

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

Yuhan Wang^{1,2,†} Zihan Li^{2,3,†} Han Liu² Simon Arberet² Martin Kraus²
 Yuyin Zhou¹ Florin-Cristian Ghesu² Dorin Comaniciu² Ali Kamen² Riqiang Gao^{2,*}
¹UC Santa Cruz ²Siemens Healthineers ³University of Washington [†]Equal contribution

Abstract

Voxel-wise dose prediction is a critical yet challenging task in radiotherapy (RT) planning, as bespoke models trained from scratch often struggle to generalize across institutions, scanners, and planning protocols. Meanwhile, large generative backbones pretrained on billion-scale visual data have learned strong geometric and structural priors that may transfer beyond their source domain. We propose **DiffKT3D**, an Any2Any 3D diffusion framework that transfers prior knowledge from pretrained video diffusion models to clinically meaningful dose prediction. To support flexible conditioning over heterogeneous clinical inputs, including CT, target and organ masks, body mask, and beam descriptors, we introduce a modality- and role-aware Any2Any formulation that avoids cross-attention while retaining a unified conditioning interface. We further apply reinforcement-learning (RL) post-training with a clinically informed Scorecard reward to align generations with institutional treatment preferences. On the GDP-HMM challenge, DiffKT3D reduces voxel-wise MAE from 2.07 to 1.93 relative to the top challenge solution, while also improving image quality and clinical preference alignment. These results show that pre-trained diffusion priors, combined with flexible multimodal conditioning and scorecard-aligned post-training, provide an effective approach to RT dose prediction.

1. Introduction

Radiotherapy (RT) is one of the most common cancer treatments. Dose prediction (DP) is an important AI task in RT planning: it estimates a 3D dose distribution from patient anatomy and machine configuration by learning from large collections of historical deliverable plans [3, 13, 15, 27]. A typical DP setting uses a planning CT scan and delineated structures, such as planning target volumes (PTVs) and organs at risk (OARs), as inputs and a reference 3D RT dose as the target. Dose predictors support several stages of the

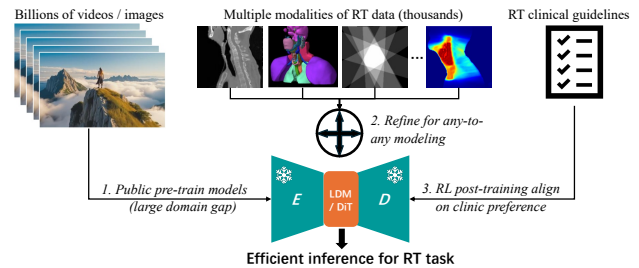


Figure 1. Illustration of the proposed DiffKT3D. We first transfer priors from diffusion models pretrained on large-scale public video or CT data, despite a substantial domain gap to radiotherapy dose prediction. These backbones are then adapted to heterogeneous RT modalities with relatively limited data, followed by RL post-training driven by guideline-derived clinical Scorecards to better align predictions with institutional planning preferences.

RT pipeline, including planning optimization [4], fluence prediction [52], leaf sequencing [14], and plan-quality improvement [17]. For precise and personalized prediction, additional conditions such as beam geometries and prescription information can also be incorporated.

Historically, DP has been formulated as a supervised regression problem, where models minimize voxel-wise errors such as MAE and MSE between predicted and reference dose [3, 15]. Recent advances in generative modeling have opened new directions for DP. Generative models such as GANs [13, 25] and diffusion models [12, 64] have shown promise in capturing complex dose distributions and improving prediction quality. However, most existing diffusion-based methods are trained from scratch, are often limited to slice-wise formulations, and rarely include a post-training stage for clinical alignment.

Foundation models have also shown strong transfer across natural language processing, computer vision, and medical imaging [2, 37, 39]. Many are pretrained with self-supervised objectives or on tasks unrelated to the target application, yet they adapt effectively through lightweight finetuning or even training-free use [21, 44, 53]. Recent studies further suggest that *general-domain* models pretrained on natural images can transfer to medical imaging despite the

*Corresponding author: riqiang.gao@siemens-healthineers.com

domain gap [21, 30, 44, 62]. Evidence on generative transfer is more limited, although recent work such as Text2CT shows that large diffusion backbones and general text encoders can be adapted to medical volume generation [18]. Existing medical work, however, has focused mainly on feature extraction backbones rather than generative backbones. This motivates our first question: *Can 3D diffusion priors trained on a distant source domain improve target-domain generation?*

A useful bridge comes from dense prediction in natural images. Recent work has repurposed large diffusion backbones for geometric prediction tasks such as monocular depth estimation, where pretrained generative priors act as structured predictors rather than pure image synthesizers [23]. RT dose fields share several of these properties: they are dense, globally constrained, anatomy-aware, and shaped by heterogeneous conditioning signals. Unlike standard image-to-image settings, however, dose planning is not a single-condition problem. Beyond CT and anatomical structures, treatment geometry materially affects the final dose field. This makes RT dose prediction a natural testbed for whether pretrained diffusion priors can be adapted to a more structured multimodal generation problem.

Preference alignment is another important consideration in generative AI [38, 40]. It is especially relevant to RT because a multidisciplinary team of oncologists, physicists, and dosimetrists designs treatment plans for individual patients, while institutional protocols can differ substantially [8, 11]. Reinforcement-learning-based post-training has emerged as an effective way to align generative models with user preferences in NLP and computer vision [9, 26, 49, 56, 59]. Its use in medical imaging, especially RT dose planning, remains limited. In this setting, preference is multifaceted and reflects clinically meaningful trade-offs between PTV coverage and OAR sparing. More importantly, the standard diffusion objective optimizes denoising quality rather than the final plan utility measured by clinical criteria. This motivates our second question: *Can we align diffusion generations with clinical preferences via RL post-training?*

As illustrated in Figure 1, we propose **DiffKT3D**, a novel **3D Diffusion** model framework addressing two critical **Knowledge Transfer** questions. Three main contributions of DiffKT3D are outlined below.

First, we adapt advanced 3D diffusion models pretrained on natural videos (Wan 2.1 [50]) or CT data (MAISI [19]) and lightly fine-tune them for dose prediction. Despite the domain gap, this transfer strategy improves both accuracy and efficiency. The benefit becomes even more pronounced when transferring across related RT sites, demonstrating stronger generalization than regression-based models.

Second, because DP involves heterogeneous inputs spanning seven modalities {CT, PTV, OAR, body, beam plate, angle plate, dose}, we introduce a modality- and role-aware

Any2Any conditioning scheme. In this framework, any modality can serve as the target while the remaining modalities act as conditions. This design supports flexible input combinations, enables cross-modal imputation, and improves robustness across diverse clinical scenarios.

