Any2Any 3D Diffusion Models with Knowledge Transfer: *A Radiotherapy Planning Study*

Supplementary Material

Supplementary Contents

A. Detailed Model Structures **2**

B. Training Details **3**
- B.1. Data Sources . . . 3
- B.2. Conditioning Modalities and Structure Selection . . . 3
- B.3. Backbone Adaptation and Supervised Training . . . 3
- B.4. Baselines and Fairness Protocol . . . 3
- B.5. RL Post-training (ScardNFT) . . . 4
- B.6. Discussion on MAISI Diffusion Prior . . . 4
- B.7. Optimization and Hyperparameters . . . 4

C. More Visual Results **6**

D. Additional Baseline Comparisons **8**

E. Extended Ablation Studies **9**
- E.1. Noise Prediction Parameterization . . . 9
- E.2. VAE Adaptation Strategies . . . 9

F. Statistical Significance Analysis **10**

A. Detailed Model Structures

As shown in Figure 1, our DiffKT3D adopts a VAE–DiT hybrid architecture [7, 10, 17]. For illustration we depict three volumetric inputs X_a, X_b, and X_g corresponding to CT, structure masks (PTV and OARs), and dose, respectively; the same pipeline applies to all available modalities. Each volume is first passed through a frozen 3D VAE encoder [7] to obtain compact latent representations. These latent grids are then patchified into token sequences and concatenated before being fed into a stack of DiT blocks [10]. The diffusion process operates entirely in this latent-token space. After denoising, the output tokens are reshaped back into latent feature maps V_a, V_b, and V_g, which are decoded by the corresponding VAE decoders to recover volumetric predictions at the original spatial resolution.

The right panel of Figure 1 details a single DiT block. Each block follows a transformer-style design [16] with self-attention and a feed-forward network (FFN), all wrapped by residual connections. In the original Wan 2.1 backbone [17], each block also contains a cross-attention layer that lets vision tokens attend to text tokens. DiffKT3D does not use any language or text conditioning, so we remove this cross-attention module and let all tokens from all volumetric modalities (CT, masks, dose, and other channels) jointly interact through the shared self-attention layers. A shared timestep–role embedding (encoding the diffusion step and whether a token is a target or a conditioning token) is processed by a small MLP to produce modulation vectors. These vectors drive FiLM-like layers [11]: we apply scale-and-shift operations to the normalized tokens before the self-attention and FFN modules, and *Scale*-only gates on the residual outputs of self-attention and FFN. This modulation allows each DiT block to dynamically control feature amplification or suppression across timesteps and roles while keeping the overall architecture lightweight and stable to train.

