), which provide a more computationally efficient alternative to attention-based transformers for long-sequence modeling, we evaluated a CNN-Mamba model while keeping the CNN encoder-decoder identical to that of the CNN-ConvLSTM model.

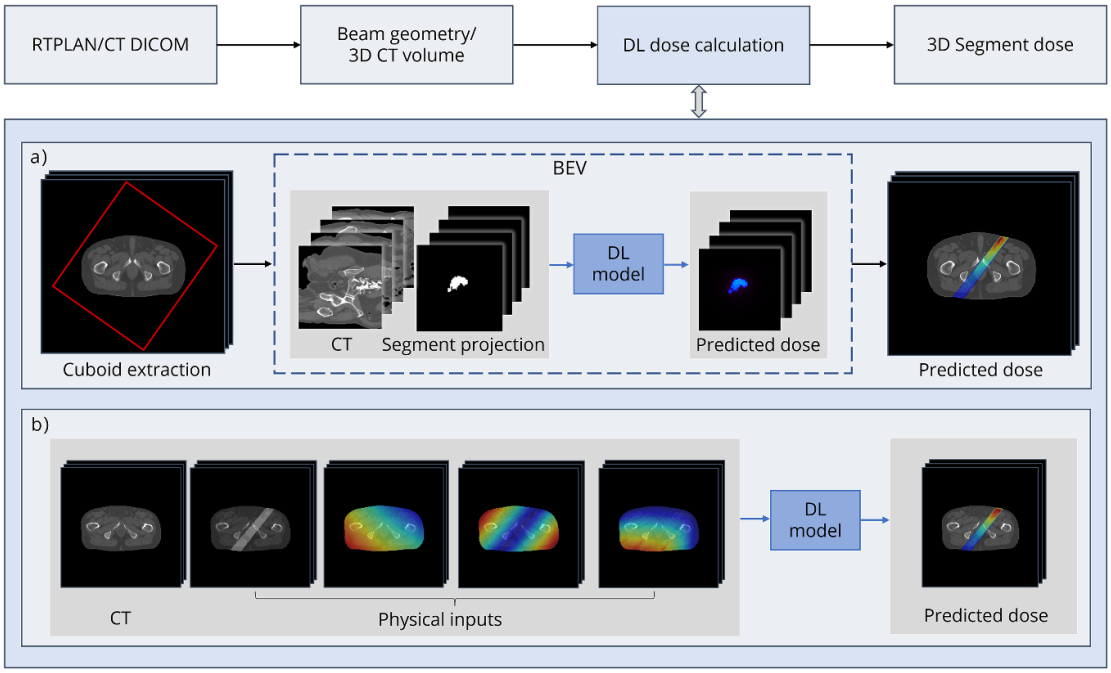

Figure 1. The workflow of DL dose calculation approaches in two coordinate representations: (a) BEV-coordinate modeling: (a, left) A beam-aligned cuboid region of interest (red), centered at the isocenter, is defined on the patient CT and resampled to the BEV grid, providing the 3D cuboid for subsequent segment projection and sequence modeling. (a, mid) A segment projection mask is generated based on the 3D CT cuboid, and a DL model is trained to predict the segment dose in BEV coordinates. (a, right) BEV-predicted dose is inverse-resampled onto the patient CT grid, enabling dose accumulation across segments and plan-level evaluation. (b) Patient-coordinate modeling: A DL model takes the 3D patient CT volume with physical inputs and predicts the segment dose on the patient CT grid.

In contrast, for the patient coordinate method (figure 1(b)), the DL model predicted dose directly on the patient CT grid from a cropped 3D CT volume together with different physical features, and the dose prediction was then restored to the full grid. We considered two input configurations on top of the CT: (i) the 3D segment projection only, and (ii) the four physical features described in Kontaxis *et al* (2020). The Unet C3D architecture was the same as that used in the BEV-based method.

2.2. Patient dataset and MC simulation

The patient dataset consisted of anonymized and implant-free planning CT images and VMAT RTPLANs from 24 prostate cancer patients treated at the Department of Radiation Oncology, LMU University Hospital. Planning CTs were acquired on a single CT scanner (Aquilion LB, Canon Medical Systems) with an original voxel size of $1 \times 1 \times 3\,\mathrm{mm}^3$, and were resampled to an isotropic $3 \times 3 \times 3\,\mathrm{mm}^3$ grid, consistent with commonly used clinical dose grids. 24 VMAT plans were generated using the Monaco treatment planning system and delivered on an Elekta Versa HD Linac equipped with an Agility 160-leaf MLC as part of clinical operation.

For the segment extraction from the VMAT plans, the continuous delivery between two consecutive control points $k \to k{+}1$ was approximated by the mid-control-point aperture. A 2D binary aperture mask $M_k \in \{0,1\}$, shown in figure 2(a), was then generated on the BEV grid covering the full field size ($40\,\mathrm{cm} \times 40\,\mathrm{cm}$ at isocenter plane), with the resolution of $1\ \mathrm{mm} \times 1\ \mathrm{mm}$. This mask was used for segment projection and as the fluence aperture in the subsequent MC simulation.

We denote by c_k the normalized cumulative meterset weight at control point k, which increases monotonically from 0 to 1. Let M represent the total Beam Meterset (monitor unit (MU)). The portion of MU delivered within the segment $k \to k+1$, denoted as M_k, is calculated as:

$$M_k = M\,(c_{k+1} - c_k)\,. \tag{1}$$

Let $\hat{D}_k$ denote the dose D for segment k. Then the plan dose for K control points can be obtained by MU-weighted accumulation:

$$D_{\mathrm{plan}} = \sum_{k=0}^{K-1} M_k \hat{D}_k. \tag{2}$$

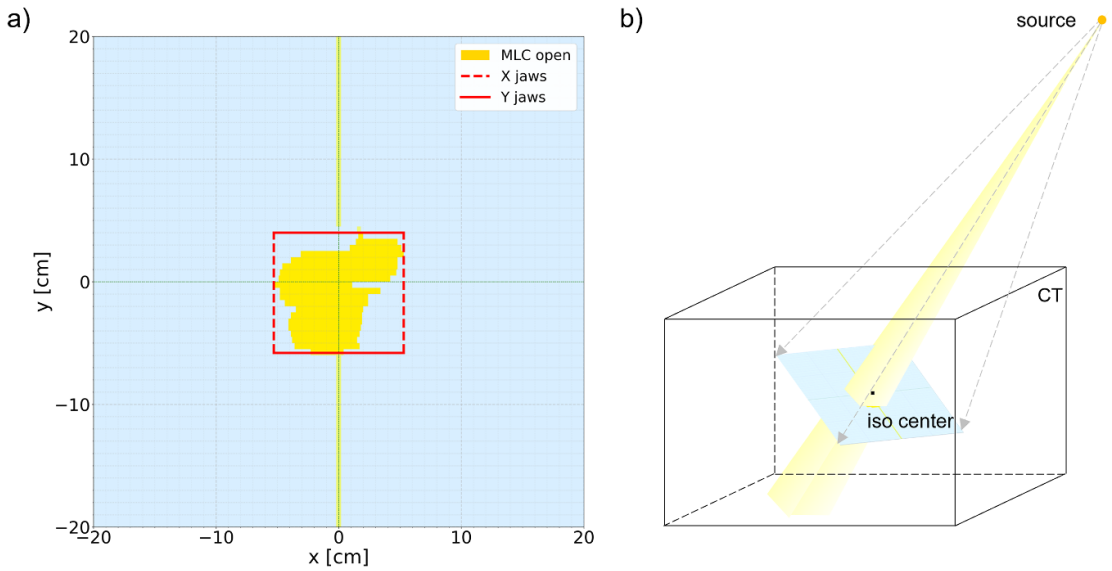

Figure 2. (a) MLC aperture on the isocenter plane within the $40 \times 40\,\mathrm{cm}^2$ field (yellow shows the open MLC region; red dashed/solid denote X/Y jaws); (b) segment dose MC simulation geometry with a point source at SAD $= 100\,\mathrm{cm}$, where the binary aperture from (a) serves as the fluence filter.

In this study, MC simulations were implemented with Geant4 (version 11.00-patch-03), using a predefined QGSP_BIC_HP_EMZ physics list. Patient CT images were converted into corresponding density and elemental composition maps through a scanner-specific calibration curve reported previously (Schmid *et al* 2015). An idealized point source was used, with the photon energy spectrum derived from the ELEKTA_PRECISE_6MV phase space files provided by the International Atomic Energy Agency (IAEA) (Capote *et al* 2006). For the segment dose MC simulation, the source-to-axis distance was set to 100 cm, the binary aperture mask served as a fluence filter on the isocenter plane with the full 40 cm × 40 cm field, as shown in figure 2(b), and the dose grid matched the patient CT spacing ($3 \times 3 \times 3\,\mathrm{mm}^3$).

For dataset partitioning, 11 patients were used for training, 3 for validation, and 10 for testing. For training, 3652 segments were extracted from 11 RTPLANs, and data augmentation shifted each MLC leaf randomly by $\delta \sim \mathcal{U}(-5\ \mathrm{mm}, 5\ \mathrm{mm})$ along its travel direction, with $\mathcal{U}$ the uniform distribution, doubling the training segment pool to 7304. For each training patient, the isocenter was shifted along the superior-inferior axis by −2, 0, +2 cm. At each isocenter, 180 gantry angles were simulated every 2° from 0° to 358°; at each angle, one segment was randomly sampled from the 7304-segment pool to generate a segment dose, yielding $11 \times 3 \times 180 = 5940$ training doses. For validation, 1167 segments were extracted from 3 patients without MLC shifting augmentation. For each patient, 180 gantry angles (0–358°) were simulated at the original isocenter; at each angle, one segment was randomly chosen from the 1167-segment pool, yielding $3 \times 1 \times 180 = 540$ segment doses. For testing, segment doses were simulated according to each plan's control point sequence, resulting in 3053 segment doses across 10 patients (216, 254, 297, 312, 214, 338, 380, 384, 311, and 347 segments, respectively). Plan doses were then reconstructed by MU-weighted accumulation of segment doses. Due to simulation time limitations (segment dose simulation with a mean runtime of approximately 25 h for $5 \times 10^6\,\mathrm{cm}^2$ histories on a single core of an Intel Xeon Gold 6354 3.00 GHz central processing unit (CPU)), photon histories were $5 \times 10^5\,\mathrm{cm}^2$ for training segment dose simulation (yielding an average statistical uncertainty of ∼5% in the $D > 10\%D_{\max}$ region), and $5 \times 10^6\,\mathrm{cm}^2$ for validation and testing segment dose simulation (yielding an average statistical uncertainty below 1.6% in the $D > 10\%D_{\max}$ region). The prescription was 60 Gy in 20 fractions. All MC plan doses were scaled so that $D_{95\%}$ planning target volume(PTV) $\geqslant 0.95 \times 60\,\mathrm{Gy} = 57\,\mathrm{Gy}$, and the same scaling factor was used for predicted plan doses. MC simulation were run on a computing cluster with 138 physical CPU cores.

