TREAT: A Unified Text-guided Conditioned Deep Learning Model for Generalized Radiotherapy Treatment Planning

Sangwook Kim[1,2,7][0000−0001−6482−9561], Yuan Gao[1,2,4,7][0009−0001−3817−5266], Thomas G. Purdie[2,8][0000−0003−4176−8457], and *Chris McIntosh[1,2,3,4,5,6,7][0000−0003−1371−1250]

[1] Peter Munk Cardiac Centre, University Health Network (UHN), Toronto, Canada
[2] Department of Medical Biophysics, University of Toronto (UofT), Toronto, Canada
[3] Department of Computer Science, UofT, Toronto, Canada
[4] Ted Rogers Centre for Heart Research, UHN, Toronto, Canada
[5] Toronto General Hospital Research Institute, UHN, Toronto, Canada
[6] Department of Medical Imaging, UofT, Toronto, Canada
[7] Vector Institute, Toronto, Canada
[8] Princess Margaret Cancer Centre, UHN, Toronto, Canada
{sangwook.kim, yuan.gao, tom.purdie, chris.mcintosh}@uhn.ca

Abstract. Deep learning has shown potential to enable automated personalized cancer treatment by automating radiotherapy treatment (RT) planning. However, generalizing RT planning across multiple protocols with deep learning remains a critical challenge due to the diversity of clinical requirements. This paper introduces TREAT: a unified **T**ext-guided **R**adiotherapy for dose pr**E**diction in **A**utomated **T**reatment planning to address these complexities. By leveraging conditional text embeddings using the CLIP text-encoder, the model dynamically adapts to protocol-specific requirements, enabling the generation of high-quality per-protocol dose distributions. We propose an efficient text-conditioning method, graph prompts pooling (GPP), to effectively leverage multiple protocol-specific prompts, and dynamic batch weighting to balance the model training using multiple datasets. We validated TREAT on five datasets—two early-stage prostate, left and right partial breast, and head-and-neck—using clinically relevant metrics: mean absolute error (MAE) of homogeneity index (HI) and dose-volume histogram (DVH). Compared to the protocol-specific model with the MAE-HI of 0.274 and the MAE-DVH of 7.46, TREAT achieves a superior performance of 0.062 and 2.87 for MAE-HI and MAE-DVH score, respectively. When compared to baseline one-hot conditioning with the MAE-HI of 0.085 and the MAE-DVH of 3.35, GPP demonstrates its efficiency in adapting prompt-based conditioning for predicting dose distributions for diverse protocols. The code is available: https://github.com/mcintoshML/TextGuided_RT.

Keywords: Dose prediction · Radiotherapy · Text-guided conditioning

*Corresponding author

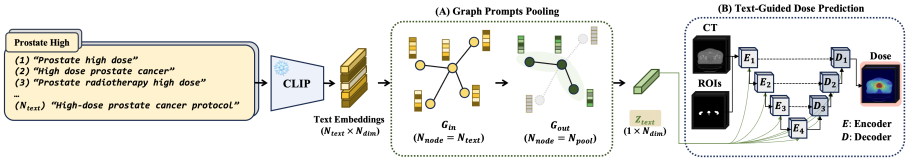

Fig. 1. The overview of TREAT. **(A) Graph Prompts Pooling**: Text prompts describing protocols (e.g., `"Prostate high dose"`) are encoded to form a graph representation, G_{in}. G_{in} undergoes pooling via a self-attention graph pooling [18] to generate G_{out} by capturing the semantic relationships among prompts, generating z_{text}. **(B) Text-Guided Dose Prediction**: z_{text} is used to condition a dose prediction model for generating dose distributions tailored to protocol-specific requirements.

1 Introduction

Radiotherapy treatment (RT) planning is a critical component in cancer care, with the goal of targeting the tumor with a prescribed radiation dose while sparing surrounding normal healthy organs. Each cancer type and anatomical site demands a specific treatment protocol with distinct dose prescriptions and fractionations. Existing automated systems adjust RT planning for individual patients within a protocol using inter-patient knowledge but require separate models per protocol, preventing inter-protocol knowledge learning, which is crucial for enhancing generalizability and enabling applicability across diverse protocols.