Third, we introduce a RL post-train, termed as **Scard-NFT**, with a new rewarding mechanism based on planning preferences captured by **Scorecard**. This method is inspired by success of DiffusionNFT [65] in text-to-image alignment and tailored for clinical-guided refinement.

We evaluate DiffKT3D on 3,732 plans from GDP-HMM and on 5,356 re-optimized plans derived from the REQUITE prostate cohort (9,088 in total) [15, 43]. The model improves MAE over top challenge solutions and also improves image quality and preference alignment. Beyond transfer from non-RT source domains, DiffKT3D adapts efficiently from head-and-neck and lung data to prostate planning with limited additional finetuning. More broadly, these results support the use of pretrained diffusion priors for clinically relevant multimodal generation in RT planning.

2. Related Work

RT Dose Prediction. Voxel-wise dose prediction has gained popularity in recent years due to advances in deep learning. Inspired by the success in image segmentation, Convolutional UNet [41] and its variants including ResUNet [10], H-DenseUNet [29], and MedNeXt [42] have been used for different cancer sites including head-and-neck [32, 35, 45, 51], lung [7, 13, 22], prostate [24, 25, 36], and esophageal cancer [60]. Although many studies use regression losses such as L1 or L2, researchers [12, 13, 25, 64] have also explored generative methods, including GANs [16] and diffusion models [20], to improve image quality. Across challenges like OpenKBP [3] and GDP-HMM [15], there is growing emphasis on generalizable models that handle multiple contexts rather than highly specialized ones. In computer vision, diffusion models often outperform GANs in complex scenarios, but state-of-the-art diffusion methods are data-hungry, computationally expensive, and mostly designed for 2D. Existing work [12, 64] trains slice-wise diffusion models, which struggle with spatial consistency across slices, highlighting the need for efficient 3D diffusion models for dose prediction.

Diffusion Priors and Any2Any Generation. Diffusion models have achieved notable success in conditional generation tasks across various modalities, initially excelling in image synthesis and later adapted for dense prediction tasks like depth estimation and segmentation [23]. Recent frameworks have further extended diffusion models to Any2Any generation, enabling unified conditional generation for arbitrary modality pairs, exemplified by Versatile Diffusion [57], UniDiffuser [5], CoDi [46], OmniGen [55], and OmniFlow [28]. Techniques like ControlNet [61], T2I-Adapter [34], and MultiDiffusion [6] further facilitate flexible integration

of diverse conditions, including edge maps and segmentation masks. Additionally, instruction-tuned models such as PixWizard [31] and joint generation-understanding models like JoDI [58] enhance practical applicability.

Most existing Any2Any and controllable diffusion frameworks, however, are developed for natural-image settings where conditions are relatively homogeneous, such as RGB images, text prompts, layouts, or edge maps. They do not directly address clinical volumetric generation, where multiple heterogeneous conditions must be fused jointly and where some conditions may be absent or unreliable at test time. RT dose prediction is especially demanding in this regard because anatomy, target definition, and beam geometry all contribute to the final field. These characteristics make it insufficient to simply concatenate all conditions into a standard conditional diffusion backbone. Instead, they motivate a unified transformer that can distinguish target and condition roles while preserving a shared multimodal latent space.

Inspired by these methodological advances, we adapt diffusion priors to radiotherapy dose prediction, leveraging their flexibility and cross-modal generalization for robust clinical outcomes.

Post-training for Diffusion Models. Standard diffusion models typically optimize voxel-level losses (e.g., MSE or MAE), which often misalign with nuanced clinical objectives. To address complex, preference-driven tasks, recent methods such as DDPO [9], Diffusion-DPO [49, 59], and DiffusionNFT [65] apply reinforcement learning (RL) strategies for preference-based fine-tuning. However, their clinical applicability remains underexplored, as general-purpose preference frameworks like ImageReward [56], HPS [54], Pick-aPic [26], and MPS [63] lack necessary interpretability. In contrast to open-domain generation, RT planning requires reward signals that remain traceable to explicit clinical criteria and can be audited by physicians and physicists. Radiotherapy dose prediction involves complex trade-offs encoded in institutional guidelines, motivating our clinically-informed RL-based post-training strategy. We introduce a Scorecard reward mechanism explicitly encoding clinical metrics, guiding diffusion generations toward institutional preferences and enhancing clinical acceptability.

3. Method

3.1. Problem Description and Motivation

RT planning involves designing a treatment strategy that precisely delivers radiation to the target region while minimizing exposure to nearby healthy tissues and organs. Dose prediction aims to generate accurate 3D dose distributions from diverse multimodal inputs (as shown in Figure 1), often under inconsistent or incomplete contouring of the PTVs and OARs. Because planning protocols vary across institutions, preferences in plan quality also differ. These variations can

be captured using clinical Scorecards [33, 47, 48], which quantitatively evaluate and compare treatment plans. Consequently, *generalizable dose prediction can be framed as a 3D generation problem conditioned on varied inputs and guided by preference-oriented objectives.*

Motivation for Methodology. Diffusion-based generative models have achieved remarkable success in computer vision, typically trained on billions of samples. In contrast, most dose prediction models rely on only hundreds or thousands of cases due to the limited availability of RT data. The DINO family has demonstrated strong feature extraction capabilities that transfer well to medical imaging even when trained on natural images, yet our work focuses on generation. We therefore explore leveraging large-scale diffusion priors from non-RT domains for dose prediction, forming our first methodological contribution.

After establishing that diffusion-based generative knowledge can effectively transfer across domains and modalities (*first contribution*), we further aim to enhance performance within the RT domain, which involves diverse multimodal inputs. To this end, we formulate dose prediction as an Any2Any conditional generation task, enabling flexible selection of any target modality τ from the set $\mathcal{M}=\{\text{ct}, \text{ptv}, \text{oar}, \text{body}, \text{beam}, \text{angle}, \text{dose}\}$, conditioned on the remaining available inputs $C=\{x_0^{(m)} \mid m \in S \subseteq \mathcal{M} \setminus \{\tau\}\}$. The target modality at the t -th step of the forward diffusion process becomes:

$$x_t^{(\tau)} = \alpha_t x_0^{(\tau)} + \sigma_t \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{I}), \quad t \sim \mathcal{U}(0, 1), \quad (1)$$

where x_0 denotes the clean data, and $\mathcal{N}$ and $\mathcal{U}$ represent Gaussian and uniform distributions, respectively. The coefficients α_t and σ_t follow a standard variance-preserving noise schedule (with $\alpha_t^2 + \sigma_t^2 = 1$), and discrete diffusion steps are linearly rescaled to $t \in [0, 1]$ for notational simplicity. This formulation allows any modality in $\mathcal{M}$ to serve as the generation target, facilitating cross-modality diffusion knowledge transfer while mitigating overfitting in low-data regimes.