2.3. BEV-coordinate processing

2.3.1. BEV resampling and inverse mapping

For the BEV-coordinate processing in figure 1(a), patient CT (in Hounsfield Unit (HU)) and segment dose were first transformed to a 3D BEV grid defined by the point source and the isocenter. The grid

was a cuboid with a full field of $40 \times 40\,\mathrm{cm}^2$ at the isocenter plane and a depth of 51.2 cm along the beam axis (± 25.6 cm centered on the isocenter plane), sufficient for general treatments. CT and dose volumes were resampled onto this grid for model training, and predicted doses were inverse-mapped to patient space. To minimize errors from sequential resampling, cubic B-spline interpolation was applied on isotropic $2 \times 2 \times 2\,\mathrm{mm}^3$ BEV voxels, yielding a BEV grid of $256 \times 200 \times 200$ voxels. The array was ordered as $(N_z, N_y, N_x) = (256, 200, 200)$, where z follows the beam axis and (x, y) span the two lateral beam-divergence directions. Both forward and inverse interpolations were run on the GPU via CuPy 13.6.0 (Okuta *et al* 2017).

2.3.2. BEV segment projection

To encode the segment geometry, the 2D binary aperture on the isocenter plane was projected onto the 3D BEV grid on the GPU via CuPy. For each BEV voxel center $\vec{p}$, we constructed a ray originating from the source position $\vec{s}$ with direction $\vec{d} = \vec{p} - \vec{s}$:

$$\vec{r}(t) = \vec{s} + t\vec{d}, \qquad t \geqslant 0. \tag{3}$$

The isocenter plane was defined by the beam-axis unit normal $\vec{n}$ and the signed source-to-plane distance d_{iso}. The ray-plane intersection was obtained by:

$$t_{\mathrm{int}} = \frac{d_{\mathrm{iso}}}{\vec{d} \cdot \vec{n}}. \tag{4}$$

The intersection point $\vec{p}_{\mathrm{int}} = \vec{s} + t_{\mathrm{int}}\vec{d}$ was then mapped to the local in-plane coordinates (u, v) aligned with the two lateral BEV axes, and the 2D aperture map was sampled at (u, v) by bilinear interpolation. Repeating over all voxels yielded a 3D segment projection mask, as shown in figure 3(a).

2.4. Patient-coordinate processing

For the patient-coordinate processing in figure 1(b), the patient CT (in HU) and segment dose volumes were directly cropped to $128 \times 192 \times 192$ at $3 \times 3 \times 3\,\mathrm{mm}^3$ voxels for model training. For additional inputs, two configurations were used: (i) a 3D segment projection (from section 2.3.2) inverse-resampled onto the patient CT grid; and (ii) a physical feature set (Kontaxis *et al* 2020) comprising the segment projection, distance-from-source (DFS), central-beamline distance (CBD), and radiological depth (RD). DFS was calculated as the Euclidean distance from the point source to each voxel. CBD represented the perpendicular distance from each voxel to the central beam axis. RD was the line integral of density along the beam direction starting at the body surface. To compute RD, CT HU were first mapped to physical density using the scanner-specific calibration curve (Schmid *et al* 2015), then a cumulative density integral was evaluated along rays parallel to the beam axis with depth initialized to zero at the surface. All physical features were computed on GPU via Cupy and cropped to $128 \times 192 \times 192$ at $3 \times 3 \times 3$ mm^3 voxels, as illustrated in figure 3(b).

2.5. Model architectures

2.5.1. BEV coordinates

For BEV coordinate methods, transformer (DoTA), Unet (C3D), and proposed CNN-ConvLSTM and CNN-Mamba models were trained and evaluated.

For DoTA, we adopted the published hyperparameters (Pastor-Serrano *et al* 2023), modifying only the number of encoder levels. Because our inputs ($256 \times 200 \times 200$) are larger than the original DoTA inputs ($96 \times 96 \times 64$), we evaluated both the original four-level downsampling configuration and a five-level downsampling variant under the same training setup. The five-level downsampling model yielded superior validation performance and was therefore used as the default configuration in our experiments.

For C3D, we followed the released configuration (Liu *et al* 2021), modifying only the number of input channels from 9 to 2 to accommodate the 3D CT and segment projection cuboid inputs.

The CNN-ConvLSTM model comprised three components, as illustrated in figures 4(a) and (b): a 2D CNN encoder, a ConvLSTM, and a 2D CNN decoder. The CNN encoder stacked four 2D convolutional layers with LeakyReLU activations (Xu *et al* 2020), compressing the two-channel BEV input of shape $(2, N_z, N_y, N_x)$ into feature maps of shape $(C, N_z, N_y/8, N_x/8)$, where $C = 64$ denotes the number of feature channels, providing a lower-resolution representation for subsequent ConvLSTM sequence modeling. Among the four convolutional layers, one used a 1-stride to enrich features, whereas the remaining three employed downsampling strides to aggregate contextual information. A single-layer ConvLSTM (Shi *et al* 2015) then modeled through-slice dependencies. At each time step t (BEV slice), the ConvLSTM module concatenated the encoded slice x_t with the previous hidden state h_{t-1}, applied a

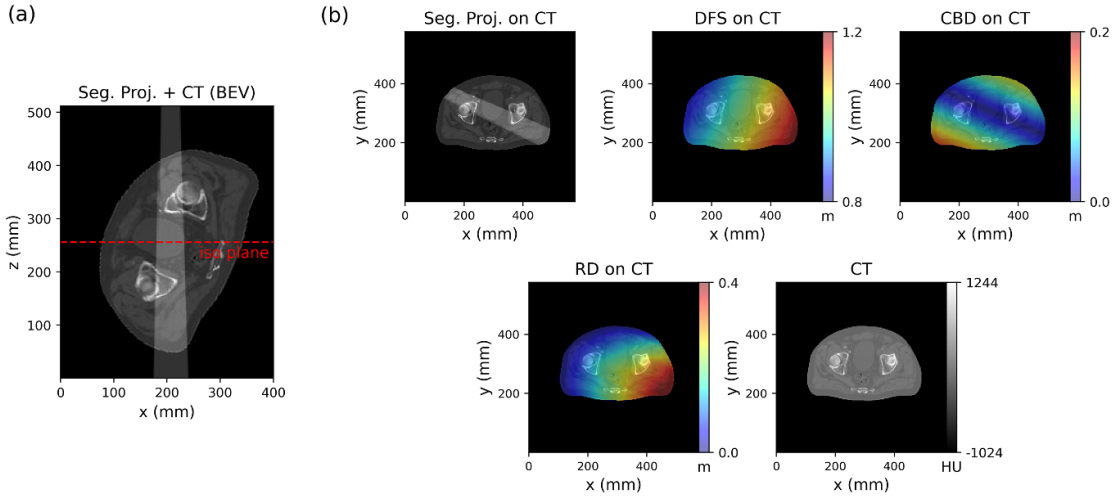

Figure 3. (a) BEV-coordinate inputs: a 3D CT cuboid covering a $40 \times 40\,\mathrm{cm}^2$ field at the isocenter plane and extending $\pm 25.6\,\mathrm{cm}$ along the beam axis about the isocenter, together with the corresponding 3D segment projection. (b) Patient-coordinate inputs: 3D physical feature volumes (segment projection, DFS, CBD, RD) and CT volumes, each of shape $128 \times 192 \times 192$ with isotropic 3 mm voxels.

3×3 2D convolution, and split the resulting $4C$-channel output into forget, input, candidate, and output gates f_t, i_t, g_t, o_t, respectively. The cell and hidden states were updated by:

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where $\odot$ is elementwise product.

The sequence of output hidden states h_t from ConvLSTM was then decoded. The CNN decoder mirrors the CNN encoder: three 2D transposed convolutional layers with LeakyReLU activations progressively upsampled the features back to the input resolution, followed by the final transposed convolution that produced a single-channel output. No skip connections were used to keep the design lightweight.

The CNN-Mamba model, illustrated in figure 4(c), shared the same 2D CNN encoder and decoder as the CNN-ConvLSTM. Then, a Spatial Mix module applied a depthwise 7×7 convolution followed by a pointwise 1×1 convolution and a LeakyReLU activation, with a residual connection. The depthwise convolution aggregated spatial neighborhood information independently per channel, while the pointwise convolution fused information across channels. This operation injected neighboring spatial context into each pixel's feature, including out-of-projection pixels, before the spatial dimensions were flattened. The enriched feature map was then projected from C to $C/2$ via a fully connected layer, passed through two stacked Mamba blocks, restored to C via a second fully connected layer, and reshaped back to $(C, N_z, N_y/8, N_x/8)$. A symmetric Spatial Mix module was applied after reshaping, redistributing the Mamba output spatially. At last, the feature map was passed to the CNN decoder to recover the full-resolution dose prediction.

2.5.2. Patient coordinates

For the patient coordinate method, we used the same C3D Unet as in the BEV-coordinate method, but with the input shape of $128 \times 192 \times 192$. Two input configurations were used with the same network: a two-channel input (3D patient CT and segment projection) and a five-channel input (3D patient CT and four physical features as described in section 2.4). To avoid ambiguity, we refer to this patient-coordinate C3D model as DeepDose-C3D.

