KINEWORLD: LATENT KINETIC WORLD MODELING FOR CONTRAST-FREE DCE-MRI SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) provides clinically informative enhancement trajectories, yet contrast-agent administration is unsuitable for some patients. Synthesizing patient-specific DCE-MRI trajectories from a pre-contrast scan requires preserving anatomical consistency while faithfully modeling tissue-specific contrast kinetics. Existing methods do not explicitly represent evolving kinetic states, often resulting in structural drift and physiologically implausible kinetic trajectories. We propose KINEWORLD, a latent kinetic world model that anchors cumulative contrast-dependent changes to the pre-contrast latent state, thereby separating kinetic evolution from time-invariant anatomy to preserve anatomical consistency. To achieve kinetic fidelity, a pharmacokinetics-inspired rollout recursively integrates learned influx and efflux velocities, while kinetics-aware retrieval supplies relevant yet diverse long-range patient history to preserve phase-dependent evolution. Relay-guarded self-rollout training further limits accumulated anatomical and kinetic errors by selectively restoring ground-truth states upon excessive deviation. Across three breast and prostate DCE-MRI datasets, KINEWORLD improves SSIM by over 11% and lesion kinetic curve CCC by over 17% relative to the strongest baselines. A blinded reader study further demonstrates substantially improved agreement with report-derived kinetic categories. Source code, pretrained weights, and our curated internal dataset will be publicly released to facilitate future research.[1]

1 INTRODUCTION

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a temporally resolved imaging technique widely used in clinical practice O'Connor et al. (2011). Following contrast-agent administration, serial MR images are acquired at multiple time points to capture the temporal evolution of lesion enhancement dynamics, underpinning accurate lesion characterization and treatment response assessment Li et al. (2024); Khalifa et al. (2014). However, contrast administration is clinically contraindicated in vulnerable patient cohorts, such as those with severe renal impairment, contrast allergies, or pregnancy Tofts & Parker (2013). This limitation motivates us to leverage world models Ding et al. (2025); Yang et al. (2025), which excel at learning and simulating complex physical dynamics, to model MRI contrast-enhancement kinetics Osuala et al. (2024); Kong et al. (2026). Specifically, a DCE-MRI world model should capture patient-specific anatomy and simulate tissue-specific enhancement dynamics, enabling individualized DCE-MRI trajectories to be synthesized from a pre-contrast scan without contrast-agent administration.

Despite its clinical potential, developing such a DCE-MRI world model poses two major challenges: *anatomical consistency* and *kinetic fidelity*. Specifically, *anatomical consistency* requires patient anatomy, particularly lesions, to remain spatially and morphologically stable across all time points, as even subtle lesion displacement or deformation may alter clinical interpretation Kong et al. (2026). More importantly, *kinetic fidelity* require the model to faithfully reproduce the temporally ordered evolution of contrast enhancement. Achieving this requires modeling dependencies and state transitions among the contrast wash-in, peak-enhancement, and wash-out phases, while accounting for lesion-specific variations in enhancement rate and delayed-phase behavior Li et al. (2024).

[1] Code is available at Anonymous Github Repository.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

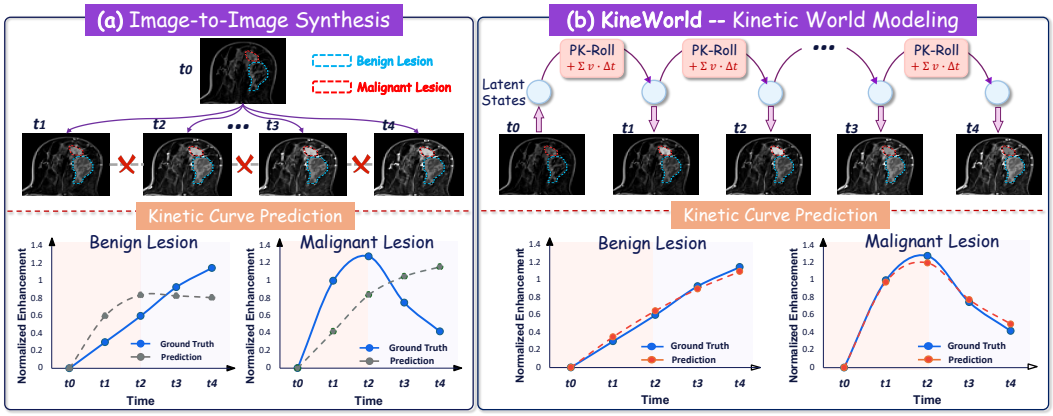

Figure 1: Framework comparison of **(a)** conventional image-to-image (I2I) synthesis and **(b)** KINEWORLD. The kinetic curves (gray: ground truth, blue: I2I, red: KINEWORLD) show that I2I exhibits premature saturation for the benign lesion and misses the characteristic wash-out of the malignant lesion. By recursively rolling forward temporally connected latent states, KINEWORLD faithfully preserves the enhancement patterns of both benign and malignant lesions.

As illustrated in Fig. 1**(a)**, existing methods typically formulate DCE-MRI synthesis as an image-to-image (I2I) translation problem Kong et al. (2026); Osuala et al. (2024), independently mapping the pre-contrast image at t_0 to each target post-contrast phase. Conditioning all predictions on the same anatomical baseline helps preserve anatomy, but treats the post-contrast phases as separate targets rather than temporal process. Consequently, these methods do not explicitly model how contrast uptake at one phase constrains its subsequent evolution. This limitation is reflected by the lesion-level kinetic curves at the bottom of Fig. 1**(a)**, which plot the mean normalized pixel intensity within each lesion region across t_0–t_4. Although the ground-truth benign lesion exhibits persistent enhancement, the I2I prediction rises prematurely and then plateaus. More notably, the ground-truth malignant lesion shows rapid wash-in, and subsequent wash-out, whereas the predicted curve follows a persistently increasing trajectory.

To address these limitations, we propose KINEWORLD, a pharmacokinetic(PK)-inspired world model for DCE-MRI trajectory synthesis, as illustrated in Fig. 1**(b)**. Our key insight is to represent tissue enhancement as a dynamical process governed by the balance between contrast-agent influx and efflux, motivated by the Tofts model Tofts (1997). Starting from the pre-contrast state, KINEWORLD autoregressively rolls forward temporally connected latent states and decodes them into DCE-MRIs. This formulation explicitly captures inter-phase dependencies and lesion-specific enhancement kinetics. We realize this framework through three technical designs and a large-scale DCE-MRI dataset spanning diverse lesion types:

PK-Roll. Rather than directly predicting target images, we model enhancement evolution in a latent space anchored by the pre-contrast anatomical state. A PK-inspired rollout operator decomposes the instantaneous kinetic velocity into learned influx and anatomy-conditioned efflux components. Integrating the predicted velocity over time keeps the patient anatomy anchored while allowing contrast-dependent components to evolve continuously.

Kinetic-aware Retrieval. Long-horizon rollout requires historical context beyond the immediately preceding state. We therefore maintain a patient-specific memory bank that stores previously generated latent states. For each transition, the model dynamically retrieves states relevant to the current kinetic phase. Maximum Marginal Relevance Carbonell & Goldstein (1998) further balances the retrieval relevance and diversity, reducing redundant retrieval during stable phases while preserving informative long-range context.

Relay-Guard Training. To mitigate error accumulation during training, we introduce a closed-loop rollout strategy in which the model is trained primarily on its own generated histories. When a generated trajectory deviates beyond a predefined trust region, Relay-Guard forwards a stop-gradient ground-truth state to the subsequent rollout under a limited intervention budget. This selective intervention interrupts severe error propagation and limits memory-bank contamination while retaining moderate prediction errors as informative on-policy training signals.

Large-scale DCE-MRI Dataset. To support the learning and evaluation of heterogeneous enhancement kinetics, we curate a large-scale breast DCE-MRI dataset comprising 3,491 DCE-MRI series, totally 21,078 MRI volumes with paired reports, and 78,116 lesion annotations spanning diverse categories. The scale and diversity provide broad coverage of clinically heterogeneous lesion enhancement trajectories, facilitating the development and evaluation of DCE-MRI world models.

Extensive evaluations on the Duke, ProstateX, and curated internal datasets demonstrate that KINEWORLD significantly outperforms state-of-the-art baselines, achieving over 11% improvements in SSIM-based anatomical consistency and over 17% gains in Concordance Correlation Coefficient (CCC) for pharmacokinetic trajectory fidelity.

2 METHOD

2.1 PRELIMINARIES

Following the influx–efflux principle of the Tofts pharmacokinetic (PK) model Tofts (1997), we define the instantaneous kinetic velocity $\boldsymbol{V}_e(t)$[2] as

$$\boldsymbol{V}_e(t) = \frac{\mathrm{d}\boldsymbol{Z}_e(t)}{\mathrm{d}t} = \boldsymbol{Z}_{\mathrm{in}}(t) - \boldsymbol{Z}_{\mathrm{ef}}(t), \tag{1}$$

where $\boldsymbol{Z}_e(t)$ denotes the current enhancement state. Its evolution is governed by two factors: 1) the *influx* $\boldsymbol{Z}_{\mathrm{in}}(t)$, which characterizes newly arriving enhancement and is inferred jointly from patient-specific anatomy and historical enhancement states; 2) the *efflux* $\boldsymbol{Z}_{\mathrm{ef}}(t) = \boldsymbol{k}_e \odot \big[\boldsymbol{Z}_e(t) - \boldsymbol{Z}_e(t_0)\big]$, where $\odot$ denotes the Hadamard product. This formula characterizes the spatial attenuation of the current enhancement residual through an anatomy-conditioned efflux coefficient $\boldsymbol{k}_e$.

