

## Medical World Model

Yijun Yang<sup>1,\*</sup>, Zhao-Yang Wang<sup>2</sup>, Qiuping Liu<sup>3</sup>, Shuwen Sun<sup>3</sup>, Kang Wang<sup>4</sup>, Rama Chellappa<sup>2</sup>,  
 Zongwei Zhou<sup>2</sup>, Alan Yuille<sup>2</sup>, Lei Zhu<sup>1,5,†</sup>, Yu-Dong Zhang<sup>3</sup>, Jieneng Chen<sup>2,†</sup>

<sup>1</sup>The Hong Kong University of Science and Technology (Guangzhou) <sup>2</sup>Johns Hopkins University

<sup>3</sup>The First Affiliated Hospital of Nanjing Medical University

<sup>4</sup>University of California, San Francisco <sup>5</sup>The Hong Kong University of Science and Technology

Project: <https://yijun-yang.github.io/MeWM> Github: <https://github.com/scott-yjyang/MeWM>

### Abstract

*Providing effective treatment and making informed clinical decisions are essential goals of modern medicine and clinical care. We are interested in simulating disease dynamics for clinical decision-making, leveraging recent advances in large generative models. To this end, we introduce the Medical World Model (MeWM), the first world model in medicine that visually predicts future disease states based on clinical decisions. MeWM comprises (i) vision-language models to serve as policy models, and (ii) tumor generative models as dynamics models. The policy model generates action plans, such as clinical treatments, while the dynamics model simulates tumor progression or regression under given treatment conditions. Building on this, we propose the inverse dynamics model that applies survival analysis to the simulated post-treatment tumor, enabling the evaluation of treatment efficacy and the selection of the optimal clinical action plan. As a result, the proposed MeWM simulates disease dynamics by synthesizing post-treatment tumors, with state-of-the-art specificity in Turing tests evaluated by radiologists. Simultaneously, its inverse dynamics model outperforms medical-specialized GPTs in optimizing individualized treatment protocols across all metrics. Notably, MeWM improves clinical decision-making for interventional physicians, boosting F1-score in selecting the optimal TACE protocol by 13%, paving the way for future integration of medical world models as the second readers.*

### 1. Introduction

Clinical decision-making is at the heart of patient care, driving outcomes and shaping the trajectory of healthcare interventions. Physicians constantly weigh the multimodal factors including medical images and patient history to deter-

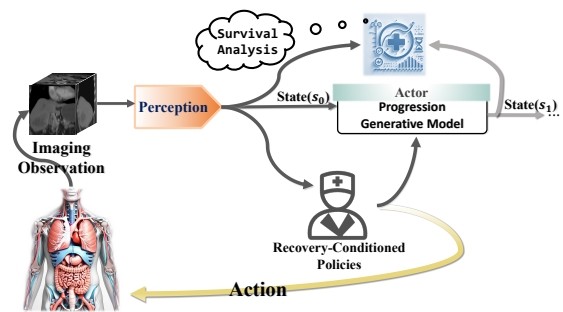


Figure 1. **Formulation of Medical World Model.** It integrates imaging observations with perception modules to form an initial state, which is then processed by a progression generative model to predict future states of disease under different treatment conditions. Recovery-conditioned policies guide treatment decisions, creating a feedback loop for optimizing clinical interventions.

mine the best course of action for each patient. Artificial intelligence (AI) models are increasingly assuming a crucial role in this process by analyzing the complex multimodal data, revealing patterns that may be difficult to detect with human observation alone, and suggesting tailored treatment strategies based on predictive analytics.

Foundation models [7] such as large language models (LLMs) [33, 57] present a new frontier in medical AI research and development. However, recent studies [26] show that LLMs, even those specifically tailored for medicine [13, 54, 55], diagnose significantly worse than clinicians and make less informed treatment decisions. This is due to several challenges. **First**, the complexity of diseases themselves, such as tumors that evolve under the influence of diverse biological and chemotherapy factors, calls for models that can adapt and account for disease progression. **Second**, clinical decision-making necessitates not only accurate predictions but also visually trackable insights that physicians can trust.

Recent breakthroughs in world models (WMs) [3, 8, 39, 45, 62] provide a promising avenue for overcoming these obstacles. By generating a predictive distribution of how

\*Work done while visiting at JHU.

†Corresponding authors.

the world states evolve, WMs mirror the way human planners imagine future scenarios and then make informed decisions via inverse dynamics [3, 18]. Although they remain largely underexplored in the medical domain, world models hold significant potential for generating clinically realistic images and simulating disease progression, which in turn can facilitate more effective and visually trackable treatment planning. Figure 1 illustrates our formulation of introducing WMs into medical scenarios and how WMs integrate these capabilities to support clinical decision-making.

In this work, we introduce Medical World Model (MeWM) to address these challenges and push the boundaries of AI-driven clinical decision support. MeWM comprises three primary components: (1) a **Policy Model** powered by vision-language architectures, which generates the potential action combos from a patient’s current state and specific clinical scenario; (2) a **Dynamics Model** that forwards and simulates tumor dynamics, predicting how tumors could progress or regress under different treatment conditions by generative modeling; (3) an **Inverse Dynamics Model** that performs survival risk analysis on the simulated post-treatment tumor, and quantitatively evaluates treatment efficacy. Beyond forward simulation, this system heuristically explores the optimal plan with the assistance of a segmentation model. By uniting these elements, MeWM delivers a holistic framework for decision-making: it can synthesize realistic post-treatment tumors that pass Turing tests against radiologists, and it outperforms specialized GPT-like models on Transarterial Chemoembolization (TACE) Protocol Exploration (over 10%↑ in F1-score).

Overall, our contributions are threefold.

- We propose the medical world models, where we develop a multimodal policy model that leverages vision-language capabilities to propose a tailored set of treatment action combos, and we design a generative dynamics model that accurately captures potential evolution of tumors, enabling forward-looking simulations for different interventions.
- We integrate an inverse dynamics model that translates these action-conditioned simulation into survival analysis metrics, thereby offering a transparent and evidence-based tool for choosing the optimal treatment protocol.
- We demonstrate a substantial leap in AI-driven decision support for interventional medicine, improving the F1-score in selecting the optimal treatment protocol by 13% and offering a compelling glimpse into the future of precision healthcare.

