Any2Any 3D Diffusion Models with Knowledge Transfer

A Radiotherapy Planning Study

Yuhan Wang[1,2,†] Zihan Li[2,3,†] Han Liu[2] Simon Arberet[2] Martin Kraus[2] Yuyin Zhou[1] Florin-Cristian Ghesu[2] Dorin Comaniciu[2] Ali Kamen[2] Riqiang Gao[2,*]

[1]UC Santa Cruz [2]Siemens Healthineers [3]University of Washington †Equal contribution *Corresponding author

Abstract

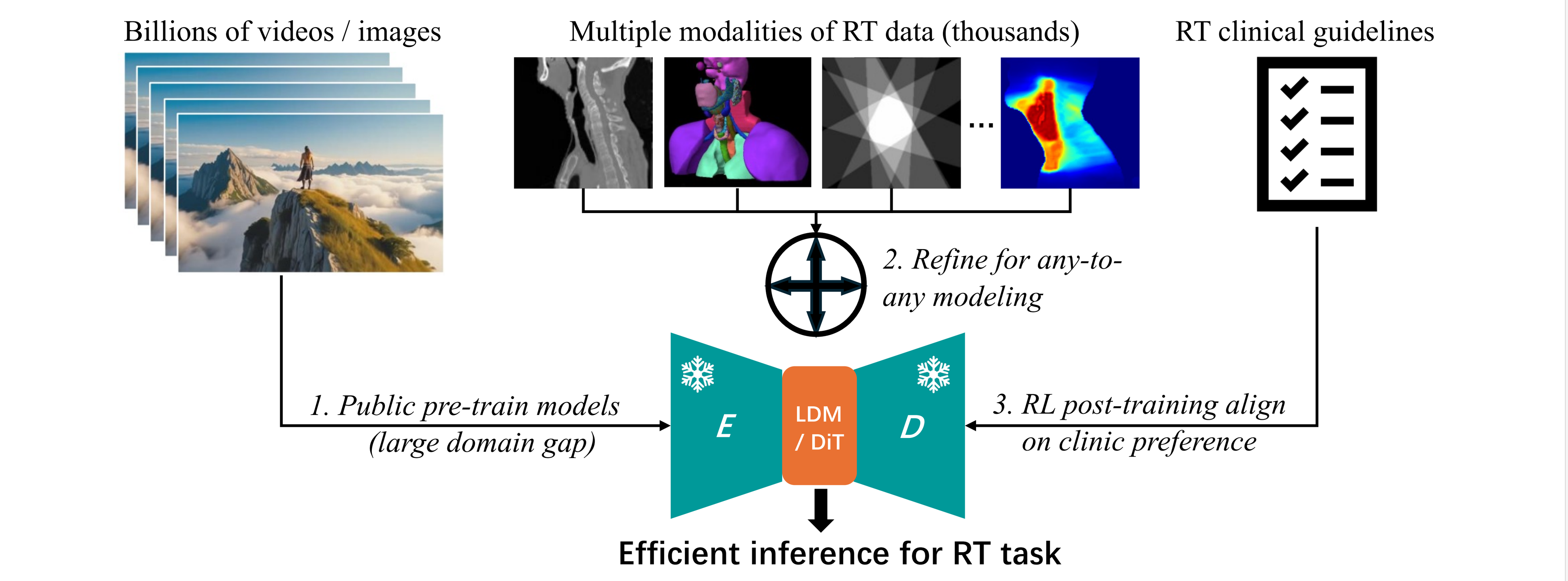
Background: Voxel-wise dose prediction is critical for practical radiotherapy planning; diffusion models have achieved remarkable success in various domains, but their potential in radiotherapy planning remains underexplored. **Method:** We introduce **DiffKT3D**, a unified Any2Any 3D diffusion framework that transfers priors from pretrained video/CT diffusion models, flexibly conditions on CT, PTV, OAR, body, beam, angle, and dose modalities, and post-trains dose generation using a clinically informed Scorecard reward. **Result:** **DiffKT3D** achieves new state-of-the-art performance on voxel-wise error and superior image quality, meanwhile **ScardNFT** improves plan preference alignment.

