Beyond Generative AI: World Models for Clinical Prediction, Counterfactuals, and Planning

Mohammad Areeb Qazi
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE
mohammad.qazi@mbzuai.ac.ae

Maryam Nadeem
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE
maryam.nadeem@mbzuai.ac.ae

Mohammad Yaqub
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE
mohammad.yaqub@mbzuai.ac.ae

Abstract

Healthcare requires AI that is predictive, reliable, and data-efficient. However, recent generative models lack physical foundation and temporal reasoning required for clinical decision support. As scaling language models show diminishing returns for grounded clinical reasoning, *world models* are gaining traction because they learn multimodal, temporally coherent, and action-conditioned representations that reflect the physical and causal structure of care. This paper reviews *World Models* for healthcare systems that learn predictive dynamics to enable multistep rollouts, counterfactual evaluation and planning. We survey recent work across three domains: (i) medical imaging and diagnostics (e.g., longitudinal tumor simulation, projection-transition modeling, and Joint Embedding Predictive Architecture i.e., JEPA-style predictive representation learning), (ii) disease progression modeling from electronic health records (generative event forecasting at scale), and (iii) robotic surgery and surgical planning (action-conditioned guidance and control). We also introduce a capability rubric: **L1** temporal prediction, **L2** action-conditioned prediction, **L3** counterfactual rollouts for decision support, and **L4** planning/control. Most reviewed systems achieve **L1–L2**, with fewer instances of **L3** and rare **L4**. We identify cross-cutting gaps that limit clinical reliability; under-specified action spaces and safety constraints, weak interventional validation, incomplete multimodal state construction, and limited trajectory-level uncertainty calibration. This review outlines a research agenda for clinically robust *prediction-first* world models that integrate generative backbones (transformers, diffusion, VAE) with causal/mechanical foundation for safe decision support in healthcare.

1 Introduction

Healthcare systems are under tremendous pressure from aging populations, chronic diseases, and shortages of workers. Populations are rapidly aging (projected 1 in 6 over 60 by 2030) [1], and global

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning.