In practice, we discretize Eq. 1 over acquisition timestamps $\{t_i\}_{i=0}^{N}$ as

$$\boldsymbol{Z}_e(t_n) = \boldsymbol{Z}_e(t_0) + \sum_{i=0}^{n-1} \boldsymbol{V}_e(t_i)\Delta t_i, \qquad \Delta t_i = t_{i+1} - t_i, \tag{2}$$

where $\boldsymbol{Z}_e(t_0)$ denotes the patient-specific pre-contrast state and provides a complete anatomical baseline. This formulation offers two advantages. *First*, accumulating contrast-dependent changes over a fixed pre-contrast baseline separates kinetic evolution from time-invariant anatomy, thereby promoting the anatomical consistency. *Second*, we shift the prediction target from raw image intensities to instantaneous kinetic velocity, replacing direct observation-to-observation mappings with Tofts PK model-inspired latent state transitions. This formulation facilitates DCE-MRI world modeling, encouraging the model to capture underlying pharmacokinetic regularities and thereby enabling fine-grained control over the direction, magnitude, and timing of enhancement evolution.

2.2 PK-ROLL: PHARMACOKINETICS-INSPIRED STATE ROLLOUT

As illustrated in Fig. 2**(a)**, we instantiate Eq. 2 by rolling out a patient-specific latent kinetic trajectory over a dense physical-time grid $\{t_i\}_{i=0}^{N}$, satisfying $\Delta t = t_{i+1} - t_i = 1s$. Our objective is to learn the transition dynamics between latent states, thereby enabling state prediction at arbitrary timestamps. The pre-contrast scan $\boldsymbol{I}_0$ serves as the anatomical baseline, with its latent representation extracted via the frozen Stable Diffusion 1.5 Rombach et al. (2022) (SD1.5) VAE encoder: $\boldsymbol{z}_0 = \mathrm{Enc}_{\mathrm{VAE}}(\boldsymbol{I}_0)$.

Kinetic State. At each time step t_i, each kinetic state is formulated as a triplet $(\hat{\boldsymbol{z}}_i, \hat{\boldsymbol{v}}_i, t_i)$, where $\hat{\boldsymbol{z}}_i$ denotes the predicted latent enhancement state (hereafter referred to as the latent state) and $\hat{\boldsymbol{v}}_i$ represents its instantaneous kinetic velocity. At each acquisition time step t_i, the corresponding ground-truth latent state is given by $\boldsymbol{z}_i = \mathrm{Enc}_{\mathrm{VAE}}(\boldsymbol{I}_i)$. Then, starting from $\hat{\boldsymbol{z}}_0 = \boldsymbol{z}_0$ and $\hat{\boldsymbol{v}}_0 = \boldsymbol{0}$, we explicitly predict the kinetic velocity and recursively update the latent state by $\hat{\boldsymbol{z}}_{i+1} = \hat{\boldsymbol{z}}_i + \Delta t \cdot \hat{\boldsymbol{v}}_{i+1} = \boldsymbol{z}_0 + \Delta t \cdot \sum_{j=0}^{i} \hat{\boldsymbol{v}}_{j+1}$.

[2]Here we present a simplified version for conciseness. A formal formulation is provided in Appendix A.1.

PK-inspired Rollout Operator. To recursively predict the kinetic velocity, we design a PK-inspired rollout operator by emulating the coupled influx–efflux dynamics in Eq. 1. Specifically, at time step t_i, we predict the latent kinetic velocity as the difference between a learned influx term and an anatomy-conditioned efflux term as

$$\hat{\boldsymbol{v}}_{i+1} = \underbrace{g_\theta(\hat{\boldsymbol{z}}_i, \mathcal{C}_i, \mathrm{PE}(t_i))}_{\text{influx}} - \underbrace{\boldsymbol{k}_e \odot (\hat{\boldsymbol{z}}_i - \boldsymbol{z}_0)}_{\text{efflux}}, \tag{3}$$

where $\mathcal{C}_i$ denotes the dynamically retrieved historical kinetic context (§ 2.3). The influx predictor $g_\theta(\cdot)$ is implemented as a Transformer block that dynamically integrates the current state and historical kinetic context through attention. $\mathrm{PE}(\cdot)$ denotes a sinusoidal temporal embedding. The anatomy-conditioned efflux coefficient is predicted from the pre-contrast latent state $\boldsymbol{k}_e = \sigma(h_\phi(\boldsymbol{z}_0))$, where $h_\phi(\cdot)$ is a lightweight ASPP module Chen et al. (2017) and $\sigma(\cdot)$ denotes the sigmoid activation. It induces washout proportional to the current enhancement residual while suppressing unstable latent drift. Detailed architectures of g_θ and h_ϕ are provided in Appendix A.2.

Decoding. The predicted latent state at time step t_i is decoded back to image space via $\hat{\boldsymbol{I}}_i = \mathrm{Dec}_{\mathrm{VAE}}(\hat{\boldsymbol{z}}_i)$, with the entire rollout supervised concurrently in both latent and image spaces by

$$\mathcal{L}_{\mathrm{recon}} = \frac{1}{N}\sum_{i=1}^{N}\left(\|\hat{\boldsymbol{z}}_i - \boldsymbol{z}_i\|_1 + \left\|\hat{\boldsymbol{I}}_i - \boldsymbol{I}_i\right\|_1\right). \tag{4}$$

Although $\mathcal{L}_{\mathrm{recon}}$ anchors predictions at the observed timestamps, it does not explicitly constrain the intermediate transitions. Autoregressive rollout may therefore exhibit non-physical temporal jitter between observations. To promote temporal continuity and penalize abrupt kinetic accelerations, we introduce a smoothness regularization over all rollout steps:

$$\mathcal{L}_{\mathrm{smooth}} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left\|\frac{\hat{\boldsymbol{v}}_{i+1} - \hat{\boldsymbol{v}}_i}{\Delta t}\right\|_1. \tag{5}$$

Since $\hat{\boldsymbol{v}}_i$ represents the first-order temporal change, this objective penalizes abrupt second-order changes and encourages a smooth latent kinetic trajectory.

2.3 KA-RETRIEVAL: LONG-RANGE KINETICS-AWARE RETRIEVAL.

Kinetic Memory Bank. Long-horizon rollout across dense time points cannot rely solely on a local preceding window, as early contrast uptake and spatial redistribution along the kinetic trajectory contain critical clues for predicting subsequent states. However, incorporating every historical state directly into the computation is computationally prohibitive and largely redundant. As presented in Fig. 2**(b)**, we therefore maintain a patient-specific memory bank $\mathcal{B}$ in a key–value form. Each entry contains a key $\boldsymbol{k}_i = [\hat{\boldsymbol{z}}_i;\, \hat{\boldsymbol{v}}_i]$, which is the concatenation of the latent enhancement and velocity, with the tuple of the kinetic state token and the timestamp $(\hat{\boldsymbol{z}}_i, t_i)$ as its value. Since the key $\boldsymbol{k}_i$ is constructed from the kinetic variables, it inherently distinguishes states with different enhancement intensities, or states sharing similar intensity but exhibiting opposite instantaneous enhancement change, such as wash-in versus wash-out phases.

Kinetics-aware Retrieval To predict the next state at t_i, the historical context is assembled by

$$\mathcal{C}_i = \underbrace{\{\boldsymbol{z}_0\}}_{\text{initial anchor}} \cup \underbrace{\mathrm{Retrieve}_K\big(\boldsymbol{q}_i,\, \mathcal{B}_{(0,\,i-L)}\big)}_{\text{retrieval from long-range memory}} \cup \underbrace{\{\hat{\boldsymbol{z}}_{i-L}, \ldots, \hat{\boldsymbol{z}}_{i-1}\}}_{\text{local history}}, \tag{6}$$

where the query $\boldsymbol{q}_i = [\hat{\boldsymbol{z}}_i;\, \hat{\boldsymbol{v}}_i]$ $(i \geq 1)$ is the same as key, and L is the local-window length. $\mathcal{B}_{(0,\,i-L)}$ indicates the tokens $\{\hat{\boldsymbol{z}}_k\}_{k=1}^{i-L-1}$ in the memory bank $\mathcal{B}$. The initial token anchors the anatomy information from the pre-contrast state, the local history preserves the short-term continuity, and the retrieved top-K tokens recover long-range evolution context.