Deep learning (DL) has enabled patient-specific treatment plans by leveraging large datasets without manual intervention [14,15,21,24,29,32]. For example, DoseNet by Kearney et al. [13] and a pyramid-based framework by Gheshlaghi et al. [9] have advanced automated RT planning for prostate and head-and-neck cancers, respectively. However, these methods are limited to specific protocols and cannot leverage inter-protocol knowledge.

Text-guided vision models have shown promise in improving medical image analysis using clinical text [8, 10, 25, 30]. Chung et al. [5] developed a diffusion model for MRI reconstruction using text prompts from patient metadata using a pre-trained text encoder, while Oh et al. [26] proposed a text-guided model for target volume segmentation in RT. However, current text-guided models focus on using prompts to condition downstream tasks but lack mechanisms to effectively pool multiple conditioning prompts, limiting their ability to fully leverage the semantically rich text embeddings and resulting in suboptimal conditioning. Liu et al. [20] presented GPT-RadPlan using GPT-4Vision to interactively tune the optimization parameters for RTP. However, a unified text-guided DL model for directly generating dose distributions, which could significantly streamline RT planning workflows, remains unexplored.

Thus, this study aims to address these gaps by proposing a protocol conditioned model that leverages inter-protocol knowledge and text prompts to guide dose prediction across multiple protocols. CT scan and segmentation protocols

Table 1. Data characteristics, including (train/validation/test) splits.

Datasets	Prescribed dose	Num patients	Regions of Interest
Prostate high	60 Gy	110 (95/5/10)	PTV, prostate, rectum, bladder, left/right femur, bowel
OpenKBP [2]	70 Gy	340 (200/40/100)	PTV70, brain stem, spinal cord, left/right parotid, esophagus, larynx, mandible
Prostate low	42.7 Gy	111 (77/11/23)	PTV, rectum, bladder, left/right femur, bowel
Partial Breast Left	26 Gy	223 (155/22/45)	PTV, Eval-TreatedVolume-L, heart, left/right lung
Partial Breast Right	26 Gy	188 (132/18/39)	PTV, Eval-TreatedVolume-L, heart, left/right lung

are often identical across RT protocols for a specific site (e.g., prostate), and thus a single non-RT-protocol-aware model cannot accurately predict dose. For simplicity, we henceforth refer to RT protocols only, and leave automated segmentation out of scope. By conditioning on the protocols, the proposed model leverages inter-protocol knowledge along with context (i.e., the specific protocol), learning from all cases across protocols rather than being limited to per-protocol cases, enabling better generalization through utilization of inter-protocol dose-feature redundancies.

Contributions: We introduce a unified **T**ext-guided **R**adiotherapy for dose pr**E**diction for **A**utomated **T**reatment planning (**TREAT**) which leverages text-guided conditioning to embed inter-protocol information, improving dose prediction across diverse protocols. We further propose **graph prompts pooling** to dynamically generate robust text embeddings and **dynamic batch weighting** to balance training losses during multi-dataset training. Our extensive experiments show that TREAT outperforms existing protocol-specific models, and the variations of TREAT with different pooling methods and text-encoders.

2 Methods and Materials

2.1 Datasets

We collected 972 patients from five RT datasets, split into 659 for training, 96 for validation, and 217 for testing (Table 1). Prostate and partial left/right breast datasets were collected from Princess Margaret Cancer Centre (Toronto, ON, Canada), and processed using Med-ImageTools [16]. OpenKBP [2] is a public RT dataset for head-and-neck cancer. We simplified the region of interests (ROIs) into two binary maps: one for organs at risk (OARs) and one for target volumes (TVs), reducing variability across protocols and improving computational efficiency. We resampled CT and ROI maps to (256, 256, 96), z-score normalized CT images, and min-max normalized target dose maps.