In addition, reinforcement learning (RL) has achieved notable success in LLM post-training and text-to-image alignment, both of which share a common goal with RT clinical Scorecards: capturing user or expert preferences. Motivated by this similarity, we adapt the DiffusionNFT framework by reformulating a clinically grounded Scorecard into an RL reward function for post-training. This provides an additional safeguard for out-of-distribution cases by explicitly aligning the generative model with clinically preferred trade-offs.

Figure 2 summarizes our approach, with detailed model structure provided in Supplementary A, and method details presented in the following subsections.

3.2. Any2Any DiT framework

Patchify Head. Our diffusion process operates entirely within a shared VAE latent space. For each modality m ,

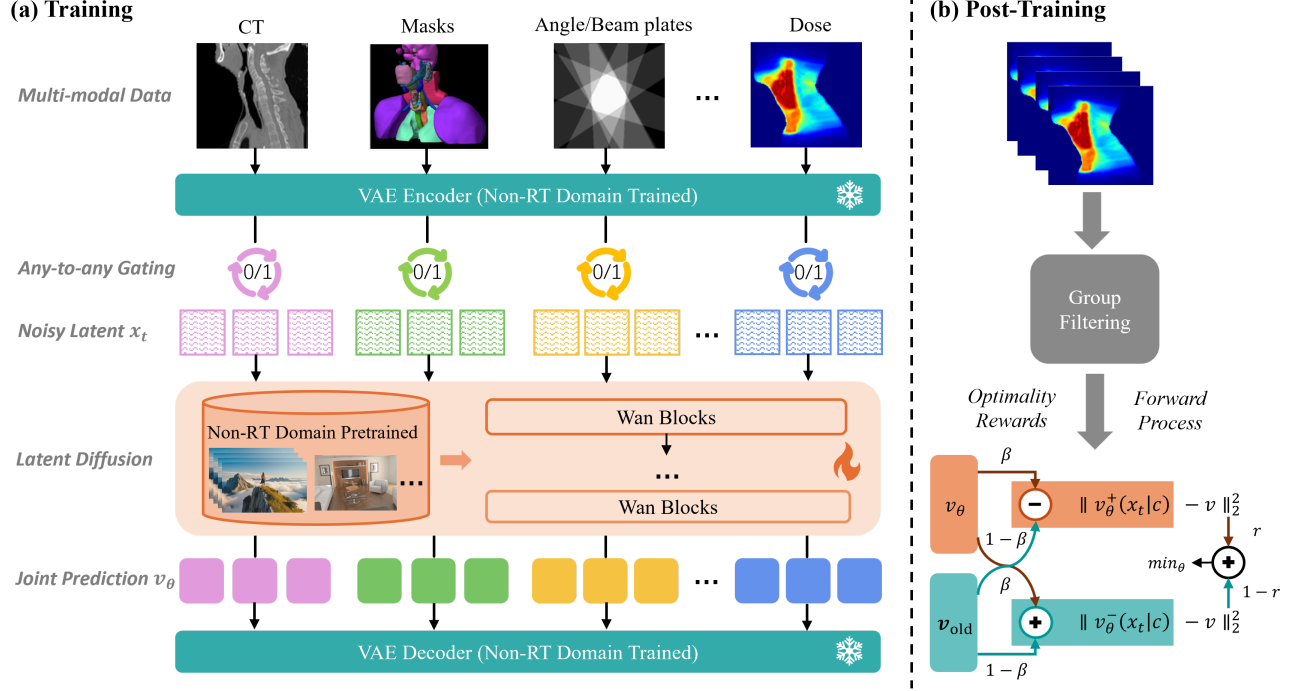


Figure 2. Training mechanism for DiffKT3D. The multimodal inputs first pass through the VAE encoder to obtain latent features. With the Any2Any gating mechanism, each modality is randomly assigned as either a condition or a target. Conditional modalities are independently encoded into patch tokens, while target modalities are combined with latent noise x_t . Each token carries a role embedding that distinguishes targets from conditions, while modality identity is conveyed by modality-specific patch heads and slot-aware 4D RoPE. The DiT jointly attends to the clean condition tokens and the noised target tokens, predicting the noise parameterization (v_θ) for the selected target modality. The VAE and DiT are pretrained on videos, and only the DiT blocks are fine-tuned. During post-training, we convert a clinically informed Scorecard into an RL reward, improving clinical preference alignment while maintaining voxel-level fidelity.

raw input volumes are first encoded into latent representations by a pretrained, frozen VAE encoder. We then reuse the 3D patch embedding block from the Wan DiT backbone and extend it to be modality-specific: a lightweight modality-specific patch embedding PE_m , implemented as a compact 3D convolution with the same structure as the original DiT patch embed, projects these latent grids (or their noised versions $x_t^{(\tau)}$ for target modalities) into tokens with hidden dimension D . After diffusion in token space, predictions are decoded back to the original voxel space via the VAE decoder. This design keeps the DiT backbone architecture unchanged while delegating modality handling to the patch-embedding blocks, allowing the backbone to operate in a unified latent token space.

Role-aware conditioning and AdaLayerNorm. We use a learnable binary role embedding $e^{\text{role}} \in \{e^{\text{tar}}, e^{\text{cond}}\}$ to tag each token as either the noised prediction target or a clean condition. This role embedding is added directly to the corresponding input token, so target/condition identity is represented at the per-token level rather than through a separate global conditioning branch. Following Wan, all transformer blocks still share a single AdaLayerNorm modulation network driven by the timestep embedding e_t . This design keeps

the backbone architecture unchanged, introduces no additional pooling or cross-attention modules, and allows the model to jointly process target and condition tokens with full self-attention.

Slot-aware 4D RoPE positional embeddings. To explicitly encode token origins (dose vs. conditions) while preserving precise 3D spatial relationships, we extend standard 3D RoPE with an additional slot axis, yielding a 4D RoPE. Tokens are arranged in slot-major order, where each slot index S corresponds to one modality (the target dose or a particular condition).