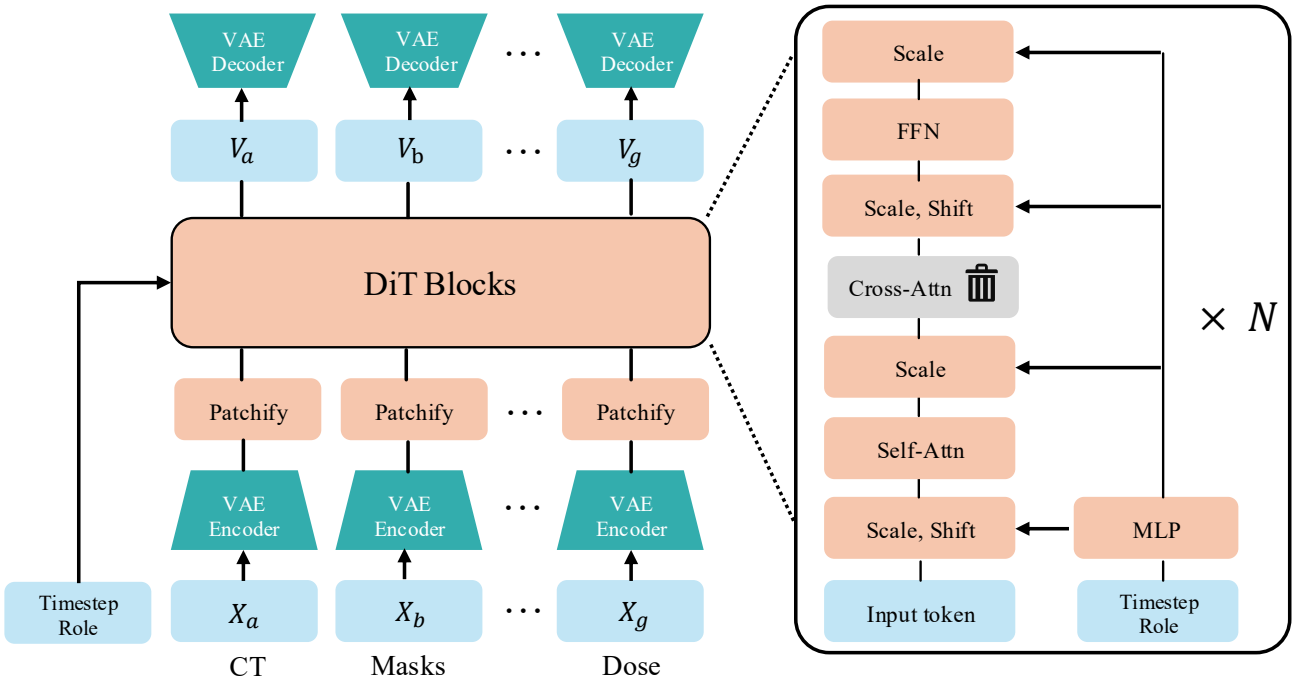

Figure 1. Architecture of the proposed VAE–DiT-based conditional diffusion model DiffKT3D. Left: multi-branch VAE–DiT pipeline for CT (X_a), structure masks (X_b), and dose (X_g) with their corresponding latent outputs $\{V_a, V_b, V_g\}$. Right: structure of a single DiT block with timestep–role modulation using FiLM-style (*Scale, Shift*) layers and residual gates (*Scale*). The Any2Any gating and noisy latent are not shown for simplification. We remove cross-attention layers from original Wan DiT blocks because DiffKT3D does not use language tokens.

B. Training Details

B.1. Data Sources

We train and evaluate DiffKT3D on the official GDP–HMM Grand Challenge dataset [2] for head-and-neck and lung cancer and on a prostate dataset derived from the REQUITE cohort [15]. Spacing, orientation, challenge-defined bounding boxes, and body-mask conventions follow the organizer protocol. For model input, we then extract a fixed $97 \times 128 \times 160$ ROI crop around the PTV isocenter so that the volume is compatible with the Wan 2.1 VAE. The REQUITE cohort cases were re-optimized in Varian Eclipse ESAPI under multiple planning configurations, yielding multiple plans per patient. CT images are clipped to $[-1000, 1000]$ HU and normalized on a per-patient basis before entering the model; all structure masks, beam plates, and angle plates are rasterized onto the same grid as the reference dose. The in-plane size 128×160 matches the challenge bounding box, while the depth of 97 voxels is chosen as $97 = 4 \times 24 + 1$ so that the downsampled latent depth satisfies the causal attention constraint in the Wan 2.1 VAE. After cropping, all modalities are linearly scaled to the range $[-1, 1]$ before being passed into the frozen VAE encoder, matching the expected input range of the pretrained backbone.

Note for REQUITE data: We thank all the contributors to the REQUITE project, including the patients, clinicians and nurses. The core REQUITE consortium consists of David Azria, Erik Briers, Jenny Chang-Claude, Alison M. Dunning, Rebecca M. Elliott, Corinne Faivre-Finn, Sara Gutiérrez-Enríquez, Kerstie Johnson, Zoe Lingard, Tiziana Rancati, Tim Rattay, Barry S. Rosenstein, Dirk De Ruysscher, Petra Seibold, Elena Sperk, R. Paul Symonds, Hilary Stobart, Christopher Talbot, Ana Vega, Liv Veldeman, Tim Ward, Adam Webb and Catharine M.L. West.