2.6. Training

For the BEV coordinate method, to preserve the DoTA and C3D architecture designs (5 and 4 downsampling operations, respectively), the input CT and segment projection cuboids were center-cropped to $256 \times 192 \times 192$ so each dimension was divisible by $2^5 = 32$ or $2^4 = 16$, while the proposed CNN-ConvLSTM, CNN-Mamba models with three downsampling operations kept using the $256 \times 200 \times 200$ input. For the patient coordinate method, all inputs had shape $128 \times 192 \times 192$. Global maximum normalization, using the maximum segment dose and CT HU value from training data, was applied to the

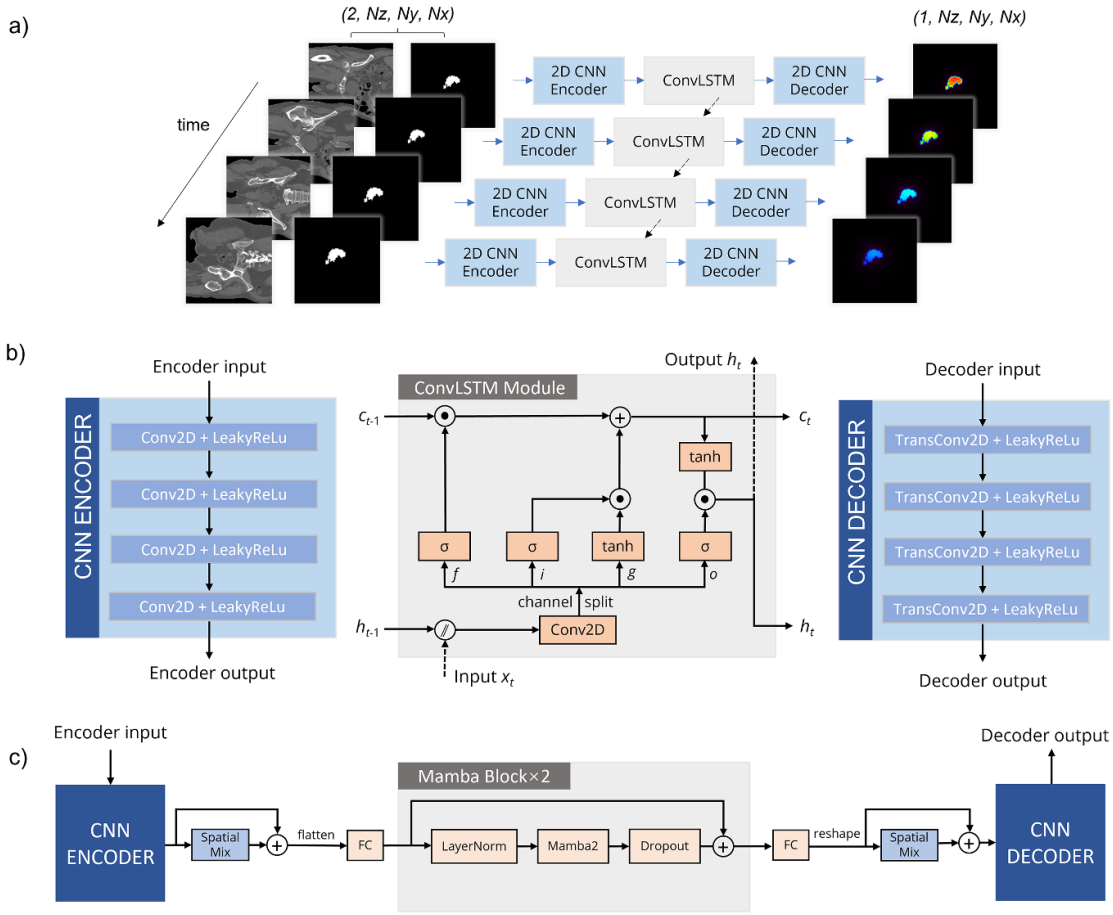

Figure 4. (a) CNN-ConvLSTM for segment dose prediction: Model inputs are CT and segment-projection cuboids arranged as a two-channel sequence $(2, N_z, N_y, N_x)$. A shared 2D CNN encodes each BEV slice, a ConvLSTM aggregates temporal context across N_z, and a 2D CNN decoder outputs per-slice dose maps, yielding a dose cuboid of shape $(1, N_z, N_y, N_x)$. (b, left) A 2D CNN encoder applies stacked `Conv2D+LeakyReLU` blocks to each time slice of the two-channel input; (b, center) a single-layer ConvLSTM performs gated updates of hidden and cell states across N_z and aggregates temporal context; (b, right) a 2D CNN decoder upsamples with `TransConv2D+LeakyReLU` blocks to recover the in-plane resolution and outputs a single-channel dose map per slice. The σ functions denote the forget, input and output gates, whereas the tanh functions denote the updating of the cell state information. The $\odot$, $\oplus$ and double-slash symbol denote element-wise multiplication, element-wise addition, and concatenation, respectively. (c) CNN-Mamba architecture: shares the 2D CNN encoder-decoder with (b); a Spatial Mix module applies a depthwise 7×7 convolution followed by a pointwise 1×1 convolution with a residual connection, encoder feature maps are then flattened, linearly projected from C to $C/2$, passed through two Mamba blocks to model through-slice dependencies, and projected back to C; a second symmetric Spatial Mix module is applied after reshaping before decoding into per-slice dose maps. The $\oplus$ denotes element-wise addition.

Table 1. Summary of input shape and voxel size across different models. Models above the dividing line use BEV coordinates, while those below use patient coordinates.

Model	Inputs	Shape	Voxel size (mm^3)
CNN-ConvLSTM	CT, seg. proj.	$256 \times 200 \times 200$	$2 \times 2 \times 2$
CNN-Mamba	CT, seg. proj.	$256 \times 200 \times 200$	$2 \times 2 \times 2$
DoTA	CT, seg. proj.	$256 \times 192 \times 192$	$2 \times 2 \times 2$
C3D	CT, seg. proj.	$256 \times 192 \times 192$	$2 \times 2 \times 2$
DeepDose-C3D	CT, seg. proj.	$128 \times 192 \times 192$	$3 \times 3 \times 3$
DeepDose-C3D	CT, four inputs	$128 \times 192 \times 192$	$3 \times 3 \times 3$

training, validation, and testing datasets. Table 1 shows an overview of the input shape and voxel size across different models.

The loss for all models was mean squared error. The C3D, DeepDose-C3D, CNN-ConvLSTM, CNN-Mamba models were implemented in PyTorch 2.8.0 using the Adam optimizer with an initial learning

rate of 10^{-3} halved if the validation loss did not decrease for 15 epochs. The CNN-Mamba implementation additionally used mamba-ssm 2.2.5 for Mamba blocks. Because transformer-based models are sensitive to hyperparameters, we adopted the published DoTA training settings: the LAMB optimizer with an initial learning rate of 10^{-3} which halved every 15 epochs. As the original DoTA implementation is publicly available in TensorFlow, we re-implemented DoTA in PyTorch 2.8.0, matching all layer-wise training parameters to the original implementation, to disentangle performance difference from different DL frameworks.

Owing to the training memory limits, we set the batch size of DoTA, C3D and DeepDose-C3D to 1; for the lightweight CNN-ConvLSTM and CNN-Mamba models, the batch size was set to 4. All models were trained on a single NVIDIA RTX A6000 GPU (48 GB) with an Intel Xeon Gold 6354 3.00 GHz CPU for one week until the validation loss stabilized. Test performance was reported for the checkpoint with the lowest validation loss.

2.7. Evaluation

After predicting segment doses with the five models, BEV-grid predictions were inverse-resampled to the patient CT grid; for patient-coordinate models, crops were restored to their original spatial positions. Segment doses were denormalized and then multiplied by their specific MU and all predicted doses were evaluated on the $3 \times 3 \times 3$ mm^3 dose grid. Considering the statistical uncertainty of segment doses (below ~1.6%) and dose grid, 3D local gamma passing rates γ_{pr} (2%/3 mm, 10%D_{max} dose cutoff) between the prediction and MC dose per segment were computed using PyMedPhys (Biggs *et al* 2022). Differences in γ_{pr} across models were also evaluated using the Friedman test with Nemenyi post hoc comparisons, and 2D dose difference maps and 1D dose profiles were used to evaluate the representative segment dose prediction case. For the full plan test, 3D local gamma pass rates γ_{pr} were calculated between the accumulated predicted dose and the MC dose across all models. Four criteria (2%/3 mm, 1%/3 mm, 2%/0 mm, and 1%/0 mm) were assessed within three regions ($D > 10\%D_{\mathrm{max}}$, the PTV, and organ at risks (OARs) (bladder and rectum)). The 3 mm distance-to-agreement was selected to align with the dose grid resolution; additionally, results at 0 mm were reported to evaluate the direct voxel-wise dose agreement. The dose volume histograms (DVHs) and selected DVH indices for representative structures were also reported: PTV ($D_{2\%}$, $D_{95\%}$), bladder ($D_{2\%}$, $V_{40\,\mathrm{Gy}}$, $V_{48\,\mathrm{Gy}}$), and rectum ($D_{2\%}$, $V_{30\,\mathrm{Gy}}$, $V_{40\,\mathrm{Gy}}$).

For the dose-calculation runtime, two components were included: (i) model inference time and (ii) pre/post-processing time. For BEV methods, preprocessing includes BEV resampling and segment projection, and postprocessing performs inverse resampling the predicted dose to the patient CT grid. For patient-coordinate methods, preprocessing included generating physical inputs, either the segment projection alone or four physical features (DFS, CBD, RD, and segment projection), and performing cropping; postprocessing restored the predicted dose to patient's original spatial position. Segment-level dose calculation times across different models were measured on a single NVIDIA RTX A6000 GPU (48 GB) paired with an Intel Xeon Gold 6354 3.00 GHz CPU. For the full-plan dose calculation, all models were executed with CUDA Graphs and automatic mixed precision enabled, using multi-batch inference to process all segments. We first investigated the effect of batch size on full plan inference time on a single NVIDIA RTX A6000 GPU (48 GB) for each model. Using the batch size that yielded the fastest inference for each model, we then evaluated cross-GPU performance by measuring total dose calculation times on a single older generation NVIDIA Quadro RTX 8000 GPU (48 GB), the NVIDIA RTX A6000 GPU (48 GB) used for training, and a latest-generation NVIDIA RTX PRO6000 Max-Q GPU (96 GB).