For each attention head, we split the channel dimension d into four sub-dimensions assigned to the slot and the three spatial axes:

$$d = d_S + d_H + d_W + d_D. \quad (2)$$

This allocation reserves a dedicated subspace for each axis, so that slot and spatial phases can be rotated independently.

For each axis $a \in \{S, H, W, D\}$, sinusoidal frequencies are precomputed as

$$\text{freqs}_a(i) = \theta_a^{-2i/d_a}, \quad i = 0, 1, \dots, \frac{d_a}{2} - 1, \quad (3)$$

where θ_a is a base period hyperparameter for axis a . Following standard RoPE parameterization, larger θ_a values

yield longer-wavelength components. In all experiments, we simply set $\theta_S = N_{\text{slots}}$ and $\theta_H = H$, $\theta_W = W$, $\theta_D = D$, so that the lowest frequencies roughly span the full extent of each axis while higher indices encode finer spatial detail.

The 4D rotary embedding for a token at coordinates (S, H, W, D) concatenates axis-wise embeddings:

$$\text{RoPE}_{4D}(S, H, W, D) = [\text{RoPE}_S(S), \text{RoPE}_H(H), \text{RoPE}_W(W), \text{RoPE}_D(D)]. \quad (4)$$

During self-attention, queries and keys are rotated as

$$\begin{aligned} Q' &= \text{RoPE}_{4D}(S, H, W, D) \circ Q, \\ K' &= \text{RoPE}_{4D}(S, H, W, D) \circ K, \end{aligned} \quad (5)$$

where $\circ$ denotes element-wise complex rotation on paired channels, and the rotated (Q', K') are then used in standard dot-product attention. The slot axis supplies a dedicated rotary phase per modality (e.g., $S=0$ for dose and $S \geq 1$ for different conditions), while the spatial axes (H, W, D) are shared across slots. This design preserves a unified full-attention pass, but encourages structured cross-slot interactions (dose $\leftrightarrow$ CT/structures/beams) without adding extra parameters or attention blocks.

3.3. v -parameterized Diffusion Objective

All diffusion and training are performed in the VAE latent/token space; for notational simplicity, x_0 , x_t , and v denote latent targets unless otherwise stated. Instead of predicting noise ε , we adopt the v -parameterization, which provides a better balance of signal-to-noise across timesteps. Given the forward process in (1), we define

$$v(x_0, \varepsilon, t) = \alpha_t \varepsilon - \sigma_t x_0. \quad (6)$$

Given $(x_t^{(\tau)}, C, t)$ the model outputs $v_\theta(x_t^{(\tau)}, C, t) \in \mathbb{R}^{H \times W \times D}$ and is trained by

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, \varepsilon, \tau, S} [\|v_\theta(x_t^{(\tau)}, C, t) - v(x_0^{(\tau)}, \varepsilon, t)\|_2^2]. \quad (7)$$

This choice is notationally convenient and improves optimization stability on 3D dose grids. In particular, from $(x_t^{(\tau)}, v, t)$ we can recover x_0 and ε exactly:

$$x_0 = \alpha_t x_t - \sigma_t v, \quad \varepsilon = \sigma_t x_t + \alpha_t v. \quad (8)$$

Thus v_θ still parameterizes the same forward process as noise prediction, but yields better-conditioned gradients across diffusion timesteps in practice.

At each training step we sample a target modality τ uniformly from M and then draw a conditioning set $S \subseteq M \setminus \{\tau\}$ according to a simple curriculum on the number of observed modalities. This ensures that any modality can act either as a target or as a condition, exposes the model to diverse conditioning patterns, and strengthens robustness to missing or incomplete inputs at inference time.

3.4. Scorecard-aligned RL Post-training

Pure diffusion training does not explicitly optimize for clinical objectives, such as precise PTV coverage or OAR sparing. To bridge this gap, we propose **ScardNFT**, an RL-based post-training approach inspired by DiffusionNFT [65], which aligns generated dose distributions with clinical guidelines via differentiable, Scorecard-based rewards.

Scorecard Reward. We define a scalar clinical reward r^{raw} from standardized plan-quality metrics. For each anatomical structure $s \in \mathcal{S}$, the Scorecard specifies one of three metric types: `DoseAtVolume`, `VolumeAtDose`, or `MeanDose`, and maps each measured metric value into a normalized structure-specific score via a piecewise-linear function $\text{score}_s(\cdot)$. Let $\phi_s(y; C)$ denote the corresponding DVH-style statistic computed from a candidate dose y under conditions C , and let $w_s \geq 0$ be a tunable importance weight. The aggregate reward is then computed as a weighted sum:

$$r^{\text{raw}}(y, C) = \sum_{s \in \mathcal{S}} w_s \text{score}_s(\phi_s(y; C)), \quad y \equiv x_0^{(\text{dose})}. \quad (9)$$

We instantiate these clinical Scorecards using established radiotherapy templates for head and neck [47] and lung [48], covering critical dose-volume histogram (DVH) points, mean dose, and ring constraints. When patient prescriptions differ from the scorecard template, we proportionally rescale PTV thresholds to the prescribed dose to maintain consistent clinical scoring criteria across cases.

Normalization & Anchors. Plan quality can vary widely across patients and sites, so raw rewards must be normalized for stable learning. Within each case, we first standardize per-structure scores and aggregate them into r^{raw} , and optionally apply site-wise normalization to account for systematic site differences. To prevent reward hacking, we introduce two anchor terms: (i) a hinge penalty enforcing strict adherence to hard clinical constraints (e.g., minimum PTV D95 or maximum OAR thresholds), and (ii) an MAE anchor relative to available reference doses, discouraging trivial reductions in overall dose magnitude. The resulting optimality probability r used for RL is clipped to $[0, 1]$:

$$r = \frac{1}{2} + \frac{1}{2} \text{clip}\left(\frac{r^{\text{raw}} - \mathbb{E}_{y \sim \pi_{\text{old}}}[r^{\text{raw}}]}{Z_C}, -1, 1\right), \quad (10)$$

where π_{old} denotes the current diffusion policy and Z_C is a running estimate of reward dispersion under conditions C . Thus $r \in [0, 1]$ behaves as a Bernoulli-style optimality probability, with $r \approx 1$ indicating clinically preferred plans and $r \approx 0$ indicating poor plans under the same inputs.

Policy Update & Final Objective. Starting from a pre-trained diffusion checkpoint, we perform clinical preference-aligned policy updates via ScardNFT. For each training case, we first draw K candidate samples from multiple independent initial noises using a deterministic ODE sampler

Table 1. Main results on GDP-HMM (validation & test). Metrics: MAE (Gy; ↓), clinical Scorecard (Score) (↑), PSNR (dB; ↑), SSIM (↑), and LPIPS (↓). The main table spans both columns; bold indicates the best in its block.