B.2. Conditioning Modalities and Structure Selection

Each patient is represented by up to seven modalities (see the modality visualization in the appendix of [2]):

$$\{\,\text{CT},\ \text{PTV},\ \text{OAR masks},\ \text{body mask},\ \text{dose},\ \text{beam plate},\ \text{angle plate}\,\}.$$

The "PTV" channel encodes the optimized planning target volumes after any site-specific post-processing. Beam and angle plates follow the official GDP–HMM implementation and provide beam geometry and gantry angle information on the same voxel grid as the dose.

To make supervision consistent across disease sites, we standardize the set of OARs used during training. As in the challenge data, we retain up to about 30 OARs for head-and-neck, and 7 OARs for lung plans. For prostate plans, we retain four OARs: bladder, rectum, femoral head (left), and femoral head (right). All masks are stored as floating-point channels and normalized jointly with the other modalities to $[-1, 1]$ before patch embedding.

B.3. Backbone Adaptation and Supervised Training

We initialize DiffKT3D from the public Wan 2.1 DiT+VAE checkpoint [17], whose DiT backbone follows the scalable diffusion transformer design of Peebles and Xie [10], and keep the VAE completely frozen throughout all experiments. On top of Wan's 3D patch embedding, we introduce seven modality-specific 3D patch-embedding heads, one per modality in the set above. Each head has the same architecture as the original Wan patch embed but uses separate parameters, mapping the latent grids (or their noised versions for target modalities) into tokens of hidden dimension D.

To support Any2Any training, we augment the backbone with: (i) a learnable binary role embedding that tags each token as either target or condition and is injected via the shared AdaLayerNorm modulator by adding it to the timestep embedding, and (ii) a 4D RoPE positional encoding that assigns rotary phases along a slot axis (modality ID) and the three spatial axes (H, W, D). These additions are lightweight and leave the Wan DiT block structure unchanged; only the DiT blocks and the new embedding layers are fine-tuned on RT data.

B.4. Baselines and Fairness Protocol

In the GDP–HMM challenge, regression baselines are the top challenge entries built on MedNeXt [13], nnU-Net [6], and latent diffusion backbones [12], and we evaluate them using the official model weights released by the organizers. Diffusion baselines include an MAISI-based conditional U-Net [3] and a conditional DiT variant that concatenates all conditioning modalities with the dose channel [10]. All methods operate on exactly the same cropped $97 \times 128 \times 160$ volumes and use the same set of modalities and OAR selection as DiffKT3D.

For the REQUITE prostate experiments, where no challenge leaderboard is available, we initialize all baselines from their publicly released GDP–HMM-trained checkpoints and fine-tune them on prostate data under the same schedule as our model: identical preprocessing, crop size, effective batch size, and number of epochs. Our internal regression and diffusion variants are also trained with the same protocol. This setup ensures that performance differences come from the model design (Any2Any conditioning, role embeddings, 4D RoPE, and post-training) rather than from data handling or compute budget.

Table 1. GDP–HMM validation and test results on the MAISI backbone with different output parameterizations. MAE is reported in Gy.

Method	Valid MAE	Test MAE	Infer time (H100)	Train time (H100)
Challenge Winner	2.03	2.07	<1 s	144 h
Ours (MAISI, x_0-pred, 1 step)	**1.89**	**1.95**	<1 s	12 h
Ours (MAISI, noise-pred, 5 steps)	10.475	11.945	1.2 s	12 h
Ours (MAISI, noise-pred, 50 steps)	10.636	11.969	5.7 s	12 h

B.5. RL Post-training (ScardNFT)

After supervised training we perform a lightweight RL-style post-training stage using the ScardNFT objective, which instantiates the DiffusionNFT formulation [19] on our clinical scorecard. For each patient, we generate candidate dose predictions with a 10-step deterministic sampler from the Flow-Matching/DPM-Solver family [8, 9], starting from multiple initial noises and evaluate each candidate using the clinically informed Scorecard together with a voxel-wise mean absolute error (MAE) anchor. Based on these scores, we construct four positive/negative sample pairs per case and optimize the DiffusionNFT-style loss described in the main text.

