D2T2- Multimodal Automated Planning for Brachytherapy Supplementary Material

Lance C. Moore, Aranyo Mitra, Ryan Truong, Karoline Kallis
Kelly Kisling, Sandra M. Meyers, Nuno Vasconcelos
University of California, San Diego
lamoore@ucsd.edu

In this supplement, we provide more details regarding the model architecture, datasets, evaluation metrics, and visualizations which complement the main paper.

1. Architecture and Training Details

Architecture: **D2T2** is built off of the standard DiT architecture, with an embedding depth E and a layer depth L. For this project, we have selected $E = 1024$ and $L = 24$, corresponding to the DiT-L configuration from the original paper. While larger models were considered, they proved to be too computationally burdensome for the available compute resources. As depicted in Figure 1, each DiT block in **D2T2** contains a self-attention layer and a cross-attention layer where the self-attention layer conducts attention between dwell location tokens and the cross-attention layer conditions the dwell location tokens on the latents of the treatment images. This structure allows dwell tokens to integrate global image context while maintaining a flexible, variable-length representation across applicator types.

adaLN Normalization: As with the original DiT model, we use adaptive layer normalization (adaLN) prior to projecting tokens into self- or cross-attention layers. In contrast to the original work, where CLIP embeddings were used in the adaLN, we instead pool the embedded dwell location tokens and the image latent tokens for the self- and cross-attention adaLN layers respectively. These pooled embeddings serve as global summaries of dwell locations and image context which enable the use of the adaLN layers without the typical CLIP embeddings, which were not developed for use with 3D medical images.

Dropout and Droppath: To regularize the model and prevent overfitting, we implement both layer-wise dropout and residual connection droppath throughout the network. Each attention layer is assigned a flat dropout rate of 0.1 and each projection layer is assigned 0.2 for a dropout probability. For each residual connection, we also apply a droppath with probability $0.2 \times \frac{L_i}{L}$ where L_i is the intermediate layer depth, and L the total layer depth as described above.

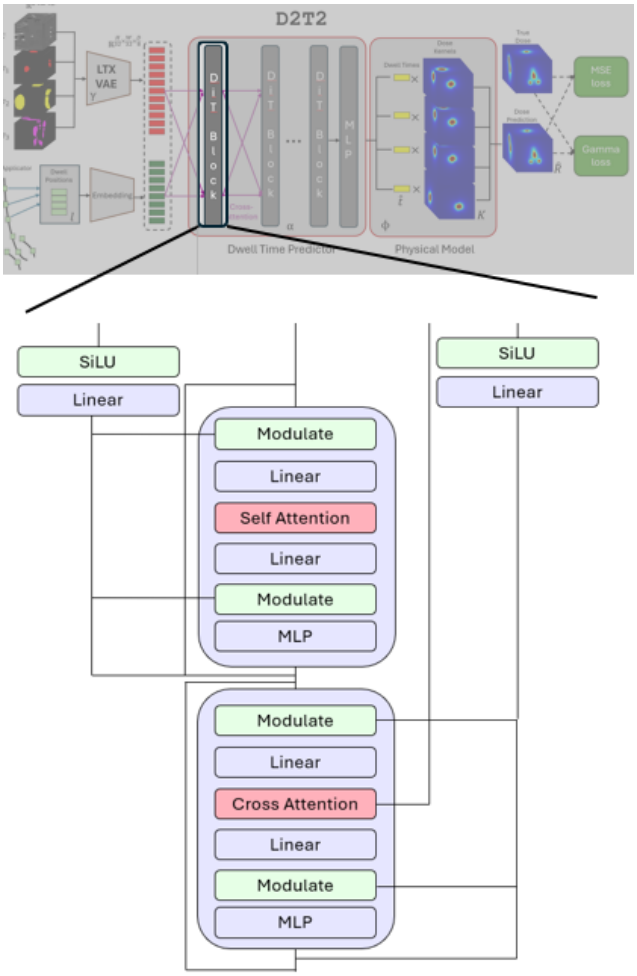

Figure 1. A single DiT block from the **D2T2** model. Self-attention between dwell location tokens allows the model to learn features related to the interaction of dwell times for each dwell location. Cross-attention conditions the dwell location tokens on global context from the latent space encoded treatment images. Each token is modulated by an adaLN normalization prior to projection into the attention heads and the MLP layer.