Family	Method	Backbone	Pretrain	Any2Any	RL	Validation					Test				
						MAE↓	Score↑	PSNR↑	SSIM↑	LPIPS↓	MAE↓	Score↑	PSNR↑	SSIM↑	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	regao (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+ScardNFT)	DiT	✓	✓	✓	1.91	138.17	32.55	0.979	0.022	1.93	137.55	32.73	0.980	0.020

(Flow/DPM-Solver family, with scheduler state snapshots for reproducibility). Each candidate dose y is evaluated by computing its reward $r^{\text{raw}}(y, C)$, which is then transformed into an optimality probability r via (10). Inspired by DiffusionNFT [65], we introduce two *implicit* policy targets:

$$\tilde{v}_{\theta}^{+} = (1-\beta)v_{\text{old}} + \beta v_{\theta}, \quad \tilde{v}_{\theta}^{-} = (1+\beta)v_{\text{old}} - \beta v_{\theta}, \quad (11)$$

where v_{old} is derived from the current model (with gradients stopped), v_{θ} is the new prediction, and $\beta \in (0, 1]$ is a small mixing coefficient controlling how aggressively we move away from the old policy. We then optimize a dual loss that increases likelihood for higher-rewarded samples and penalizes lower-rewarded ones:

$$\mathcal{L}_{\text{NFT}} = \mathbb{E} \left[r \|\tilde{v}_{\theta}^{+} - v\|_2^2 + (1-r) \|\tilde{v}_{\theta}^{-} - v\|_2^2 \right], \quad (12)$$

where v denotes the ground-truth target from Eq. (6). The final training objective balances voxel-level diffusion consistency and clinical preference alignment:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{NFT}}(\theta) + \lambda \mathcal{L}_{\text{diff}}(\theta), \quad (13)$$

with $\lambda > 0$ controlling the strength of the RL-style update. **Why this works.** Intuitively, $\mathcal{L}_{\text{diff}}$ preserves voxelwise fidelity to historical plans, while $\mathcal{L}_{\text{NFT}}$ reshapes the conditional diffusion score field, assigning higher likelihood to dose configurations that satisfy clinical Scorecards under identical conditions C . The v -parameterization ensures well-scaled gradients across timesteps (α_t, σ_t) , which we find empirically stabilizes ScardNFT updates, especially in high signal-to-noise ratio regimes.

4. Experiments

4.1. Setup: Datasets, Splits, Metrics, and Training

We evaluate on the complete GDP-HMM Grand Challenge dataset [1], comprising official *training* (2,878 plans), *validation* (356 plans), and *test* (498 plans) splits for head-and-neck (HaN) and lung cancer sites, and on a prostate dataset derived from the REQUITE cohort [43], comprising *training*

(5,100 plans) and *test* (256 plans) after mask augmentation and plan re-optimization using the Eclipse Script API. Unless explicitly stated otherwise, voxelwise evaluations are conducted within the body mask.

Preprocessing. All cases are first mapped to the official challenge voxel spacing and orientation. For model input, we then extract a fixed ROI crop that is compatible with the Wan VAE. CT images undergo intensity clipping to $[-1000, 1000]$ HU, patient-wise z-score normalization, and mask binarization. Beam and angle plates are rasterized onto the same grid. Before patch embedding, each modality is min-max scaled to $[-1, 1]$. Detailed resampling, ROI cropping, and normalization are provided in the supplementary material. The beam and angle plates follow the GDP-HMM preprocessing protocol and are generated consistently with [13].

Primary metrics. We evaluate performance using the following metrics: (i) voxelwise mean absolute error (MAE), computed within the body mask with a 5 Gy threshold following the challenge protocol [15]; (ii) clinically-informed plan quality Scorecards [47, 48], integrating key PTV and OAR metrics into a single scalar score; and (iii) standard image quality metrics including PSNR, SSIM, LPIPS, Dice, and 2D slice-level FID.

Training details. We utilize the Any2Any DiT architecture described in Sec. 3.2. The backbone is Wan 2.1 (1.3B parameters), paired with VAE 2.1. Training proceeds through three distinct stages: (A) Any2Any pretraining with uniform target sampling and curriculum masking, (B) dose-only fine-tuning, and (C) ScardNFT post-training, balancing losses $\mathcal{L}_{\text{NFT}}$ and $\mathcal{L}_{\text{diff}}$ via a tunable hyperparameter λ . Further training details are provided in supplementary B.

4.2. Main Results on GDP-HMM Dataset

The GDP-HMM benchmark contains a broad set of strong regression-based challenge entries built on MedNeXt, nnUNet, and LDM backbones (Table 1). These models represent the best supervised pipelines available but remain constrained by voxel-level training. To form a complete comparison spectrum, we additionally include (i) a conditional diffusion U-Net adapted from the original MAISI backbone,

Table 2. Prediction parameterization for the supervised Any2Any model before ScardNFT. The final 10-step ScardNFT result is reported for reference.

Method	Pred Type	Steps	MAE↓	Score↑
Ours (Any2Any)	x_0 -pred	1	2.45	117.64
Ours (Any2Any)	v -pred	1	2.12	133.59
Ours (Any2Any)	v -pred	10	1.90	136.22
Ours (Any2Any+ScardNFT)	v -pred	10	1.91	138.17

using an x_0 -prediction objective with joint fine-tuning of the MAISI VAE decoder (see Supplementary B.6), and (ii) our own conditional DiT baseline that simply concatenates all conditioning modalities with dose for prediction.

Our Any2Any design yields an improvement in performance over both regression-style models and the concatenation-based diffusion baseline. Beyond numerical gains, this indicates that (1) jointly modeling all modalities in a unified diffusion space and (2) separating “role” (target vs. condition) through explicit embeddings are both essential for robust cross-modal dependency learning. Adding ScardNFT introduces consistent improvements in clinical alignment with minimal change in voxel-level fidelity, confirming that RL-guided updates reshape preference behavior rather than the underlying reconstruction quality.

Single-step prediction analysis. To better understand diffusion behavior in dose prediction, we compare x_0 -pred and v -pred in the supervised Any2Any model before ScardNFT (Table 2) and report the final 10-step ScardNFT result for reference. While x_0 -pred directly regresses the clean signal, the v -pred approach performs better in the single-step regime. Iterative refinement from 1 to 10 steps further improves the supervised model, and ScardNFT provides an additional gain in clinical preference alignment.