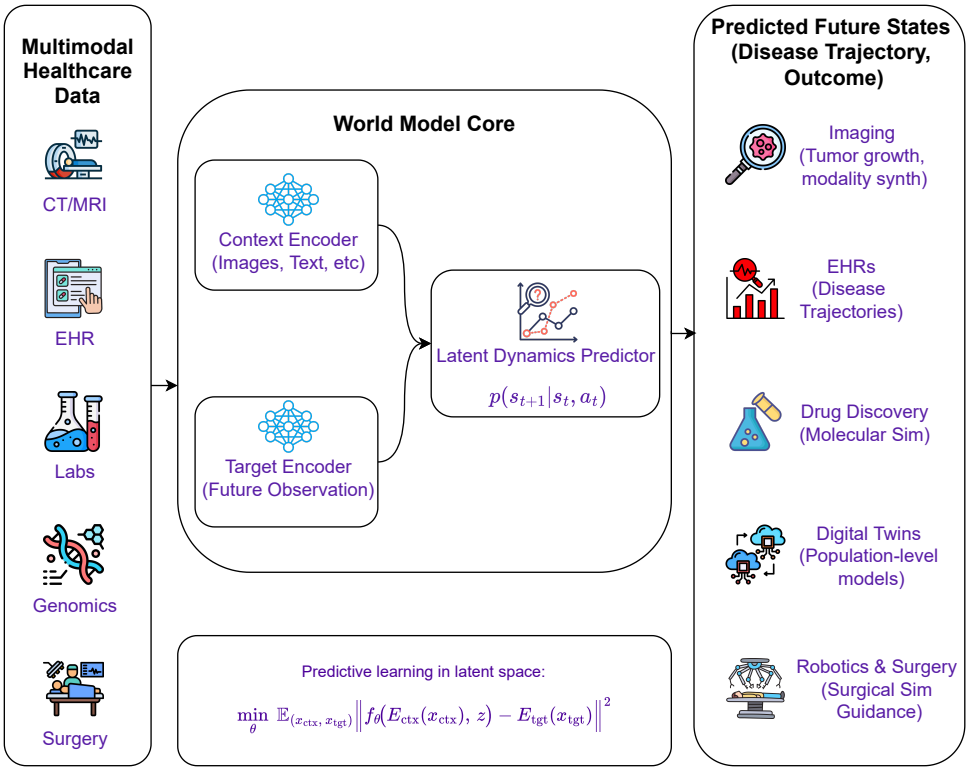

Figure 1: Conceptual schematic of world models for healthcare. Multimodal clinical inputs are encoded into a latent state; a latent dynamics predictor models transitions $p(s_{t+1} \mid s_t, a_t)$; predicted futures support downstream tasks across imaging, EHR trajectories, drug discovery, surgical robotics, and digital twins. The JEPA-style objective (bottom) trains a predictor f_θ to map context embeddings $E_{\text{ctx}}(x_{\text{ctx}})$ and latent dynamics z to match target embeddings $E_{\text{tgt}}(x_{\text{tgt}})$.

health workers are estimated to exceed 10 million by 2030 [1]. This results in ever-growing clinical data, underscoring the importance of AI-driven solutions to improve the quality and efficiency of care. During the past decade, AI in healthcare has evolved from traditional statistical models to deep learning and, more recently, large generative models. Deep neural networks delivered breakthroughs in medical imaging and diagnostics, leading to a broader use in fields such as radiology, pathology, and similar areas[2, 3, 4, 5, 6, 7]. Large Language Models (LLMs) based on transformer architectures now demonstrate strong performance on clinical NLP tasks (e.g., near-expert USMLE performance) [8]. In parallel, image-generative models have emerged as powerful tools; diffusion models synthesize realistic MRI/CT for augmentation and anonymization [9, 10, 11], while VAEs and related frameworks generate synthetic patient data across modalities [12]. These advances illustrate how transformers, diffusion models, and VAEs are reshaping research and practice, enabling improved prediction, data augmentation, and discovery.

However, generative models are insufficient for high-stakes medicine. They lack grounding in the physical, spatial, and causal structure of clinical reality, and can produce plausible but incorrect outputs ("hallucinations") with dangerous consequences [8]. This motivates a prediction-first, world-based alternative in which the model learns *predictive dynamics* - often formalized as $p(s_{t+1} \mid s_t, a_t)$ or via future-latent predictive objectives (e.g. JEPA) [13]. In this paradigm, a *world model* (WM) is an explicit generative model of state dynamics that supports internal simulation, counterfactual evaluation, and planning. Early works show that agents can learn compact latent 'worlds' and train

policies inside them before successful transfer to the real world [14], and recent medical AI has begun to explore such ideas in clinical imaging and surgical simulation [15].

This paper presents the first focused review of *world models in healthcare*, covering medical imaging and diagnostics, disease progression modeling, and robotic surgery/surgical planning. To compare heterogeneous methods, we introduce a capability rubric: **L1** temporal prediction; **L2** action-conditioned prediction; **L3** counterfactual rollouts for decision support; **L4** planning/control. We analyze how these works adapt concepts from generative modeling and model-based reasoning, assess the extent of current progress, and identify open challenges. Because WMs in medicine are still in an early phase, a systematic synthesis can guide research efforts, highlight opportunities, and clarify the trajectory of this emerging area of clinical AI.

2 Background and Foundations

A WM in machine learning refers to a system that learns an internal generative model of the dynamics of the environment, enabling the prediction of future states given current observations and potential actions. Formally, a WM approximates the transition distribution $p(s_{t+1} \mid s_t, a_t)$, where s_t denotes the state at time t and a_t is an action taken by the agent as shown in figure 1. This contrasts with discriminative or purely generative models that focus on classification or static sample generation; WMs emphasize learning predictive dynamics that allow simulation, counterfactual reasoning, and planning. The idea has roots in reinforcement learning (RL), most notably in Sutton's Dyna architecture [16], which combined real experience with updates from a learned model to accelerate policy learning. The concept was revitalized in deep learning through a WM framework [17], demonstrating that compact latent representations of environments, learned with generative recurrent neural networks, could serve as internal simulators where agents learn policies 'in their dreams'.