3. Results

3.1. Segment dose prediction results

Table 2 presents the local γ_{pr} (2%/3 mm, $D > 10\%D_{\mathrm{max}}$), model inference and pre/post-processing times, and model inference GPU memory cost for per-segment dose prediction across different models in the test dataset (3053 segments).

Among BEV coordinate methods, the γ_{pr} comparison results demonstrated that CNN-ConvLSTM, CNN-Mamba, and DoTA achieved accurate dose predictions, with mean γ_{pr} values exceeding 94%, while C3D performed slightly worse, yielding mean γ_{pr} values of approximately 92%. For dose calculation runtime, under a fixed pre/post-processing overhead of 28 ms, CNN-ConvLSTM and CNN-Mamba clearly benefited from their lightweight architectures, achieving model inference times of 51 ms and 39 ms, respectively, while requiring only 2.9 GB of GPU memory per segment.

Table 2. The local γ_{pr} (2%/3 mm, $D > 10\%D_{max}$), dose calculation runtime, and single batch inference GPU memory cost for per-segment dose prediction across different models in the test dataset. Models above the dividing line use BEV coordinates, while those below use patient coordinates.

Model	Inputs	γ_{pr} (mean ± SD, min)	Runtime (ms)		Inf. mem (GB)
			Inf.	Pre/Post	
CNN-ConvLSTM	CT, seg. proj.	95.60 ± 2.11 (85.97)	51	28	2.9
CNN-Mamba	CT, seg. proj.	95.70 ± 2.07 (88.22)	39	28	2.9
DoTA	CT, seg. proj.	94.27 ± 3.29 (73.20)	270	28	12.8
C3D	CT, seg. proj.	91.94 ± 4.44 (69.24)	462	28	14.2
DeepDoseC3D	CT, seg. proj.	86.32 ± 11.21 (29.10)	232	30	7.2
DeepDoseC3D	CT, four inputs	91.16 ± 7.88 (30.80)	236	120	9.0

For patient-coordinate methods, DeepDose-C3D using only CT and segment-projection inputs achieved a mean γ_{pr} of approximately 86%. Incorporating additional physical inputs (DFS, CBD, RD) increased the mean γ_{pr} to 91%, but raised the pre/post-processing time from 30 ms to 120 ms. However, the larger standard deviation and lower worst-case γ_{pr} indicate that DeepDose-C3D remains less robust than the BEV-coordinate models. Given its substantially lower performance compared to the multi-input configuration, the DeepDose-C3D model relying solely on CT and segment-projection inputs is omitted from the remaining results.

Figure 5 shows a typical segment dose prediction case (assessed by γ_{pr}) in the test dataset, across four BEV models and the DeepDose-C3D (CT + four physical inputs) model. BEV models tended to show dose differences concentrated at high-dose edges with red/blue alternation, while DeepDose-C3D exhibited a mild in-field negative bias, with less pronounced edge errors. The five models showed local γ_{pr} (2%/3 mm, $D > 10\%D_{max}$) higher than 90%.

Figure 6 presents an outlier DeepDose-C3D (CT + four physical inputs) segment dose prediction from the test set, compared against the same-segment predictions of CNN-ConvLSTM, CNN-Mamba, DoTA, and C3D. DeepDose-C3D exhibited a visible negative bias in the high-dose region, causing γ_{pr} to be lower than 40%. In contrast, BEV models still had residuals at high-dose edges and exhibited mostly $\Gamma < 1$.

Figure 7 further shows patient specific boxplots for γ_{pr} of predicted segment doses from CNN-ConvLSTM, CNN-Mamba, DoTA, C3D, and DeepDose-C3D models in the test dataset. Similar to the results in table 2, DeepDose-C3D showed more low outliers compared with BEV models in three of ten test patients. The Friedman test indicated a significant overall difference among the five models, subsequent Nemenyi post-hoc comparisons also showed all pairwise differences significant at $\alpha = 0.05$ (all off diagonal $p < 0.05$). Figure 8 shows dose profiles of all models for the segment dose shown in figure 5. In the dose regions of D> $1\%D_{max}$, all models align closely with the MC ground truth from dose profiles.

3.2. Plan dose prediction results

Table 3 shows the local γ_{PR} of accumulated plan dose (described in section 2.2) predictions evaluated in three regions from the five models across 10 test patients. All models achieved γ_{PR} (1%/3 mm, $D > 10\%D_{max}$) higher than 99%. Performance differences become more pronounced under stricter criteria (1%/0 mm), where CNN-ConvLSTM, CNN-Mamba and DoTA generally outperform the other models, especially in the PTV region. Figure 9 presents a nominal plan dose prediction case (P016) from the five models. Both BEV models and DeepDose-C3D performed with excellent accuracy. The predicted dose distributions from all models show good agreement with the MC reference, with isodose contours that generally match. The corresponding dose-difference maps showed deviations within 1.5 Gy or $2\%D_{max}$ over most regions, and the γ_{PR} (1%/3 mm, $D > 10\%D_{max}$) were all above 99%.

Figure 10 further compares the DVHs of the MC reference and the predicted plan doses from the five models for patient P016. The DVH curves for the PTV, bladder, rectum, and body overlap well. Consistent with the DVH visual agreement, table 4 shows that across the 10 test patients all DVH metrics (PTV $D_{2\%}$, $D_{95\%}$; bladder $D_{2\%}$, V_{40}, V_{48}; rectum $D_{2\%}$, V_{30}, V_{40}) differed from the MC reference by less than 0.5 Gy or 0.5%.

3.3. Full plan dose calculation runtime

For the full-plan dose calculation, we first determined an efficient batch size for each architecture by performing multi-batch inference on a representative test case (P016, 254 segments) on an RTX A6000

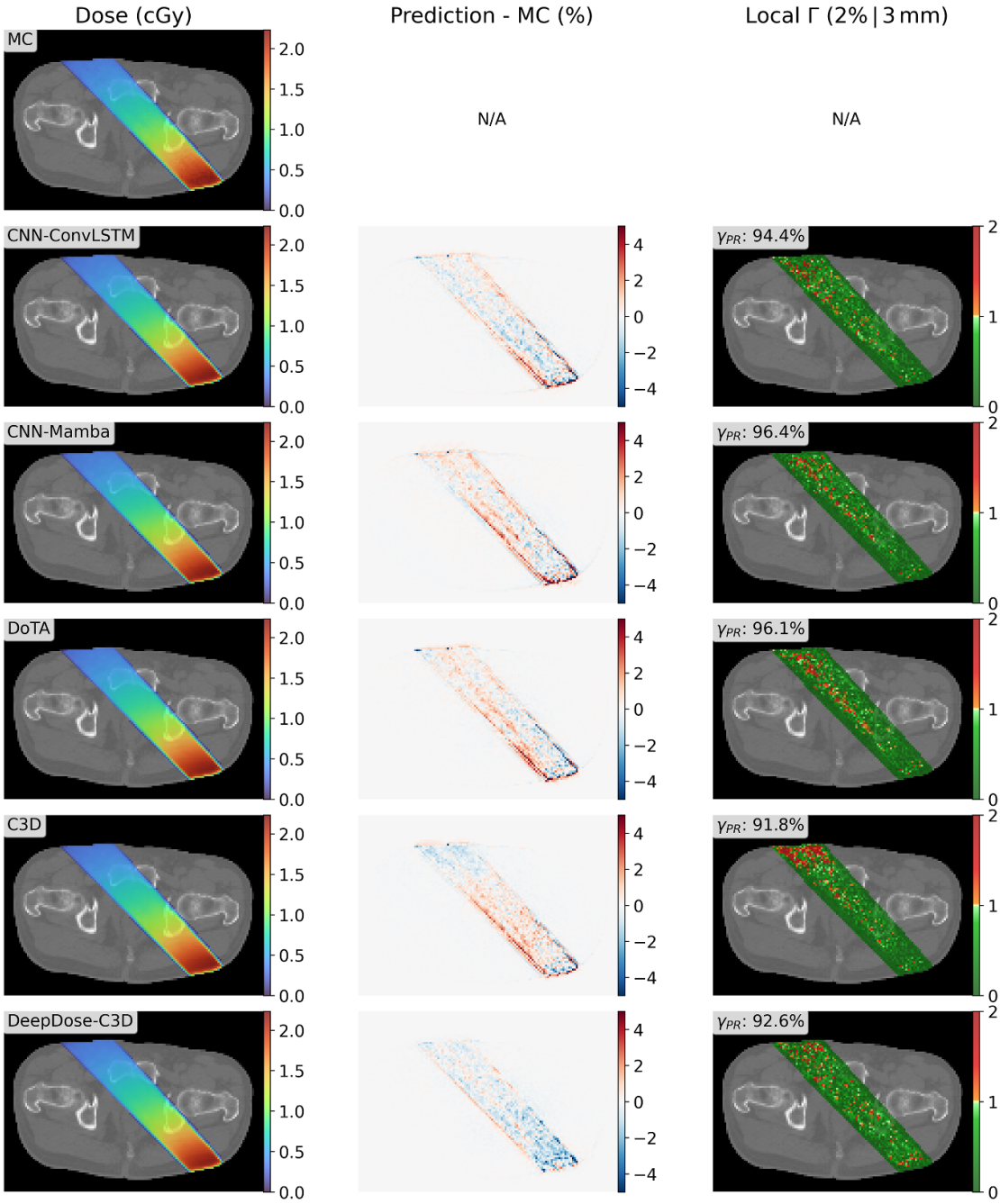

Figure 5. A typical segment dose prediction case in the test dataset on a representative axial slice. Left: dose map (cGy). Middle: prediction-MC percentage dose difference map with a shared $\pm 5\%$ scale. Right: local Γ maps at 2%/3 mm with $D > 10\%D_{\max}$. Rows: MC (reference; N/A for dose difference and Γ map), CNN-ConvLSTM, CNN-Mamba, DoTA, C3D, and DeepDose-C3D (CT + four physical inputs).