## 2. Related Work

### 2.1. Generative World Modeling

World models [23, 39] aim to simulate dynamic environments by predicting future states and rewards based on cur-

rent observations and actions. Originally developed for constrained settings like Atari games [24], their ability to model state transitions has been extended to real-world scenarios through joint learning of policies and world models, improving sample efficiency in simulated robotics [53], real-world robots [63] and autonomous driving [29, 61]. While early world models focused on simple state transitions, modern approaches integrate structured action-object relationships [59] and multi-modal conditioning [5, 21]. For instance, Du *et al.* [17] present long-horizon video plans by synergising vision-language models and text-to-video models. Luo *et al.* [46] propose to ground video models to continuous action by leveraging video-guided goal-conditioned exploration to learn a goal-conditioned policy. In embodied decision-making, Lu *et al.* [45] enables agents to imaginatively explore the world with high generation quality and exploration consistency using video generative models. However, there is still no work investigating the applicability of world modeling in medical image analysis and clinical decision-making.

### 2.2. Tumor Synthesis

Tumor synthesis has emerged as an attractive research topic across various medical imaging modalities, such as CT [11, 47, 67], MRI [4, 31, 64], and endoscopic videos [12, 41]. There are also many works on synthesizing non-cancerous lesions including chest CT synthesis [6, 47, 67, 68], and diabetic lesion synthesis in retinal images [15, 71]. Recent studies focus on improving the realism of synthetic tumors in the liver, kidney and pancreas [11, 30, 38] by leveraging the large generative models like diffusion models [28, 52, 56]. While these methods are conditioned only on shape masks, Li *et al.* [43] propose text-driven tumor synthesis by descriptive reports and conditional diffusion models. However, most of these works implemented tumor synthesis as a data augmentation to improve tumor detection tasks. They overlook its potential to empower clinical decision-making in treatment planning. In this work, we delve into the relatively unexplored field of tumor dynamics simulation by generating post-treatment tumors using pre-treatment scans and treatment actions.

### 2.3. Prognosis and Clinical Decision-making

Post-treatment prognosis in medical imaging is essential for evaluating therapy effectiveness, predicting disease recurrence, and guiding further clinical decisions. CT is widely used to assess structural and functional changes in tumors following interventions such as surgery, chemotherapy, radiation therapy, transarterial chemoembolization (TACE), and immunotherapy [25, 32, 40, 44, 65, 66]. Lee *et al.* [40] employ a CNN-based model to predict the post-treatment survival of patients with hepatocellular carcinoma (HCC) using CT images and clinical information. In addition,

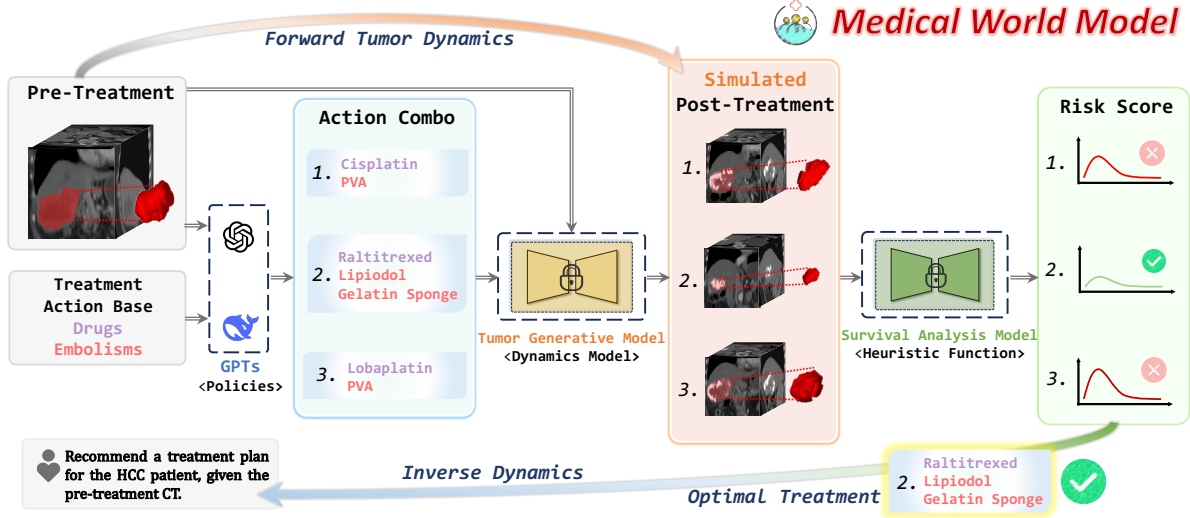


Figure 2. **Overview of TACE Protocol Exploration by Medical World Model.** (1) GPTs (Policy Model): construct the TACE action combos by the observation of pre-treatment CT, integrating clinical guidelines and policies. (2) Tumor Generative Model (Dynamics Model): simulates post-treatment tumor based on different TACE intervention protocols, predicting treatment outcomes. (3) Survival Analysis Model (Heuristic Function): assesses risk scores from both simulated post-treatment CT and pre-treatment CT to determine the most effective TACE protocol. Note that the 3D tumor masks (colored in red) can be extracted using a well-trained segmentation network (as Assistant Model). The framework enables visually trackable protocol optimization by iterating between clinical policy guidance, generative modeling, and survival analysis.

LLMs are increasingly being explored to assist in clinical decision-making [9, 26, 36, 42]. However, little attention has been given to applying LLMs for post-treatment prognosis. They did not leverage the feedback from survival analysis to achieve prompt intervention as well.

### 3. Medical World Models

**Overall Framework.** As shown in Fig. 2, our MeWM takes a visual observation of pre-treatment CT  $x_0$ , a language treatment goal  $g$  to simulate the future state and explore the best treatment protocol. **Policy model** (§ 3.1) acquires the descriptive observation based on the visual state, and constructs a set of treatment protocols by the language goal and clinical guidelines. To perform the exploration, given the pre-treatment CT and an action combo, the **dynamics model** (§ 3.2) predicts the concrete resulting state, *i.e.*, generating post-treatment CT. Finally, **inverse dynamics model** (§ 3.3) driven by Heuristic Function predicts the risk score from pre-treatment CT and simulated post-treatment CT with tumor masks from Assistant Model, to effectively prune branches in search and heuristically determine the optimal solution.