2.2 Model Architecture

We utilized Swin-Transformer 3D U-Net [4](Swin U-Net) as the backbone[†]. Swin U-Net is designed to capture hierarchical and spatial features from the inputs, while being effectively adjustable when utilizing conditional inputs to adapt to different conditions dynamically. Swin U-Net employs the Swin-Transformer (Swin-T), which uses shifted window attention to model long-range dependencies in volumetric imaging efficiently. The inputs to TREAT consist of a single-channel CT, x_c, and a two-channel ROI, x_r, as shown in Fig. 2. All encoder and decoder blocks are conditioned on ROI inputs to maximize their influence. The encoder module extracts hierarchical features from CT and ROI inputs while integrating text conditioning to adapt to different contexts. It uses patch merging layers, P^{merge}, to reduce spatial resolution and increase channel dimensionality. Convolution blocks (Conv), in the encoder consist of two sequential convolution layers, each followed by layer normalization (LN) and ReLU. Attention blocks use shifted window-based multi-head self-attention, $Attn_{self}$, and feed-forward layers with LN, incorporating cross-attention, $Attn_{cross}$, between x_c and x_r to enhance anatomical integration. The decoder mirrors the encoder with condition blocks, Conv, and Attention blocks, with the skip connections from corresponding encoder blocks for feature propagation. In the decoder, TREAT uses patch-expanding layers instead of merging. The final decoder block processes the output through two convolution layers with a ReLU activation.

2.3 Graph Prompts Pooling

We introduce **Graph Prompts Pooling (GPP)** to generate adaptive text representations for conditioning the dose prediction model. The motivation behind GPP stems from optimizing text representations for different protocols. Unlike existing graph-based prompt-tuning methods [7, 19], GPP instead directly leverages protocol-specific prompt embeddings to initialize a graph where nodes represent text embeddings, and edges encode cosine similarity to pool the best-combined embeddings. While the edge attributes remain non-trainable, GPP updates node features dynamically during training to refine text representations. As shown in Fig. 2(A), we employ Self-Attention Graph Pooling[‡] [18] to iteratively refine the graph by alternating between graph convolution and pooling operations, pooling the most informative nodes based on attention scores. After each pooling step, global graph representations are extracted and combined to generate the final text embedding, z_{text}, using fully connected layers, FC. z_{text} then dynamically re-weighting feature channels of x_c utilizing squeeze-and-excitation [12], based on attention scores calculated from x_c and z_{text}.

2.4 Dynamic Batch Weighting

Dynamic Batch Weighting (DBW) assigns loss weights dynamically within mini-batches during multi-dataset training with gradient accumulation. Unlike