However, standard top-K retrieval evaluates candidate entries independently, neglecting inter-item redundancy within the retrieved set. For instance, consecutive states in stationary regimes exhibit minimal variation, causing unconstrained top-K selection to yield highly redundant tokens and

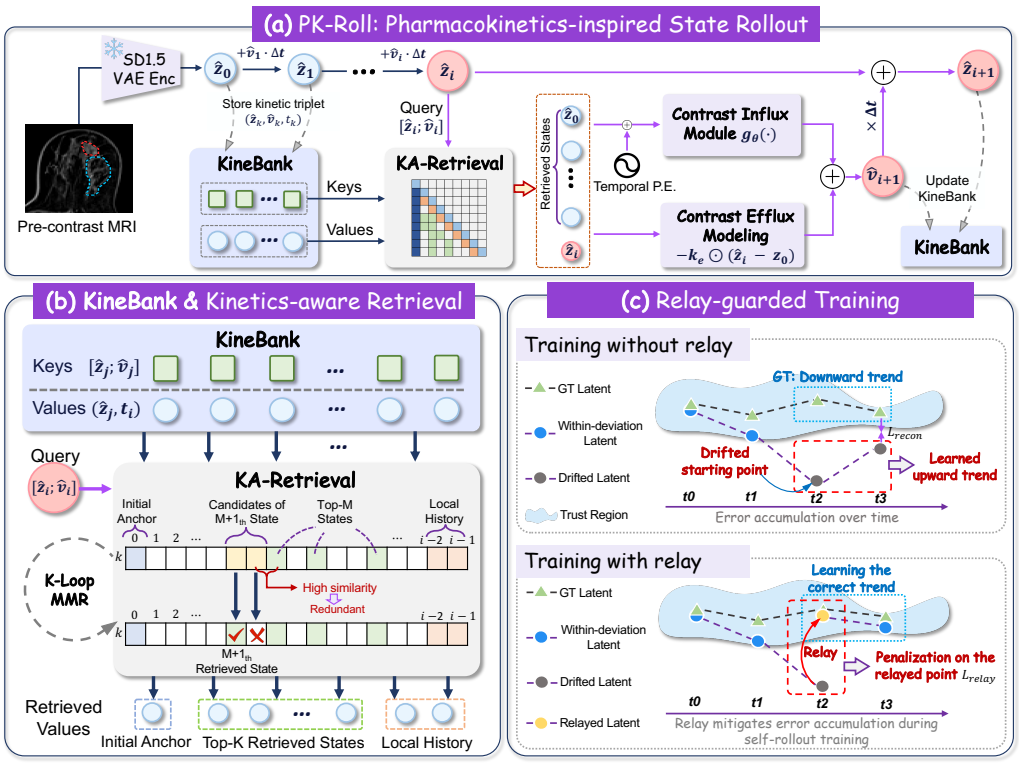

Figure 2: Overview of KINEWORLD. **(a)** Anchored by the pre-contrast scan, *PK-Roll* recursively evolves the latent state $\hat{z}_i$ in latent space. **(b)** *KineBank* stores trajectory history, where *KA-Retrieval* queries relevant and informative historical context for each step. **(c)** *Relay-guarded Training* intervenes when trajectory drift exceeds the trust region, stabilizing self-rollout optimization.

severely degrade effective context capacity. To enforce token diversity, we construct the long-range memory $\mathcal{S}$ iteratively via Maximum Marginal Relevance (MMR) Carbonell & Goldstein (1998). Starting from $\mathcal{S} = \varnothing$, entries are added one at a time,

$$i^* = \arg\max_{i \notin \mathcal{S}} \Big[\underbrace{\cos(\boldsymbol{q}_k, \boldsymbol{k}_i)}_{\text{relevance}} - \lambda_m \cdot \underbrace{\max_{j \in \mathcal{S}} \cos(\boldsymbol{k}_i, \boldsymbol{k}_j)}_{\text{redundancy}} \Big], \qquad \mathcal{S} \leftarrow \mathcal{S} \cup \{i^*\}, \tag{7}$$

until K entries are collected. Since K is small in practice, the sequential rounds incur negligible computation overhead. From the second round onward, every candidate pays a redundancy tax before entering the memory set. The max operator makes this tax decisive, as a candidate is penalized for duplicating *any* previously selected entry and receives no exemption for differing from the rest. The coefficient λ_m balances relevance against diversity. Crucially, because the redundancy term is measured in the same kinetic key space as relevance, the penalty is kinetics-aware by design, eliminating the need for hand-designed detectors or thresholds. Pseudo code is in Appendix A.3.

2.4 RELAY-GUARDED TRAINING

Error accumulation and relay. The rollout proceeds in an on-policy manner, where every step operates on states generated by the model itself. A local error may therefore enter the memory bank and propagate through subsequent predictions, as in Fig. 2**(c)**. Because ground-truth images are available only at discrete acquisition times, and training on ground-truth histories alone does not reproduce the compounding errors of self-rollout, we introduce *relay-guarded self-rollout training*. The model conditions on its generated history, while a ground-truth state is reintroduced only when the trajectory exits a predefined trust region. We call each such correction a *relay*.

Deviation-triggered relays. The deviation of a rollout prediction is measured by $\mathrm{dev}_i = \|\hat{\boldsymbol{z}}_i - \boldsymbol{z}_i\|_1 / \|\boldsymbol{z}_i\|_1$. A relay triggers when the deviation exceeds threshold τ_d. To ensure the model learns from the errors causing each reset, the triggered deviations are penalized via the relay loss:

$$\mathcal{L}_{\mathrm{relay}} = \frac{1}{N} \sum_{i=1}^{N} \rho_i \cdot \mathrm{dev}_i, \tag{8}$$

where $\rho_i = \mathbb{1}[\mathrm{dev}_i > \tau_d]$ indicates a relay event. If triggered, the deviated states are corrected via

$$\hat{\boldsymbol{z}}_i \leftarrow (1-\rho_i)\hat{\boldsymbol{z}}_i + \rho_i \,\mathrm{sg}(\boldsymbol{z}_i), \quad \hat{\boldsymbol{v}}_i \leftarrow (1-\rho_i)\hat{\boldsymbol{v}}_i + \rho_i \,\mathrm{sg}(\bar{\boldsymbol{v}}_i), \tag{9}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Since true velocities are unobservable, we approximate them via the backward secant $\bar{\boldsymbol{v}}_i = (\boldsymbol{z}_i - \boldsymbol{z}_{i-1})/(t_i - t_{i-1})$. Concurrently, corrupted states are removed from KINEWORLDand replaced with ground-truth entries. This prevents future predictions from compensating for the current state error, as illustrated in Fig. 2**(c)** bottom.

2.5 TRAINING OBJECTIVE AND SCHEDULE

Objective. At each image acquisition point, the losses are evaluated on the pre-relay prediction. If triggered, a relay is applied only afterward to initialize the subsequent rollout. The overall training objective is $\mathcal{L} = \mathcal{L}_{\mathrm{recon}} + \mathcal{L}_{\mathrm{smooth}} + \mathcal{L}_{\mathrm{relay}}$, where all terms are equally weighted in our experiments.

Horizon curriculum. To gradually expose the model to longer self-generated histories, we define $H(e)$ as the number of post-contrast phases included from each case at training epoch e. During warm-up, $H(e)$ increases from 1 post-contrast phase to all E available post-contrast phases

$$H(e) = \left\lceil 1 + (E-1)\min\left(\frac{e}{E_{\mathrm{warm}}}, 1\right)\right\rceil, \tag{10}$$

where e is the zero-indexed training steps, E_{warm} is the warm-up duration, and $\lceil\cdot\rceil$ denotes the ceiling function that maps the scheduled horizon to the nearest upper integer.

3 EXPERIMENTS

3.1 DATASETS AND IMPLEMENTATION DETAILS

Datasets We evaluate our framework on three DCE-MRI datasets. (1) **Duke** Saha et al. (2021) dataset comprises 922 breast examinations with 3–4 post-contrast phases, with *only* malignant lesion masks. We use the official train/test split. (2) **Internal dataset** [3], comprising 3,491 breast DCE-MRI series with 5–7 post-contrast phases, paired reports, and pathology-confirmed malignant and benign lesion masks, split 8:1 for training and testing. (3) **PROSTATEx** Armato III et al. (2018) comprises 343 prostate examinations with lesion-center coordinates. We also use 8:1 train/test split. For each dataset, 10% of the training partition is held out for validation and hyperparameter selection, while the test set remains untouched.

Implementations For KA-Retrieval, we set the local window size $L = 2$, the retrieval top $K = 6$, and the MMR trade-off coefficient $\lambda_m = 0.4$. The relay-guarded threshold is set to $\tau_d = 0.1$. The model is optimized by AdamW with a batch size of 8, a learning rate of 1×10^{-4}, and zero weight decay for 10 epochs. The warm-up duration is $E_{\mathrm{warm}} = 1$ epoch. All models are trained on a NVIDIA H20 GPU (80GB). The data preprocessing is provided in Appendix A.4.

Baselines We compare our framework with two complementary families of baselines. The *image-to-image* group comprises EditAR Mu et al. (2025), ControlNet Zhang et al. (2023), T2I Mou et al. (2024), CCNet Osuala et al. (2024), and CEK-World Kong et al. (2026). The *video-generation* group includes AnimateDiff Guo et al. (2024), DynamiCrafter Xing et al. (2024a), and Latent Video Diffusion Models (LVDM) He et al. (2022), which are adapted to generate the post-contrast sequence conditioned on the pre-contrast image. All methods are tuned with the same efforts, using identical data preprocessing and augmentations. Details are provided in Appendix A.5.

[3] Approved under Clinical Trial No. XXXX.