#### 3.1. Policy Model

Vision-language models [10, 70] have emerged as a powerful source of prior knowledge about the clinical world, providing rich information about how to complete promising treatment from large-scale internet data and clinical guidelines. Based on TACE clinical guidelines, we set

up the exploration configurations, including all potential chemotherapy drugs (*e.g.*, Raltitrexed, Cisplatin) and embolism materials (*e.g.*, Lipiodol, Gelatin Sponge). The two parts constitute the action base, which provides possible TACE protocols for Generative Dynamics Models as conditions. Then, we adopt a pre-trained large multimodal model (LMM), *e.g.*, GPT-4o, to serve as policies. Given a high-level goal  $g$  (*e.g.*, “What TACE treatment protocols are recommended for a patient diagnosed with hepatocellular carcinoma (HCC) given the pre-treatment CT?”), the policy model  $\pi_{\text{VLM}}(x_0, g)$  extracts the visual observation and tumor context from the given pre-treatment CT  $x_0$  to prompt the proper Transarterial Chemoembolization (TACE) actions. To constrain the excessively large tree search in the action base, we further prompt the Large Language Reasoning Model, *i.e.*, Deepseek-R1 [22], to refine the drug set and embolism set by the clinical policies, whose final cardinalities are  $D$  and  $E$ , respectively.

#### 3.2. Dynamics Model

**Radiotherapy Report Extraction and Generation.** While most existing studies focus on human-authored radiology reports, we aim to address radiotherapy reports to extract more comprehensive information on treatment protocols. However, raw radiotherapy reports pose significant challenges due to noise and fragmented information, which hinder controlled tumor synthesis. To mitigate these issues, we propose a two-stage text preprocessing framework consisting of data cleaning and augmentation. In the first stage,

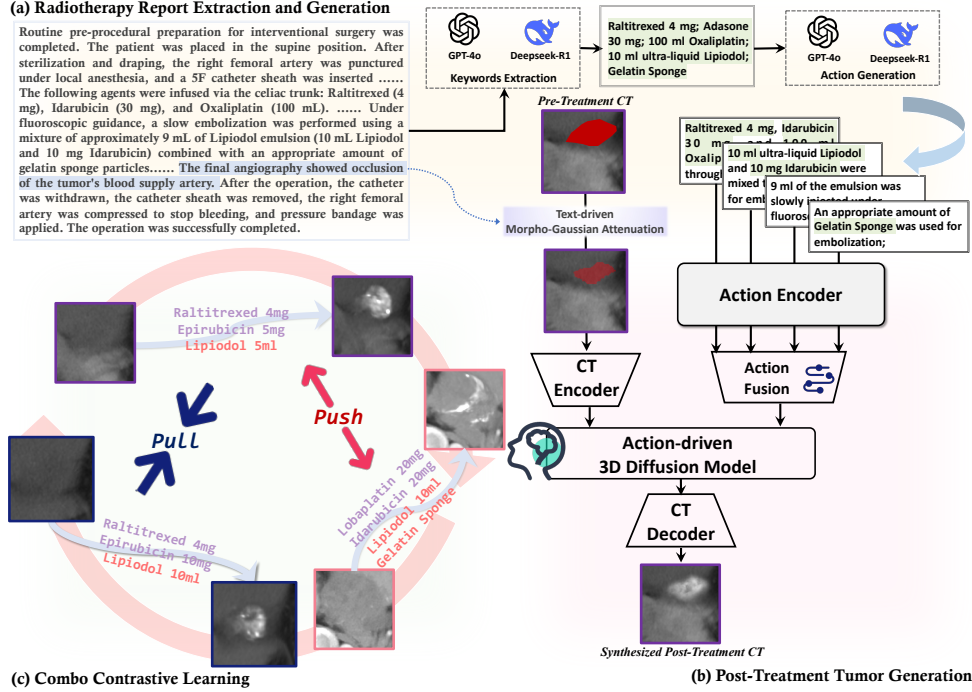


Figure 3. **Dynamics Model based on Tumor Generative Model.** The training framework consists of three parts: (a) *Radiotherapy Report Extraction and Generation*: GPT-4o and Deepseek-R1 extract key treatment details from radiotherapy reports and generate corresponding TACE surgical actions. (b) *Post-Treatment Tumor Generation*: An Action-driven 3D Diffusion Model is conditioned by fused action embeddings and attenuated CT features to generate post-treatment tumors that simulate treatment outcomes. (c) *Combo Contrastive Learning (CCL)*: The model learns from treatment variations by pushing apart dissimilar combos and pulling together similar ones, improving its ability to generate realistic and action-aware post-treatment tumor appearances.

we perform keyword extraction by aggregating the outputs of both GPT-4o and DeepSeek-R1, focusing on key entities such as drugs, embolic agents, and their corresponding dosages. In the second stage, we leverage the same tools for text generation, constructing a structured core action description based on the extracted keywords. This approach enhances the consistency and informativeness of processed reports, facilitating downstream tasks in tumor synthesis and treatment analysis.

**Post-Treatment Tumor Generation.** We adopt Latent Diffusion Models (LDMs) [52] for latent feature extraction from 3D Pre-treatment CT volumes and integrate textual action embedding for controlled tumor synthesis. Each 3D Post-treatment CT volume  $x_1 \in \mathbb{R}^{H \times W \times D}$  is encoded into a lower dimensional latent representation  $z_1 = \mathcal{E}(x_1)$  using a 3D VQGAN autoencoder [19]. In the latent space, following the spirit of DiffTumor [11], we define a diffusion process that progressively adds noise to the latent representation  $z_1$  over discrete time steps  $t = 1, \dots, T$ . Given a pair of tumor-present pre-treatment CT volume  $x_0$  and the mask of its tumor region  $m_0$ , we condition the denoising model on the masked pre-treatment latent representation  $z'_0 = \mathcal{E}(m'_0 \odot x_0)$  where  $m'_0$  is the attenuated mask from  $m_0$  by our proposed *Text-driven Morpho-Gaussian Attenuation*. Specifically, to mimic the effects of TACE treatment,

the process begins with occlusion assessment on radiotherapy reports. The textual descriptions (e.g., *occluded*, *reduced*, *disappear*) are extracted and analyzed to determine the attenuation level  $l \in \{1, 2, 3, 4\}$ , where a higher value corresponds to better tumor curative effects. Then, morphological erosion and dilation with the adaptive kernel by  $l$  are applied to  $m_0$ , simulating occlusion-induced tumor structural dynamics. Simultaneously, adaptive Gaussian blurring is employed to exhibit the characteristics of heterogeneous intensity changes due to lipiodol deposition, necrotic transformation, and reduced perfusion. The final attenuated mask  $m'_0$  is computed by the three steps, ensuring a smooth transition between tumor and organ tissues. Note that this attenuation is only used during training.