To keep this stage efficient and stable, we avoid full-parameter fine-tuning. Instead, we insert rank-64 LoRA adapters into all self-attention and feed-forward layers of the DiT backbone and update only these adapters together with the small modulation networks, keeping the original Wan weights frozen [5]. This post-training is only a small fraction of overall training cost but yields improved clinical preference alignment reported in the main paper without sacrificing voxel-level accuracy.

B.6. Discussion on MAISI Diffusion Prior

MAISI [3] is originally trained to generate CT images, which is fundamentally different from the task of generating dose. To evaluate whether MAISI can serve as a viable backbone for dose prediction, we ported our training pipeline to the public MAISI latent diffusion model, replacing the Wan VAE+DiT with the MAISI VAE+UNet backbone while keeping the same data preprocessing and optimization schedule as DiffKT3D. Table 1 reports performance on the official validation and test sets.

With an x_0-prediction objective and joint fine-tuning of the MAISI VAE decoder, our re-implementation improves MAE over the challenge top-1 model while using substantially less training time and without task-specific model crafting. *This provides additional evidence, beyond the main Wan-based experiments, that diffusion priors learned from a large-gap source domain can transfer effectively to a target domain.*

However, we find that switching the same MAISI backbone to a noise-prediction objective while freezing the VAE decoder causes performance to collapse: both validation and test MAE degrade by over $5\times$. Increasing the number of sampling steps from 5 to 50 does not recover performance (Table 1).

Our closer look indicates that the failure is driven by precision rather than optimization. After standardizing doses, the ground-truth dose maps are normalized to a fixed $[0, 1]$ range and then linearly rescaled to $[0, 70]$ Gy. In contrast, the decoded dose from noise-predicted MAISI latents often occupies a wider and case-dependent range, roughly $[0, 1.05]$–$[0, 1.2]$, corresponding to $[0, 74]$–$[0, 84]$ Gy after rescaling. In other words, the maximum of the decoded range is no longer pinned at 1 but drifts between about 1.05 and 1.2 across patients. This drifting dynamic range makes it difficult for a single regression head to align voxel intensities across patients, especially near the clinically important 60–70 Gy region, and explains why MAISI behaves well for x_0-prediction with a jointly trained decoder but degrades sharply for the noise-parameterization under a frozen-decoder regime.

Since DiffKT3D is explicitly designed around a frozen VAE and Any2Any conditioning, we therefore adopt the Wan 2.1 VAE and DiT backbone [10, 17] rather than MAISI, and include the MAISI-based model as a baseline for DiffKT3D.

B.7. Optimization and Hyperparameters

The base DiffKT3D models (before ScardNFT post-training) are obtained from a two-stage supervised training pipeline: Stage A Any2Any pretraining followed by Stage B dose-only fine-tuning. Both supervised stages use the same Flow-Matching objective with a v-prediction parameterization in the latent/token space. We train for 100 epochs on eight B200 GPUs with data parallelism, using a per-GPU batch size of 1 (effective batch size 8). The model is optimized with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, no weight decay) and a constant learning rate of 1×10^{-4} after a linear warm-up of 500 steps. We use 1,000 training timesteps with a flow-shift parameter of 3.0, and enable timestep embedding with an additional σ-embedding. Training is performed in bfloat16 mixed precision with gradient accumulation disabled and gradient clipping at a global norm

of 0.1. We do not use classifier-free guidance or dropout in the conditioning pathways (CFG dropout probability set to 0). The subsequent ScardNFT post-training stage uses the same optimizer but only updates the LoRA adapters and modulation MLPs described in Sec. B.5.

C. More Visual Results

Figure 2 illustrates the input representation used by DiffKT3D for one head-and-neck patient with two clinically acceptable plans. Each row corresponds to one plan for the same anatomy. From left to right, we visualize the structure masks (PTV and OARs), the body/CT information, the rasterized beam plate, the angle plate encoding beam directions, and the resulting dose distribution.