1. Introduction & Motivation

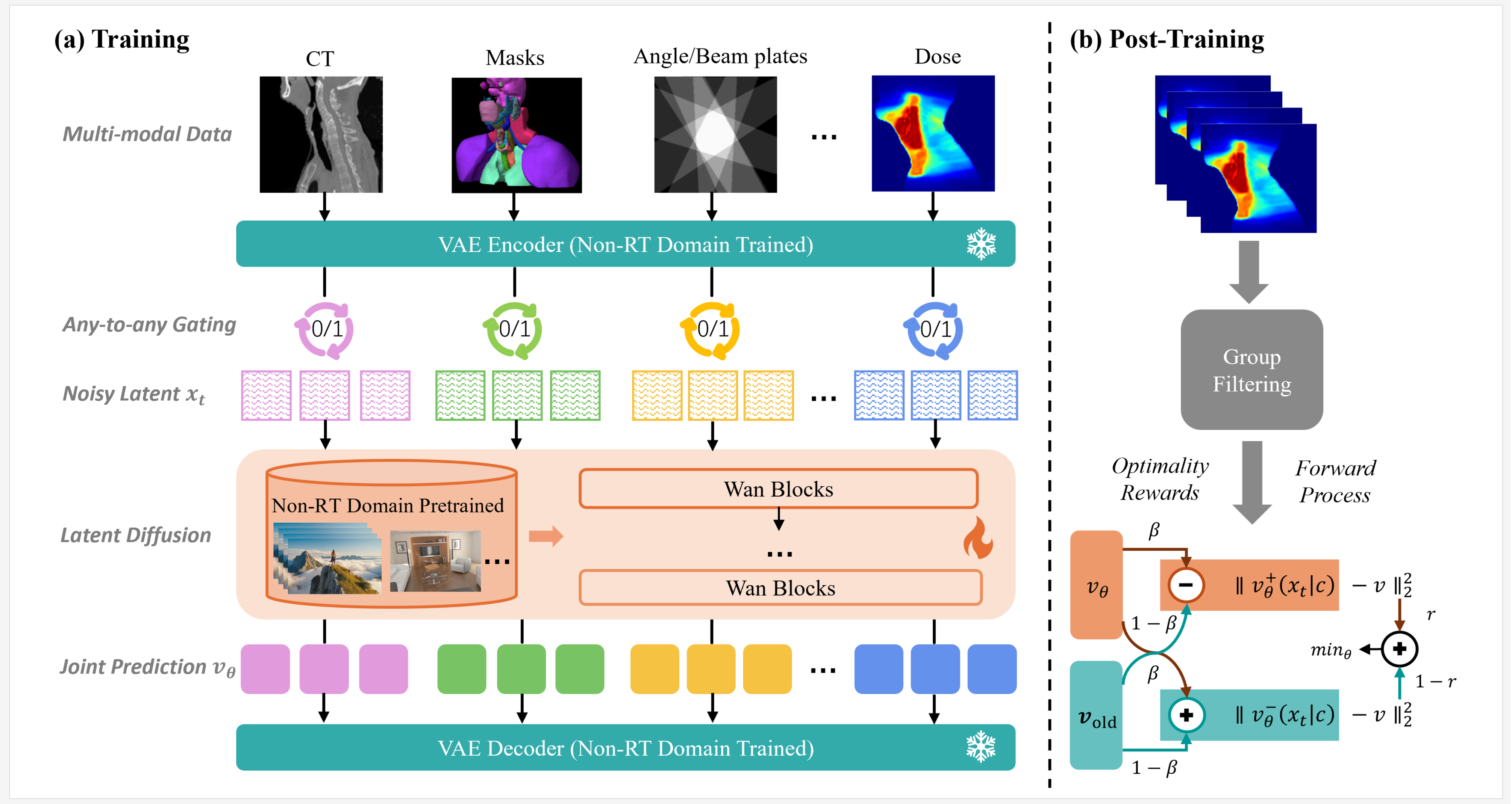
- RT dose prediction maps patient anatomy and machine settings to a 3D dose distribution for downstream planning and quality assurance.
- Strong supervised baselines optimize voxel losses, but clinical acceptability also depends on PTV coverage, OAR sparing, and institutional Scorecard preferences.
- General-domain diffusion models provide reusable 3D generative priors; the key question is whether these priors survive the domain gap to RT dose planning.