4.3. Results on REQUITE Prostate

We further assess transfer by fine-tuning checkpoints pre-trained on GDP-HMM (head-and-neck and lung) directly on the REQUITE prostate dataset. As shown in Table 3 and Figure 3(b), our Any2Any diffusion model converges faster and achieves better accuracy than strong regression baselines. Moreover, adopting a best-of-10 sampling strategy provides additional gains, indicating further headroom through stochastic decoding.

4.4. Ablations and Supporting Studies

Detailed ablation studies in Table 4 and Figure 3(a) clarify the contributions of key components and design choices. Introducing pretraining substantially accelerates convergence compared with training from scratch and yields a better final model, confirming the value of transfer from pretrained diffusion backbones. Transitioning from conditional diffusion to the unified Any2Any training paradigm further improves performance, demonstrating the benefit of flexible

Table 3. Comparisons on REQUITE-Prostate. We report MAE (Gy; ↓), PSNR (dB; ↑), SSIM (↑), and LPIPS (↓). Both ours and baselines are pretrained with GDP-HMM and fine-tuned on prostate data. † denotes best-of-10 oracle selection.

Method	MAE (Gy)↓	PSNR (dB)↑	SSIM↑	LPIPS↓
Yasin (AAPM '25 Chal.)	1.48	34.53	0.956	0.026
tyxiong123 (AAPM '25 Chal.)	1.37	34.74	0.957	0.023
rcgao (AAPM '25 Chal.)	1.53	33.51	0.953	0.031
PVmed (AAPM '25 Chal.)	1.52	33.54	0.950	0.030
SKLSDE-BH (AAPM '25 Chal.)	1.65	32.74	0.944	0.034
MedVision (AAPM '25 Chal.)	1.48	34.42	0.949	0.032
Ours (Any2Any)	1.01	36.80	0.963	0.012
Ours (Any2Any)†	0.97	37.09	0.965	0.011

Table 4. Component ablations on the validation set. The three removal rows are coupled ablations: when role embeddings, full attention, or modality-specific patch embeddings are removed, 4D RoPE is also disabled to keep the token-identity interface consistent.

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
w/o Role Emb (and w/o 4D RoPE)	✓	✓	—	✓	✓	—	2.01 / 134.04
w/o Full Attention (and w/o 4D RoPE)	✓	✓	—	—	✓	—	2.15 / 130.80
w/o Patch Embed (and w/o 4D RoPE)	✓	✓	✓	✓	—	—	2.02 / 134.71
- 4D RoPE	✓	✓	✓	✓	✓	—	1.96 / 135.27
Ours+ScardNFT	✓	✓	✓	✓	✓	✓	1.91 / 138.17

Table 5. Single-modality prediction under the remaining-1 (predict-one) setting. *What this table shows:* each listed modality is predicted from all the others. CT uses FID; segmentation-like modalities use Dice; Dose and Beam Plate use MAE only. Angle Plate is omitted because it is used as an auxiliary geometry representation rather than a standalone evaluation target.

Method	Output Modality	FID↓	Dice↑	MAE↓
Ours (Any2Any)	CT	1.47	—	—
Ours (Any2Any)	PTV mask	—	72.13	—
Ours (Any2Any)	OAR mask	—	54.63	—
Ours (Any2Any)	Body mask	—	95.09	—
Ours (Any2Any)	Dose	—	—	1.91
Ours (Any2Any)	Beam Plate	—	—	1.54

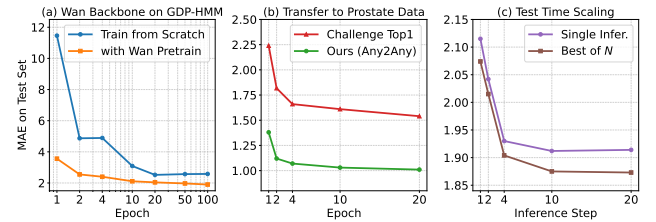


Figure 3. MAE vs. training epochs / inference steps. The single figure contains three subplots: (left) pretrain vs. from-scratch across epochs, (middle) model-transfer finetuning curve, and (right) test-time scaling (single-sample vs. best-of-10).

cross-modality modeling. The remaining removal rows are coupled ablations: when role embeddings, full attention, or modality-specific patch embeddings are removed, we also disable 4D RoPE to keep the slot/role interface consistent.

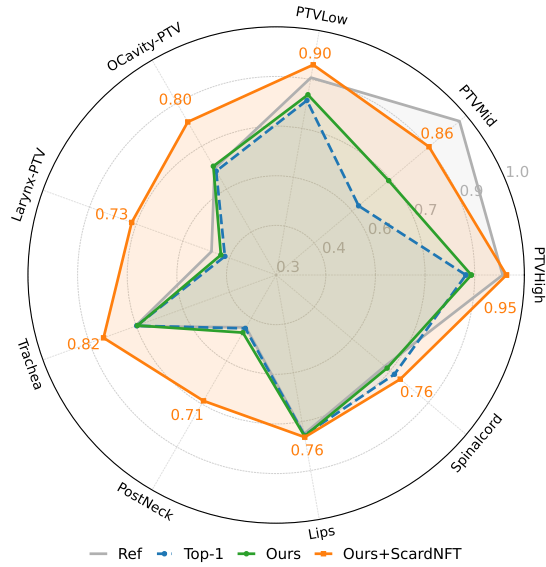


Figure 4. Per-structure scorecard value comparison of head-and-neck plans. The plot contrasts reference, the challenge Top-1 baseline, our diffusion model, and our RL-enhanced variant (Ours+ScardNFT), showing how reinforcement learning improves alignment with institutional planning objectives.

Under this setting, removing role information degrades both voxelwise accuracy and clinical score, replacing full attention with causal attention causes the largest drop, and removing modality-specific patch embeddings also hurts performance. Disabling 4D RoPE alone leads to an additional decline, showing that explicit slot-aware positional encoding remains useful even when the other components are kept. Finally, ScardNFT improves clinical alignment while preserving voxelwise accuracy.

Figure 3 further illustrates training efficiency and transferability. Figure 3(a) shows the efficiency gain from Wan pretraining compared with training from scratch. Figure 3(b) demonstrates rapid adaptation to the prostate dataset while outperforming the top GDP-HMM challenge solution. Figure 3(c) highlights the benefit of best-of-10 inference over single-sample inference.

As illustrated in Figure 4, ScardNFT post-training improves clinical alignment across key anatomical structures. Compared with the baseline without ScardNFT and the top regression method, the ScardNFT variant achieves higher clinical scores for PTV coverage and OAR sparing while maintaining nearly identical voxelwise MAE performance (Table 1). These gains suggest that scorecard-aligned post-training can improve plan quality assessment and related downstream use cases.