GPU (figure 11). The highlighted points indicate the batch size yielding the shortest total inference time, and these per-model batch sizes were then fixed for the cross-GPU evaluation. Table 5 shows the average per-plan inference time and pre/post-processing time on 10 test patients (average 305 segments per plan) across three GPUs (Quadro RTX8000, RTX A6000, and RTX PRO6000 Max-Q). CNN-ConvLSTM was consistently the fastest model, followed by CNN-Mamba, whereas DoTA, C3D, and DeepDose-C3D were substantially slower and, for DeepDose-C3D, additionally dominated by pre/post-processing because of its multiple physical input generation.

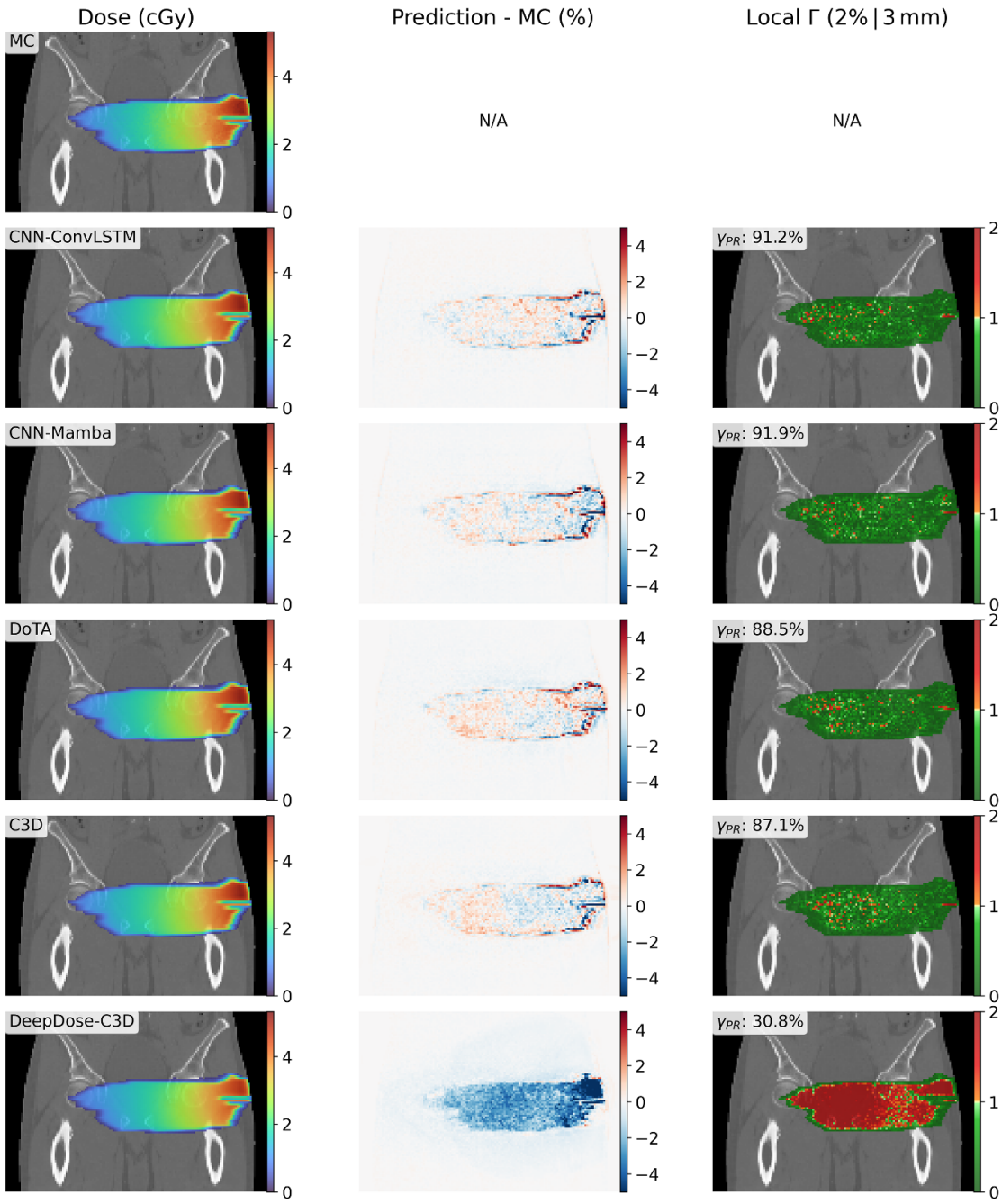

Figure 6. An outlier segment dose prediction case from DeepDose-C3D in the test dataset on a representative coronal slice. Left: dose map (cGy). Middle: prediction-MC percentage dose difference map with a shared $\pm 5\%$ scale. Right: local Γ maps at 2%/3 mm with $D > 10\%D_{\max}$. Rows: MC (reference; N/A for dose difference and Γ map), CNN-ConvLSTM, CNN-Mamba, DoTA, C3D, and DeepDose-C3D (CT + four physical inputs).

4. Discussion

This work presented a comparison study of DL models for photon segment dose prediction under a unified dataset, preprocessing, and evaluation protocol, covering both BEV and patient-coordinate representations and multiple network architectures. All five evaluated models achieved high overall accuracy, with average local γ_{PR} (2%/3 mm, $D > 10\%D_{\max}$) values consistently above 90% for segment doses and γ_{PR} (1%/3 mm, $D > 10\%D_{\max}$) above 99% for full plan doses. For the segment dose prediction, BEV methods (CNN-ConvLSTM, CNN-Mamba, DoTA) showed very similar dosimetric performance, with C3D performing slightly worse. In contrast, the patient-coordinate DeepDose-C3D required additional physics-based inputs (DFS, CBD, RD) to approach similar accuracy but still exhibited larger variability and worse outliers (similar to the report from DeepDose (Kontaxis *et al* 2020)). For the full plan dose

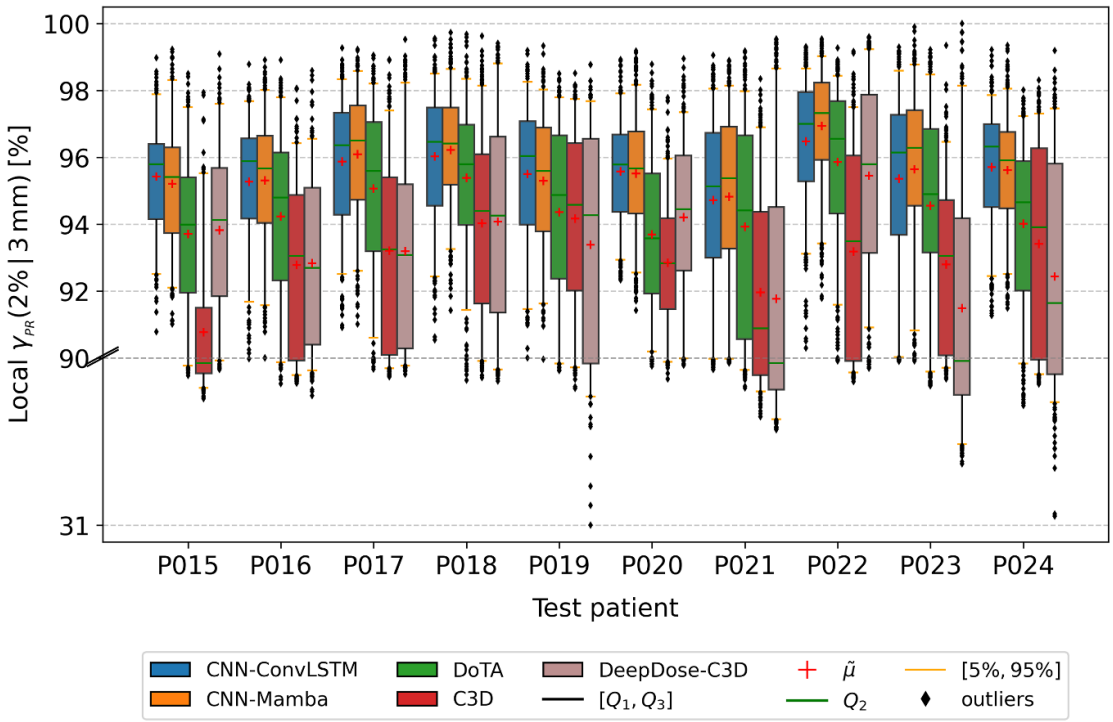

Figure 7. Patient specific boxplots for local γ_{PR} (2%/3 mm, $D > 10\%D_{max}$) of predicted segment dose from CNN-ConvLSTM, CNN-Mamba, DoTA, C3D, and DeepDose-C3D (CT + four physical inputs) models in the test dataset. The red cross marks the mean, and the green line marks the median. The box spans the 25th–75th percentile range. Orange whiskers extend to the 5th and 95th percentiles, and black points indicate outliers. For readability, the *y*-axis region below 90% is compressed using a broken scale.