Weight Decay, Learning Rate, and Optimizer: We

train **D2T2** with a flat 1×10^{-4} learning rate and the AdamW optimizer for all experiments. Additionally, we apply different weight decay for different components of the network, with no decay applied to bias, norm layers, or positional embeddings and a decay of 0.1 applied otherwise.

Positional Encoding: For the dwell location tokens, we apply a standard sinusoidal positional encoding as is typical for many transformer models. For the image tokens, we apply a learned positional embedding after encoding into the LTX latent space and projecting to the model embedding dimension with the concatenated image classification.

Image Class Token: To help the model distinguish anatomical structures, each encoded binary contour mask receives a learnable class token before projection. Contours in the dataset were manually grouped into semantic classes (e.g., OARs vs. tumor targets), and each class is associated with its own token. This improves cross-attention conditioning by providing semantic context.

Training Setup: We use a multi-node Distributed Data Parallel (DDP) training strategy with two nodes each with eight first-generation 32GB Habana Gaudi accelerators to train **D2T2**. We use 4 training samples per processor, resulting in a global batch size of 64 samples per training iteration. We trained all models in this work using a mixed BF16 precision.

Code: The source code for **D2T2** is available on Github at: https://github.com/UCSD-Health-QUIVER/D2T2.git

2. Dataset

Export Methods: To export the large amounts of clinical imaging data for the dataset, we used the PyESAPI (Python Eclipse Scripting Application Programming Interface) open source package, a python extension of the standard Eclipse Scripting API that is C#-based. The python API version allows for customizable data extraction endpoints absent in the C# version, and was thus preferred. This package, in conjunction with the SimpleITK python package, allows for the export of imaging data in the NiFTI data format, which maintains important clinical information such as anisotropic spacing and image origin.

Image Preprocessing: For each plan, we identify the center of mass of the 3D coordinates for all dwell locations in the plan and crop a 128^3 voxel ($1mm$ isotropic spacing) field of view for the CT and contour images centered on that point. The original CT and contour images have dimensions of 512 x 512 x z_n voxels where, z_n denotes the total number of CT image slices for a plan. CT intensity data are first clipped at -1000 and 2000 Hounsfield Units (HU) and then normalized to 0 and 1. The image spacing, direction and origin are extracted from the cropped CT image to be used in mapping dose kernels to their respective dwell positions.

Dose Kernels: For each distinct dose kernel in our dataset, we export the dose distribution for a single dose kernel image $K(v)$ from the treatment planning system. For our systems, all plans before January 31, 2020, used a centered Varisource Iridium 192 (Varian Medical Systems, Inc., Palo Alto, CA) radioactive source while all plans after this date used a Bravos Iridium 192 (Varian Medical Systems, Inc., Palo Alto, CA) radioactive source. These sources emit gamma radiation, which is standard for HDR brachytherapy. This dose kernel, once loaded, can be used for creating the cumulative dose distribution by mapping to all dwell positions and does not need to be repeatedly loaded. For each applicator, dwell locations were extracted along with the dwell location specific 4D transformation matrices, which include the translation vector and the 3D rotation matrices required to spatially orient the individual dose kernels to each dwell location. The single extracted dose kernel was then mapped by translating to each of the dwell locations in the CT space using the aforementioned translation vector. Next, using a PyTorch package, at each dwell location the mapped dose kernels were rotated to match the orientation of the applicator. For this, a list of dose kernel images was created, with the kernel translated to each dwell position, and converted to a PyTorch tensor. Then for each dose kernel image in this tensor, the corresponding rotation matrices were used to rotate around the center using the sampling grid in physical space while preserving spatial metadata such as origin, direction and spacing. The mapped and rotated dose kernels were then resampled to CT image dimensions and scaled by dividing the dose kernel data by the dose kernel air kerma strength, the cumulative dwell times, and the dose grid scaling, and then multiplying by the patient air kerma strength.

Optimizer Inputs: Based on the dose predictions generated from the model [see Section 3 - Brachytherapy, Prior learning-based prediction], a Python-based COBYLA optimizer is used to generate corresponding dwell times, where the mean squared error difference between the predicted dose and the calculated dose corresponding to dwell times at each step of the optimizer is minimized. The optimizer takes the patient air kerma strength, predicted dose image and the overall dose distribution obtained from the cumulative mapping of the dose kernel to each dwell position as inputs.