Also, we condition the denoising model on the generated textual action. Given the action combo  $a = \{a_1, \dots, a_H\}$ , each sub-action, respectively, undergoes encoding through a CLIP [51] text encoder  $\phi(\cdot)$  followed by linear projection  $\sigma_1(\cdot)$ , enabling dimension reduction to a latent clinical concept space. To enhance the semantics of action conditions in  $D$  different drug and  $E$  embolism keywords, we introduce learnable concept embeddings  $c$ , which extract keyword representations from the given action combo. This explicit pharmaceutical grounding enables precise modeling of therapeutic components while maintaining robustness to



context variations. The final action condition  $\tau(a)$  is the fusion of holistic text and concept embeddings by fully connected layers  $\sigma_2$ , where  $[\cdot]$  denotes concatenation operation:

$$\tau(a) = \sigma_2([\sigma_1(\phi(a_1)), \dots, \sigma_1(\phi(a_H))], c], \quad (1)$$

The training objective of diffusion model is as follows:

$$\mathbb{E}_{z_1, \epsilon \sim \mathcal{N}(0,1), t} \left[ \left\| \epsilon - \epsilon_\theta \left( z_t, z'_0, m_0, \tau(a), t \right) \right\|_2^2 \right], \quad (2)$$

where  $\epsilon_\theta(\cdot, t)$  is a 3D U-Net with interleaved self-attention layers and convolutional layers [11, 28, 50] that predict the noise given the input variable and conditions.

**Combo Contrastive Learning.** We adopt a contrastive learning strategy that aligns action combos with tumor evolution to enhance the realism and discrimination of post-treatment tumor synthesis. Given a pre-treatment anchor pair  $(x_0, m_0)$ , along with an action combo  $a$ , the goal is to generate a post-treatment CT  $\hat{x}$  by generative model  $f_{DM}(\cdot)$ .

For each anchor, positive samples,  $\hat{x}^+ = f_{DM}(x_0^+, m_0^+, a^+)$ , are defined as synthetic post-treatment CT from another pair but its action combo contains the same drug/embolism keywords. Negative samples,  $\hat{x}^- = f_{DM}(x_0, m_0, a^-)$ , in contrast, are generated using the same pair but action combos with diverse keywords, leading to distinct tumor evolution patterns. This contrastive loss is incorporated into the tumor generative model:

$$\mathbb{E} \left[ -\log \frac{\exp(\text{sim}(\hat{x}, \hat{x}^+)/\delta)}{\sum \exp(\text{sim}(\hat{x}, \hat{x}^-)/\delta) - \exp(\text{sim}(\hat{x}, \hat{x}^+)/\delta)} \right], \quad (3)$$

where  $\text{sim}(\cdot)$  denotes cosine similarity,  $\delta$  is the temperature scaling factor. This contrastive learning strategy ensures that tumors simulated from similar treatment protocols exhibit consistent attenuation effects, while those derived from distinct protocols remain differentiable.

### 3.3. Inverse Dynamics Model

Inverse Dynamics Model, which empowers our full framework, aims to infer the most effective treatment strategy by analyzing relationships between pre-treatment conditions, intervention actions, and expected post-treatment outcomes. We unfold its essence in three aspects: (1) Assistant Model; (2) Heuristic Function and (3) TACE Protocol Exploration.

**Assistant Model.** To better discriminate the tumor in synthesized post-treatment CT  $\hat{x}$ , we introduce a tumor segmentation model as Assistant Model  $H_{\text{seg}}(\cdot)$ . Post-treatment tumors are characterized by heterogeneous high-intensity regions due to *calcification/lipiodol* deposition, irregular shapes reflecting *necrotic tissue* changes, and reduced or absent contrast enhancement in viable tumor areas, in contrast to traditional pre-treatment CT tumors. Thus, we adapt a pre-trained nnUNet-based [34] model to this post-treatment context by finetuning it on our ground truth pairs

of post-treatment CT and mask. Given the well-trained Assistant Model, the simulated post-treatment CT  $\hat{x}$  from Tumor Generative Model is processed for the segmentation of liver and tumor. The post-treatment CT with the predicted mask  $\hat{m}$  is subsequently utilized for survival analysis.

**Heuristic Function.** We use survival analysis model to implement a heuristic function  $H_{\text{surv}}(x_0, m_0, \hat{x}, m_0, g)$ , which quantifies the efficacy under the specific TACE action combo by the output risk score. Inspired by DeepSurv [35], we utilize a 3D convolution-based model structure, the 3D ResNet (MC3) [58], as the feature extractor of survival analysis model. Given the pre-treatment pair  $(x_0, m_0)$  and simulated post-treatment pair  $(\hat{x}, \hat{m})$ , we extract their concatenated CT and mask, respectively, and bidirectionally align the semantics of pre- and post-treatment by Cross-Attention Transformer [37]. After that, we adopt an attention-based aggregator to fuse pre- and post-features, followed by fc layers to determine the risk score. The action combo with a lower risk score should bring greater efficacy for the patient. Note that, for training, we leverage multi-task learning strategy, *i.e.*, CoxPH [35] and OS regression, to improve the generalization of survival analysis.

**TACE Protocol Exploration.** Given a combination of the proposed models above, we are able to predict TACE protocol from any Pre-treatment CT  $x$  by a language treatment goal  $g$ . To reason the optimal action combo, we propose to search for a list of actions to reach  $g$ , corresponding to finding a treatment plan consisting of both drug and embolism components, which optimizes:

$$\hat{x}_{1:H}^* = \arg \min_{\hat{x}_{1:H} \sim \pi_{\text{VLM}}, f_{DM}} H_{\text{surv}}(x_0, m_0, \hat{x}, H_{\text{seg}}(\hat{x}), g), \quad (4)$$

where  $H = H_d + H_e$ .

With this objective in mind, we exhibit a tree-search exploration procedure. Our exploration algorithm initializes a set of  $B$  parallel protocol beams. We sample the potential action space composed of  $D$  drugs and  $E$  embolisms and clinical rules using  $\pi_{\text{VLM}}$ . The clinical rules are introduced to prune unreasonable branches, *e.g.*, concomitant use of multiple platinum-based agents is contraindicated due to the risk of cumulative toxicity and myelosuppression. We sequentially explore the two parts to ensure that TACE protocol contains both drugs and embolism. For each current action combo, we synthesize  $T$  post-treatment tumors from  $f_{DM}(x_0, m_0, a)$  to obtain a more reliable simulation. We then use our heuristic function  $H_{\text{surv}}(x_0, m_0, \hat{x}, \hat{m}, g)$  with the assistance of  $H_{\text{seg}}(\hat{x})$  to select the generated tumor with the best average survival score of  $T$  replicas among  $D$  or  $E$  actions. After every step of extending all beams, we discard the beam with the worst survival score and replace its action combo with the best beam. To prevent cumulative toxicity and organ dysfunction, we prohibit over-exploration by drug horizon  $H_d$  and embolism horizon  $H_e$ . Our final action combo is taken from the beam with the best survival