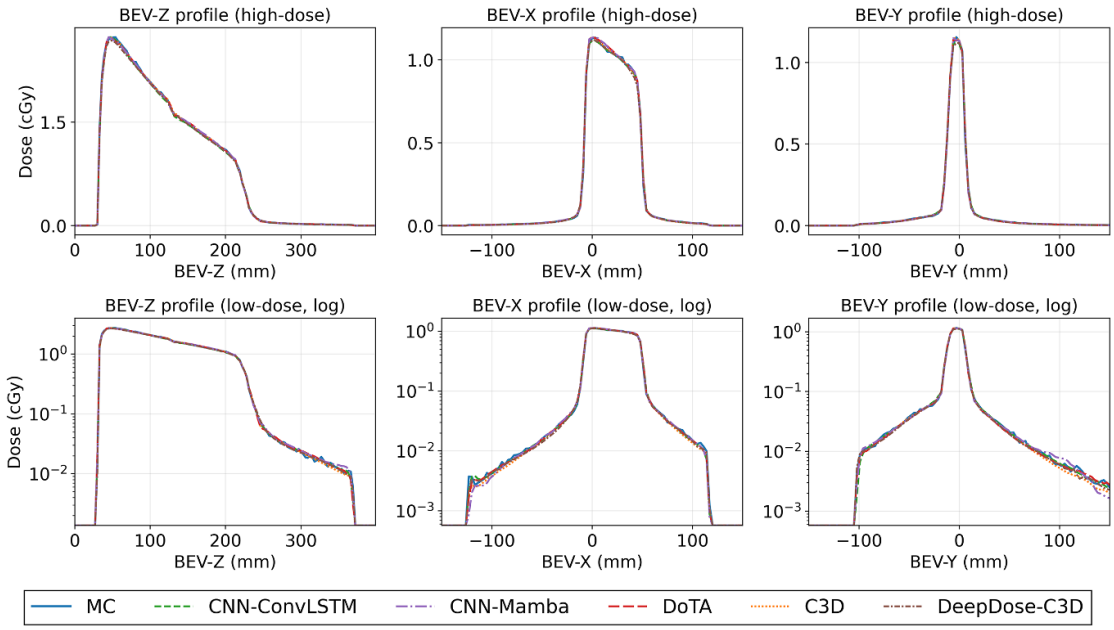

Figure 8. Comparison of 1D dose profiles between the MC segment dose and predictions from five models along the *Z* (beam direction), *X*, and *Y* (lateral directions) axes in the BEV coordinate for the representative segment dose shown in figure 5. All profiles intersect the isocenter. The top row uses a linear scale to illustrate high-dose regions, whereas the bottom row utilizes a logarithmic scale to highlight low-dose regions.

prediction, both methods showed convincing accuracy from γ_{PR} (1%/3 mm) and DVH analyses, all models reproduced MC-based target coverage and OAR sparing within about 0.5 Gy or 0.5% points, suggesting that the remaining discrepancies are clinically small.

Table 3. Mean local γ_{PR} (2%/3 mm, 1%/3 mm, 2%/0 mm, 1%/0 mm) for full-plan dose evaluated in three regions ($D > 10\%\ D_{max}$, PTV, and OARs including Bladder and Rectum) across 10 test patients for each of the five models. Values are presented as mean ± SD (%).

$D > 10\%D_{max}$				
Model	2%/3 mm	1%/3 mm	2%/0 mm	1%/0 mm
CNN-ConvLSTM	99.91 ± 0.05	99.62 ± 0.12	89.57 ± 2.17	67.73 ± 3.56
CNN-Mamba	99.83 ± 0.08	99.46 ± 0.18	86.79 ± 2.22	63.22 ± 3.20
DoTA	99.90 ± 0.05	99.58 ± 0.13	88.96 ± 2.17	64.90 ± 3.33
C3D	99.83 ± 0.08	99.26 ± 0.31	85.99 ± 2.97	58.99 ± 5.64
DeepDose-C3D	99.82 ± 0.10	99.07 ± 0.67	87.43 ± 6.34	57.03 ± 10.94
PTV				
Model	2%/3 mm	1%/3 mm	2%/0 mm	1%/0 mm
CNN-ConvLSTM	99.94 ± 0.06	98.57 ± 0.92	99.73 ± 0.15	90.43 ± 3.08
CNN-Mamba	99.95 ± 0.07	99.03 ± 0.73	99.60 ± 0.28	90.59 ± 2.38
DoTA	99.95 ± 0.06	98.41 ± 0.82	99.54 ± 0.25	87.85 ± 3.32
C3D	99.91 ± 0.11	96.52 ± 3.64	99.30 ± 0.61	82.26 ± 10.62
DeepDose-C3D	99.73 ± 0.54	94.00 ± 8.82	98.62 ± 1.83	76.15 ± 19.86
Bladder & Rectum				
Model	2%/3 mm	1%/3 mm	2%/0 mm	1%/0 mm
CNN-ConvLSTM	99.99 ± 0.01	99.88 ± 0.10	88.83 ± 5.65	69.66 ± 7.83
CNN-Mamba	99.99 ± 0.01	99.88 ± 0.08	87.30 ± 6.02	66.78 ± 7.36
DoTA	99.99 ± 0.01	99.68 ± 0.37	88.77 ± 4.93	66.42 ± 7.63
C3D	99.99 ± 0.01	99.23 ± 1.09	84.53 ± 6.75	59.68 ± 12.83
DeepDose-C3D	99.98 ± 0.02	99.47 ± 0.54	93.21 ± 5.57	69.35 ± 15.30

From the efficiency perspective, the proposed lightweight BEV models (CNN-ConvLSTM and CNN-Mamba) achieved per-segment inference times of 51 ms and 39 ms, respectively, with only 2.9 GB GPU memory, compared with inference times of 270–462 ms and 12–14 GB GPU memory for DoTA and C3D. Additional physics inputs in DeepDose-C3D increased pre/post-processing overhead from 30 ms to 120 ms per segment, whereas BEV models operated with a fixed 28 ms overhead using only CT and segment projections. After optimizing batch size for each model, plan-level dose calculation runtimes on modern GPUs were reduced to a few seconds per plan (e.g. total dose calculation in the order of 5–6 s on an RTX PRO6000 Max-Q for $\sim$ 305 segments) for CNN-ConvLSTM and CNN-Mamba. It should be noted that, although CNN-Mamba exhibited faster single-batch inference, CNN-ConvLSTM achieved shorter total plan inference times after the batch size exceeded 1. Overall, these findings suggest that BEV-coordinate modeling with lightweight sequence models such as CNN-ConvLSTM or CNN-Mamba offer a favorable balance of accuracy, robustness, and computational cost for fast photon dose calculation.

While GPU MC methods typically require tens of seconds for full plan calculations, a recent GPU MC engine for photon dose calculation (Liu *et al* 2025) has demonstrated second-level performance, achieving $> 99.23\%$ γ pass rates at 3%/3 mm in $\sim$3 s ($> 99.73\%$ γ pass rates at 2%/3 mm in $\sim$15.4 s) for VMAT plans on a single RTX 4080 GPU. In this context, our DL models were evaluated under more stringent criteria (local 1%/3 mm γ for full-plan doses). The CNN-ConvLSTM and CNN-Mamba models achieved average γ_{PR} values of 99.6% and 99.5%, respectively, indicating accuracy comparable to that of Liu *et al* at tighter thresholds. For online real-time adaptive radiotherapy workflows, however, conventional MC dose calculation must operate on CT images, so additional steps such as in-house synthetic CT generation and image registration are required when only on-board images (e.g. Magnetic resonance imaging (MRI)) are available, adding latency and potential sources of error. In contrast, DL-based methods have shown feasibility of predicting dose directly from MRI (Li *et al* 2025, Xiao *et al* 2025), thereby bypassing synthetic CT generation and offering a pathway closer to truly real-time adaptive radiotherapy.

Regarding some recent advanced DL photon dose methods, Fan *et al.* proposed DeepBEVDose, a 2D BEV U-Net mapping CT and fluence to dose, achieving average global 3%/3 mm γ passing rates above 95% for plan doses, with the model inference speed of 1 s per-beam, demonstrating the potential for online IMRT optimization (Fan *et al* 2025). While our DL framework designed for VMAT, its