Table 1: Analysis of **anatomical consistency** on three datasets. **Bold** and underlined values indicate the best and second-best results, respectively.

Method	Duke Dataset				Internal Dataset				ProstateX Dataset			
	PSNR↑	SSIM↑	LPIPS↓	rMSE↓	PSNR↑	SSIM↑	LPIPS↓	rMSE↓	PSNR↑	SSIM↑	LPIPS↓	rMSE↓
Image-to-Image-Based Methods												
ControlNet Zhang et al. (2023)	19.74	0.5221	0.2620	298.8	18.89	0.4973	0.2812	309.2	21.86	0.7004	0.2578	227.4
T2I Mou et al. (2024)	17.25	0.4332	0.2915	300.6	17.01	0.4222	0.3033	314.5	19.19	0.6787	0.2893	217.3
CCNet Osuala et al. (2024)	21.82	0.4142	0.3128	293.4	20.79	0.4025	0.2899	336.70	22.64	0.6550	0.2666	174.9
EditAR Mu et al. (2025)	20.64	0.4171	0.3318	288.1	20.12	0.4116	0.3468	292.6	23.76	0.6892	0.3048	166.7
CEKWorld Kong et al. (2026)	21.28	0.5603	0.2623	238.8	21.10	0.6216	0.2847	255.3	24.09	0.7589	0.2611	86.09
Video-Generation-Based Methods												
LVDM He et al. (2022)	20.06	0.5210	0.2879	268.1	19.22	0.4678	0.3031	289.0	23.26	0.7020	0.2714	110.58
AnimateDiff Guo et al. (2024)	21.81	0.5219	0.2820	240.9	19.13	0.5630	0.2907	266.2	23.75	0.7134	0.2793	127.43
DynamiCrafter Xing et al. (2024a)	21.06	0.5414	0.2622	245.7	19.98	0.6035	0.2983	238.9	24.08	0.7412	0.2493	98.29
KINEWORLD	**26.06***	**0.7568***	**0.2245***	**164.5***	**24.61***	**0.7407***	**0.2453***	**178.0***	**26.10***	**0.8440***	**0.2149***	**50.44***

* indicates statistical significance ($p < 0.05$) over the top competing baseline under a two-sided paired Student's t-test on per-subject scores.

3.2 ANATOMICAL CONSISTENCY

Reliable kinetic modeling requires contrast evolution without introducing spurious anatomical variations. We therefore evaluate the agreement between each generated MRI and its ground truth using PSNR and rMSE for intensity correspondence, SSIM for structural similarity, and LPIPS for perceptual similarity. Detailed metric definitions are provided in Appendix A.6. As shown in Table 1, KINEWORLD achieves the best performance across all metrics on all datasets, improves SSIM by over 11%. I2I methods generally rank second, likely because directly conditioning each target phase on the pre-contrast image provides a short pathway for preserving anatomy. In contrast, KINEWORLD initializes the PK-Roll trajectory from the pre-contrast latent state and retains it as an anatomical anchor while autoregressively evolving subsequent states. This design jointly preserves anatomy and models kinetic progression, limiting structural drift throughout the rollout, as further evidenced by the qualitative error maps in Fig. 3.

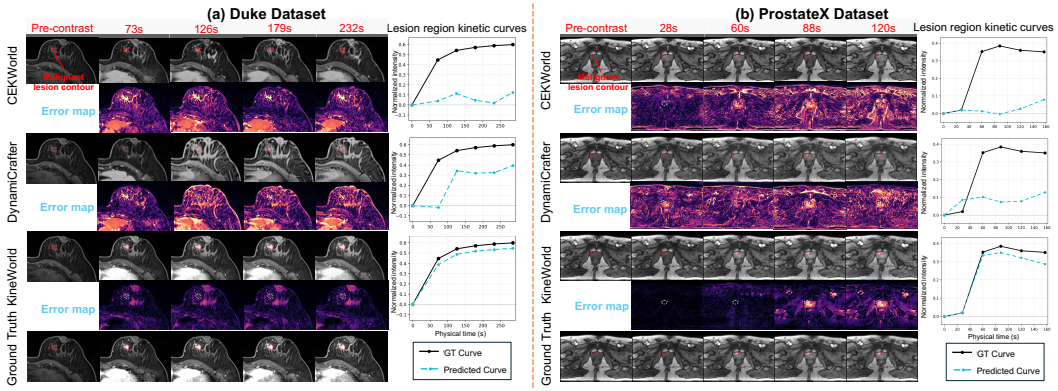

Figure 3: Visualization on **(a)** Duke and **(b)** ProstateX test sets. Generated images and error maps at different time points, and the corresponding lesion-region kinetic curves are presented.

3.3 LESION-LEVEL KINETIC FIDELITY

Metrics Evaluation. We assess kinetic fidelity by comparing generated and ground-truth lesion kinetic curves, *defined as the mean intensity within each lesion over time*. CCC, rMSE, and S-MAE quantify trajectory agreement, intensity error, and temporal-rate error, respectively. Details are provided in Appendix A.7. As shown in Table 2, KINEWORLD substantially outperforms all baselines across datasets and metrics, improving CCC by over 17%. The gains are partly attributable to PK-Roll, which progressively integrates latent kinetic velocities to reflect the cumulative dynamics of contrast influx and efflux. In addition, KineBank and KA-retrieval condition each transition on relevant yet diverse historical states, preserving patient-specific dependencies across the wash in/out phases. Consequently, as presented in Fig. 3, the generated lesion kinetic curves more closely match the ground truth in enhancement magnitude, temporal slope and peak timing.

Table 2: Analysis of **kinetic fidelity** on three datasets. **Bold** and underlined values indicate the best and second-best results, respectively.

Method	Duke Dataset			Internal Dataset			ProstateX Dataset		
	CCC↑	rMSE↓	S-MAE↓	CCC↑	rMSE↓	S-MAE↓	CCC↑	rMSE↓	S-MAE↓
Image-to-Image-Based Methods									
ControlNet Zhang et al. (2023)	0.2368	258.4	198.3	0.2074	287.6	181.2	0.3615	126.4	105.3
T2I Mou et al. (2024)	0.1247	314.6	241.7	0.0983	346.2	223.4	0.2741	157.3	132.8
CCNet Osuala et al. (2024)	0.2813	239.7	183.4	0.2546	265.1	167.8	0.4078	114.7	97.63
EditAR Mu et al. (2025)	0.1762	289.3	218.9	0.1537	318.4	204.7	0.3159	142.8	119.5
CEKWorld Kong et al. (2026)	0.3275	217.8	166.2	0.2968	241.9	151.3	0.4682	101.56	86.71
Video-Generation-Based Methods									
LVDM He et al. (2022)	0.4873	167.5	128.6	0.4481	186.4	117.3	0.6847	69.43	59.82
AnimateDiff Guo et al. (2024)	0.5638	149.3	112.7	0.5174	168.2	101.4	0.7562	58.71	50.36
DynamiCrafter Xing et al. (2024a)	0.6417	126.8	106.5	0.6029	143.6	97.26	0.8014	46.58	41.73
KINEWORLD	**0.8015***	**76.26***	**57.89***	**0.7833***	**80.75***	**49.14***	**0.9412***	**27.21***	**25.44***

* indicates statistical significance ($p < 0.05$) over the top competing baseline under a two-sided paired Student's t-test on per-subject scores.

Reader Study. To assess whether the synthesized MRI trajectories preserve clinically meaningful lesion kinetics, we conduct a blinded reader study on the internal test set. Each lesion is matched to its MRI report by laterality, anatomical location, and lesion index, and its reported kinetic description is mapped to one of six predefined curve categories using a fixed rubric to derive the reference label.[4] Two senior radiologists, each with over 10 years of experience, independently evaluate non-overlapping patient subsets while blinded to the generation method, MRI report, and reference label. Agreement with the reference labels is assessed using exact agreement, Macro-F1, and Cohen's κ, with each metric macro-averaged across readers.

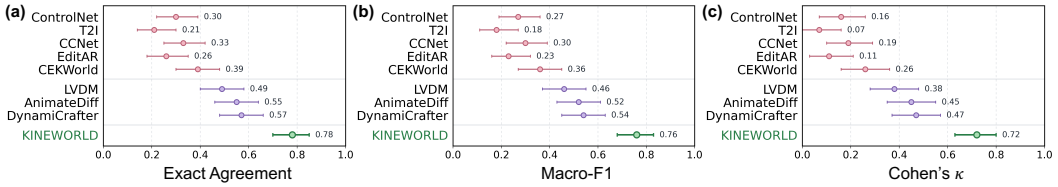

Figure 4: Reader study on the internal test set. Horizontal bars show the **(a)** exact agreement, **(b)** Macro-F1, **(c)** Cohen's κ with report-derived kinetic labels. Higher values indicate better agreement.

3.4 Model Analysis

The role of the efflux term. The efflux term in Eq. 3 is essential for modeling contrast washout. Removing it degrades performance on the Duke and PROSTATEx datasets (Fig. 5**(a,b)**), with larger declines in the late phases (i.e., acquisition phases after phase 2), where washout becomes more prominent; and in CCC, which is more sensitive to kinetic fidelity than SSIM. This phase- and metric-specific degradation supports the intended physiological role of the efflux term. The smaller yet consistent SSIM decline further suggests that the anatomy-conditioned efflux factor $\boldsymbol{k}_e$ retains spatial context. The results are further supported by Fig. 5**(c)**, where $\boldsymbol{k}_e$ map is strongly activated around lesions and perfusion-related vascular regions, suggesting that $\boldsymbol{k}_e$ map captures anatomically grounded, physiologically meaningful washout dynamics.