---

**Algorithm 1** TACE Protocol Exploration with MeWM

---

```
1: Input: Pre-treatment CT  $x_0$ , Pre-treatment tumor mask  $m_0$ , Language treatment goal  $g$ 
2: Functions: VLM Policy Model  $\pi_{\text{VLM}}$ , Dynamics Model  $f_{\text{DM}}$ , Heuristic Function  $H_{\text{surv}}$ , Assistant Model  $H_{\text{seg}}$ 
3: Hyperparameters: Drug Actions factor  $D$ , Embolism Actions factor  $E$ , Tumor Generative factor  $T$ , Protocol Beams  $B$ , Drug horizon  $H_d$ , Embolism horizon  $H_e$ 
4:  $\text{plans} \leftarrow [[x_0] \ \forall \ i \in \{1 \dots B\}]$  # Initialize B Different TACE Protocol Beams
5:  $\text{drug}_{1:D}, \text{embo}_{1:E}, \text{rule} \leftarrow \pi_{\text{VLM}}(x_0, g)$  # Generate  $D$  Different Drug,  $E$  Embolism Actions, Clinical Rules
6: for  $h = 1 \dots H_d$  do
7:   for  $b = 1 \dots B$  do
8:      $\text{tumors} \leftarrow [f_{\text{DM}}(x, \text{drug}_i) \text{ for } j \text{ in } (1 \dots T) \text{ for } i \text{ in } (1 \dots D) \text{ if } \text{rule}]$  # Generate tumors from  $x$  and plans[b] under rule
9:      $\text{plans}[b].\text{append}(\text{argmin}(\text{tumors}, H_{\text{surv}}, H_{\text{seg}}))$  # Add Tumor with Lowest Risk to Plan
10:   end for
11:    $\text{max\_idx}, \text{min\_idx} \leftarrow \text{argmax}(\text{plans}, H_{\text{surv}}, H_{\text{seg}}), \text{argmin}(\text{plans}, H_{\text{surv}}, H_{\text{seg}})$ 
12:    $\text{plans}[\text{max\_idx}] \leftarrow \text{plans}[\text{min\_idx}]$  # Periodically Replace the Plan with High Risk
13: end for
14: for  $h = 1 \dots H_e$  do
15:   for  $b = 1 \dots B$  do
16:      $\text{tumors} \leftarrow [f_{\text{DM}}(x, \text{embo}_i) \text{ for } j \text{ in } (1 \dots T) \text{ for } i \text{ in } (1 \dots E) \text{ if } \text{rule}]$  # Generate tumors from  $x$  and plans[b] under rule
17:      $\text{plans}[b].\text{append}(\text{argmin}(\text{tumors}, H_{\text{surv}}, H_{\text{seg}}))$  # Add Tumor with Lowest Risk to Plan
18:   end for
19:    $\text{max\_idx}, \text{min\_idx} \leftarrow \text{argmax}(\text{plans}, H_{\text{surv}}, H_{\text{seg}}), \text{argmin}(\text{plans}, H_{\text{surv}}, H_{\text{seg}})$ 
20:    $\text{plans}[\text{max\_idx}] \leftarrow \text{plans}[\text{min\_idx}]$  # Periodically Replace the Plan with High Risk
21: end for
22:  $\text{plan} \leftarrow \text{argmin}(\text{plans}, H_{\text{surv}}, H_{\text{seg}})$  # Return Plan with Lowest Risk
```

---

score and adopted as TACE protocol for the patient. The pseudocode of our method is also provided in Algorithm 1.

## 4. Experiments

**HCC-TACE In-house Dataset.** We collect a large repository of 338 longitudinal paired pre- and post-treatment CT scans with well-annotated liver/tumor masks and clinical records, such as TACE radiotherapy reports as gold action and Overall Survival (OS) time. We split the training set (validation set included) and testing sets in the 9:1 ratio.

**HCC-TACE-Seg Public Dataset [49].** For external validation, we use patients from HCC-TACE-Seg public dataset referring to a single-institution collection with confirmed HCC treated at The University of Texas MD Anderson Cancer Center. We conduct data curation and preprocessing to collect 78 cases containing pre-treatment CT, post-treatment CT, TACE Gold Action, and OS time. We use 80% cases to fine-tune and validate MeWM and leave 20% cases for the exploration evaluation.

### 4.1. Evaluation on Generation Quality

**Visual Turing Test (Human Evaluation).** We conduct an action-driven Visual Turing Test on 240 CT scans of post-treatment tumors, where 120 scans contain real post-treatment tumors, and 120 scans contain synthetic post-treatment tumors generated by different tumor synthesis models. Three radiologists (R1-R3) participated in this study, independently evaluating five groups of 48 CT scans each and classifying them as either real or synthetic. It

is important to note that the radiologists' evaluations are based on whether the synthetic tumor closely resembles a post-treatment tumor, which typically contains a mixture of lipiodol deposition, necrotic, and viable tumor regions, distinguishing it from ordinary pre-treatment tumors. The test results are summarized in Table 1. The sensitivity scores of all radiologists remain high (above 91%), demonstrating their ability to correctly identify real post-treatment tumors. However, specificity scores vary among the methods, indicating different levels of realism in the synthetic tumors. Notably, our method MeWM achieves the lowest specificity scores (79.17% for R1, 70.83% for R2, and 75.00% for R3), suggesting that a large proportion of synthetic tumors generated by our approach are mistaken as real. This indicates superior realism compared to other methods such as SynTumor [30], Pixel2Cancer [38], DiffTumor [11], and Text-toMorph [43]. Figure 4 illustrates examples from the test, where a real tumor is compared with synthetic tumors that radiologists correctly or incorrectly classified. This highlights synthetic tumors closely resemble real ones.

**Perceptual Evaluation.** We perform perceptual evaluation using FID and LPIPS scores, where lower values indicate better simulation quality. Our method achieves the best FID (0.71) and LPIPS (0.6120), demonstrating the highest fidelity in synthetic tumor generation. These results confirm that MeWM effectively synthesizes realistic post-treatment tumors, making it more challenging for radiologists to distinguish between real and synthetic cases.

Table 1. **Action-driven Visual Turing Test** involves three radiologists (R1-R3) each evaluating five groups of 48 CT scans each, with 24 real post-treatment tumors and 24 synthetic post-treatment tumors from a tumor generative model, respectively. They were tasked with categorizing each CT scan as either real or synthetic. A higher sensitivity score indicates better discriminative ability of radiologists, while a lower specificity score indicates a higher number of synthetic tumors being identified as real. We also provide perceptual evaluation using FID and LPIPS compared to corresponding real post-treatment scans. Lower FID and LPIPS indicate better simulation results.

