Advanced Conditioning Strategies for Diffusion Models for Radiation Therapy Dose Prediction

Victor Cobilean, MS[1]; Harindra S. Mavikumbure, PhD[1]; Devin Drake, MS[1]; Swagat Das, MS[1]; Ibtisam Almajnooni, MS[2]; Elisabeth Weiss, MD[2]; Lulin Yuan, PhD[2]; Milos Manic, PhD[1]

[1] College of Engineering, Virginia Commonwealth University; [2] School of Medicine, Virginia Commonwealth University

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

The purpose of this study is to improve diffusion-based radiotherapy dose prediction by conditioning the model with pre-trained 3D and segmentation encoders, enhancing volumetric consistency and reducing reliance on manual anatomical annotations.

A DDPM-based model predicts 3D dose distributions from slice-wise 2D generations. Conditioning information is provided by: 1) a 2D encoder extracting CT and PTV features, 2) a pre-trained MedSAM encoder providing anatomical context through cross-attention, and 3) a pre-trained 3D VAE supplying volumetric context and slice-position information. The model is trained to predict diffusion noise at each timestep.

Evaluated on 115 lung cancer treatment plans, the proposed method achieved the best performance, with a Mean Absolute Error of 3.37 ± 0.16 Gy and a Volume Score of 4.33 ± 0.47%. The model better preserved high-frequency dose details and beam-direction characteristics. Ablation studies demonstrated consistent improvements from both conditioning encoders.

Incorporating pre-trained 3D and segmentation encoders into the diffusion framework improves dose prediction accuracy, preserves volumetric consistency, and reduces dependence on manual organ-at-risk segmentations.

CONTACT

Victor Cobilean
Virginia Commonwealth University
cobileanv@vcu.edu

INTRODUCTION

The primary contribution of this study is a novel conditioning strategy that **significantly enhances the performance of base diffusion models**, achieving results competitive with state-of-the-art methods while preserving high-frequency details and the directional characteristics of radiation beams, which are essential for clinical applicability. The proposed strategy leverages representations from auxiliary pre-trained models: 1) a **3D Variational Autoencoder (VAE)** to ensure volumetric consistency, and 2) a **MedSAM encoder [1]** to capture anatomical context. The proposed framework effectively addresses the challenge of maintaining global spatial coherence in slice-wise predictions. In addition, it reduces the dependence on manual organ-at-risk annotations.

METHODS AND MATERIALS

Our approach is based on the **Denoising Diffusion Probabilistic Model (DDPM) [2],** where the goal of the **iterative denoising process** is to estimate the noise at each step. Figure1 shows the neural network architecture, which predicts noise from the input 2D CT and planning target volume (PTV) slices, combined with features extracted from pre-trained models. The model backbone is a **U-Net**. Local structural features are encoded from the **CT and PTV slices** using a **2D structure encoder** and incorporated **via element-wise addition.** Anatomical features are extracted using a **MedSAM [1]** segmentation encoder and integrated through **cross-attention layers.** To ensure volumetric consistency, global spatial features obtained from a **pre-trained 3D VAE** and **learnable slice-position embeddings** are fused and injected using **Adaptive Group Normalization [3]**. The network is trained to minimize noise prediction error using a mean absolute error (MAE) loss. Spatial and pixel-level augmentations are applied during training to reduce overfitting. This study used **115 de-identified lung cancer radiotherapy cases** with planning CT images and clinically approved dose plans acquired under an IRB-approved protocol. Treatment plans included **IMRT and VMAT** dose plans, with prescribed doses ranging from **30–66 Gy** (median: **60 Gy in 30 fractions**). The dataset was randomly split into training, validation, and test sets using a 70:10:20 ratio.