[†]https://github.com/1152545264/SwinUnet3D

[‡]https://github.com/inyeoplee77/SAGPool

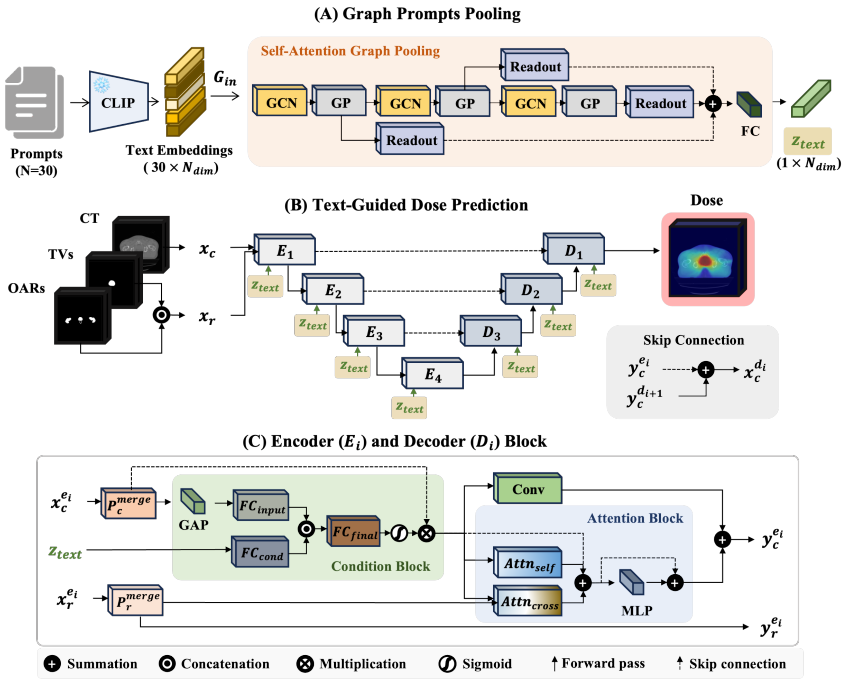

Fig. 2. Model architecture of TREAT. **(A) Graph Prompts Pooling**: Text prompts are encoded into embeddings ($30 \times N_{dim}$) using CLIP text-encoder, where $N_{dim} = 512$, and processed through Self-Attention Graph Pooling to generate the final text embedding z_{text} ($1 \times N_{dim}$). **(B) Text-Guided Dose Prediction**: The encoder-decoder processes CT, x_c, and ROI, x_r, inputs, conditioned on z_{text}, with skip connections ensuring effective feature propagation. **(C) Encoder and Decoder Blocks**: These include a Condition Block, integrating z_{text} via global average pooling (GAP) and fully connected layers, FC, and an Attention Block, which applies self- and cross-attention to enhance CT-ROI integration. Best viewed in colour.

fixed weighted sampling, which uses static weights based on the inverse of dataset size across the entire training, DBW adjusts weights based on the dataset distribution within each mini-batch, mitigating the impact of dominant datasets. We used a batch size of 2 and gradient accumulation of 4, thus DBW calculates weights across 8 combined samples per mini-batch. This ensures that no single dataset dominates the optimization process, allowing balanced learning from all datasets, regardless of their size or sampling frequency.

2.5 Implementation Details

We used a L1 loss, calculated only for the voxels within ROIs to exclude non-critical regions where dose prediction is trivial. We applied min-max normalization using first-epoch average losses as max value per dataset, ensuring balanced

dataset contributions. We used AdamW with an initial learning rate of 10^{-4} and an exponential learning rate scheduler with γ of 0.96. We used PyTorch for the model implementation using an NVIDIA L40S GPU.

3 Experiments and Results

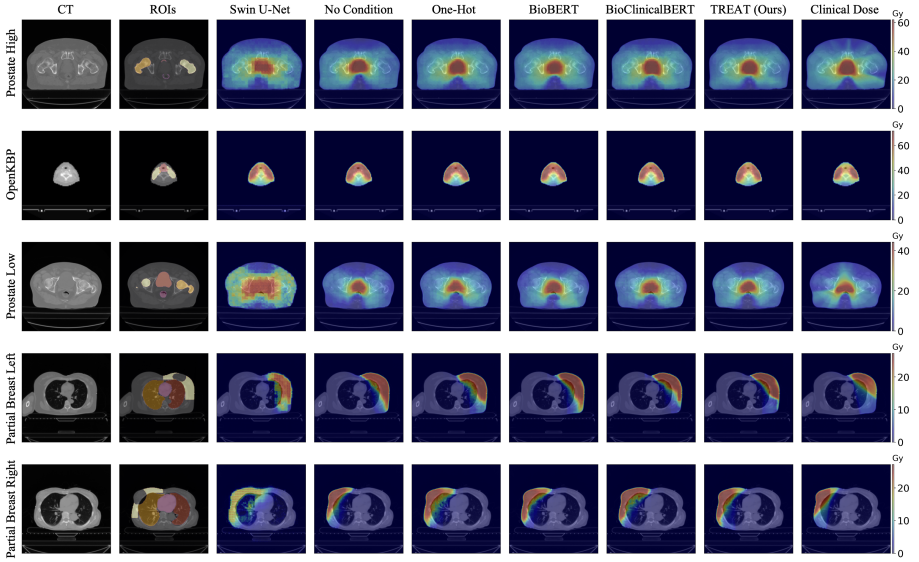

Fig. 3. Qualitative comparison of dose prediction results across different conditioning methods and datasets. Columns show CT and ROI inputs, predicted dose distributions from various models and clinical dose distributions.