We provide additional qualitative examples in Figure 3 to complement the quantitative results in the main paper. For three representative patients from the head-and-neck, lung, and prostate cohorts, we display the CT slice with contoured PTV and OARs, followed by the ground-truth dose distribution, the prediction from DiffKT3D, and the prediction from the challenge top-1 baseline. Below each row we plot the DVHs of the target and selected OARs, and we also show voxel-wise absolute error maps between the predictions and the ground-truth dose. These examples illustrate that, across disease sites, DiffKT3D better preserves PTV coverage while reducing hot spots in nearby OARs and body compared with the top-1 baseline.

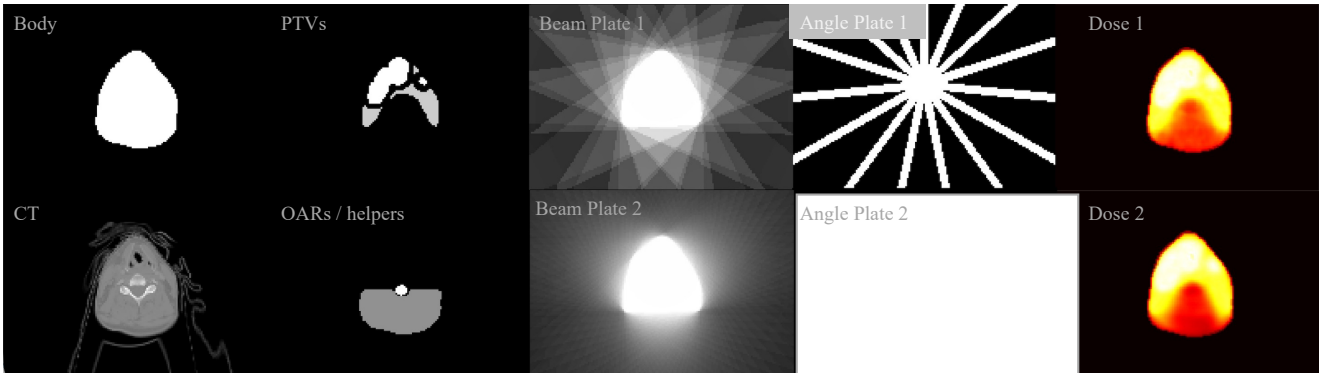

Figure 2. Example of the per-plan input modalities for one head-and-neck cancer patient with two different treatment plans (IMRT and VMAT). We show structure masks, body/CT, beam plates, angle plates, and the corresponding dose distributions.