Methods	R1			R2			R3			Perceptual metrics	
	sensitivity	specificity ↓	accuracy	sensitivity	specificity ↓	accuracy	sensitivity	specificity ↓	accuracy	FID ↓	LPIPS ↓
SynTumor [30]	100.0	95.83	97.92	87.50	95.83	91.67	100.0	95.83	97.92	3.33	0.6832
Pixel2Cancer [38]	95.83	100.0	97.92	91.67	95.83	93.75	100.0	100.0	100.0	3.34	0.6831
DiffTumor [11]	100.0	91.67	95.83	95.83	87.50	91.67	100.0	87.50	93.75	1.40	0.7660
TextoMorph [43]	100.0	91.67	95.83	91.67	83.33	87.50	95.83	87.50	91.67	1.03	0.9111
MeWM (Ours)	100.0	<b>79.17</b>	89.58	91.67	<b>70.83</b>	81.25	91.67	<b>75.00</b>	83.33	<b>0.71</b>	<b>0.6120</b>

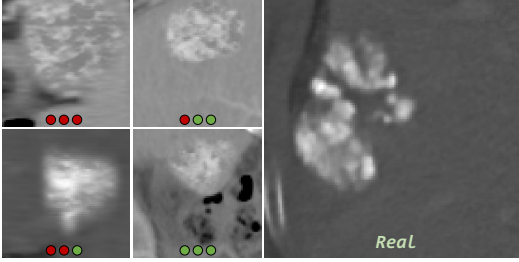


Figure 4. **Examples of Visual Turing Test.** We present one real tumor alongside examples of synthetic tumors that were correctly and incorrectly identified. A **red dot** indicates the radiologist classified the post-treatment tumor as synthetic, while a **green dot** signifies it was identified as real.

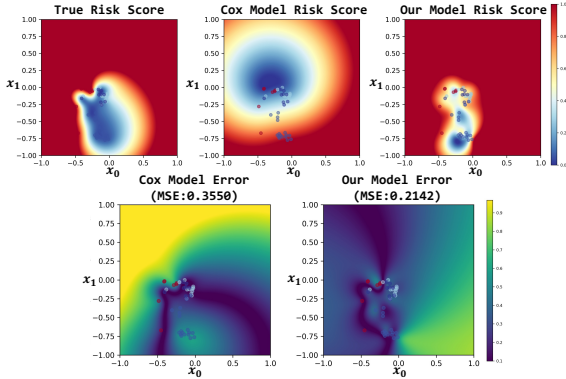


Figure 5. **Performance of heuristic function on survival analysis.** The first three heatmaps show the true risk distribution, Cox model predictions, and our heuristic function predictions. The last two depict prediction errors, with lower MSE (0.2142) for our model compared to the Cox model (0.3550), demonstrating improved accuracy in capturing localized risk patterns.

## 4.2. Survival Analysis

In Figure 5, we evaluate survival risk regression between the popular Cox Proportional Hazards model [20] and our heuristic function model on the HCC-TACE-Seg dataset. The true risk distribution (left) is estimated using the Nelson-Aalen estimator [14]. The Cox model fails to accurately distinguish between high- and low-risk samples from low-dimensional deep features, resulting in an overly smoothed risk distribution. In contrast, our model produces a risk map that better aligns with the true distribution, ef-

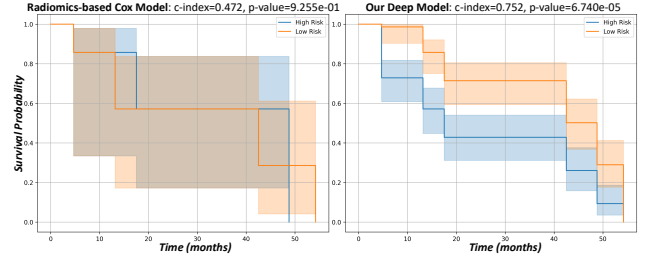


Figure 6. **Kaplan-Meier Survival Curves: Radiomics-based Cox Model [60] vs. our deep model.** The left shows the survival curves predicted by the Cox model based on Radiomics features. The right presents the survival curves from our model based on deep features, which achieves a significantly higher c-index of 0.752 and a log-rank p-value of  $6.74e-5$ , demonstrating a stronger ability to distinguish between high- and low-risk groups. Shaded areas represent confidence intervals.

fectively capturing variations in risk levels. Error analysis shows higher Mean Square Error (MSE), *i.e.*, 0.3550 for Cox and lower MSE, 0.2142 for our model, indicating superior accuracy. Figure 6 further presents Kaplan-Meier survival curves comparing risk stratification performance between the Radiomics-based Cox model and our deep model. These results demonstrate that our heuristic function better estimates survival risks, reduces prediction errors, and captures complex patterns beyond the capabilities of the Cox model and radiomics features.

## 4.3. Results on TACE Protocol Exploration

**Evaluation Strategy.** For treatment planning evaluation, we utilize four metrics in Table 2: (1) *F1-score*: harmonizes Precision and Recall, balancing redundancy and omissions; (2) *Jaccard Index*: measures prediction overlap with gold actions, emphasizing category-level alignment; (3) *Precision*: reflects recommendation purity, penalizing incorrect or redundant drugs/embolisms; (4) *Recall*: captures therapeutic coverage, highlighting critical omissions.

**Partial Observation Misleads GPTs.** For Multi-modal Large Language Models (*e.g.*, GPT-4o, MedGPT, HuatuoGPT-Vision), they are prompted with pre-treatment CT slices and allowed to predict the action combo from the given action set. These inferior results (over -10% in F1-

Table 2. **TACE Protocol Exploration Evaluation on HCC-TACE in-house dataset and public dataset.** F1-score, Jaccard index, Precision, and Recall are computed between the predicted action combo and gold action. MeWM significantly advances multimodal GPTs in exploring optimal individualized treatment protocol across all metrics, even comparable to interventional physicians.