We evaluated dose prediction performance using clinically relevant metrics: MAE of Homogeneity Index (HI) [11], Dose-Volume Histogram (DVH) [2, 14], and voxel-wise MAE within the foreground of CT inputs. We employed DVH metrics for in-house datasets from the clinical treatment guidelines from our institution.

Experiment 1: TREAT vs. Protocol-specific models: We evaluated TREAT against protocol-specific models; note that we trained the protocol-specific models in Table 2 using our training datasets and present the average performance across all datasets in Table 1. Compared to protocol-specific models, TREAT achieved superior performance with MAE-HI, MAE-DVH, and MAE-Voxel of 0.0621, 2.874, and 1.235, respectively. This demonstrates that, unlike protocol-specific models, TREAT improves the adaptability for different protocols by effectively integrating protocol-specific textual conditioning generated from GPP to leverage inter-protocol information. We also compared TREAT

Table 2. Performance of protocol-specific and unified dose prediction models. We show the average scores across all datasets, presented as mean ± std. The best scores are in bold, and second-best are underlined. ↓ indicates lower scores are better. TREAT consistently outperformed across the majority of datasets and evaluation metrics.

Models	Unified	MAE-HI ↓	MAE-DVH ↓	MAE-Voxel ↓
U-Net 3D [6]		0.1243 ± 0.1178	4.442 ± 4.873	1.542 ± 0.622
V-Net [23]		0.2609 ± 0.2076	7.700 ± 7.939	4.455 ± 0.872
DoseNet [13]		0.1348 ± 0.1299	6.054 ± 5.891	1.814 ± 0.768
Residual U-Net 3D [3]		0.1085 ± 0.1067	4.742 ± 5.422	1.717 ± 0.608
HD U-Net [24]		0.1586 ± 0.1199	7.018 ± 7.395	2.877 ± 0.792
C3D [21]		0.1394 ± 0.1204	6.496 ± 6.205	1.916 ± 0.891
Attention U-Net [27]		0.1353 ± 0.0944	4.400 ± 4.466	1.455 ± 0.545
Swin U-Net 3D [4]		0.2739 ± 0.2250	7.464 ± 9.117	1.544 ± 0.639
Dose-PYFER [9]		0.1044 ± 0.1359	8.249 ± 10.489	1.954 ± 1.333
No Condition	✓	0.0909 ± 0.1239	4.466 ± 5.288	1.488 ± 0.571
One-Hot	✓	<u>0.0847</u> ± 0.1190	<u>3.353</u> ± 3.933	<u>1.269</u> ± 0.496
TREAT (Ours)	✓	**0.0621** ± 0.0836	**2.874** ± 3.611	**1.235** ± 0.552

with **No Condition**, a unified model without any conditioning, and **One-Hot** encoding methods, finding that TREAT outperformed both across all metrics.

Experiment 2: Ablation experiments: We conducted three ablation experiments: **(A1) Primary components**, **(A2) Text encoders**, and **(A3) Pooling methods** shown in Table 3. **(A1)** To validate the importance of each component, we analyzed their individual contributions. The results showed that each component enhanced dose prediction accuracy across all metrics. **(A2)** We assessed the impact of four different text encoders; CLIP [28], BioBERT [17], BioClinicalBERT [1], and PubmedBERT [31]. We generated text embeddings, z_{text}, using the text-encoders. CLIP outperformed others across all metrics, highlighting its flexibility for conditional dose prediction. **(A3)** We compared the performance of four prompt pooling strategies—Random, Average, Multi-Layer Perceptron (MLP), and GPP. GPP achieved the lowest errors across all metrics. While MLP merges prompts dynamically, it discards original token features, limiting its ability to capture token relationships for robust conditioning. We used one randomly sampled prompt per case during training to expose the model to all prompts, and a single representative prompt per dataset during testing.