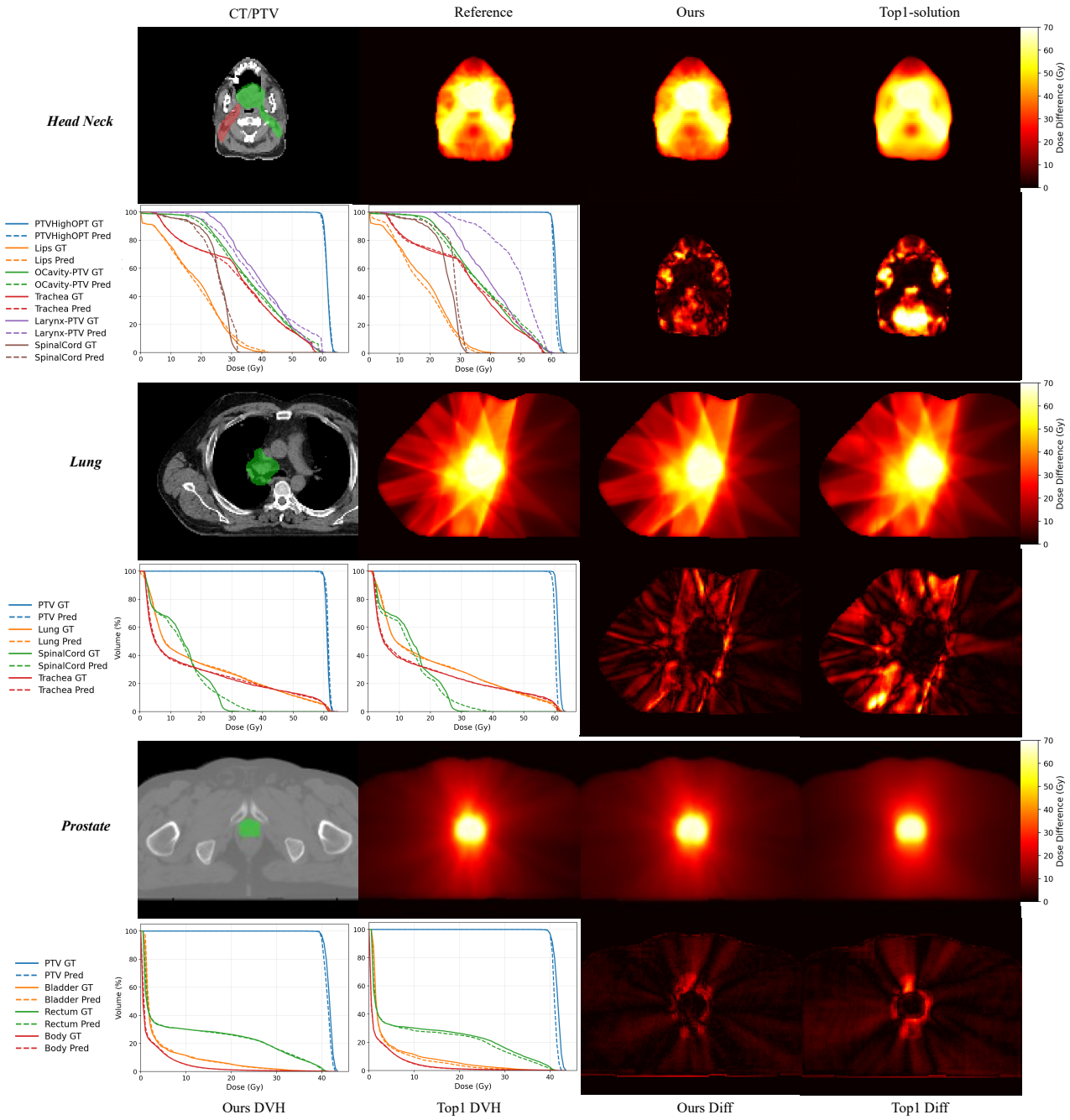

Figure 3. Qualitative comparison on representative head-and-neck, lung, and prostate cases. For each case we show CT with delineated PTV/OARs, ground-truth dose, DiffKT3D prediction, and the challenge top-1 baseline, together with DVHs and voxel-wise absolute error maps. Color bars are in Gy.

D. Additional Baseline Comparisons

To provide a broader assessment of DiffKT3D against alternative diffusion-based conditioning strategies, we evaluate three additional baselines on the GDP–HMM dataset: (i) a *3D ControlNet* [18] that applies ControlNet-style conditioning branches to the Wan 2.1 DiT backbone, (ii) a *2D slice-wise diffusion* model following prior RT dose prediction works [1] that processes each axial slice independently with a 2D diffusion backbone, and (iii) a *LoRA-only* variant that replaces full DiT fine-tuning with rank-64 LoRA adapters under the same Any2Any paradigm. Table 2 reports performance alongside the main models from the paper, together with per-case inference time and peak GPU memory on a single H100.

Table 2. Extended comparison on GDP–HMM (validation set). All diffusion models use 10-step sampling. Inference time and peak GPU memory are measured per case on a single H100 GPU.

Method	MAE↓	Score↑	Time (s)↓	Mem (GB)↓
Challenge Top-1 (regression)	2.03	134.26	3.19	3.08
Ours (MAISI, CT pretrain)	1.89	135.89	3.16	2.68
Ours (Conditional DiT)	2.07	135.41	8.35	6.67
Ours (Any2Any, full fine-tuning)	1.90	136.22	16.04	8.48
Ours (Any2Any + ScardNFT)	1.91	138.17	16.42	8.70
3D ControlNet	2.42	125.79	17.65	10.24
2D slice-wise diffusion	2.14	132.90	17.95	32.40
LoRA on Wan DiT + Any2Any	2.26	132.44	16.42	8.70

Analysis. The 3D ControlNet approach, while effective in natural image domains, performs substantially worse (MAE 2.42, Score 125.79) when applied to heterogeneous RT modalities. We attribute this to the fundamental mismatch between the ControlNet design—which assumes a single conditioning modality of the same domain as the generation target—and the RT setting where multiple structurally diverse modalities (CT, binary masks, beam plates, angle encodings) must jointly guide generation. The ControlNet copy-branch architecture cannot flexibly distinguish between these heterogeneous inputs, and its additional parameters increase memory overhead without compensating performance gains.