Qualitative examples in Figure 5 further illustrate model robustness. Compared with the leading regression baseline, our Any2Any+ScardNFT model produces dose distributions with better target conformity in head-and-neck cases, fewer artifacts in lung cases, and less oversmoothing in head-and-

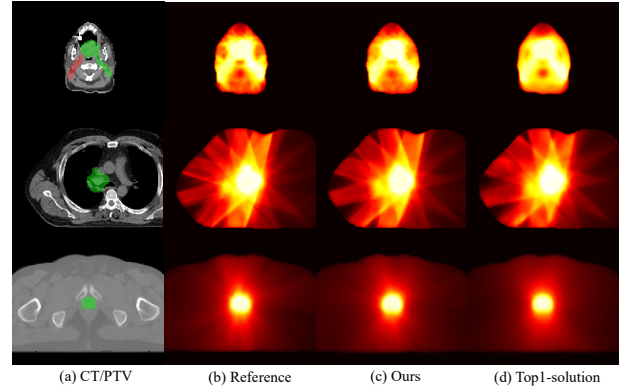


Figure 5. Qualitative predictions on GDP-HMM and REQUITE (top: head-and-neck, middle: lung, bottom: prostate). Columns: Input CT, reference dose, our Any2Any+ScardNFT, Top-1 baseline. Our method achieves superior target conformity.

neck and prostate cases. More visualizations are provided in Supplementary C.

Table 5 assesses robustness under the remaining-1 prediction scenario, where each listed modality is predicted from all other modalities. The model performs well across imaging (CT), segmentation (PTV, OAR, body mask), and dose-related outputs, supporting its cross-modal generative capability.

5. Conclusion and Discussion

Conclusion. We presented DiffKT3D, an Any2Any 3D diffusion framework for voxel-wise radiotherapy dose prediction and cross-modal imputation. By integrating pretrained diffusion priors with modality-aware conditioning and scorecard-aligned reinforcement learning, DiffKT3D outperformed strong regression and diffusion baselines on GDP-HMM and REQUITE.

Discussion. Our results suggest that pretrained diffusion priors can transfer effectively to clinically relevant multimodal generation, while RL-based post-training can be adapted to explicit treatment-planning criteria. Together, these findings support DiffKT3D as a practical framework for RT dose prediction under heterogeneous conditions.

This combination is attractive in RT, where planning data are expensive and optimization targets vary across institutions. In such settings, pretrained priors and clinically grounded post-training offer a scalable way to improve robustness and preference alignment.

Limitation and Future Work. DiffKT3D remains computationally expensive because it uses a large Wan backbone with full 3D attention. Future work will study lighter backbones, sparse attention, token pruning, and distillation. The training objective may also benefit from dose-specific losses such as DVH-based terms [13, 22, 36] and weighted MAE [15]. More broadly, the Any2Any design may extend to other RT planning tasks, such as leaf sequencing.

References

- [1] AAPM Grand Challenge Organizers. Generalizable dose prediction for heterogeneous multi-cohort and multi-site radiotherapy planning (gdp-hmm) grand challenge. <https://www.aapm.org/GrandChallenge/GDP-HMM/>, 2025. Accessed: 2025-10-24. 6
- [2] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023. 1
- [3] Aaron Babier, Binghao Zhang, Rafid Mahmood, Kevin L. Moore, Thomas G. Purdie, Andrea L. McNiven, and Timothy C. Y. Chan. Openkbp: The open-access knowledge-based planning grand challenge. *arXiv preprint arXiv:2011.14076*, 2020. 1, 2
- [4] Aaron Babier et al. Openkbp-opt: An international and open-source framework for plan optimization in knowledge-based planning. *arXiv preprint arXiv:2202.08303*, 2022. 1
- [5] Fan Bao et al. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*, 2023. 2
- [6] Omer Bar-Tal et al. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 2
- [7] Ana María Barragán-Montero, Dan Nguyen, Weiguo Lu, Mu Han Lin, Roya Norouzi-Kandalan, Xavier Geets, Edmond Sterpin, and Steve Jiang. Three-dimensional dose prediction for lung imrt patients with deep neural networks: robust learning from heterogeneous beam configurations. *Medical Physics*, 2019. 2
- [8] Søren M. Bentzen, Louis S. Constine, Joseph O. Deasy, Abraham Eisbruch, Andrew Jackson, Lawrence B. Marks, Randall K. Ten Haken, and Ellen D. Yorke. Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. *International Journal of Radiation Oncology Biology Physics*, 2010. 2
- [9] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2, 3
- [10] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 2
- [11] Gary A. Ezzell, Jay W. Burmeister, Nesrin Dogan, Thomas J. LoSasso, John G. Mechalakos, Dimitris Mihailidis, Aimee Molineu, Jatinder R. Palta, Chester R. Ramsey, Brian J. Salter, Jianguo Shi, Ping Xia, Cedric X. Yu, and Ying Xiao. IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM task group 119. *Medical Physics*, 2009. 2
- [12] Zhenghao Feng et al. Diffdp: Radiotherapy dose prediction via a diffusion model. *arXiv preprint arXiv:2307.09794*, 2023. 1, 2
- [13] Riqiang Gao, Bin Lou, Zhoubing Xu, Dorin Comaniciu, and Ali Kamen. Flexible-c^m gan: Towards precise 3d dose prediction in radiotherapy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 715–725, 2023. 1, 2, 6, 8
- [14] Riqiang Gao, Florin-Cristian Ghesu, Simon Arberet, Shahab Basiri, Esa Kuusela, Martin Kraus, Dorin Comaniciu, and Ali Kamen. Multi-agent reinforcement learning meets leaf sequencing in radiotherapy. *arXiv preprint arXiv:2406.01853*, 2024. 1
- [15] 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. 1, 2, 6, 8
- [16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2
- [17] Mary P Gronberg, Beth M Beadle, Adam S Garden, Heath Skinner, Skylar Gay, Tucker Netherton, Wenhua Cao, Carlos E Cardenas, Christine Chung, David T Fuentes, et al. Deep learning-based dose prediction for automated, individualized quality assurance of head and neck radiation therapy plans. *Practical radiation oncology*, 13(3):e282–e291, 2023. 1
- [18] Pengfei Guo, Can Zhao, Dong Yang, Yufan He, Vishwesh Nath, Ziyue Xu, Pedro RAS Bassi, Zongwei Zhou, Benjamin D Simon, Stephanie Anne Harmon, et al. Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model. *arXiv preprint arXiv:2505.04522*, 2025. 2
- [19] 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. 2
- [20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 2
- [21] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11375–11385, 2024. 1, 2
- [22] Gourav Jhanwar, Navdeep Dahiya, Parmida Ghahremani, Masoud Zarepisheh, and Saad Nadeem. Domain knowledge driven 3d dose prediction using moment-based loss function. *Physics in Medicine & Biology*, 67(18):185017, 2022. 2, 8
- [23] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*, 2023. 2
- [24] Vasant Kearney, Jason W Chan, Samuel Haaf, Martina Descovich, and Timothy D Solberg. Dosenet: a volumetric dose prediction algorithm using 3d fully-convolutional neural networks. *Physics in Medicine & Biology*, 63(23):235022, 2018. 2