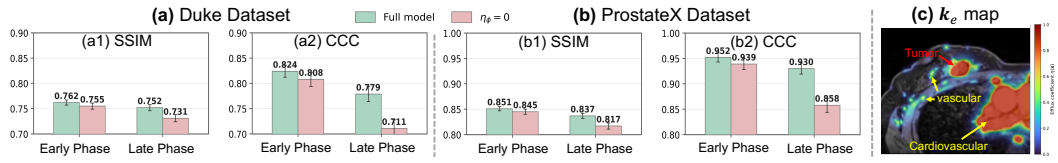

Figure 5: Efflux term analysis on **(a)** Duke and **(b)** ProstateX dataset. **(c)** Visualization of $\boldsymbol{k}_e$.

Effectiveness and Efficiency of KA-Retrieval. Effective long-horizon rollout requires informative historical context without excessive computation. As shown in Fig. 6**(a1–4,b)**, KA-Retrieval

[4]The curve categories, mapping rubric, reference labels, and study interface are detailed in Appendix A.9.

outperforms rule-based samplings, confirming that its gains arise from selecting predictive kinetic states rather than merely enlarging the context. Increasing the MMR λ_m promotes more diverse retrieval by replacing redundant nearby states with heterogeneous kinetic history as presented in Fig. 6**(a5–7,b)**. However, excessive diversity compromises relevance and degrades performance (**a8,b**). We therefore set $\lambda_m = 0.4$, which performs consistently well across metrics and datasets. With λ_m fixed, as shown in Fig. 6**(c)**, performance saturates beyond retrieval top-$K = 6$ while computation continues to increase. We therefore set $K = 6$ for optimal efficiency-accuracy trade-off.

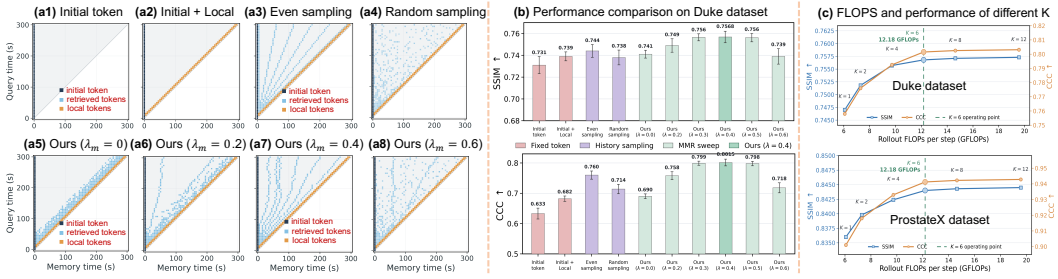

Figure 6: Analysis of KA-Retrieval. Retrieved-token maps of **(a1–a4)** rule-based strategies and **(a5–a8)** ours KA-Retrieval under different $\lambda_m (K = 6)$. Each point is a selected state. **(b)** Performance comparison between retrieval strategies. **(c)** Performance trade-off for different retrieval budgets.

Adaptive Relay for Stable Self-Rollout. Stable long-horizon rollout requires correcting severe deviations without over-relying on ground-truth states. As shown in Fig. 7**(a)**, errors accumulate rapidly without relays, whereas always relaying ($\tau_d = 0$) reduces drift but limits exposure to self-generated histories. Relay-Guard instead resets only when the trajectory exits the trust region, yielding the lowest late-phase errors while preserving predominantly on-policy training. Moreover, the relay trigger ratio decreases steadily after warm-up (Fig. 7**(b)**), showing that the model progressively transitions from corrective guidance to stable autonomous rollout.

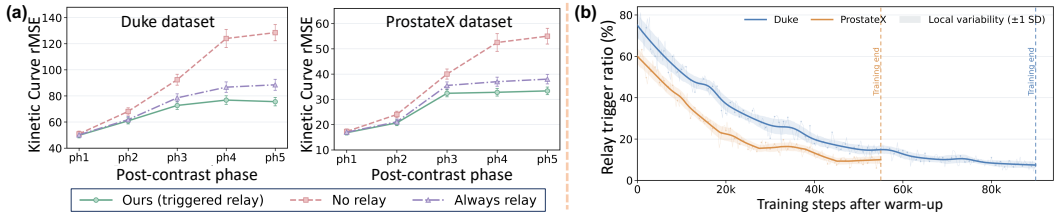

Figure 7: Ablation of relay-guarded training. **(a)** Phase-wise lesion-enhancement rMSE on the Duke and PROSTATEx datasets. **(b)** Relay-trigger ratio after horizon warm-up.

Other analysis. Additional hyperparameter sensitivity analysis is provided in Appendix A.11.

4 Related Work

Appendix A.12 reviews related work on DCE-MRI synthesis, controllable image/video generations, and world models. Unlike prior methods, KineWorldmodels patient-specific contrast evolution as a PK-inspired latent state rollout.

5 Conclusion

We presented KineWorld, a latent kinetic world model for synthesizing patient-specific DCE-MRI trajectories from a pre-contrast scan. By combining pharmacokinetics-inspired state rollout, long-range kinetic-aware retrieval, and relay-guarded training, KineWorld preserves anatomy consistency while faithfully modeling contrast kinetics. Experiments across three datasets demonstrate the superior performance, and clinical potential of KineWorld.

AI USE STATEMENT

In this work, we used generative AI tools to **(i)** edit author-written text for grammar, clarity, style, and readability, **(ii)** assist with writing and refining the implementation of the web-based reader study platform. The study design and evaluation protocol were determined entirely by the authors without AI assistance. We did not use generative AI tools for any other research tasks classified as requiring disclosure under the ICLR 2027 AI Policy. Such tasks were either performed entirely by the authors without AI assistance or were not applicable to this work. In particular, generative AI tools were not used to generate research ideas, scientific claims, methods, experimental data, analyses, citations, or interpretations of results. All AI-assisted textual edits were manually reviewed by the authors. The reader study platform was also tested by the radiologists to verify its functionality and ensure that its implementation conformed to the intended study design. We take responsibility for the final content of this work, including any text, code, or artifacts assisted with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, model architecture, training and inference procedures, and evaluation protocol in the main paper. Additional implementation details, including data preprocessing and augmentation procedures, dataset splits, hyperparameters, computational settings, and reader study design, are provided in the appendix and supplementary materials. To facilitate reproducibility and future research, we will publicly release the source code, pretrained model weights, and the curated internal dataset upon publication, applicable regulations, and regional policies.

ETHICS STATEMENT

This work uses an internally curated dataset and includes a reader study involving human participants. The study protocol was approved by the Ethics Committee of `[Hospital Name]`, and registered under Clinical Trial No. `[XXXX]`. *The identifying information has been anonymized for double-blind review and will be disclosed in the final version upon publication.* All participants were informed of the purpose and procedure of the study, the use of their responses for research, and their right to withdraw, and informed consent was obtained before participation. Participants were compensated in accordance with the `[Hospital Name]` guidelines. We did not collect or release directly identifying information, and the study results are reported only in aggregate. Any participant-derived data released with this work will be de-identified and shared only to the extent permitted by participant consent, applicable regulations, and institutional policies.

The proposed methods and released resources are intended for research purpose solely. We will release only de-identified and anonymized data for which redistribution is permitted and will provide documentation describing its provenance, curation process, intended uses, and limitations. All datasets and reader study in this work may reflect biases arising from the data sources, vendor types, and participant populations. Thus, all users should evaluate their suitability, potential biases, and privacy implications before deployment in real-world or high-stakes settings.

REFERENCES

Samuel G Armato III, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, John S Kirby, Nicholas Petrick, Karen Reddington, Maryellen L Giger, Kenny Cha, Maciej A Mazurowski, et al. Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of Medical Imaging*, 5(4):044501, 2018.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15791–15801, 2025.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George B. van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 2206–2240. PMLR, 2022.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 247–253, 1998.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 2025.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 2026a.

Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. Clarity: Medical world model for guiding treatment decisions by simulating context-aware disease trajectories. In *European Conference on Computer Vision*, pp. 202–221. Springer, 2026b.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. LongLive-RAG: A general retrieval-augmented framework for long video generation. *arXiv preprint arXiv:2606.02553*, 2026.

Michael Ingrisch and Steven Sourbron. Tracer-kinetic modeling of dynamic contrast-enhanced MRI and CT: A primer. *Journal of Pharmacokinetics and Pharmacodynamics*, 40(3):281–300, 2013.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28991–29017, 2025.