The 2D slice-wise diffusion baseline achieves reasonable voxelwise accuracy (MAE 2.14) but suffers from inter-slice inconsistency and significantly worse clinical Scorecard alignment (132.90). Processing each axial slice independently discards 3D spatial context that is crucial for dose conformality, particularly in complex head-and-neck geometries where dose gradients span many slices. Additionally, this approach requires substantially more GPU memory (32.40 GB vs. 8.70 GB for our full model) due to the need to process all slices sequentially and stitch results.

The LoRA-only variant (MAE 2.26, Score 132.44) demonstrates that parameter-efficient fine-tuning alone is insufficient to fully adapt the pretrained Wan backbone for the RT domain when used throughout the entire training pipeline. Full fine-tuning of the DiT blocks during the main supervised training stage remains essential for closing the large domain gap between natural video and medical dose data. However, as discussed in Section B.5, LoRA proves effective for the lightweight ScardNFT post-training stage, where it stabilizes RL updates without degrading the well-trained base model.

Computational context. In a clinical RT planning workflow, optimization-based planning typically requires 5–30 minutes per case depending on complexity and beam arrangement. In this context, the 16-second inference time of DiffKT3D—even with 10-step sampling—is well within practical deployment thresholds. Dose prediction serves as an initialization or quality-assurance tool rather than the final deliverable, and faster single-step inference ($\sim$2 s) can be used for interactive exploration when speed is preferred over peak accuracy. We note that further runtime reductions are achievable through CUDA/C++ deployment optimization and model distillation, which we leave for future work.

E. Extended Ablation Studies

We present two additional ablation studies that complement the analyses in the main paper: (i) the effect of noise-prediction (ϵ-prediction) parameterization, and (ii) the impact of VAE adaptation strategies.

E.1. Noise Prediction Parameterization

Table 2 in the main paper compares x_0-prediction and v-prediction under the Any2Any framework. Here we additionally evaluate ϵ-prediction (noise prediction), which is the most common parameterization in standard diffusion literature [4], to provide a complete picture of prediction-type choices.

Table 3. Comparison of prediction parameterizations on GDP–HMM (validation).

Prediction Type	Steps	MAE↓	Score↑
x_0-pred	1	2.45	117.64
ϵ-pred	1	2.16	133.09
v-pred	1	2.12	133.59
v-pred	10	1.90	136.22
ϵ-pred	10	1.93	137.82
v-pred	10	**1.91**	**138.17**

Analysis. Both ϵ-pred and v-pred substantially outperform x_0-pred in the single-step regime, confirming that direct signal regression is suboptimal for the diffusion formulation adopted by DiffKT3D. At 10 steps, v-pred achieves the strongest overall result in this comparison, while ϵ-pred remains competitive. This trend is consistent with observations in the fast diffusion sampling literature [14], where v-parameterization improves training stability and sample quality under few-step inference. We therefore adopt v-pred as the default for all main experiments.

E.2. VAE Adaptation Strategies

DiffKT3D keeps the pretrained Wan 2.1 VAE entirely frozen during training. Here we evaluate whether adapting the VAE decoder—via LoRA or full fine-tuning—could reduce the reconstruction gap introduced by the domain shift from natural video to medical dose distributions.

Table 4. Effect of VAE adaptation on GDP–HMM (validation). All variants use the same Any2Any DiT with v-pred and 10-step sampling, followed by ScardNFT post-training.