Since 2018, WM research has expanded significantly, with architectures such as Dreamer [18], SimCore [19], and MuZero [20] demonstrating the advantages of model-based reasoning in complex and high-dimensional domains. More recently, the Joint Embedding Predictive Architecture (JEPA) has been proposed as a unifying framework for predictive learning [13]. JEPA emphasizes learning representations by predicting future latent embeddings rather than raw observations, making it particularly suitable for high-dimensional data such as video or multimodal medical records. These developments reflect a broader trajectory: from early model-based RL with hand-crafted features, to latent-variable generative models capable of internal simulation and planning.

WMs are closely linked with advances in large generative architectures. Variational autoencoders (VAE) [21] enable compression of high-dimensional inputs (e.g., imaging data) into structured latent spaces that can be rolled forward in time. Diffusion models [22] offer powerful generative capabilities with uncertainty quantification, increasingly applied to imaging tasks, including the synthesis of medical data. Transformers [23] excel at modeling long-range dependencies and multimodal fusion, making them natural backbones for WMs in clinical settings where data span temporal, textual, and visual modalities. In this sense, WMs can be viewed as an extension of generative AI. Instead of generating only text or images, they generate trajectories of states and outcomes, providing a predictive simulator for decision support.

Healthcare is a particularly compelling domain for WMs. Clinical problems are inherently multimodal, temporal, and causal: disease trajectories unfold over time, interventions alter patient states, and data sources range from imaging to electronic health records (EHRs) to genomics. Traditional discriminative models, even large generative ones, struggle to capture these dynamics in a way that supports clinical reasoning. WMs can integrate multimodal patient data into a unified latent state, simulate how that state evolves under different interventions, and thus enable critical what-if analyses for decision making. By grounding predictions in learned representations of medical reality, WMs have the potential to overcome the brittleness and lack of physical grounding observed in current generative AI systems [13, 24].

3 Findings and Discussion

Recently, several studies have pushed toward world-model–style learning in healthcare. For clarity, we group the literature by application area: Medical Imaging and Diagnostics, Disease Progression

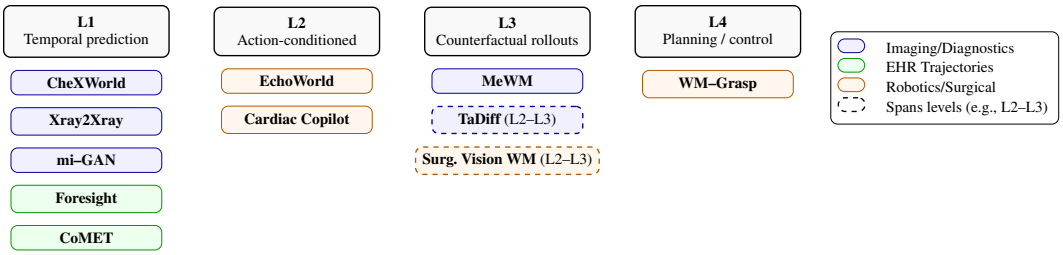

Figure 2: Capability map of reviewed papers across four levels: L1 (temporal prediction), L2 (action-conditioned prediction), L3 (counterfactual rollouts for decision support), L4 (planning/control). Colors denote domains (Imaging, EHR, Robotics). Dashed borders indicate methods spanning adjacent levels (e.g., TaDiff and Surgical Vision WM at L2–L3).

Modeling (EHR), and Robotic Surgery & Surgical Planning, and briefly summarize each study's contribution in Table 1.