Experiment 3: Significance of text conditioning: To evaluate the importance of prompt components, we removed critical (e.g., site or prescription) and non-critical (e.g., filler words) elements from input prompts[§]. For example, when removing critical words from the prompt `"prostate high dose"`, it becomes `"cancer dose"`. Table 4 shows dynamic pooling methods, MLP and GPP, heavily rely on critical information for generating the semantically plausible conditioning, with a huge performance drop of -74.64% relative difference (RD) for GPP, when critical words were removed. In contrast, removing non-critical words

[§]All prompts can be found in https://github.com/mcintoshML/TextGuided_RT

Table 3. Results of the three ablation experiments: A1 (primary components), A2 (text encoders), and A3 (pooling methods). N_C refers to No Condition.

Exp	Models	MAE-HI ↓	MAE-DVH ↓	MAE-Voxel ↓
A1	N_C	0.1103 ± 0.1271	6.312 ± 6.984	2.575 ± 1.033
	CLIP	0.1047 ± 0.1257	4.471 ± 4.960	1.562 ± 0.638
	CLIP+GPP	0.0845 ± 0.1138	3.563 ± 4.770	1.573 ± 0.759
	CLIP+GPP+DBW	**0.0621** ± 0.0836	**2.874** ± 3.611	**1.235** ± 0.552
A2	BioBERT [17]	0.0950 ± 0.1324	4.245 ± 5.072	1.389 ± 0.549
	BioClinicalBERT [1]	0.0890 ± 0.1191	4.186 ± 4.892	1.387 ± 0.577
	PubmedBERT [31]	0.0671 ± 0.0917	3.034 ± 3.684	**1.220** ± 0.492
	CLIP [28]	**0.0621** ± 0.0836	**2.874** ± 3.611	1.235 ± 0.552
A3	Random	0.0752 ± 0.1072	3.679 ± 4.167	1.319 ± 0.546
	Average	0.0820 ± 0.1195	3.736 ± 4.256	1.356 ± 0.543
	MLP	0.0774 ± 0.1116	3.262 ± 3.845	1.256 ± 0.523
	GPP	**0.0621** ± 0.0836	**2.874** ± 3.611	**1.235** ± 0.552

Table 4. MAE-DVH of four different pooling methods when critical and non-critical text components were systematically removed from the input prompts. RD(%) stands for relative difference of **Removed** compared to the **Original**.

Variations	Pooling	Dynamic	Original ↓	Removed ↓	RD(%)
Critical removed	Random		3.679 ± 4.167	3.799 ± 4.267	-3.262
	Average		3.736 ± 4.256	3.851 ± 4.406	-3.078
	MLP	✓	3.262 ± 3.845	7.879 ± 8.387	-141.54
	GPP (Ours)	✓	2.874 ± 3.611	5.019 ± 5.617	-74.64
Non-Critical removed	Random		3.679 ± 4.167	3.686 ± 4.227	-0.190
	Average		3.736 ± 4.256	3.732 ± 4.271	0.107
	MLP	✓	3.262 ± 3.845	3.255 ± 3.868	0.215
	GPP (Ours)	✓	2.874 ± 3.611	2.871 ± 3.614	0.104

had negligible impact, 0.104% RD for GPP, demonstrating its ability to focus on meaningful textual conditions while ignoring irrelevant information.

4 Discussion and Conclusion

In this study, we proposed a unified dose prediction model for RT planning. Our experiments on five different datasets demonstrated that our proposed approach outperformed protocol-specific models and other variations.