Contributions from the paper:

- Transfer video/CT diffusion priors into 3D RT generation.
- Train a modality- and role-aware Any2Any DiT over seven clinical modalities.
- Align generated dose with clinical Scorecards through RL post-training.



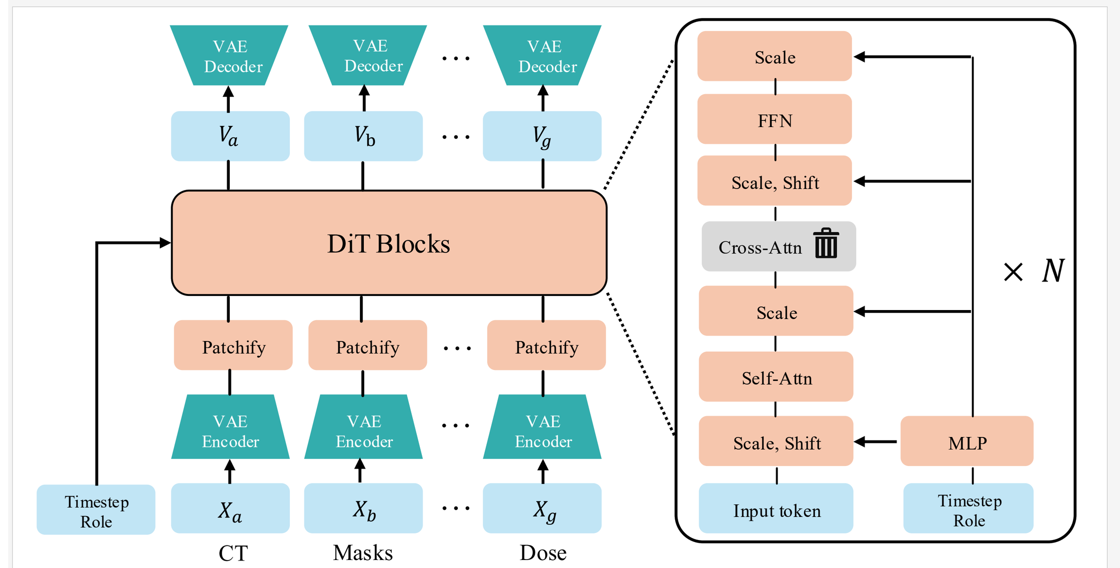
2. Methodology:Any2Any DiT Framework



- Volumes are encoded into a shared VAE latent space and projected by modality-specific 3D patch heads.
- Domain embeddings identify each modality; role embeddings mark target versus condition tokens.
- Clean condition tokens and noised target tokens interact through full self-attention, avoiding extra cross-attention modules.
- 4D RoPE adds a slot axis to spatial coordinates, preserving modality identity and 3D anatomy.
- **ScardNFT** improves per-structure Scorecard values, aligning dose predictions with institutional PTV coverage and OAR sparing preferences while preserving voxel MAE.

DiffKT3D combines 1) transferable diffusion priors for other domains, 2) Any2Any conditioning, and 3) clinical preference alignment with RL-post training in one 3D generation framework.

3. Dataset and Experimental Setup



Adjusted Wan backbone in our Any2Any Diffusion Model (DiffKT3D).

- **GDP-HMM:** 2,878 train, 356 validation, and 498 test plans from head-and-neck and lung sites.
- **REQUITE prostate:** 5,100 train and 256 test plans for cross-site transfer after Eclipse re-optimization.
- **Training:** Any2Any pretraining, dose-only fine-tuning, then **ScardNFT** post-training.
- **Metrics:** Mean Absolute Error (MAE), Scorecard, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Dice coefficient, and Fréchet Inception Distance (FID).