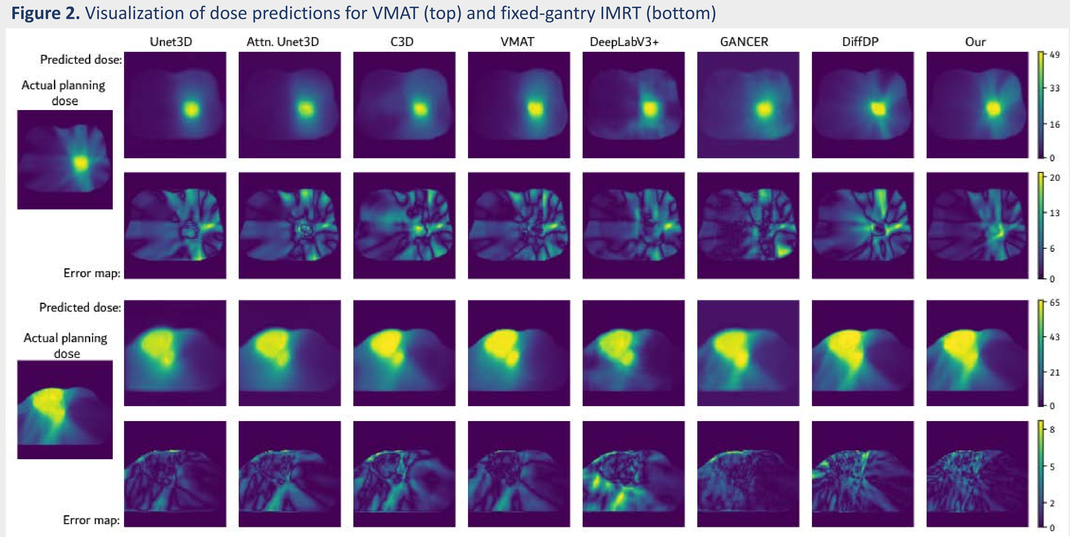
Table 1. Comparison with SOTA models

Metric	Unet3D	Attn. Unet3D	C3D	DeepLabV3+	VMAT	GANCER	DiffDP	Our (Diff-Seg3D)
MAE ↓	3.541 ± 0.262	3.395 ± 0.258	3.406 ± 0.265	3.864 ± 0.218	3.560 ± 0.171	3.917 ± 0.166	3.639 ± 0.261	**3.374 ± 0.158**
DVH Score ↓	6.428 ± 1.156	6.175 ± 1.337	6.227 ± 1.334	6.221 ± 1.227	**5.418 ± 1.173**	6.306 ± 0.848	6.463 ± 0.678	5.607 ± 1.084
PSNR ↑	25.968 ± 0.474	26.193 ± 0.581	**26.319 ± 0.383**	25.092 ± 0.324	25.340 ± 0.242	25.065 ± 0.225	25.509 ± 0.477	26.218 ± 0.427
SSIM ↑	0.848 ± 0.017	0.866 ± 0.011	**0.872 ± 0.023**	0.821 ± 0.009	0.846 ± 0.010	0.817 ± 0.011	0.815 ± 0.020	0.833 ± 0.015
Dose Score ↓	5.185 ± 0.921	4.944 ± 0.937	5.012 ± 1.069	5.062 ± 0.874	**4.407 ± 0.849**	5.594 ± 0.498	5.573 ± 0.235	4.685 ± 0.898
Volume Score ↓	5.382 ± 0.949	5.523 ± 1.735	5.668 ± 1.872	5.022 ± 1.172	4.460 ± 0.332	4.623 ± 0.540	4.505 ± 0.915	**4.330 ± 0.473**

Table 2. Ablation study

2D enc. input	3D CT enc.	Segm. enc.	MAE	DVH Score	PSNR	SSIM	Dose Score	Volume Score
$X_{CT}^{(i)}, X_{PTV}^{(i)}, X_{OAR}^{(i)}$	✓	✓	**3.370 ± 0.125**	**5.334 ± 0.654**	**26.273 ± 0.530**	**0.835 ± 0.005**	**4.489 ± 0.495**	**4.343 ± 0.585**
	✓		3.399 ± 0.278	5.660 ± 0.972	26.177 ± 0.410	0.829 ± 0.015	4.947 ± 0.805	4.407 ± 0.771
		✓	3.645 ± 0.397	6.459 ± 0.783	25.434 ± 0.908	0.818 ± 0.025	5.680 ± 0.780	4.885 ± 1.142
			3.737 ± 0.346	6.564 ± 0.767	25.042 ± 0.493	0.795 ± 0.024	5.590 ± 0.828	4.601 ± 1.018
$X_{CT}^{(i)}, X_{PTV}^{(i)}$	✓	✓	**3.374 ± 0.158**	**5.607 ± 1.084**	**26.218 ± 0.427**	**0.833 ± 0.015**	**4.685 ± 0.898**	**4.330 ± 0.473**
	✓		3.485 ± 0.313	5.754 ± 0.843	25.976 ± 0.405	0.826 ± 0.020	5.025 ± 0.513	4.537 ± 0.850
		✓	3.693 ± 0.152	6.462 ± 0.977	25.323 ± 0.480	0.808 ± 0.015	5.731 ± 0.788	4.752 ± 0.917
			3.694 ± 0.242	6.410 ± 0.821	25.208 ± 0.365	0.799 ± 0.012	5.537 ± 0.685	4.618 ± 1.152
$X_{CT}^{(i)}$	✓	✓	**7.165 ± 1.266**	**13.947 ± 0.902**	**20.769 ± 0.902**	**0.726 ± 0.037**	**15.429 ± 1.586**	**11.100 ± 2.295**
	✓		7.386 ± 1.355	15.072 ± 0.431	20.140 ± 1.384	0.711 ± 0.034	15.935 ± 1.982	11.183 ± 2.837
		✓	7.231 ± 1.152	14.831 ± 0.461	20.494 ± 1.330	0.716 ± 0.020	15.764 ± 1.389	11.198 ± 2.521
			7.341 ± 1.347	15.926 ± 1.791	20.172 ± 1.444	0.720 ± 0.030	16.786 ± 2.519	11.431 ± 2.962