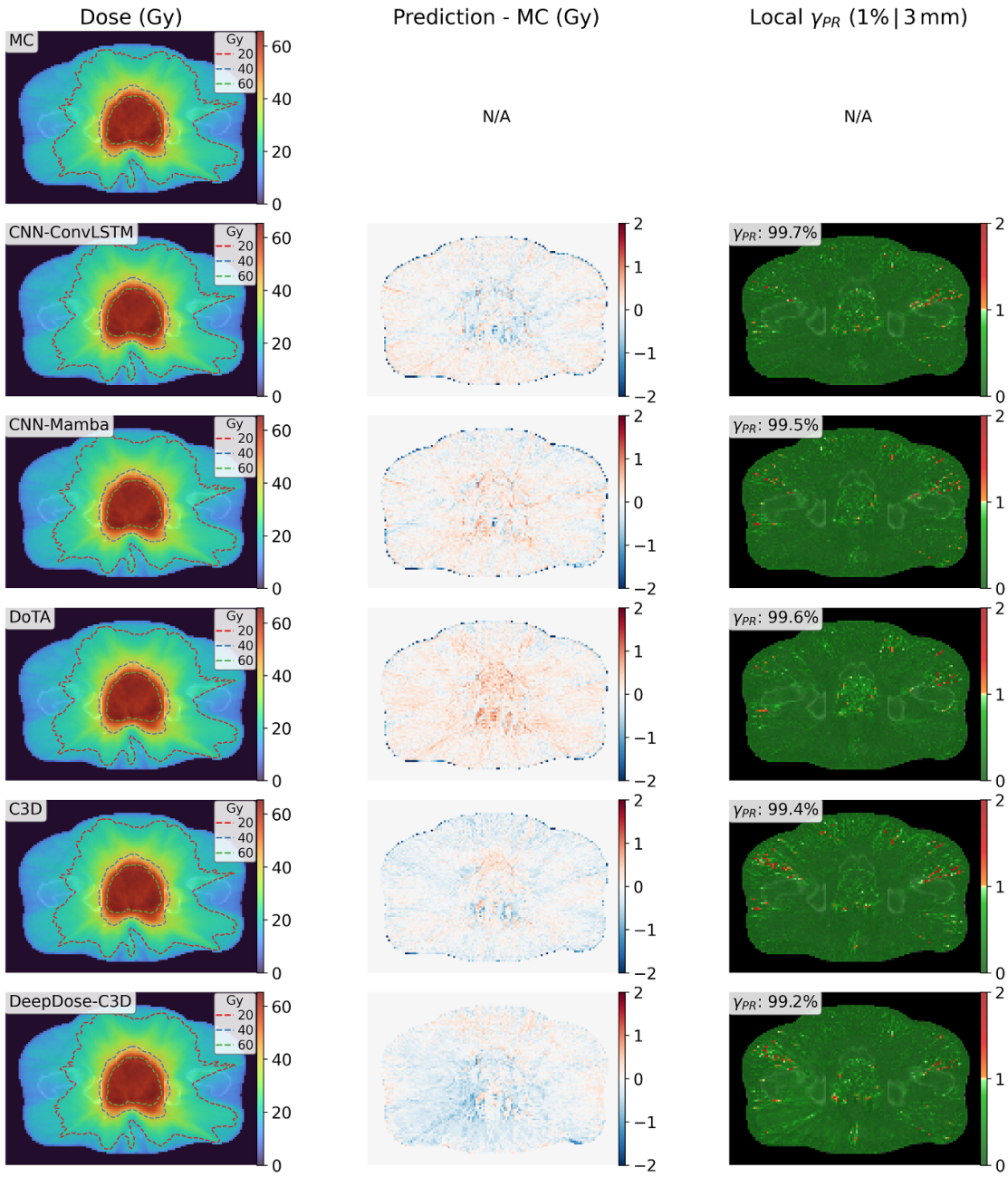

Figure 9. A nominal plan dose prediction case (P016) in the test dataset on a representative axial slice. Left: dose map (Gy) with the 20, 40, and 60 Gy isodose lines. Middle: prediction-MC dose difference map with ± 2 Gy. Right: local Γ maps at 1%/3 mm with $D > 10\%D_{\max}$. Rows: MC (reference; N/A for dose difference and Γ map), CNN-ConvLSTM, CNN-Mamba, DoTA, C3D, and DeepDose-C3D (CT + four physical inputs).

BEV formulation is compatible with IMRT segment dose prediction, suggesting a path toward a unified BEV-based framework in the future. Witte *et al.* developed a 2D convolutional recurrent network for VMAT dose calculation, achieving global 2%/2 mm γ pass rates of 99.6% with an average runtime of 8 s per plan (Witte and Sonke 2024). In contrast to our BEV resampling, their method models table, mid-gantry, and collimator rotations as rigid CT transformations, and predicts spline coefficients to reconstruct the dose, offering a more lightweight geometry encoding and a potential extension direction. Yan *et al.* focused on online quality assurance, encoding fluence maps into the CT domain with explicit delivery geometry and using a 3D U-Net for secondary dose verification, reaching 3%/2 mm γ pass rates of 99.9% with runtimes of 15 ms per plan (Yan *et al* 2024). The method enables ultra-fast full-plan verification but is limited to accumulated dose and does not provide segment-wise predictions. As these three methods are not open-source and target slightly different problem settings, we did not include them in our quantitative comparison, but regard them as complementary and relevant directions for future

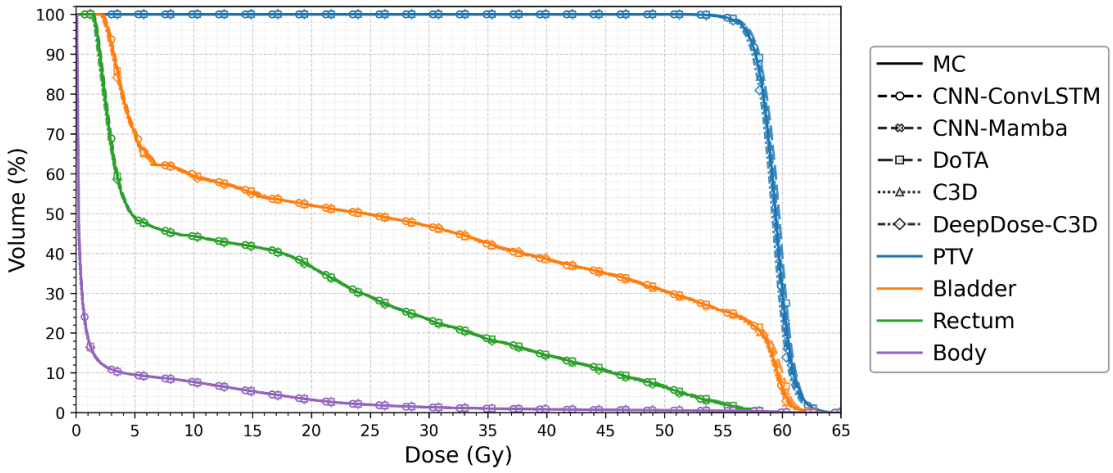

Figure 10. DVH comparison between the MC reference and the five models (CNN-ConvLSTM, CNN-Mamba, DoTA, C3D, and DeepDose-C3D) for a nominal patient case (P016).

Table 4. Plan DVH metrics (mean±SD across 10 test patients) after PTV $D_{95\%}$ normalization to 0.95 × 60 Gy.

	PTV		Bladder			Rectum		
Model	$D_{2\%}$ (Gy)	$D_{95\%}$ (Gy)	$D_{2\%}$ (Gy)	V_{40} (%)	V_{48} (%)	$D_{2\%}$ (Gy)	V_{30} (%)	V_{40} (%)
MC	63.6 ± 1.5	57.0 ± 0.0	60.1 ± 1.7	24.4 ± 10.2	19.9 ± 9.1	59.0 ± 2.0	39.9 ± 9.8	29.0 ± 7.0
CNN-ConvLSTM	63.4 ± 1.5	56.9 ± 0.1	60.0 ± 1.7	24.3 ± 10.1	19.9 ± 9.2	58.9 ± 1.9	39.9 ± 9.9	29.0 ± 7.1
CNN-Mamba	63.6 ± 1.5	57.1 ± 0.1	60.2 ± 1.7	24.4 ± 10.1	20.0 ± 9.1	59.1 ± 2.0	40.0 ± 9.9	29.0 ± 7.2
DoTA	63.5 ± 1.6	57.1 ± 0.2	60.2 ± 1.8	24.3 ± 10.1	20.0 ± 9.2	59.0 ± 2.0	40.0 ± 9.9	29.0 ± 7.1
C3D	63.2 ± 1.6	56.8 ± 0.2	59.9 ± 1.9	24.2 ± 10.1	19.8 ± 9.1	58.8 ± 1.9	39.8 ± 9.8	28.8 ± 7.0
DeepDose-C3D	63.4 ± 1.7	56.9 ± 0.3	60.0 ± 1.8	24.3 ± 10.2	19.9 ± 9.2	58.9 ± 2.1	40.0 ± 9.8	29.0 ± 7.1

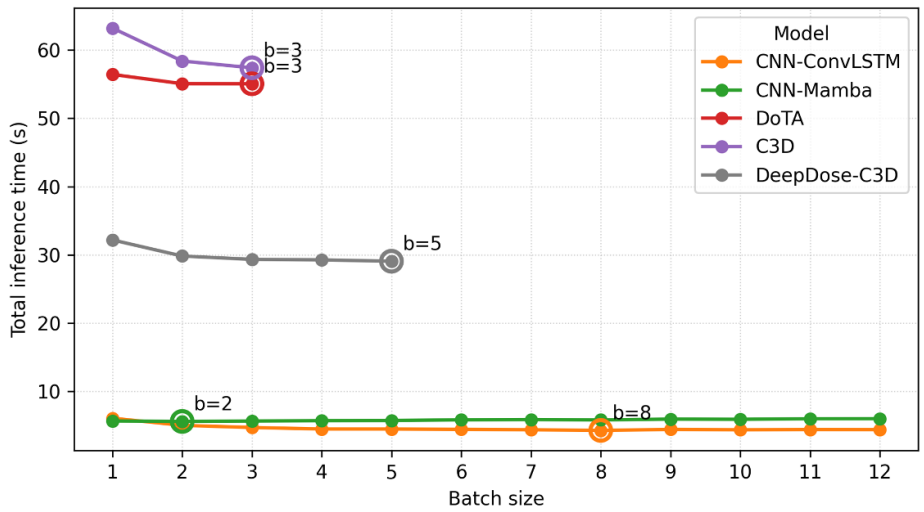

Figure 11. Total inference time as a function of batch size for the five models on a single test plan (P016, 254 segments) using an RTX A6000 GPU (48 GB). Batch size was increased up to the GPU-memory limit for each model. The highlighted markers indicate the batch size yielding the shortest total inference time for each model. For visualization, CNN-ConvLSTM and CNN-Mamba are shown up to batch size 12, as the total inference time varies only marginally beyond this value.

Table 5. Total dose calculation time on three different GPUs. Average inference and pre/post-processing time per plan for 10 test patients (average 305 segments per plan).

	Quadro RTX8000 (48 GB)		RTX A6000 (48 GB)		RTX PRO6000 Max-Q (96 GB)	
Model	Inf. (s)	Pre/post (s)	Inf. (s)	Pre/post (s)	Inf. (s)	Pre/post (s)
CNN-ConvLSTM	7.5	7.1	5.0	5.7	2.4	3.1
CNN-Mamba	20.8	7.1	6.5	5.7	3.1	3.1
DoTA	107.9	7.1	65.2	5.7	30.5	3.1
C3D	108.8	7.1	67.3	5.7	35.6	3.1
DeepDose-C3D	54.4	45.8	35.4	36.9	16.6	18.8

work. Nevertheless, global γ comparisons against three other recent DL segment dose calculation studies (Pastor-Serrano *et al* 2023, Liang *et al* 2024, Schneider *et al* 2025), along with patient-specific boxplots for global γ passing rates (2%/2 mm, $D > 10\%D_{\max}$) of the test segment doses, are provided in the supplementary materials to allow for reference comparison.

At last, although our MC simulations incorporated realistic photon beam characteristics including MLC leaf and jaw positions from clinical VMAT plans and energy spectrum from the IAEA ELEKTA_PRECISE_6MV phase space, they do not correspond to a commissioned clinical MC engine. Consequently, systematic discrepancies remain between our reference doses and those obtained in a complete clinical workflow that includes machine-specific commissioning. For eventual clinical use in rapid plan re-optimization or independent secondary dose calculation, additional work on phase-space correction (Martins *et al* 2019), more comprehensive MLC modeling (e.g. tongue-and-groove, and interleaf leakage) and explicit consideration of collimator rotation within the Geant4 MC code, or alternatively segment dose extraction directly from clinically commissioned dose engines, will be required. Experimental validation using physical measurements (e.g. diode array or film) is also essential for ultimate clinical commissioning. Moving forward, we plan to experimentally validate the best-performing model to confirm its clinical deliverability. Another limitation of this study is that all models were trained and evaluated on a single-site dataset; thus, their generalizability to broader anatomical sites, diverse imaging vendors, and multi-institutions requires further investigation. Since the present models input raw HU values, dosimetric accuracy is inherently tied to the scanner-specific HU-to-density calibration of the training data; to ensure portability for future multi-institutional deployment, models trained with site-specific calibrated density inputs rather than raw HU values will be recommended.