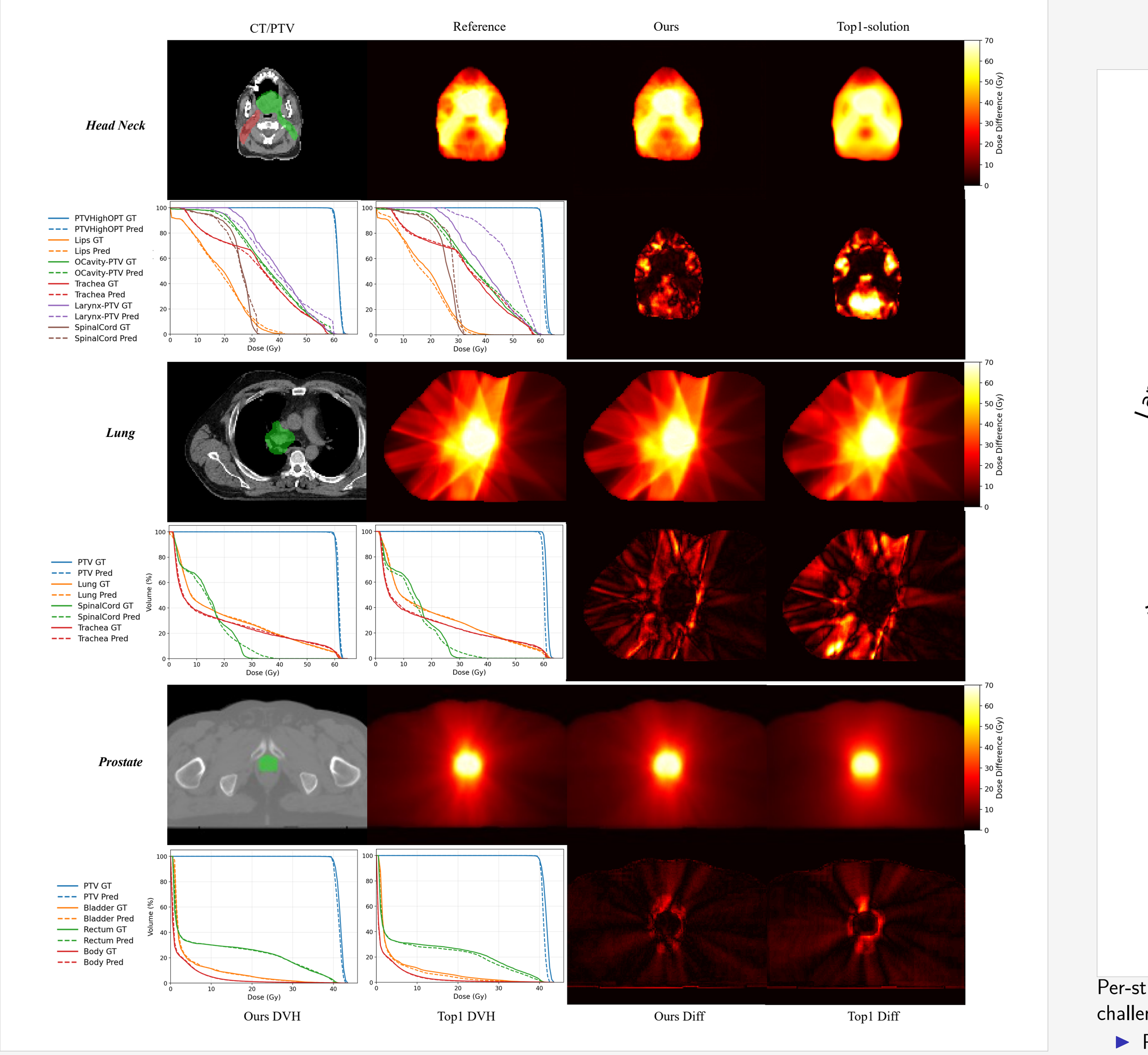
Setup Notes:

- Validate the value of general-domain diffusion priors for RT dose prediction.
- Validate the flexibility of Any2Any conditioning for both final dose prediction and cross-modal imputation.
- Validate the clinical preference alignment of **ScardNFT** for improving Scorecard metrics without sacrificing voxel-level accuracy.