Dataset Composition: Table 1 summarizes the statistics of the brachytherapy dataset used in this work. Each of the 1716 patients were randomly assigned to training, validation, or testing splits. Between different treatment sites we notice large amounts of variability, with the difference in prescribed dose in breast and cervical plans in the training fold as an example at 4.73 and 6.82 respectively. This reflects the inherent variability of brachytherapy treatment plans, where different treatment sites require different amounts of prescribed radiation and changes in planning overall. Despite this native variability in our data, we

Fold	Site	Count Plans	Mean±STD Prescribed Dose (%)	Mean±STD AKS (U)	Median±IQR Organ Contours (#)	Median±IQR Dwell Locations (#)
Train	Breast	353	4.73 ±1.88	28526.54 ±7006.92	3 ±1	80 ±48
	Cervical	1245	6.82 ±0.68	30508.54 ±7034.66	3 ±0	24 ±22
	Endometrial	858	5.96 ±0.71	31862.90 ±6649.40	3 ±3	40 ±59
	Other	729	6.44 ±2.66	32061.31 ±7114.31	3 ±1	41 ±62
	Vaginal	188	5.94 ±0.66	31864.19 ±6866.04	3 ±1	81 ±66
	All	3373	6.25 ±1.62	31056.78 ±7025.78	3 ±0	39 ±58
Validation	Breast	55	4.73 ±1.82	27528.20 ±6297.87	2 ±1	73 ±55
	Cervical	385	6.79 ±0.71	29220.63 ±7191.81	3 ±0	22 ±22
	Endometrial	155	6.23 ±0.86	32229.06 ±6934.67	3 ±3	34 ±55
	Other	179	6.31 ±2.03	32674.29 ±8598.99	3 ±2	43 ±67
	Vaginal	36	5.94 ±0.69	30578.10 ±6064.66	3 ±1	85 ±52
	All	810	6.40 ±1.34	30504.95 ±7563.80	3 ±0	32 ±54
Test	Breast	75	4.73 ±1.87	25891.26 ±6832.35	3 ±1	66 ±55
	Cervical	372	6.72 ±0.70	30942.48 ±8001.44	3 ±0	26 ±24
	Endometrial	158	6.04 ±0.69	32895.83 ±7220.80	2 ±3	30 ±60
	Other	173	6.76 ±4.04	32167.40 ±7285.94	3 ±2	44 ±52
	Vaginal	28	5.77 ±0.40	32588.01 ±6429.47	3 ±0	120 ±40
	All	806	6.38 ±2.12	31175.45 ±7757.81	3 ±0	36 ±55
	Grand Total	4989	6.30 ±1.67	30986.36 ±7239.35	3 ±0	37 ±57

Table 1. **Dataset statistics across training, validation, and test folds.** For each anatomical site and dataset split, we report the mean ± standard deviation (STD) and median ± interquartile range (IQR) of key clinical parameters: prescription dose (Rx), air-kerma strength (AKS), number of organ-at-risk (OAR) contours, and number of dwell locations.

also notice a consistent trend of similarity between training folds, with the average prescribed dose for breast equal to 4.73 for all three folds as an example. This suggests that the dataset partition maintained the underlying distribution of plan characteristics.

3. Additional Results

Site	D2T2 GPR (%) ↑	OPT GPR (%) ↑
Breast	**79.84** ± 22.10	74.20 ± 29.93
Cervical	86.17 ± 17.64	**88.26** ± 17.20
Endometrial	**83.74** ± 20.86	83.06 ± 22.60
Other	**81.45** ± 21.29	79.72 ± 25.31
Vaginal	61.57 ± 34.37	**67.69** ± 35.40

Table 2. **Site-specific GPR comparison.** Mean ± std across test plans. (3%/3mm, calculated analytically)

Gamma Pass Rate: The overall gamma pass rate (GPR) is shown in Table 2. The **D2T2** model outperforms the OPT model for all sites but cervical, for which OPT was specifically designed, and vaginal, where many treatments used a cylinder applicator with no contours at all, making the plans particularly challenging, and none of the models work well.

Relative Dwell Time Error. The absolute MAE for dwell times can be difficult to interpret without a broader clinical context. As such, we have included a relative dwell time error assessment in table 3 for **D2T2** compared to the OPT model.