- [25] Vasant Kearney, Jason W. Chan, Tianqi Wang, Alan Perry, Martina Descovich, Olivier Morin, Sue S. Yom, and Timothy D. Solberg. DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. *Scientific Reports*, 2020. 1, 2
- [26] Yuval Kirstain et al. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. 2, 3
- [27] Xiaoyan Kui, Fang Liu, Min Yang, Hao Wang, Canwei Liu, Dan Huang, Qinsong Li, Liming Chen, and Bei Zou. A review of dose prediction methods for tumor radiation therapy. *Meta-Radiology*, 2(1):100057, 2024. 1
- [28] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. *arXiv preprint arXiv:2412.01169*, 2024. 2
- [29] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 2018. 2
- [30] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023. 2
- [31] Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Huan Teng, Junlin Xie, Yu Qiao, Peng Gao, et al. Pixwizard: Versatile image-to-image visual assistant with open-language instructions. *arXiv preprint arXiv:2409.15278*, 2024. 3
- [32] Shuolin Liu, Jingjing Zhang, Teng Li, Hui Yan, and Jianfei Liu. Technical note: A cascade 3d u-net for dose prediction in radiotherapy. *Medical Physics*, 48(11):7132–7141, 2021. 2
- [33] Anthony Magliari, Ryan Clark, Lesley Rosa, and Sushil Beriwal. Hn-sib-bpi: A single click, sub-site specific, dosimetric scorecard tuned rapidplan model created from a foundation model for treating head and neck with bilateral neck. *Medical Dosimetry*, 50(1):63–69, 2025. 3
- [34] Chong Mou et al. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2
- [35] Dan Nguyen, Xun Jia, David Sher, Mu-Han Lin, Zohaib Iqbal, Hui Liu, and Steve Jiang. 3d radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture. *Physics in Medicine & Biology*, 2019. 2
- [36] Dan Nguyen, Rafe McBeth, Azar Sadeghnejad Barkousaraie, Gyanendra Bohara, Chenyang Shen, Xun Jia, and Steve Jiang. Incorporating human and learned domain knowledge into training deep neural networks: A differentiable dose-volume histogram and adversarial inspired framework for generating pareto optimal dose distributions in radiation therapy. *Medical Physics*, 47(3):837–849, 2020. 2, 8
- [37] Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1
- [38] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 2
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1
- [40] Rafael Rafailov, Archit Sharma, Eric Mitchell, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023. 2
- [41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2015. 2
- [42] 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. 2
- [43] Petra Seibold, Adam Webb, Miguel E. Aguado-Barrera, David Azria, Celine Bourcier, Muriel Brengues, Erik Briers, Renee Bultjnk, 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. 2, 6
- [44] Xinrui Song, Xuanang Xu, and Pingkun Yan. Dino-reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 608–617. Springer, 2024. 1, 2
- [45] Mumtaz Hussain Soomro, Victor Gabriel, Leandro Alves, Hamidreza Nourzadeh, and Jeffrey V. Siebers. Deepdosenet: A deep learning model for 3d dose prediction in radiation therapy. *arXiv preprint arXiv:2111.00077*, 2021. 2
- [46] Zineng Tang et al. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023. 2
- [47] Varian Medical Affairs. Bilateral head&neck 70/63/56gy (hn-sib-bpi) [rapidplan]. <https://medicalaffairs.varian.com/hn-sib-bpi-rapidplan-vmat2>, 2024. Accessed: 2024-10-19. 3, 5, 6
- [48] Varian Medical Affairs. Lung – conventional 60gy (nrg lu-004 / atkins km 2021). <https://medicalaffairs.varian.com/lung-conventional-vmat2>, 2024. Accessed: 2024-10-19. 3, 5, 6
- [49] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of*

- the *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8238, 2024. [2](#), [3](#)
- [50] 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](#)
- [51] Bin Wang, Lin Teng, Lanzhuji Mei, Zhiming Cui, Xuanang Xu, Qianjin Feng, and Dinggang Shen. Deep learning-based head and neck radiotherapy planning dose prediction via beam-wise dose decomposition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 575–584. Springer, 2022. [2](#)
- [52] Wentao Wang, Yang Sheng, Chunhao Wang, Jiahao Zhang, Xinyi Li, Manisha Palta, Brian Czito, Christopher G Willett, Qiuwen Wu, Yaorong Ge, et al. Fluence map prediction using deep learning models—direct plan generation for pancreas stereotactic body radiation therapy. *Frontiers in artificial intelligence*, 3:68, 2020. [1](#)
- [53] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical SAM adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. [1](#)
- [54] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 2023. [3](#)
- [55] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. [2](#)
- [56] Jiazhen Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023. [2](#), [3](#)
- [57] Xingqian Xu et al. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. [2](#)
- [58] Yifeng Xu, Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. Jodi: Unification of visual generation and understanding via joint modeling. *arXiv preprint arXiv:2505.19084*, 2025. [3](#)
- [59] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihao Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8941–8951, 2024. [2](#), [3](#)
- [60] Jingjing Zhang, Shuolin Liu, Hui Yan, Teng Li, Ronghu Mao, and Jianfei Liu. Predicting voxel-level dose distributions for esophageal radiotherapy using densely connected network with dilated convolutions. *Physics in Medicine & Biology*, 65(20):205013, 2020. [2](#)
- [61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. [2](#)
- [62] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. [2](#)
- [63] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8018–8027, 2024. [3](#)
- [64] Yiwen Zhang et al. Dosediff: Distance-aware diffusion model for dose prediction in radiotherapy. *arXiv preprint arXiv:2306.16324*, 2023. [1](#), [2](#)
- [65] 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. [2](#), [3](#), [5](#), [6](#)