Medical Imaging and Diagnostics: Imaging is a natural testbed for learned dynamics because anatomical states evolve and interventions alter future observations. Medical World Model (MedWM) formulates treatment planning as an action-conditioned simulation: A vision language policy proposes interventions and a generative dynamics model predicts posttreatment tumor states for protocol selection [25]. Mi-GAN predicts future 3D brain MRI volumes to model Alzheimer's progression from a baseline scan using a multi-information GAN [26]. CheXWorld uses a JEPA-style predictive objective with context/target encoders and a predictor to learn radiograph representations that capture local anatomy, global layout, and domain variation [27]. Xray2Xray learns projection-transition dynamics at all acquisition angles to encode latent 3D chest volume for downstream risk and diagnostic tasks [28]. Treatment-aware diffusion (TaDiff) conditions longitudinal MRI generation and tumor segmentation in therapy to forecast diffuse glioma evolution under alternative treatment plans [29].

Disease Progression Modeling: Longitudinal EHRs are event streams in which generative models simulate trajectories and forecast future outcomes. Foresight is a generative transformer that converts clinical text to coded concepts and auto-regressively forecasts future disorders, procedures, substances, and findings in large hospital cohorts [30]. Generative Medical Event Models Improve with Scale (CoMET) trains decoder-only transformers on billions of medical events (Epic Cosmos), showing scaling laws and improved multitask predictions while simulating patient-timeline event sequences [31].

Robotic Surgery & Surgical Planning: Guidance and control require models that couple the visual state with action-conditioned dynamics. EchoWorld pre-trains a motion-aware WM for echocardiography that encodes anatomy and the effect of probe motion, reducing plane-guidance error on large-scale ultrasound data [32]. Cardiac Copilot introduces a WM 'Cardiac Dreamer' whose latent spatial features provide a navigation map for real-time probe guidance, improving navigation error on clinical scans [33]. Surgical Vision World Model (SurgWM) adapts action-controllable video generation (latent action inference + dynamics) to synthesize controllable surgical video from unlabeled data [34].

Figure 2 summarizes the distribution of the reviewed works along a capability ladder from **L1** (temporal prediction) to **L4** (planning/control). The literature is concentrated at **L1–L2**: radiograph and longitudinal MRI papers emphasize future prediction in observation or latent space (e.g., future-latent predictive objectives or generation conditioned by therapy), while EHR models largely auto-regress event sequences without explicit actions. **L3–L4** capabilities counterfactual rollouts used for decision support and closed-loop control are comparatively rare, emerging primarily in treatment-aware imaging simulators and robotic/surgical systems. This pattern aligns with the arguments that clinically useful systems must move from sample generation to *prediction-first, world-grounded* modeling of dynamics and interventions (for example, learning $p(s_{t+1} \mid s_t, a_t)$ or future latent prediction) rather than language-only next-token modeling [24, 13].

Domain effects are visible in the map. **Imaging/diagnostics** spans L1–L3: predictive representation learning supports transfer (L1), projection/longitudinal dynamics enable interventional "what-if"

Table 1: Representative works at the intersection of large generative models and world modeling in healthcare. Short names are shown in bold; citations refer to the bibliography.

Year	Paper	Category	Description
2025	**MeWM** [25]	Imaging & Planning	Action-conditioned 3D generator simulates post-treatment tumor from CT/EHR for protocol selection.
2025	**CoMET** [31]	EHR Trajectories	Scalable generative event model forecasting multi-horizon clinical timelines.
2021	**mi–GAN** [26]	Imaging (Neuro)	Multi-information GAN predicts future 3D MRI to model Alzheimer's progression.
2025	**CheXWorld** [27]	Radiography	JEPA-style latent prediction for local anatomy, global layout, and domain shifts.
2025	**EchoWorld** [32]	Ultrasound Guidance	Motion-aware world modeling under probe motion improves echo plane guidance.
2025	**Surg-VM** [34]	Surgical Video Sim	Action-controllable surgical video generation with latent actions for training.
2024	**WM-Grasp** [35]	Surgical Grasping	World-model RL for general surgical grasping robust to object variation/disturbance.
2024	**Cardiac Copilot** [33]	Ultrasound Guidance	World-model ("Cardiac Dreamer") encodes cardiac spatial structure for probe navigation.
2025	**Xray2Xray** [28]	Radiography (Vol)	World model learns 3D volumetric context by modeling projection transitions.
2023	**TaDiff** [29]	Neuro MRI	Treatment-aware diffusion predicts longitudinal MRIs and tumor masks under therapies.
2024	**Foresight** [30]	EHR Forecasting	Generative transformer forecasting future medical events from patient timelines.