Site	Model	DT MAE (%) ↓
Breast	**D2T2**	**1.67**±0.84
	OPT	2.02 ± 0.9
Cervical	**D2T2**	**1.51**±0.82
	OPT	2.64 ± 1.08
Endometrial	**D2T2**	**2.77** ±2.66
	OPT	4.26 ± 3.99
Other	**D2T2**	**1.84** ±1.6
	OPT	3.14 ± 3.48
Vaginal	**D2T2**	**0.99** ±0.33
	OPT	1.14 ±0.38
Overall	**D2T2**	**1.83** ±1.6
	OPT	2.96 ± 2.62

Table 3. **Relative (%) dwell time error.** Mean absolute error (MAE) for predicted dwell times in percentage of total clinical dwell time.

4. Gamma Loss Function

Definition and parameters The gamma index is a spatially relaxed error measure, which is frequently used in radiation oncology to compare dose distributions. We use the pymedphys package for all computed gamma index calculations in this work. For the parameters of the gamma index, we use $\Delta D = 3\%$ and $\Delta d = 3mm$, as these are the typical selections for these parameters in brachytherapy literature. Additionally, we select a global normalization value

Synthetic Images	MAE ↓
8000	$(1.30 \pm 5.58) \times 10^{-2}$

Table 4. **Gamma prediction performance on synthetic data.** Mean absolute error (MAE) of the gamma predictor model τ, evaluated on 8000 synthetically generated dose distributions held out from training and validation.

equal to each plan's prescribed dose which is also standard practice. For speed of calculation, we additionally select an interpolation factor of 5, which is half the recommended number.

Comparison to Synthetic Data To demonstrate the accuracy of the trained gamma prediction model τ, we first compare the predicted gamma index to the calculated gamma index for synthetically generated plans which were held out from datasets used for training and validation of τ. The overall MAE in gamma are displayed in Table 4. These dose distributions were generated using the same method as used to generate $\mathcal{D}'_{\mathcal{S}/\mathcal{M}/\mathcal{L}}$, but with new plans not seen in training or validation.

Comparison to Predicted Plans We also compare the calculated gamma to that predicted by model τ for the predictions of the best performing **D2T2** ablation, where $\lambda = 0.05$. A visual comparison is shown in figure 2, where we have plotted 2D slices from the τ predicted gamma and the algorithmically calculated gamma. The predicted gamma index images demonstrate a smoother and more continuous gamma when compared to the more jagged calculated gamma. We note here that the τ model is designed to be used as a complementary loss for model training, and the assessments here are meant to demonstrate that it accurately approximates the gamma index function. We do not suggest that this model replace analytical gamma index calculations in clinical practice.

5. Discussion

D2T2 demonstrates that a large-scale transformer architecture can directly predict dwell time vectors across diverse brachytherapy anatomical treatment sites without needing time-intensive iterative optimization methods. By conditioning dwell-location tokens on encoded 3D image latents and integrating a physics based layer, the model is trained to predict clinically comparable plans in seconds, instead of minutes or hours. Figure 3 visualizes dose distribution predictions by **D2T2** across anatomical sites and compares them to the ground truth and previous baseline methods. In most cases, the dose prediction by **D2T2** is clearly more similar to the groundtruth than that produced by the baseline.

Limitations: Despite these promising findings, several limitations remain. First, **D2T2** is trained and evaluated on data from a single institution and therefore reflects local practices, which may not generalize to other institutions. Second, the cropped 128 voxel field of view may hinder the model in learning global contextual information, which could improve performance at a cost of higher computational burden. However, brachytherapy dose is highly localized, and the selected field of view likely contains most clinically relevant information, and is consistent with contemporary brachytherapy literature. Third, while the Gamma prediction model τ provides an efficient differentiable surrogate for gamma index evaluation, it was trained only on synthetically generated brachytherapy dose distributions, and would likely not generalize to other treatment modalities without modification and re-training. Finally, we implicitly assume that the ground truth plans from our dataset are of ideal quality, which is likely not true. While these plans were produced by experienced physicians, there are some limitations. First, the time constraints of clinical planning practice can result in trade-offs between a perfect future plan, and a good-enough plan which is deliverable now. Indeed, this is one of the main motivators of the model. While plan quality may vary between clinicians, this is the closest one can get to a ground truth. Since plan quality depends on numerous criteria, it is also difficult to produce a globally optimal plan. Second, the procedure currently used in clinical practice, where physicians adjust dose distributions while attempting to achieve an optimal solution that delivers sufficient dose to a target and a limited amount of dose to surrounding organs, is not ideal. In fact, an optimal solution which meets all clinical metrics may not even be achievable for a given patient during treatment. A better procedure would be to provide the physicians with a set of metrics-based knobs and a visualization of the resulting dose produced by the physical model used in **D2T2**. We plan to investigate this possibility in future work, as well as extend the model to multi-institutional datasets and increase the control of predictions through user prompts.