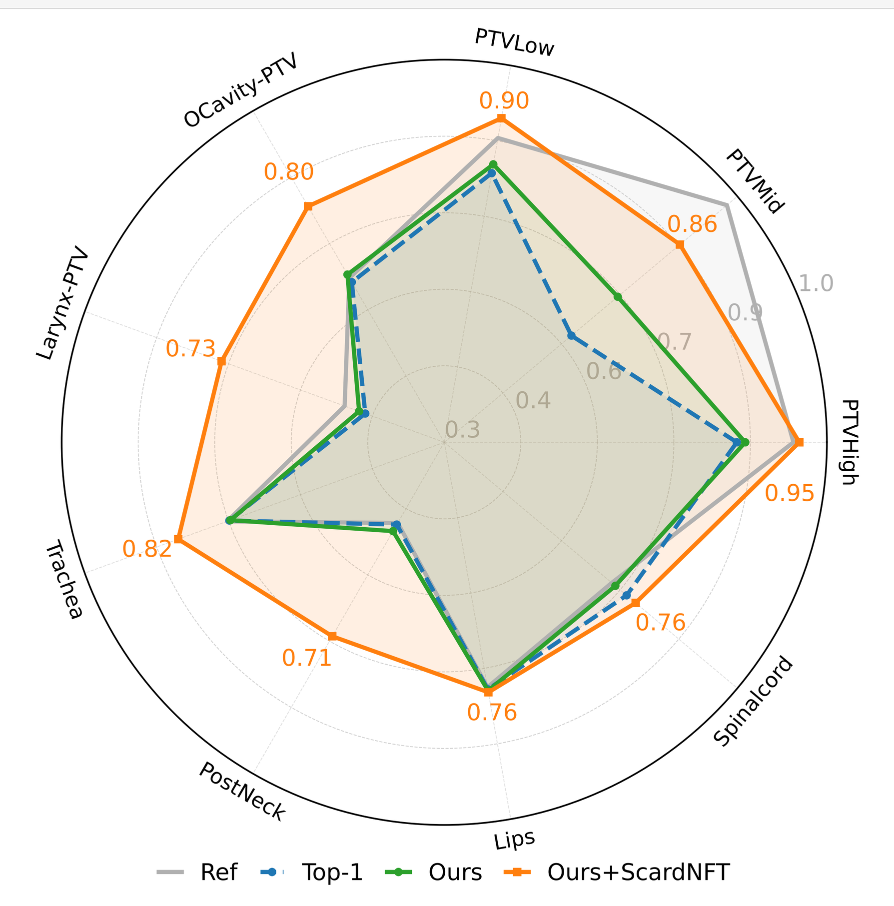
4. Main Results

Main results on GDP-HMM Challenge data(validation & test). Metrics: MAE (Gy; ↓), clinical Scorecard (Score) (↑), PSNR (dB; ↑), SSIM (↑), and LPIPS (↓).