simulation (L2–L3), yet few studies close the loop with planning. **EHR** work concentrates at L1, reflecting strong timeline forecasting on scale but limited action semantics and limited counterfactual validation. **Robotics/surgical planning** naturally advances to L2 and beyond because actions (probe/tool motion) are explicit; instances of L3 appear where controllable video generation is used for simulated outcomes and L4 is reached when learned dynamics drive control policies. Across domains, large generative backbones (transformers, diffusion, VAE) are common, but the defining step toward "world modeling" is the *use of learned dynamics for rollouts under interventions*—not generative fluency per se.

Two cross-cutting gaps explain why many methods remain at L1–L2. First, **action semantics and constraints** are underspecified outside robotics: therapy, protocol, dose, or timing are rarely formalized as action variables with units and safety limits. Second, **the interventional validity** is weakly measured: simulated futures may be realistic, but incorrect about treatment effects. Progress on these fronts, together with multimodal state construction, trajectory-level uncertainty calibration, and decision-aligned evaluation, will be decisive in advancing toward clinically reliable L3–L4 WMs [24].

4 Future Work

Although world–modeling approaches are beginning to deliver predictive representations and simulation capabilities in healthcare, advances are needed in various aspects. Here, we present the gaps and outline directions to address them.

- **From L1 to L2: formalize actions.** Define clinically meaningful action spaces per domain: imaging (protocol/angle/contrast), EHR (drug, dose, timing), robotics (tool pose / force) —- and encode safety constraints within the model and interface of implementation.
- **From L2 to L3: establish counterfactual correctness.** Evaluate "what-if" rollouts for interventional validity via multi-site holdouts, matched cohorts or natural experiments, and clinician adjudication; favor decision-centric endpoints (e.g., survival, adverse events) over perceptual or token-level metrics.
- **From L3 to L4: close the loop.** Couple learned dynamics with model-based RL/MPC and reliable off-policy evaluation; stage prospective, small-N pilots prior to deployment, with uncertainty-aware planning and explicit abstention policies.
- **Multimodal integrated state at scale.** Learn unified latent states fusing imaging, EHR, waveforms, and genomics under missingness and irregular sampling; characterize scaling laws linking data/model size to rollout fidelity and decision value.
- **Trajectory-level uncertainty and robustness.** Provide calibrated predictive distributions over entire trajectories (e.g., conformal bands for longitudinal imaging/EHR), stress-test

out-of-distribution shifts (scanner/site, demographics), and assess subgroup fairness of *recommended actions*.

- **Causal and mechanistic grounding.** Combine learned dynamics with causal identification (adjustment sets, instrumental variables, transportability checks) and embed mechanistic priors (projection geometry, tumor growth or PK–PD, anatomy/physics) to improve extrapolation and safety.
- **Evaluation and reporting standards.** Adopt shared capability benchmarks aligned with the ladder in Figure 2; standardize horizon lengths, action definitions, counterfactual protocols, and ablations isolating action-conditioning and planning components.
- **Tooling, privacy, and governance.** Release reference implementations and lightweight simulators (imaging acquisition, EHR counterfactual sandboxes, surgical video) for reproducibility; define governance for simulation-based recommendations with privacy-preserving training and audited synthetic data.
- **Efficiency and deployability.** Use distillation, amortized planning, and latent-space MPC for real-time robotics and bedside decision support; report latency, compute, and energy alongside accuracy.