Fahmi Khalifa, Ahmed Soliman, Ayman El-Baz, Mohamed Abou El-Ghar, Tarek El-Diasty, Georgy Gimel'farb, Rosemary Ouseph, and Amy C Dwyer. Models and methods for analyzing dce-mri: A review. *Medical Physics*, 2014.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Lingdong Kong, Jiayu Zhu, Jingwei Yi, Weitao Li, Yunyun Chen, Jingchao Shen, Bingchen Hou, Yan Wang, Jie Zhang, Chenglong Zhang, Jie Tian, Xiahai Zhuang, and Guangyao Dai. MRI contrast enhancement kinetics world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Christiane K. Kuhl, Peter Mielcareck, Susanne Klaschik, Claudia Leutner, Eva Wardelmann, Jürgen Gieseke, and Hans H. Schild. Dynamic breast MR imaging: Are signal intensity time course data useful for differential diagnosis of enhancing lesions? *Radiology*, 211(1):101–110, 1999.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941, 2023.

Alex Lamb, Anirudh Goyal Goyal, Ying Zhang, Saizheng Zhang, Aaron C. Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022.

Xin Li, Wei Huang, and James H Holmes. Dynamic contrast-enhanced (dce) mri. *Magnetic Resonance Imaging Clinics*, 2024.

Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *Robotics: Science and Systems*, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

Jiteng Mu, Nuno Vasconcelos, and Xiaolong Wang. Editar: Unified conditional generation with autoregressive models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7899–7909. IEEE, 2025.

James P. B. O'Connor, Paul S. Tofts, Kenneth A. Miles, Laura M. Parkes, Gerard Thompson, and Alan Jackson. Dynamic contrast-enhanced imaging techniques: Ct and mri. *The British Journal of Radiology*, 84(Spec. Issue 2):S112–S120, 2011.

Masashi Okada and Tadahiro Taniguchi. DreamingV2: Reinforcement learning with discrete world models without reconstruction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 985–991, 2022.

Richard Osuala, Daniel M Lang, Preeti Verma, Smriti Joshi, Apostolia Tsirikoglou, Grzegorz Skorupko, Kaisar Kushibar, Lidia Garrucho, Walter HL Pinaya, Oliver Diaz, et al. Towards learning contrast kinetics with multi-condition latent diffusion models. In *International conference on medical image computing and computer-assisted intervention*, pp. 713–723. Springer, 2024.

Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23178–23191, 2022.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

Sadhana Ramanarayanan et al. DCE-Diff: Diffusion model for synthesis of early and late dynamic contrast-enhanced MR images from non-contrast multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4091–4100, 2024.

Harry Robertshaw, Han-Ru Wu, Alejandro Granados, and Thomas C. Booth. World model for AI autonomous navigation in mechanical thrombectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 680–690, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635. PMLR, 2011.

Tankred Saanum, Peter Dayan, and Eric Schulz. Simplifying latent dynamics with softly state-invariant world models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 38355–38382, 2024.

Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Jingxi Weng, Elizabeth H Cain, Connie E Kim, Sujata V Ghate, Ryan Walsh, and Maciej A Mazurowski. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. *The Cancer Imaging Archive*, 10:5, 2021.

Paul S. Tofts. Modeling tracer kinetics in dynamic Gd-DTPA MR imaging. *Journal of Magnetic Resonance Imaging*, 7(1):91–101, 1997.

Paul S Tofts and Geoff JM Parker. Dce-mri: Acquisition and analysis techniques. *Clinical perfusion MRI: Techniques and Applications*, 2013.

Paul S. Tofts, Gunnar Brix, David L. Buckley, Jeffrey L. Evelhoch, Elizabeth Henderson, Michael V. Knopp, Henrik B. W. Larsson, Ting-Yim Lee, Nina A. Mayr, Geoffrey J. M. Parker, Rolf E. Port, Jeff Taylor, and Robert M. Weisskoff. Estimating kinetic parameters from dynamic contrast-enhanced T1-weighted MRI of a diffusable tracer: Standardized quantities and symbols. *Journal of Magnetic Resonance Imaging*, 10(3):223–232, 1999.

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.

Lilian C Wang, Wendy B DeMartini, Savannah C Partridge, Sue Peacock, and Constance D Lehman. Mri-detected suspicious breast lesions: predictive values of kinetic features measured by computer-aided evaluation. *American Journal of Roentgenology*, 193(3):826–831, 2009.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. DynamiCrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024a.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024b.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the baton: Trajectory-relayed on-policy distillation. *arXiv preprint arXiv:2607.26057*, 2026.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8319–8329, 2025.

Ying Yang, Zhengyao Lv, Yujia Zeng, Tianlin Pan, Haofan Wang, Yueming Lyu, Binxin Yang, Hubery Yin, Chen Li, Jing Lyu, et al. Stableworld: Towards stable and consistent long interactive video generation. *arXiv preprint arXiv:2601.15281*, 2026a.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. UniSim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1389–1399, 2023.

Zefan Yang, Ge Wang, James Hendler, Mannudeep K. Kalra, and Pingkun Yan. X-WIN: Building chest radiograph world model via predictive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6920–6930, June 2026b.

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*, pp. 25993–26003, 2025a.

Yang Yue, Yulin Wang, Chenxin Tao, Pan Liu, Shiji Song, and Gao Huang. CheXWorld: Exploring image world modeling for radiograph representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20778–20788, June 2025b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

A Appendix

A.1 Formal Connection Between PK-Roll and the Tofts Model

This section formalizes the structural connection between PK-Roll and the classical Tofts pharmacokinetic model. PK-Roll adopts the influx–efflux state-transition structure of the Tofts model as a pharmacokinetic inductive bias; it does not estimate physical contrast-agent concentrations or pharmacokinetic parameters.

Classical Tofts Model. The Tofts model describes the exchange of a low-molecular-weight contrast agent between blood plasma and the extravascular extracellular space (EES). Taking the pre-contrast concentration $C_t(0)$ as the baseline, the tissue contrast-agent concentration evolves as

$$\frac{\mathrm{d}C_t(t)}{\mathrm{d}t} = \underbrace{K^{\mathrm{trans}} C_p(t)}_{\text{contrast influx}} - \underbrace{k_{\mathrm{ep}} C_t(t)}_{\text{contrast efflux}}, \tag{11}$$

where $C_p(t)$ is the plasma contrast-agent concentration described by the temporal arterial input function (AIF), and $C_t(t)$ is the tissue concentration. The transfer constant K^{trans} governs transport from plasma into the EES, whereas k_{ep} governs contrast-agent return from the EES to plasma. Integrating Eq. 11 gives

$$C_t(t) = C_t(0) + \int_0^t \left[K^{\mathrm{trans}} C_p(\tau) - k_{\mathrm{ep}} C_t(\tau)\right] \mathrm{d}\tau. \tag{12}$$

Thus, the enhancement state accumulates over time according to the balance between incoming contrast and tissue-dependent efflux.

From Concentration Kinetics to Latent Dynamics. To transfer this state-evolution principle to the learned latent space, we define the contrast-dependent latent residual as

$$\hat{\boldsymbol{r}}_i = \hat{\boldsymbol{z}}_i - \boldsymbol{z}_0, \tag{13}$$

where $\boldsymbol{z}_0$ encodes the pre-contrast anatomy and $\hat{\boldsymbol{z}}_i$ is the predicted latent state at time t_i. Only $\hat{\boldsymbol{r}}_i$ evolves during the rollout, while $\boldsymbol{z}_0$ remains a fixed anatomical anchor. This separation prevents the efflux process from attenuating time-invariant anatomical information.

Since the AIF $C_p(t)$ is a temporal function and ground-truth PK parameters are unavailable, PK-Roll replaces the physical influx $K^{\mathrm{trans}} C_p(t)$ with a learned, history-dependent temporal function

$$\boldsymbol{Z}_{\mathrm{in}} = g_\theta(\hat{\boldsymbol{z}}_i, \mathcal{C}_i, \mathrm{PE}(t_i)), \tag{14}$$

where $\mathcal{C}_i$ contains relevant local and long-range kinetic history retrieved from the patient-specific memory bank. The spatially adaptive efflux coefficient $\boldsymbol{k}_e$ is predicted from the anatomical information via

$$\boldsymbol{k}_e = \sigma(h_\phi(\boldsymbol{z}_0)). \tag{15}$$

Following the influx–efflux structure of the Tofts model, the latent kinetic velocity over $[t_i, t_{i+1}]$ is

$$\hat{\boldsymbol{v}}_{i+1} = \underbrace{g_\theta(\hat{\boldsymbol{z}}_i, \mathcal{C}_i, \mathrm{PE}(t_i))}_{\text{learned influx}} - \underbrace{\boldsymbol{k}_e \odot \hat{\boldsymbol{r}}_i}_{\text{anatomy-conditioned efflux}}. \tag{16}$$

The structural correspondence can therefore be summarized as

$$\begin{aligned} C_t(t_i) &\longleftrightarrow \hat{\boldsymbol{r}}_i, \\ k_{\mathrm{ep}} C_t(t_i) &\longleftrightarrow \boldsymbol{k}_e \odot \hat{\boldsymbol{r}}_i, \\ \frac{\mathrm{d}C_t(t_i)}{\mathrm{d}t} &\longleftrightarrow \hat{\boldsymbol{v}}_{i+1}, \\ K^{\mathrm{trans}} C_p(t_i) &\longleftrightarrow g_\theta(\hat{\boldsymbol{z}}_i, \mathcal{C}_i, \mathrm{PE}(t_i)). \end{aligned} \tag{17}$$

Discrete Latent Rollout. Analogous to the temporal integration in Eq. 12, PK-Roll updates the latent enhancement residual as

$$\hat{\boldsymbol{r}}_{i+1} = \hat{\boldsymbol{r}}_i + \Delta t_i \cdot \hat{\boldsymbol{v}}_{i+1}, \qquad \Delta t_i = t_{i+1} - t_i, \tag{18}$$

and reconstructs the complete latent state by

$$\hat{\boldsymbol{z}}_{i+1} = \boldsymbol{z}_0 + \hat{\boldsymbol{r}}_{i+1}. \tag{19}$$

Equivalently, the state at t_n is

$$\hat{\boldsymbol{z}}_n = \boldsymbol{z}_0 + \sum_{i=0}^{n-1} \Delta t_i \cdot \hat{\boldsymbol{v}}_{i+1}. \tag{20}$$

This formulation preserves anatomical consistency by retaining $\boldsymbol{z}_0$ as a fixed baseline and accumulating only contrast-dependent residuals. Meanwhile, the influx–efflux transition, physical-time integration, and history-conditioned influx support kinetic fidelity by modeling the direction, magnitude, and temporal dependence of enhancement evolution.