Family	Method	Backbone	Pretrain	Any2Any	RL	Validation MAE↓	Validation Score↑	Validation PSNR↑	Validation SSIM↑	Validation LPIPS↓	Test MAE↓	Test Score↑	Test PSNR↑	Test SSIM↑	Test LPIPS↓
Regression	Yasin (AAPM '25 Chal.)	MedNeXt	–	–	–	2.03	134.26	31.50	0.972	0.037	2.07	134.81	32.06	0.974	0.033
Regression	tyxiong123 (AAPM '25 Chal.)	MedNeXt	–	–	–	2.18	133.95	29.89	0.929	0.067	2.20	134.10	30.62	0.925	0.064
Regression	rcgao (AAPM '25 Chal.)	MedNeXt	–	–	–	2.05	133.40	31.24	0.972	0.051	2.08	133.62	31.77	0.972	0.048
Regression	PVmed (AAPM '25 Chal.)	MedNeXt	–	–	–	2.15	133.78	31.40	0.971	0.034	2.17	133.99	31.53	0.970	0.031
Regression	SKLSDE-BH (AAPM '25 Chal.)	nnUnet	–	–	–	2.25	133.02	31.40	0.972	0.040	2.27	133.20	31.34	0.968	0.039
Regression	MedVision (AAPM '25 Chal.)	LDM	–	–	–	2.19	134.50	31.54	0.970	0.029	2.21	134.72	31.66	0.973	0.030
Diffusion	MAISI + Ours	LDM	✓	–	–	**1.89**	135.89	32.02	0.975	0.028	1.95	135.23	32.13	0.976	0.026
Diffusion	Ours (Conditional)	DiT	✓	–	–	2.07	135.41	31.88	0.973	0.031	2.12	134.60	32.01	0.974	0.029
Diffusion	Ours (Any2Any)	DiT	✓	✓	–	1.90	136.22	32.43	0.977	0.025	**1.93**	135.36	32.60	0.978	0.023
Diffusion	Ours (Any2Any+NFT)	DiT	✓	✓	✓	1.91	**138.17**	**32.55**	**0.979**	**0.022**	**1.93**	**137.55**	**32.73**	**0.980**	**0.020**



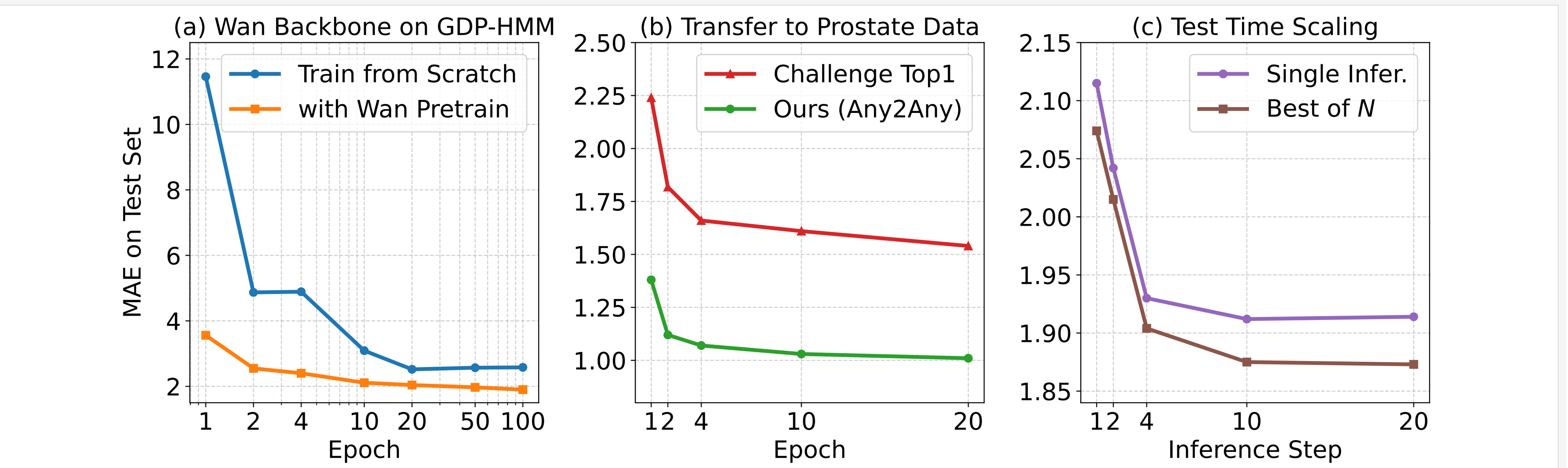
Qualitative comparison on representative head-and-neck, lung, and prostate cases. For each case we show CT with delineated PTV/OARs, ground-truth dose, DiffKT3D prediction, and the challenge top-1 baseline, together with DVHs and voxel-wise absolute error maps.



Per-structure scorecard (↑) value comparison of head-and-neck plans. The plot contrasts reference, the challenge Top-1 baseline, our diffusion model, and our RL-enhanced variant (Ours+ScardNFT).

- Post-training with **ScardNFT** improves clinical preference alignment across multiple structures, not just a single outlier endpoint.
- The improvement in Scorecard metrics is not at the cost of voxel-level accuracy, as MAE remains stable after RL post-training.