VAE Strategy	MAE↓	Score↑
Frozen VAE (default)	1.91	138.17
VAE Decoder LoRA	1.90	138.24
VAE Decoder fine-tuning	1.89	138.39
Full VAE fine-tuning	2.54	121.76

Analysis. Decoder-only adaptation yields marginal improvements: LoRA on the decoder reduces MAE by 0.01 Gy, and full decoder fine-tuning reduces it by 0.02 Gy while slightly improving the clinical Score. These gains are modest because the Wan VAE's latent space already provides a sufficiently expressive representation, and the DiT backbone compensates for residual distributional differences through its fine-tuned generation process.

In contrast, full VAE fine-tuning (encoder + decoder) dramatically degrades performance (MAE 2.54, Score 121.76). This failure occurs because modifying the encoder destroys the pretrained latent space structure that the DiT backbone relies on, effectively negating the benefit of transfer learning. The encoder-side perturbation causes a distribution mismatch between the latent codes produced during training and those expected by the frozen DiT weights from pretraining.

Based on these results, we adopt the frozen-VAE strategy as the default. It preserves the pretrained latent space, avoids the risk of catastrophic drift from encoder adaptation, and adds no additional training cost. In settings where marginal gains are desired, decoder-only LoRA offers a safe middle ground with negligible overhead.

F. Statistical Significance Analysis

To quantify whether the performance improvements of DiffKT3D over the challenge top-1 baseline are statistically significant, we conduct paired t-tests on per-patient MAE and Scorecard values across the GDP–HMM test set.

Setup. For each of the 498 test patients, we compute: (i) the voxelwise MAE within the body mask, and (ii) the clinical Scorecard value aggregating PTV coverage and OAR sparing metrics. We then perform two-sided paired t-tests comparing our final model (Any2Any + ScardNFT) against the challenge top-1 regression baseline.

Results. Both tests yield $p < 10^{-3}$, confirming that the improvements in MAE (2.07 $\rightarrow$ 1.93 Gy) and Scorecard (134.81 $\rightarrow$ 137.55) are statistically significant and not attributable to random variation across patients.

Discussion on single-step vs. multi-step inference. While the aggregated Scorecard difference between single-step v-pred (133.59) and the post-trained 10-step v-pred model (138.17) may appear moderate in absolute terms, the Scorecard aggregates metrics across 30+ regions of interest (ROIs). Many organs distant from the tumor contribute similar scores regardless of sampling depth, which can mask substantial local improvements. For clinically critical structures near the target—where precise dose gradients directly impact treatment quality—the per-organ score differences can be substantial (e.g., differences of 0–12 points for individual OARs). This observation supports the use of multi-step refinement in clinical deployment, where localized dosimetric accuracy in high-gradient regions is paramount.

References

[1] Zhenghao Feng et al. Diffdp: Radiotherapy dose prediction via a diffusion model. *arXiv preprint arXiv:2307.09794*, 2023. 8

[2] Riqiang Gao, Mamadou Diallo, Han Liu, Anthony Magliari, Jonathan Sackett, Wilko Verbakel, Sandra Meyers, Rafe Mcbeth, Masoud Zarepisheh, Simon Arberet, Martin Kraus, Florin C. Ghesu, and Ali Kamen. Automating RT planning at scale: High quality data for AI training. *arXiv preprint arXiv:2501.11803*, 2025. 3

[3] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441. IEEE, 2025. 3, 4

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 9

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 4

[6] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 3

[7] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2014. 2

[8] Yaron Lipman, Ricky T. Q. Chen, and Heli Ben-Hamu. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[9] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022. 4

[10] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 2, 3, 4

[11] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[13] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, and Klaus H. Maier-Hein. Mednext: Transformer-driven scaling of convnets for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023. 3

[14] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022. 9

[15] Petra Seibold, Adam Webb, Miguel E. Aguado-Barrera, David Azria, Celine Bourgier, Muriel Brengues, Erik Briers, Renee Bultijnck, Patricia Calvo-Crespo, Ana Carballo, et al. Requite: a prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer. *Radiotherapy and Oncology*, 138:212–224, 2019. 3

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[17] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 4

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 8

[19] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations (ICLR)*, 2026. Oral. 4