Scope of the Connection. The correspondence in Eq. 17 is structural rather than numerical. The learned latent states have no physical concentration units, and the outputs of g_θ and $\boldsymbol{k}_e$ should not be interpreted as estimates of the AIF, K^{trans}, or $\boldsymbol{k}_{\text{ep}}$. Instead, the Tofts model provides three inductive principles for PK-Roll: **(1)** enhancement evolves through an instantaneous kinetic velocity; **(2)** this velocity is governed by the balance between patient-specific influx and anatomy-conditioned efflux; and **(3)** the complete enhancement trajectory is obtained by temporally integrating these velocity increments. PK-Roll generalizes these principles into spatially adaptive, history-aware latent transitions suitable for patient-specific DCE-MRI synthesis.

A.2 Influx and Efflux Network Design

Motivation. Building on the influx–efflux decomposition established in Eq. 11, PK-Roll instantiates the two kinetic processes using complementary network components. The influx term represents the instantaneous contrast-driving effect on the current enhancement state. We model this effective influx directly in the latent space by jointly conditioning on the current state, patient-specific kinetic history, and physical time. This multi-source and history-aware dependency motivates a Transformer-based transition network that can adaptively aggregate temporally distributed contextual information.

In contrast, the efflux term regulates the kinetic evolution in proportion to the current tissue state, with its strength governed by tissue-dependent properties. Because these properties can vary spatially across lesions, surrounding tissues, and anatomical regions, we employ an anatomy-conditioned ASPP network to extract multi-scale spatial context from the pre-contrast representation. Its output modulates the current latent state to produce a spatially adaptive and state-dependent efflux. Together, the Transformer-based influx branch and the ASPP-based efflux branch form the PK-inspired transition operator illustrated in Fig. 8.

Transformer-Based Influx Network. At rollout step t_i, the influx term function is formulated as $g_\theta(\hat{\boldsymbol{z}}_i, \mathcal{C}_i, \text{PE}(t_i))$, where $\hat{\boldsymbol{z}}_i$ is the current latent enhancement state, $\mathcal{C}_i$ is the historical context retrieved from the patient-specific kinetics memory bank, and $\text{PE}(t_i)$ encodes the physical acquisition time. The influx term predicts each transition from the current state and its preceding patient-specific trajectory. As illustrated in Fig. 8 **(a)**, this history-dependent formulation naturally motivates a Transformer architecture Vaswani et al. (2017), whose attention mechanism can dynamically aggregate relevant states across the preceding kinetic context.

The current state and the retrieved historical states are represented as spatial patch tokens and projected into a common 256-dimensional embedding space. Each token is augmented with a temporal embedding. The resulting token sequence is processed by three stacked Transformer layers. Through self-attention, the network adaptively relates the current state to the most relevant historical states in $\mathcal{C}_i$. The output tokens associated with the current state are subsequently reshaped to construct the transition output of g_θ. Consequently, the predicted transition is conditioned not only on the present enhancement state but also on the patient-specific trajectory leading to it.

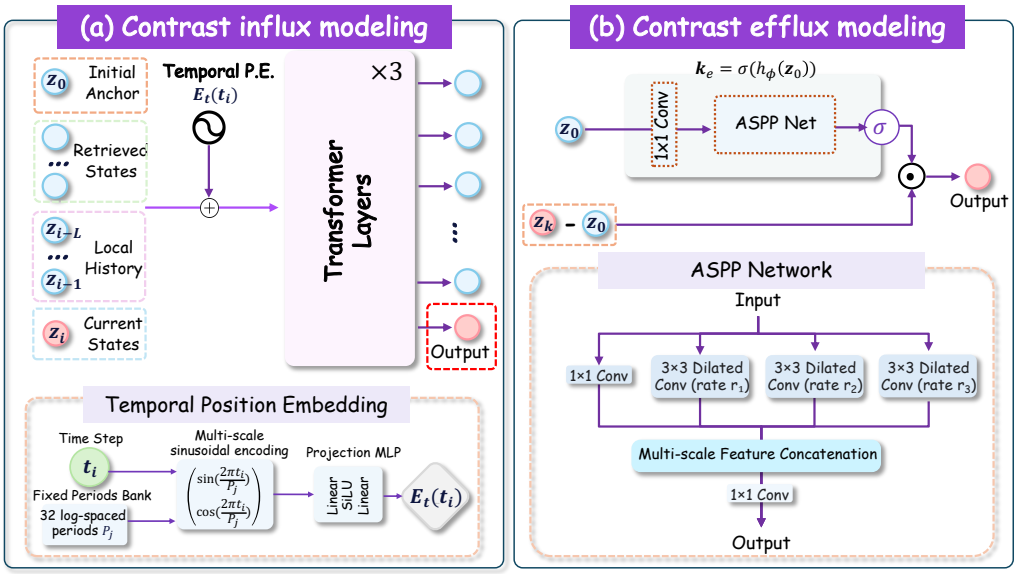

Figure 8: Detailed network architectures of the influx and efflux modules. **(a)** is the Transformer-based influx modeling network. **(b)** is the ASPP-based efflux modeling network.

To represent acquisition time continuously, rather than using a discrete phase index, we construct a multi-scale Fourier encoding of the physical timestamp. Specifically, we define $N_f = 32$ fixed, logarithmically spaced periods $P_j \in [2, 1024]$ seconds and compute

$$\boldsymbol{\varphi}(t_i) = \left[\sin\left(\frac{2\pi t_i}{P_j}\right), \cos\left(\frac{2\pi t_i}{P_j}\right)\right]_{j=1}^{N_f} \in \mathbb{R}^{64}. \tag{21}$$

The Fourier features are then projected into the Transformer embedding space using a two-layer MLP

$$\mathrm{PE}(t_i) = \boldsymbol{W}_2\ \mathrm{SiLU}(\boldsymbol{W}_1\boldsymbol{\varphi}(t_i))\,, \tag{22}$$

where $\boldsymbol{W}_1 : \mathbb{R}^{64} \to \mathbb{R}^{256}$ and $\boldsymbol{W}_2 : \mathbb{R}^{256} \to \mathbb{R}^{256}$ are learnable linear projections, and SiLU denotes the activation function Elfwing et al. (2018). The Fourier basis remains fixed, whereas the projection layers are optimized jointly with the network. This continuous physical-time representation enables the Transformer to distinguish states acquired at different timestamps and to infer different transitions for otherwise similar current states according to their preceding kinetic trajectories.

Anatomy-Conditioned Efflux Network. As illustrated in Fig. 8**(b)**, the efflux modulation should account for spatially heterogeneous tissue responses across lesions, surrounding tissues, and anatomical structures. Estimating such a modulation requires both local tissue information and broader anatomical context. We therefore adopt a lightweight atrous spatial pyramid pooling (ASPP) network Chen et al. (2017), which aggregates multi-scale spatial context without reducing the resolution of the latent representation.

Given the pre-contrast anatomical representation $\boldsymbol{z}_0$, we first apply a 1×1 convolutional projection $\boldsymbol{u}_0 = \mathrm{Conv}_{1\times1}^{\mathrm{pre}}(\boldsymbol{z}_0)$. This projection aligns the input channels and produces a compact feature representation for subsequent multi-scale processing.

The projected features $\boldsymbol{u}_0$ are then passed through four parallel ASPP branches: one 1×1 convolutional branch and three 3×3 atrous convolutional branches with dilation rates $(r_1, r_2, r_3) = (1, 2, 4)$ as $\boldsymbol{b}_0 = \mathrm{Conv}_{1\times1}^{\mathrm{branch}}(\boldsymbol{u}_0)$, $\boldsymbol{b}_1 = \mathrm{Conv}_{3\times3}^{r=1}(\boldsymbol{u}_0)$, $\boldsymbol{b}_2 = \mathrm{Conv}_{3\times3}^{r=2}(\boldsymbol{u}_0)$, $\boldsymbol{b}_3 = \mathrm{Conv}_{3\times3}^{r=4}(\boldsymbol{u}_0)$. Here, the branch with $r = 1$ captures local anatomical patterns, whereas the branches with $r = 2$ and $r = 4$ progressively enlarge the receptive field to incorporate surrounding tissue and broader anatomical context.