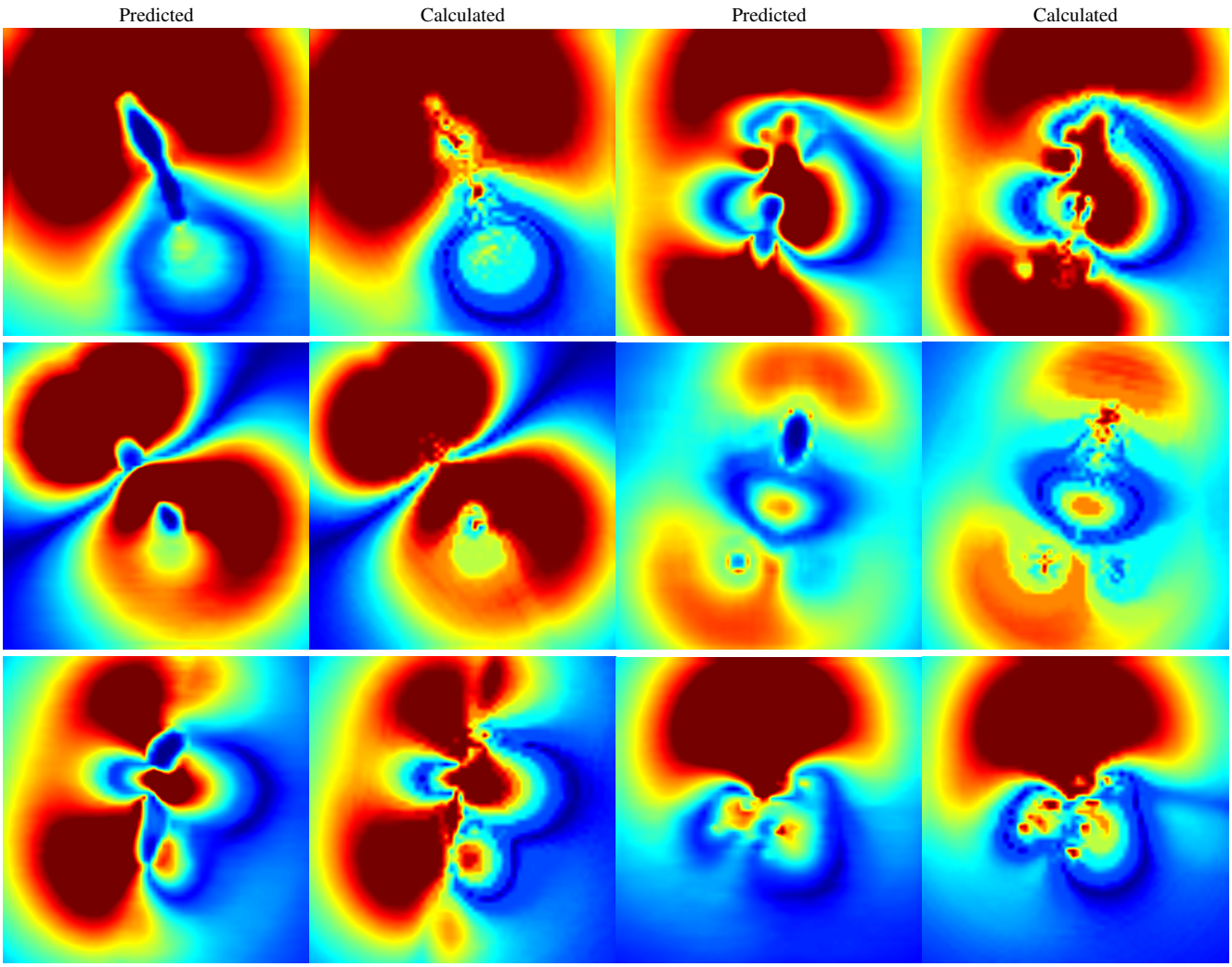

Figure 2. A 2D slice comparison between predicted and calculated gamma for randomly selected plans from the **D2T2** test set predictions.