5. Ablation Study and Analysis



Part 1. MAE vs. training epochs / inference steps. The single figure contains three subplots: (left) pretrain vs. from-scratch across epochs, (middle) model-transfer finetuning curve, and (right) test-time scaling (single vs. best-of-n).

- Figure (a) clearly illustrates the substantial efficiency gains from Wan pretraining compared to training from scratch, reaching lower MAE values with significantly fewer epochs.
- Figure (b) demonstrates the strong adaptability when transferring to the prostate dataset, rapidly converging and consistently outperforming the top GDP–HMM challenge solution.
- Figure (c) highlights the benefit of a best-of-n inference strategy, providing a clear accuracy improvement over single inference.

Part 2. Component ablations on validation set.

Variant	Pre.	Any2Any	RoleEmb	FullAttn	PatchEmb	4D RoPE	MAE↓/Score↑
From scratch (baseline)	–	–	–	–	–	–	2.58 / 119.82
+ Pretrain	✓	–	–	–	–	–	2.07 / 135.41
+ Any2Any	✓	✓	✓	✓	✓	✓	1.90 / 136.22
- Role Emb	✓	✓	–	✓	✓	–	2.01 / 134.04
- Full Attention	✓	✓	✓	–	✓	–	2.15 / 130.80
- Patch Embed	✓	✓	✓	✓	–	–	2.02 / 134.71
- 4D RoPE	✓	✓	✓	✓	✓	–	1.96 / 135.27
Ours+ScardNFT	✓	✓	✓	✓	✓	✓	**1.91 / 138.17**

Part 3. Any2Any prediction under different prediction types and sampling steps. The v-pred are the default predictions used in our experiments.

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
ϵ-pred	10	1.93	137.82
v-pred	10	**1.91**	**138.17**

Analysis

- **Pretraining matters:** Compared with training from scratch, pretrained diffusion priors converge faster and reach substantially better final MAE and Scorecard performance.
- **Any2Any helps:** Moving from standard conditional diffusion to the unified Any2Any setup yields another clear gain, showing the value of flexible cross-modality modeling.
- **Core design choices are necessary:** Removing role embeddings, full attention, or modality-specific patch embeddings consistently hurts both voxel accuracy and clinical scores.
- **Role embeddings and 4D RoPE serve different purposes:** Role embeddings mark which modalities are conditions or targets, while 4D RoPE helps distinguish different modality inputs across space and slots.
- **ScardNFT improves alignment:** RL post-training raises clinical preference scores without sacrificing MAE, improving protocol alignment rather than only optimizing voxel loss.

6. Discussion and Takeaways

Conclusion

DiffKT3D is an Any2Any 3D diffusion framework for voxel-wise dose prediction and cross-modal imputation. By combining pretrained diffusion priors, modality-aware conditioning, and Scorecard-guided RL post-training, it outperforms strong regression and diffusion baselines on GDP–HMM and REQUITE.

Discussion

- **Diffusion priors:** Public CT and video diffusion backbones transfer well to RT dose modeling and reduce training cost.
- **RL post-training:** Clinical scorecards can be turned into reward signals to improve preference alignment beyond voxel accuracy alone.
- **Any2Any design:** A unified conditional architecture supports flexible multi-modal generation beyond dose prediction.

Limitations & Future Work

- Full 3D attention with Wan 1.3B is still computationally expensive; future work will study lighter backbones, sparse attention, token pruning, and distillation.
- The training objective can be improved with dose-specific terms such as DVH-based losses and weighted MAE.
- Next steps include broader clinical validation and extending Any2Any generation to other RT planning stages.

Main References, Acknowledgements, Disclaimer

[1] Wan: Open and advanced large-scale video generative models. arXiv:2503.20314, 2025.
[2] Automating RT Planning at Scale: High Quality Data For AI Training, arXiv:2501.11803, 2025.
[3] DiffusionNFT: Online Diffusion Reinforcement with Forward Process. ICLR 2026.

Acknowledgement: 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.