We observed that a unified model without any conditioning still outperformed protocol-specific models. This shows that even without providing additional context, the model can utilize the visual relationship between CTs and ROIs across protocols, enabling the generalizable dose prediction without requiring additional context. In addition, one-hot encoding improved performance, showing the significance of conditioning for protocol-specific encoding. While one-hot encoding differentiates protocols, it cannot capture richer semantic relationships between them as shown in Table 2. TREAT leverages pooled textual features using GPP,

transferred from the prior-knowledge from the CLIP text-encoder which can capture semantic similarities (e.g., `"prostate high dose"` is closer to `"prostate low dose"` than `"head and neck cancer"`), enhancing inter-protocol information for dose prediction. Moreover, TREAT generalizes across prompt variations by leveraging GPP to learn inter-protocol structure, rather than relying on specific token formulations. This offers a significant advantage over one-hot conditioning, enabling TREAT to handle inter-protocol knowledge under varying clinical conditions, which is crucial for generalizable automated RT planning.

Future work will focus on converting predicted doses into clinically deliverable formats and validating them in prospective studies. This is essential for ensuring the practical adoption of TREAT in RT planning workflows [22]. Removing critical words from prompts resulted in performance drops, highlighting the importance of text conditioning and the potential of TREAT for zero-shot dose prediction using free-form clinical text. Interestingly, the text-encoders tailored to biomedical texts still underperformed compared to CLIP. It is potentially due to the relatively simple prompts used in TREAT, thus exploring the use of complex clinical context (e.g., patient metadata) still remains a future work. While our loss aligns with clinical goals, adding non-ROI losses may further improve safety and is planned for future work.

In conclusion, we introduced TREAT, which leverages GPP for integrating inter-protocol textual information and DBW for balanced multi-dataset training. TREAT outperformed traditional protocol-specific models, proving its scalability and efficiency in improving clinical workflows and patient care across diverse protocols, paving the way for future integration into clinical practice.

Acknowledgments. This work was supported by the CIHR (No. 183757). CM is funded by NSERC (No. RGPIN-2022-05117, DGECR-2022-00137) and holds the Chair in Medical Imaging at the Joint Department of Medical Imaging at UHN, and the Department of Medical Imaging at the UofT. SK is funded by the doctoral fellowship from Data Sciences Institute at UofT.

Disclosure of Interests. CM and TGP receive royalties from RaySearch Laboratories in relation to ML RT treatment planning.

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Babier, A., Zhang, B., Mahmood, R., Moore, K.L., Purdie, T.G., McNiven, A.L., Chan, T.C.: Openkbp: the open-access knowledge-based planning grand challenge and dataset. Medical Physics **48**(9), 5549–5561 (2021)
3. Bhalerao, M., Thakur, S.: Brain tumor segmentation based on 3d residual u-net. In: International MICCAI brainlesion workshop. pp. 218–225. Springer (2019)
4. Cai, Y., Long, Y., Han, Z., Liu, M., Zheng, Y., Yang, W., Chen, L.: Swin unet3d: a three-dimensional medical image segmentation network combining vision transformer and convolution. BMC medical informatics and decision making **23**(1), 33 (2023)