5 Conclusion

This review aimed to clarify what constitutes a *world model* for healthcare and review recent work within a capability-oriented rubric. We distinguished world modeling from generic generative AI by its focus on *predictive dynamics*, often formalized as $p(s_{t+1} \mid s_t, a_t)$ or future-latent prediction that enable rollouts, counterfactual evaluation, and planning. Surveying diagnostic imaging, EHR disease progression, and robotic surgery, we found that most methods currently achieve **L1–L2** capabilities (temporal prediction, action-conditioned prediction), with fewer instances of **L3** counterfactual decision support and rare **L4** closed-loop planning/control. The field has focused around large generative backbones (transformers, diffusion, VAEs) paired with latent dynamics learning (e.g., JEPA-style objectives), but the defining ingredient is the *use of learned dynamics* to reason under interventions rather than generative fluency alone.

Our synthesis highlights the central obstacles to progress: underspecified clinical *action spaces* (and associated safety constraints) outside robotics; limited evaluation of *interventional validity* for simulated futures; incomplete construction *multimodal state* in EHR, imaging and signals; and lack of standardized, decision-aligned evaluation protocols. Addressing these gaps is essential for moving beyond accurate forecasting toward reliable counterfactual reasoning and clinically actionable planning. WMs that integrate principled action semantics, calibrated trajectory uncertainty, and causal/mechanistic grounding can elevate healthcare AI from pattern recognition to faithful simulation and safe decision support. As community benchmarks converge on a capability ladder (L1–L4), and as open tooling, governance, and privacy-preserving training mature, we anticipate a shift toward **L3–L4** systems that meaningfully assist clinicians: simulating treatment alternatives, guiding imaging and procedures, and closing the loop in safety-critical settings. Realizing this vision will require sustained collaboration across machine learning, clinical science, and health systems engineering; the works reviewed here provide the scaffolding on which the next generation of clinically robust WMs will be built.

References

[1] World Health Organization. *World report on ageing and health*. World Health Organization, 2015.

[2] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud A. A. Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[3] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[4] Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[5] Qazi Mohammad Areeb, Mohammad Nadeem, Roobaea Alroobaea, Faisal Anwer, et al. Helping hearing-impaired in emergency situations: A deep learning-based approach. *IEEE Access*, 10:8502–8517, 2022.

[6] Qazi Mohammad Areeb and Mohammad Nadeem. Deep learning based hand gesture recognition for emergency situation: A study on indian sign language. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, pages 33–36. IEEE, 2021.

[7] Mohammad Areeb Qazi, Mohammed Talha Alam, Ibrahim Almakky, Werner Gerhard Diehl, Leanne Bricker, and Mohammad Yaqub. Multi-task learning approach for unified biometric estimation from fetal ultrasound anomaly scans. In *International Conference on Medical Imaging and Computer-Aided Diagnosis*, pages 52–61. Springer, 2023.

[8] Syed Arman Rabbani, Mohamed El-Tanani, Shrestha Sharma, Syed Salman Rabbani, Yahia El-Tanani, Rakesh Kumar, and Manita Saini. Generative artificial intelligence in healthcare: Applications, implementation challenges, and future directions. *BioMedInformatics*, 5(3):37, 2025.

[9] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023.

[10] Anees Ur Rehman Hashmi, Ibrahim Almakky, Mohammad Areeb Qazi, Santosh Sanjeev, Vijay Ram Papineni, Jagalpathy Jagdish, and Mohammad Yaqub. Xreal: Realistic anatomy and pathology-aware x-ray generation via controllable diffusion model. *arXiv preprint arXiv:2403.09240*, 2024.

[11] Mohammed Talha Alam, Raza Imam, Mohammad Areeb Qazi, Asim Ukaye, Karthik Nandakumar, and Abu Dhabi. Introducing sdice: An index for assessing diversity of synthetic medical datasets. *arXiv preprint arXiv:2409.19436*, 2024.