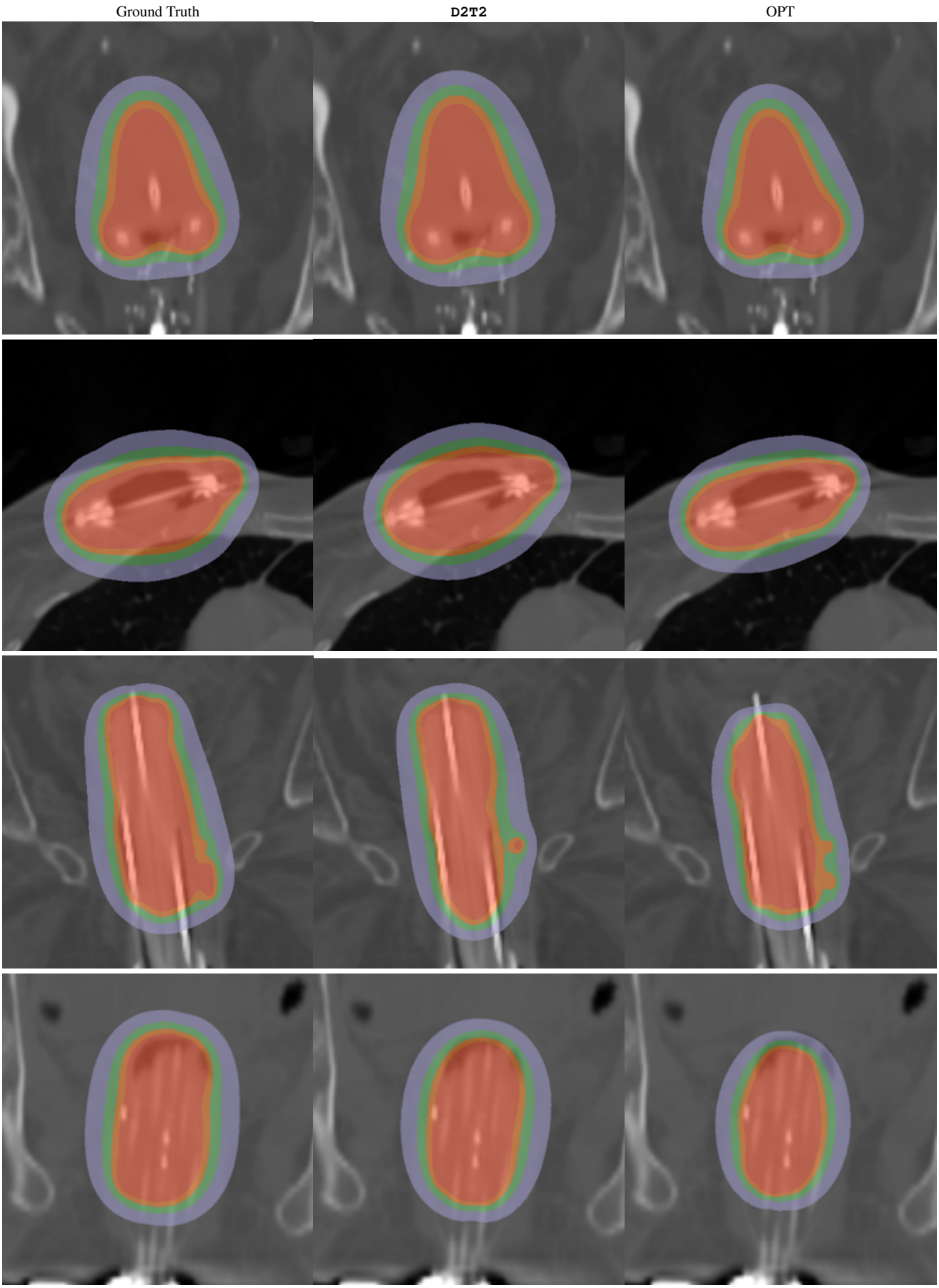

Figure 3. 2D sagittal slices from test set dose distributions with the ground truth, **D2T2**, and OPT model in the left, center and right columns respectively. Dose color wash corresponds to increasing isodose levels with purple, green, orange and red used to designate the 75/100/125/150% isodose levels respectively. The rows depict plans for cervical, breast, endometrial, and vaginal plans, in order from top to bottom.