Methods	In-house dataset				Public dataset			
	F1-score↑	Jaccard↑	Precision↑	Recall↑	F1-score↑	Jaccard↑	Precision↑	Recall↑
Physician w/ Pre-CT	48.81	38.44	46.67	54.67	71.43	63.10	66.67	78.57
Physician w/ MeWM	61.51 (+13%)	49.89	60.89	65.44	80.00 (+9%)	73.81	76.16	85.71
Qwen2.5-VL [2]	37.09	24.49	34.44	41.83	47.14	34.40	53.57	42.86
GPT-4o [33]	41.97	27.81	35.93	52.78	44.29	32.74	57.14	38.10
Claude-3.7-sonnet [1]	40.93	28.55	45.83	37.78	44.76	33.81	64.29	35.71
CT2Rep [27]	27.75	17.21	30.83	25.83	43.61	28.57	53.57	37.50
MedGPT [16, 48]	37.51	25.57	32.78	45.21	47.14	40.48	50.00	45.24
HuatuoGPT-Vision [69]	40.13	29.08	40.11	42.28	52.62	42.26	54.76	51.19
<b>MeWM(Ours)</b>	<b>52.38</b>	<b>38.59</b>	<b>63.06</b>	<b>46.17</b>	<b>64.08</b>	<b>48.45</b>	<b>72.62</b>	<b>58.93</b>

Table 3. **Ablation studies of TACE protocol exploration on both datasets.** “AM” denotes Assistant Model, while “CCL” denotes Combo Contrastive Learning. Two components significantly contribute to better exploring the optimal treatment.

Metrics ↑	In-house dataset		Public dataset	
	w/o AM	w/o CCL	w/o AM	w/o CCL
F1-score	49.13 (-3.9)	50.97 (-1.4)	60.03 (-4.1)	62.90 (-1.2)
Jaccard	35.40 (-3.2)	36.76 (-1.8)	45.36 (-3.1)	46.97 (-1.5)
Precision	56.39 (-6.7)	60.57 (-2.5)	75.00 (+2.4)	70.10 (-2.5)
Recall	45.22 (-1.0)	45.36 (-0.8)	52.38 (-5.6)	56.72 (-2.2)

score) to MeWM demonstrate that it tends to make deficient planning relying solely on vision-language models and their commonsense reasoning. This also validates the necessity of simulation from pre-treatment to post-treatment.

**MeWM as A Clinical Decision-support Tool.** MeWM demonstrates significant potential in augmenting the capabilities of radiologists and physicians, underscoring its clinical relevance in optimizing TACE planning. Reliance solely on pre-treatment CT often results in partial observation and suboptimal targeting due to heterogeneous pathological conditions. By incorporating MeWM’s recommended protocol, clinical decision-making is markedly enhanced, yielding performance improvements of 12.70, 11.45, 14.22, and 10.77 in F1-score, Jaccard, Precision, and Recall on our dataset. MeWM facilitates accurate tumor localization and enables predictive assessment of post-embolization outcomes, thereby reducing procedural uncertainty. Moreover, its synthetic post-treatment CT projections help anticipate embolization efficacy, optimize TACE distribution, and mitigate non-target embolization risks, contributing to enhanced therapeutic precision and individualized strategies. MeWM serves as a critical decision-support tool in interventional oncology, bridging anatomical imaging with functional assessment for meaningful clinical outcomes. As shown in Figure 7, interventional physicians refine TACE protocols by MeWM intervention, aligning treatment with expert practices.

**Ablation Study.** As shown in Table 3, we ablate the effectiveness of Assitant Model and Combo Contrastive

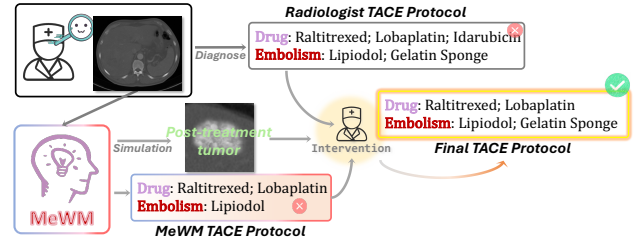


Figure 7. **Example of MeWM intervention in clinical applications.** The radiologist initially proposes a TACE protocol with Raltitrexed, Lobaplatin, Idarubicin, and embolization using Lipiodol and Gelatin Sponge. MeWM simulates a protocol with Raltitrexed, Lobaplatin, and Lipiodol. After intervention, the optimized protocol removes Idarubicin but restores Gelatin Sponge, aligning with gold action.

Learning (CCL) in TACE Protocol Exploration. The results demonstrate that both the Assistant Model and Combo Contrastive Learning (CCL) contribute significantly to its performance. Removing the assistant model, which provides the location information of tumors for heuristic function, leads to a critical drop in F1-score (52.38→49.13) on our dataset, as well as on public dataset. Similarly, omitting CCL reduces F1-score (e.g., 52.38→50.97), indicating that CCL enhances the model’s discrimination on action units. Overall, MeWM achieves the best results across all metrics, even outperforming radiologists in some areas.

## 5. Conclusion

We present Medical World Model, which marks a step toward AI-driven precision medicine by simulating disease evolution and optimizing clinical strategies. By bridging generative modeling with medical decision-making, it enables a deeper understanding of treatment outcomes and refines intervention planning. The advancements of MeWM in tumor synthesis and survival analysis set the stage for future AI systems that seamlessly integrate with clinical workflows, driving the next generation of longitudinal data-driven healthcare.



**Acknowledgement:** This work was supported by the Guangdong Science and Technology Department (No. 2024ZDZX2004), the Guangzhou Municipal Science and Technology Project (No.2024312139), the Lustgarten Foundation for Pancreatic Cancer Research, the Patrick J. McGovern Foundation Award, and the National Institutes of Health (NIH) under Award Number R01EB037669.

## References

- [1] AI Anthropic. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 2024. 8
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 8
- [3] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025. 1, 2
- [4] Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023. 2
- [5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2
- [6] Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier van der Sluijs, Małgorzata Polacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay S Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, pages 1–13, 2024. 2
- [7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1
- [8] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 1
- [9] Felix Busch, Philipp Prucker, Alexander Komenda, Sebastian Ziegelmayer, Marcus R Makowski, Keno K Bressem, and Lisa C Adams. Multilingual feasibility of gpt-4o for automated voice-to-text ct and mri report transcription. *European Journal of Radiology*, 182:111827, 2025. 3
- [10] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024. 3
- [11] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11147–11158, 2024. 2, 4, 5, 6, 7
- [12] Tong Chen, Shuya Yang, Junyi Wang, Long Bai, Hongliang Ren, and Luping Zhou. Surgsora: Decoupled rgb-d-flow diffusion model for controllable surgical video generation. *arXiv preprint arXiv:2412.14018*, 2024. 2
- [13] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023. 1
- [14] Enrico Colosimo, Flávio Ferreira, Maristela Oliveira, and Cleide Sousa. Empirical comparisons between kaplan-meier and nelson-aalen survival function estimators. *Journal of Statistical Computation and Simulation*, 72(4):299–308, 2002. 7
- [15] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017. 2
- [16] Michael D Moor. Medgpt, 2025. Accessed: March 7, 2025. 8
- [17] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 2
- [18] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. In *ICML*, 2024. 2
- [19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4
- [20] John Fox and Sanford Weisberg. Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, 2002, 2002. 7
- [21] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 2
- [22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3
- [23] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 2
- [24] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 2