5. Conclusions

This comparison study shows that both BEV- and patient-coordinate DL methods can achieve accurate photon plan dose calculation, with BEV-based approaches demonstrating more robust performance in segment dose calculation. Besides, BEV methods with lightweight sequence models (CNN-ConvLSTM and CNN-Mamba) achieved comparable accuracy with substantially lower computational cost and second-level plan runtimes, providing a more advantageous accuracy-efficiency trade-off for fast photon dose calculation.

Acknowledgments

The work of Fan Xiao was supported by China Scholarship Council (No. 202308440107), and this project was supported by Deutsche Forschungsgemeinschaft (No. 469106425).

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors. The code for the models and methods presented will be available at: https://github.com/LMUK-RADONC-PHYS-RES/DL-segment-dose-calculation.

Conflict of interest

The Department of Radiation Oncology of the University Hospital of LMU Munich has research agreements with Brainlab, Elekta and ViewRay.

ORCID iDs

Fan Xiao 0000-0002-7502-0730
Niklas Wahl 0000-0002-1451-223X
Guillaume Landry 0000-0003-1707-4068

References

Bai T, Wang B, Nguyen D and Jiang S 2021 Deep dose plugin: towards real-time Monte Carlo dose calculation through a deep learning-based denoising algorithm *Mach. Learn.: Sci. Technol.* **2** 025033
Biggs S *et al* 2022 Pymedphys: a community effort to develop an open, python-based standard library for medical physics applications *J. Open Source Softw.* **7** 4555
Capote R, Jeraj R, Ma C, Rogers D W, Sánchez-Doblado F, Sempau J, Seuntjens J and Siebers J 2006 Phase-space database for external beam radiotherapy. Summary report of a consultants' meeting
Chen M, Lin J, Park Y, Lin M-H, Pompos A, Godley A and Lu W 2025 Automatic, machine-agnostic, convolution-based beam and fluence modeling for Monte Carlo independent dose calculation *Med. Phys.* **52** e17822
Cheng B, Xu Y, Li S, Ren Q, Pei X, Men K, Dai J and Xu X G 2023 Development and clinical application of a GPU-based Monte Carlo dose verification module and software for 1.5 T MR-LINAC *Med. Phys.* **50** 3172–83
Fan J, Zhu X, Wang J, Men K, Dai J, Hu W and Liu Z 2025 A novel dose calculation system implemented in image domain *Med. Phys.* **52** 5039–50
Green O L *et al* 2018 First clinical implementation of real-time, real anatomy tracking and radiation beam control *Med. Phys.* **45** 3728–40
Gu A and Dao T 2024 Mamba: linear-time sequence modeling with selective state spaces *1st Conf. on Language Modeling*
Hissoiny S, Ozell B, Bouchard H and Després P 2011 GPUMCD: a new GPU-oriented Monte Carlo dose calculation platform *Med. Phys.* **38** 754–64
Hussein M, Heijmen B J, Verellen D and Nisbet A 2018 Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations *Br. J. Radiol.* **91** 20180270
Keall P J, El Naqa I, Fast M F, Hewson E A, Hindley N, Poulsen P, Sengupta C, Tyagi N and Waddington D E 2025 Critical review: real-time dose-guided radiation therapy *Int. J. Radiat. Oncol. Biol. Phys.* **122** 787–801
Kontaxis C, Bol G, Lagendijk J and Raaymakers B 2015 A new methodology for inter-and intrafraction plan adaptation for the MR-linac *Phys. Med. Biol.* **60** 7485
Kontaxis C, Bol G, Lagendijk J and Raaymakers B 2020 DeepDose: Towards a fast dose calculation engine for radiation therapy using deep learning *Phys. Med. Biol.* **65** 075013
Li M, Winterhalter C, Li X, Safai S, Lomax A and Zhang Y 2025 A proof-of-concept study of direct magnetic resonance imaging-based proton dose calculation for brain tumors via neural networks with Monte Carlo-comparable accuracy *Phys. Imaging Radiat. Oncol.* **35** 100806
Li X, Zhang L, Yang J and Teng F 2024 Role of artificial intelligence in medical image analysis: A review of current trends and future directions *J. Med. Biol. Eng.* **44** 231–43
Li Y, Ding S, Wang B, Liu H, Huang X and Song T 2021 Extension and validation of a GPU-Monte Carlo dose engine gDPM for 1.5 T MR-LINAC online independent dose verification *Med. Phys.* **48** 6174–83
Liang B, Xia W, Wei R, Xu Y, Liu Z and Dai J 2024 A deep learning-based dose calculation method for volumetric modulated arc therapy *Radiat. Oncol.* **19** 141
Liu S, Zhang J, Li T, Yan H and Liu J 2021 A cascade 3D U-Net for dose prediction in radiotherapy *Med. Phys.* **48** 5574–82
Liu Z, Wang Y, Han Y, Hu P, Zheng C, Yan B and Yang Y 2025 A GPU-accelerated Monte Carlo dose engine for external beam radiotherapy *Med. Phys.* **52** e17899
Lombardo E *et al* 2024 Real-time motion management in MRI-guided radiotherapy: current status and AI-enabled prospects *Radiother. Oncol.* **190** 109970
Martins J C, Saxena R, Neppl S, Alhazmi A, Reiner M, Veloza S, Belka C and Parodi K 2019 Optimization of phase space files from clinical linear accelerators *Phys. Med.* **64** 54–68
Neishabouri A, Bauer J, Abdollahi A, Debus J and Mairani A 2025 Real-time adaptive proton therapy: an AI-based spatio-temporal mono-energetic dose calculation model (CC-LSTM) *Comput. Biol. Med.* **188** 109777
Okuta R, Unno Y, Nishino D, Hido S and Loomis C 2017 Cupy: a numpy-compatible library for nvidia gpu calculations *Proc. of Workshop on Machine Learning Systems (LearningSys) in The Thirty-First Annual Conf. on Neural Information Processing Systems (NIPS)*
Onizuka R, Araki F and Ohno T 2018 Monte Carlo dose verification of VMAT treatment plans using Elekta Agility 160-leaf MLC *Phys. Med.* **51** 22–31
Otto K 2008 Volumetric modulated arc therapy: IMRT in a single gantry arc *Med. Phys.* **35** 310–7
Panettieri V, Barsoum P, Westermark M, Brualla L and Lax I 2009 AAA and PBC calculation accuracy in the surface build-up region in tangential beam treatments. Phantom and breast case study with the Monte Carlo code PENELOPE *Radiother. Oncol.* **93** 94–101
Paschal H M, Kabat C N, Papaconstadopoulos P, Kirby N A, Myers P A, Wagner T D and Stathakis S 2022 Monte Carlo modeling of the Elekta Versa HD and patient dose calculation with EGSnrc/BEAMnrc *J. Appl. Clin. Med. Phys.* **23** e13715
Pastor-Serrano O, Dong P, Huang C, Xing L and Perkó Z 2023 Sub-second photon dose prediction via transformer neural networks *Med. Phys.* **50** 3159–71

Rabe M, Kurz C, Thummerer A and Landry G 2025 Artificial intelligence for treatment delivery: image-guided radiotherapy *Strahlenther. Onkol.* **201** 283–97

Renaud M-A, Roberge D and Seuntjens J 2015 Latent uncertainties of the precalculated track Monte Carlo method *Med. Phys.* **42** 479–90

Schmid S, Landry G, Thieke C, Verhaegen F, Ganswindt U, Belka C, Parodi K and Dedes G 2015 Monte Carlo study on the sensitivity of prompt gamma imaging to proton range variations due to interfractional changes in prostate cancer patients *Phys. Med. Biol.* **60** 9329

Schneider M, Gutwein S, Mönnich D, Gani C, Fischer P, Baumgartner C F and Thorwarth D 2025 Development and comprehensive clinical validation of a deep neural network for radiation dose modelling to enhance magnetic resonance imaging guided radiotherapy *Phys. Imaging Radiat. Oncol.* **33** 100723

Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K and Woo W-c 2015 Convolutional LSTM network: A machine learning approach for precipitation nowcasting *Advances in Neural Information Processing Systems* p 28

Song T, Zhou L and Li Y 2023 Cross-engine transformation-based fast dose calculation for MRI-Linac online treatment planning *Med. Phys.* **50** 2429–37

Witte M and Sonke J-J 2024 A deep learning based dynamic arc radiotherapy photon dose engine trained on Monte Carlo dose distributions *Phys. Imaging Radiat. Oncol.* **30** 100575

Xiao F, Cai J, Zhou X, Zhou L, Song T and Li Y 2022 TransDose: a transformer-based UNet model for fast and accurate dose calculation for MR-LINACs *Phys. Med. Biol.* **67** 125013

Xiao F, Radonic D, Wahl N, Delopoulos N, Thummerer A, Corradini S, Belka C, Dedes G, Kurz C and Landry G 2025 Deep learning-based synthetic-ct-free photon dose calculation in mr-guided radiotherapy: a proof-of-concept study *Med. Phys.* **52** e70106

Xu J, Li Z, Du B, Zhang M and Liu J 2020 Reluplex made more practical: Leaky relu *2020 IEEE Symp. on Computers and Communications (ISCC)* (IEEE) pp 1–7

Yan S, Maniscalco A, Wang B, Nguyen D, Jiang S and Shen C 2024 Quality assurance for online adaptive radiotherapy: a secondary dose verification model with geometry-encoded u-net *Mach. Learn.: Sci. Technol.* **5** 045013