[12] Mahmoud Ibrahim, Yasmina Al Khalil, Sina Amirrajab, Chang Sun, Marcel Breeuwer, Josien Pluim, Bart Elen, Gökhan Ertaylan, and Michel Dumontier. Generative ai for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in biology and medicine*, 189:109834, 2025.

[13] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[14] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

[15] Saurabh Koju, Saurav Bastola, Prashant Shrestha, Sanskar Amgain, Yash Raj Shrestha, Rudra PK Poudel, and Binod Bhattarai. Surgical vision world model. *arXiv preprint arXiv:2503.02904*, 2025.

[16] Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pages 216–224. Elsevier, 1990.

[17] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[18] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[19] Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping belief states with generative environment models for rl. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[24] Yann LeCun. Beyond language models: Yann lecun's world models and the future of ai in healthcare. `https://www.onhealthcare.tech/p/beyond-language-models-yann-lecuns`, 2024. Accessed: 2025-08-27.

[25] Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical world model: Generative simulation of tumor evolution for treatment planning. *arXiv preprint arXiv:2506.02327*, 2025.

[26] Y. Zhao, B. Ma, P. Jiang, D. Zeng, X. Wang, and S. Li. Prediction of Alzheimer's disease progression with multi-information generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 25(3):711–719, 2021.

[27] Yang Yue, Yulin Wang, Chenxin Tao, Pan Liu, Shiji Song, and Gao Huang. Chexworld: Exploring image world modeling for radiograph representation learning. *arXiv preprint arXiv:2504.13820*, 2025. Also presented at CVPR 2025.

[28] Zefan Yang, Xinrui Song, Xuanang Xu, Yongyi Shi, Ge Wang, Mannudeep K. Kalra, and Pingkun Yan. Xray2xray: World model from chest x-rays with volumetric context. *arXiv preprint arXiv:2506.19055*, 2025.

[29] Q. Liu, I. Vartholomeos, Bradley J. MacIntosh, Edvard Grønning, Per Brandal, Knut Østertun Geier, Kyrre M. Emblem, Enrique Fuster-Garcia, Carla López-Mateu, et al. Treatment-aware diffusion probabilistic model for longitudinal MRI generation and diffuse glioma growth prediction. *arXiv preprint arXiv:2309.05406*, 2023.

[30] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

[31] Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale. *arXiv preprint arXiv:2508.12104*, 2025.

[32] Yang Yue, Yulin Wang, Haojun Jiang, Pan Liu, Shiji Song, and Gao Huang. Echoworld: Learning motion-aware world models for echocardiography probe guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25993–26003, June 2025.

[33] Haojun Jiang, Zhenguo Sun, Ning Jia, Meng Li, Yu Sun, Shaqi Luo, Shiji Song, and Gao Huang. Cardiac copilot: Automatic probe guidance for echocardiography with world model. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15001 of *Lecture Notes in Computer Science*, pages 190–199. Springer Nature Switzerland, 2024.

[34] S. Koju, S. Park, M. Gil, H. Gao, M. Bennamoun, et al. Surgical vision world model. *arXiv preprint arXiv:2503.02904*, 2025.

[35] Guangyao Lin, Xinyue Yan, Yuzhou Hu, Xiangchen Xie, Shuxin Wang, Shijian Song, and Max Q.-H. Meng. World models for general surgical grasping. *arXiv preprint arXiv:2405.17940*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction reflect the paper's contributions and scope

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: Given the nature of the review, we only focused on providing a brief insight into the field. So limitations are not applicable.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Please check the abstract and first paragraph of the introduction. Every theoretical result is either cited or given with proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: Given the nature of the work, reproducibility is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Given the nature of the work, access to data and code is not applicable.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Given the nature of the review work, the experimental setting/details are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Given the nature of the review work, the experimental statistical significance is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Did not require any computation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conducted the review work keeping the NeurIPS Code of Ethics in consideration.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Since this is a review work, there are no positive or negative impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Since this is a review work, no data or model is involved.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The papers are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets introduced in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There were no experiments or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There were no human subject involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM was used just for paraphrasing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.