- [25] Amr Hagag, Ahmed Gomaa, Dominik Kornek, Andreas Maier, Rainer Fietkau, Christoph Bert, Yixing Huang, and Florian Putz. Deep learning for cancer prognosis prediction using portrait photos by stylegan embedding. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 198–208. Springer, 2024. 2
- [26] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024. 1, 3
- [27] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–486. Springer, 2024. 8
- [28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 5
- [29] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2
- [30] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7422–7432, 2023. 2, 6, 7
- [31] Pu Huang, Dengwang Li, Zhicheng Jiao, Dongming Wei, Bing Cao, Zhanhao Mo, Qian Wang, Han Zhang, and Dinggang Shen. Common feature learning for brain tumor mri synthesis by context-aware generative adversarial network. *Medical Image Analysis*, 79:102472, 2022. 2
- [32] Xiaoyu Huang, Yong Huang, Ping Li, and Kai Xu. Ct-based deep learning predicts prognosis in esophageal squamous cell cancer patients receiving immunotherapy combined with chemotherapy. *Academic Radiology*, 2025. 2
- [33] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 8
- [34] Fabian Isensee, Paul F Jaeger, Simon AA 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. 5
- [35] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deep-surv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18:1–12, 2018. 5
- [36] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, Hae Park, et al. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024. 3
- [37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5
- [38] Yuxiang Lai, Xiaoxi Chen, Angtian Wang, Alan Yuille, and Zongwei Zhou. From pixel to cancer: Cellular automata in computed tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2024. 2, 6, 7
- [39] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 1, 2
- [40] Kyung Hwa Lee, Jungwook Lee, Gwang Hyeon Choi, Ji-hye Yun, Jiseon Kang, Jonggi Choi, Kang Mo Kim, and Namkug Kim. Deep learning-based prediction of post-treatment survival in hepatocellular carcinoma patients using pre-treatment ct images and clinical data. *Journal of Imaging Informatics in Medicine*, pages 1–12, 2024. 2
- [41] Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–240. Springer, 2024. 2
- [42] Jia Li, Zichun Zhou, Han Lyu, and Zhenchang Wang. Large language models-powered clinical decision support: enhancing or replacing human expertise?, 2025. 3
- [43] Xinran Li, Yi Shuai, Chen Liu, Qi Chen, Qilong Wu, Pengfei Guo, Dong Yang, Can Zhao, Pedro RAS Bassi, Daguang Xu, et al. Text-driven tumor synthesis. *arXiv preprint arXiv:2412.18589*, 2024. 2, 6, 7
- [44] Junhao Liang, Weisheng Zhang, Jianghui Yang, Meilong Wu, Qionghai Dai, Hongfang Yin, Ying Xiao, and Lingjie Kong. Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer. *Nature Machine Intelligence*, 5(4):408–420, 2023. 2
- [45] Taiming Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. Generative world explorer. *arXiv preprint arXiv:2411.11844*, 2024. 1, 2
- [46] Yunhao Luo and Yilun Du. Grounding video models to actions through goal conditioned exploration. *arXiv preprint arXiv:2411.07223*, 2024. 2
- [47] Fei Lyu, Mang Ye, Jonathan Frederik Carlsen, Kenny Erleben, Sune Darkner, and Pong C Yuen. Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. *IEEE Transactions on Medical Imaging*, 42(3):797–809, 2022. 2
- [48] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 8
- [49] Ali Morshid, Khaled M Elsayes, Ahmed M Khalaf, Mohab M Elmohr, Justin Yu, Ahmed O Kaseb, Manal Hassan, Armeen Mahvash, Zhihui Wang, John D Hazle, et al. A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. *Radiology: Artificial Intelligence*, 1(5):e180021, 2019. 6

- [50] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 5
- [51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4
- [52] 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*, pages 10684–10695, 2022. 2, 4
- [53] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *Conference on Robot Learning*, pages 1332–1344. PMLR, 2023. 2
- [54] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. 1
- [55] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025. 1
- [56] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2
- [57] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [58] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 5
- [59] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 2
- [60] Joost van Griethuysen et. al. Radiomics-based cox model, 2025. Accessed: March 7, 2025. 7
- [61] X Wang, Z Zhu, G Huang, X Chen, and J Drivedreamer Lu. Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 2
- [62] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jia-gang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72, 2024. 1
- [63] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023. 2
- [64] Zhaohu Xing, Sicheng Yang, Sixiang Chen, Tian Ye, Yijun Yang, Jing Qin, and Lei Zhu. Cross-conditioned diffusion model for medical image to image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–211. Springer, 2024. 2
- [65] Jiawen Yao, Yu Shi, Kai Cao, Le Lu, Jianping Lu, Qike Song, Gang Jin, Jing Xiao, Yang Hou, and Ling Zhang. Deepprognosis: Preoperative prediction of pancreatic cancer survival and surgical margin via comprehensive understanding of dynamic contrast-enhanced ct imaging and tumor-vascular contact parsing. *Medical image analysis*, 73:102150, 2021. 2
- [66] Jiawen Yao, Yu Shi, Le Lu, Jing Xiao, and Ling Zhang. Deepprognosis: Preoperative prediction of pancreatic cancer survival and surgical margin via contrast-enhanced ct imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 272–282. Springer, 2020. 2
- [67] Qingsong Yao, Li Xiao, Peihang Liu, and S Kevin Zhou. Label-free segmentation of covid-19 lesions in lung ct. *IEEE transactions on medical imaging*, 40(10):2808–2819, 2021. 2
- [68] Wenfang Yao, Chen Liu, Kejing Yin, William Cheung, and Jing Qin. Addressing asynchronicity in clinical multimodal fusion via individualized chest x-ray generation. *Advances in Neural Information Processing Systems*, 37:29001–29028, 2025. 2
- [69] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023. 8
- [70] Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, pages 1–26, 2024. 3
- [71] Yi Zhou, Xiaodong He, Shanshan Cui, Fan Zhu, Li Liu, and Ling Shao. High-resolution diabetic retinopathy image synthesis manipulated by grading and lesions. In *International conference on medical image computing and computer-assisted intervention*, pages 505–513. Springer, 2019. 2