Figure 2. Visualization of dose predictions for VMAT (top) and fixed-gantry IMRT (bottom)



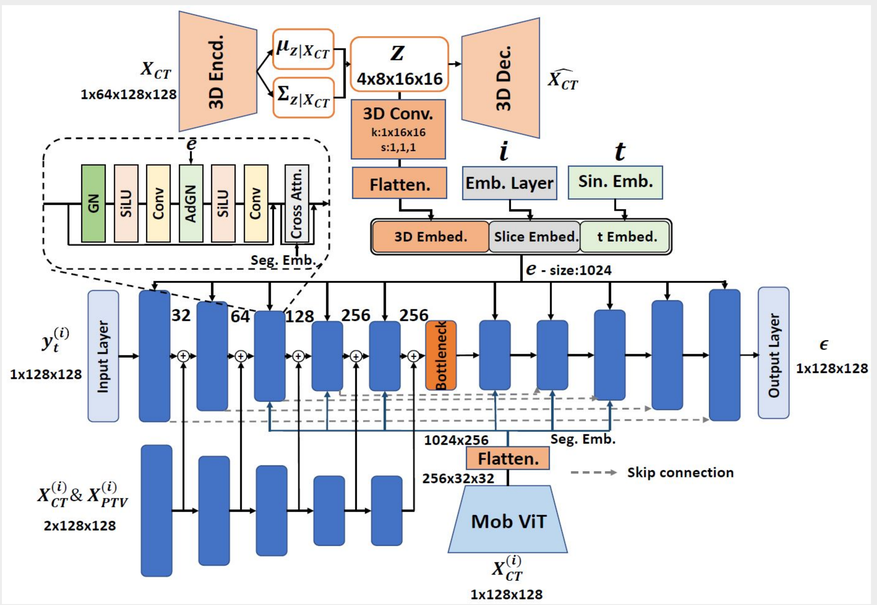
RESULTS

Table 1 compares Diff-Seg3D with several state-of-the-art dose prediction models. Our method achieved the **lowest MAE** (**3.374 ± 0.158 Gy**) and **Volume Score** (**4.330 ± 0.473%**), while maintaining competitive DVH and Dose Scores. These results demonstrate improved voxel-level accuracy and volumetric consistency without requiring OAR segmentation masks at inference.

The ablation study in Table 2 shows that **both the 3D VAE and MedSAM** conditioning modules **contribute to performance.** The best results were obtained when both conditioning strategies were used together, highlighting the complementary benefits of volumetric and anatomical guidance.

As shown in Figure 2, Diff-Seg3D **preserves high-frequency dose gradients** and **radiation beam directionality** while generating spatially consistent dose distributions that closely match the ground truth.

Figure 1. The architecture of the proposed method



DISCUSSION

The results highlight the importance of strong structural conditioning, where combining 3D volumetric features with segmentation-derived anatomical cues **improves spatial consistency** and **alignment with patient anatomy.**
The model is particularly effective at capturing **sharp dose transitions** at tumor boundaries, which are critical for treatment quality. Importantly, both the MedSAM encoder and the volumetric encoder **can benefit from pretraining on larger datasets**, potentially yielding **more robust and transferable representations**, especially in settings where radiotherapy dose data are limited.

CONCLUSION

Diff-Seg3D integrates volumetric and anatomical conditioning to enhance diffusion-based dose prediction. The framework achieves competitive performance with state-of-the-art methods, preserves clinically important dose distributions, and reduces dependence on expert-annotated OAR masks. These findings demonstrate its potential for robust and practical automated radiotherapy treatment planning.

REFERENCES

1. Ma, J., He, Y., Li, F., Han, L., You, C. and Wang, B., 2024. Segment anything in medical images. Nature communications, 15(1), p.654.
2. Ho, J., Jain, A. and Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, pp.6840-6851.
3. Dhariwal, P. and Nichol, A., 2021. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34, pp.8780-8794.