5. Chung, H., Lee, D., Wu, Z., Kim, B.H., Bouman, K.L., Ye, J.C.: Contextmri: Enhancing compressed sensing mri through metadata conditioning. arXiv preprint arXiv:2501.04284 (2025)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 424–432. Springer (2016)
7. Fang, T., Zhang, Y., Yang, Y., Wang, C., Chen, L.: Universal prompt tuning for graph neural networks. Advances in Neural Information Processing Systems **36** (2024)
8. Feng, C.M.: Enhancing label-efficient medical image segmentation with text-guided diffusion models. In: MICCAI. pp. 253–262. Springer (2024)
9. Gheshlaghi, T., Nabavi, S., Shirzadikia, S., Moghaddam, M.E., Rostampour, N.: A cascade transformer-based model for 3d dose distribution prediction in head and neck cancer radiotherapy. Physics in Medicine & Biology (PMB) **69**(4), 045010 (2024)
10. Guo, Y., Zeng, X., Zeng, P., Fei, Y., Wen, L., Zhou, J., Wang, Y.: Common vision-language attention for text-guided medical image segmentation of pneumonia. In: MICCAI. pp. 192–201. Springer (2024)
11. Helal, A., Omar, A.: Homogeneity index: effective tool for evaluation of 3dcrt. Pan Arab Journal of Oncology **8**(2), 20–4 (2015)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
13. Kearney, V., Chan, J.W., Haaf, S., Descovich, M., Solberg, T.D.: Dosenet: a volumetric dose prediction algorithm using 3d fully-convolutional neural networks. PMB **63**(23), 235022 (2018)
14. Kim, S., Khalifa, A., Purdie, T.G., McIntosh, C.: Multi-task learning for automated contouring and dose prediction in radiotherapy. Physics in Medicine and Biology (2025)
15. Kim, S., Purdie, T.G., McIntosh, C.: Cross-task attention network: Improving multi-task learning for medical imaging applications. In: MICCAI Workshop on Foundation Models for General Medical AI. pp. 119–128. Springer (2023)
16. Kim, S., Kazmierski, M., Qu, K., Peoples, J., Nakano, M., Ramanathan, V., Marsilla, J., Welch, M., Simpson, A., Haibe-Kains, B.: Med-imagetools: An open-source python package for robust data processing pipelines and curating medical imaging data. F1000Research **12**, 118 (2024)
17. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
18. Lee, J., Lee, I., Kang, J.: Self-attention graph pooling. In: International conference on machine learning. pp. 3734–3743. pmlr (2019)
19. Lee, J., Yang, W., Kang, J.: Subgraph-level universal prompt tuning. arXiv preprint arXiv:2402.10380 (2024)
20. Liu, S., Pastor-Serrano, O., Chen, Y., Gopaulchan, M., Liang, W., Buyyounouski, M., Pollom, E., Le, Q.T., Gensheimer, M., Dong, P., et al.: Automated radiotherapy treatment planning guided by gpt-4vision. arXiv preprint arXiv:2406.15609 (2024)
21. Liu, S., Zhang, J., Li, T., Yan, H., Liu, J.: A cascade 3d u-net for dose prediction in radiotherapy. Medical physics **48**(9), 5574–5582 (2021)
22. McIntosh, C., Conroy, L., Tjong, M.C., Craig, T., Bayley, A., Catton, C., Gospodarowicz, M., Helou, J., Isfahanian, N., Kong, V., et al.: Clinical integration of

machine learning for curative-intent radiation treatment of patients with prostate cancer. Nature medicine **27**(6), 999–1005 (2021)
23. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
24. Nguyen, D., Jia, X., Sher, D., Lin, M.H., Iqbal, Z., Liu, H., Jiang, S.: 3d radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture. PMB **64**(6), 065020 (2019)
25. Ni, R., Fyles, A., Haibe-Kains, B., Rink, A.: Multi-modal hr-ctv segmentation: Leveraging a large language model in cervical brachytherapy. In: AAPM 66th Annual Meeting & Exhibition. AAPM (2024)
26. Oh, Y., Park, S., Byun, H.K., Cho, Y., Lee, I.J., Kim, J.S., Ye, J.C.: Llm-driven multimodal target volume contouring in radiation oncology. Nature Communications **15**(1), 9186 (2024)
27. Osman, A.F., Tamam, N.M.: Attention-aware 3d u-net convolutional neural network for knowledge-based planning 3d dose distribution prediction of head-and-neck cancer. Journal of applied clinical medical physics **23**(7), e13630 (2022)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
29. Wang, B., Teng, L., Mei, L., Cui, Z., Xu, X., Feng, Q., Shen, D.: Deep learning-based head and neck radiotherapy planning dose prediction via beam-wise dose decomposition. In: MICCAI. pp. 575–584. Springer (2022)
30. Zeng, X., Zeng, P., Cui, J., Li, A., Liu, B., Wang, C., Wang, Y.: Abp: Asymmetric bilateral prompting for text-guided medical image segmentation. In: MICCAI. pp. 54–64. Springer (2024)
31. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI p. AIoa2400640 (2024)
32. Zhang, Y., Li, C., Zhong, L., Chen, Z., Yang, W., Wang, X.: Dosediff: distance-aware diffusion model for dose prediction in radiotherapy. IEEE Transactions on Medical Imaging (2024)