The branch features are concatenated along the channel dimension and fused through a final 1×1 convolution:

$$h_\phi(\boldsymbol{z}_0) = \mathrm{Conv}_{1\times1}^{\mathrm{fuse}}\left(\mathrm{Concat}\left[\boldsymbol{b}_0,\, \boldsymbol{b}_1,\, \boldsymbol{b}_2,\, \boldsymbol{b}_3\right]\right). \tag{23}$$

Finally, a pointwise sigmoid maps the fused features to the anatomy-conditioned modulation field as

$$\boldsymbol{k}_e = \sigma(h_\phi(\boldsymbol{z}_0)), \qquad \boldsymbol{k}_e \in (0,1). \tag{24}$$

The sigmoid ensures that the learned modulation is positive and bounded. Because $\boldsymbol{k}_e$ is derived from the fixed pre-contrast representation, it provides a consistent patient-specific anatomical modulation field throughout the rollout.

A.3 Pseudo Code of Kinetics-aware Retrieval Algorithm

Algorithmic description. Algorithm 1 summarizes the kinetics-aware context construction at rollout step i. The initial state $\boldsymbol{z}_0$ is always retained as an anatomical anchor, while the most recent L generated states provide local temporal context. From the remaining long-range history, MMR iteratively selects up to K states by balancing their relevance to the current kinetic query against their redundancy with previously selected states. The resulting long-range memory is concatenated with the initial anchor and local history to form $\mathcal{C}_i$ for predicting the next state.

Algorithm 1 Kinetics-aware retrieval for context construction

Require: Rollout step i; initial latent state $\boldsymbol{z}_0$; memory bank $\mathcal{B}$; generated states $\{\hat{\boldsymbol{z}}_j\}_{j\le i}$; kinetic embeddings $\{\hat{\boldsymbol{v}}_j\}_{j\le i}$; local-window length L; retrieval budget K; diversity weight λ_m
Ensure: Historical context $\mathcal{C}_i$

Construct the query $\boldsymbol{q}_i \leftarrow [\hat{\boldsymbol{z}}_i; \hat{\boldsymbol{v}}_i]$
Define the memory keys $\boldsymbol{k}_j \leftarrow [\hat{\boldsymbol{z}}_j; \hat{\boldsymbol{v}}_j]$
Define the long-range candidate set $\mathcal{P}_i \leftarrow \{j \mid 1 \le j \le i - L - 1\}$
if $|\mathcal{P}_i| \le K$ **then**
    $\mathcal{S} \leftarrow \mathcal{P}_i$ ▷ Retain all available candidates if $K >$ len(KineBank)
else
    $\mathcal{S} \leftarrow \varnothing$
    while $|\mathcal{S}| < K$ **do**
        for all $j \in \mathcal{P}_i \setminus \mathcal{S}$ **do**
            if $\mathcal{S} = \varnothing$ **then**
                $r_j \leftarrow 0$
            else
                $r_j \leftarrow \max_{\ell\in\mathcal{S}} \cos(\boldsymbol{k}_j, \boldsymbol{k}_\ell)$
            end if
            $s_j \leftarrow \cos(\boldsymbol{q}_i, \boldsymbol{k}_j) - \lambda_m r_j$
        end for
        $j^\star \leftarrow \arg\max_{j\in\mathcal{P}_i\setminus\mathcal{S}} s_j$
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{j^\star\}$
    end while
end if
Sort $\mathcal{S}$ in chronological order
Retrieve the long-range history $\mathcal{R}_i \leftarrow [\hat{\boldsymbol{z}}_j]_{j\in\mathcal{S}}$
Collect the local history $\mathcal{H}_i \leftarrow [\hat{\boldsymbol{z}}_{\max(1,i-L)}, \ldots, \hat{\boldsymbol{z}}_{i-1}]$
Assemble the context $\mathcal{C}_i \leftarrow [\boldsymbol{z}_0; \mathcal{R}_i; \mathcal{H}_i]$ ▷ $\mathcal{C}_i$ does not include $\boldsymbol{z}_i$
return $\mathcal{C}_i$

A.4 Data Preprocessing and Data Augmentation

Data Preprocessing Following Kong et al. (2026), all post-contrast phases were anatomically registered to the corresponding pre-contrast images, and each slice was resized to 256×256 pixels. Segmentation masks underwent the same spatial transformations. For each patient, pixel intensities were clipped to $[0, P_{99}]$, where P_{99} denotes the 99th percentile, and then linearly scaled to $[-1, 1]$.

Data Augmentation During training, we employed lesion-aware slice sampling and tumor-centered cropping to capture kinetic heterogeneity across lesion regions. Tumor-positive slices accounted for 50% of the sampled slices, half of which were cropped around the lesion. The crop side length was randomly set to 3–5 times the lesion extent, with center jitter of up to $\pm 10\%$ of the crop size. The same crop was applied consistently across all DCE phases and resized to 256×256 pixels to preserve temporal alignment.

To ensure a fair comparison, all baseline models and our approach are trained using the exact same data preprocessing and augmentation pipelines described above.

A.5 Baseline Adaptation and Implementation Details

All baselines are initialized from their official pretrained weights and trained on the same preprocessed DCE-MRI data using identical training, validation, and test splits. For the image-to-image baselines, each post-contrast phase is generated independently from the pre-contrast image. Following Kong et al. (2026), the elapsed acquisition time relative to the pre-contrast phase is provided as a textual condition in the `HH:MM:SS` format. To minimize differences attributable to the generative backbone, all image-to-image methods are adapted using the same frozen SD 1.5 backbone Rombach et al. (2022).

For the video-generation baselines, the pre-contrast image is used as the initial conditioning frame, and the subsequent post-contrast phases are generated as a sequence. We preserve each method's original architecture and temporal position encoding while replacing its input formulation with pre-contrast conditioning. Text prompts are fixed to empty strings to exclude additional semantic information.

For a fair comparison, all baselines are optimized under comparable training schedules and hyperparameter tuning protocols, and with the same tuning efforts.

A.6 Evaluation Metrics for Anatomical Consistency

We evaluate anatomical consistency by comparing each generated post-contrast image $\hat{\boldsymbol{I}}_t$ with its corresponding ground-truth image $\boldsymbol{I}_t$. Let Ω denote the image domain, $N = |\Omega|$ the number of pixels, and L the intensity dynamic range used for evaluation. The mean squared error at phase i is

$$\mathrm{MSE}_i = \frac{1}{N} \sum_{\boldsymbol{x} \in \Omega} \left(\hat{\boldsymbol{I}}_i(\boldsymbol{x}) - \boldsymbol{I}_i(\boldsymbol{x}) \right)^2 . \tag{25}$$

We derive the peak signal-to-noise ratio and root mean squared error as

$$\mathrm{PSNR}_i = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}_i} \right), \qquad \mathrm{rMSE}_i = \sqrt{\mathrm{MSE}_i}. \tag{26}$$

PSNR measures overall signal reconstruction quality, whereas rMSE directly quantifies pixel-level intensity deviation.

Structural similarity (SSIM) compares local luminance, contrast, and structural information between generated and reference images Wang et al. (2004). Learned perceptual image patch similarity (LPIPS) measures their distance in deep feature space and therefore captures perceptual discrepancies that may not be reflected by pixel-wise errors Zhang et al. (2018). Higher PSNR and SSIM indicate better anatomical preservation, while lower rMSE and LPIPS indicate smaller intensity and perceptual errors. All metrics are computed phase-wise against the corresponding ground truth and averaged over the evaluated post-contrast images.

A.7 Evaluation Metrics for Time–Intensity Curves

We evaluated the agreement between the predicted and reference regional intensity curves within three-dimensional tumor regions. Before evaluation, both the predicted and reference images were represented in the original image intensity space. Therefore, none of the metrics described below were computed in the normalized $[-1, 1]$ image domain.

Let R denote a three-dimensional tumor region, $p \in R$ denote a voxel within that region, and $k \in \{0, \ldots, K\}$ denote the dynamic imaging phase, where $k = 0$ corresponds to the pre-contrast DCE image. The reference and predicted regional mean intensities at phase k are defined as

$$s_k^* = \frac{1}{|R|} \sum_{p \in R} I_k^*(p), \qquad \hat{s}_k = \frac{1}{|R|} \sum_{p \in R} \hat{I}_k(p), \tag{27}$$

where $I_k^*(p)$ and $\hat{I}_k(p)$ denote the reference and predicted voxel intensities in the original image space, respectively, and $|R|$ is the number of voxels in R. The averaging is performed directly over all voxels in the three-dimensional tumor region, rather than by first computing separate averages for individual slices.

To describe contrast enhancement relative to the pre-contrast image, the reference and predicted baseline-relative intensity curves are defined as

$$c_k^* = s_k^* - s_0^*, \qquad \hat{c}_k = \hat{s}_k - s_0^*. \tag{28}$$

The predicted curve is anchored to the reference pre-contrast intensity such that

$$\hat{c}_0 = c_0^* = 0. \tag{29}$$

Thus, c_k^* and $\hat{c}_k$ represent changes in the regional mean intensity relative to the same reference pre-contrast baseline. Let t_k denote the physical acquisition time of phase k, expressed in a consistent time unit. The actual acquisition times are used in all time-dependent calculations, thereby supporting nonuniform intervals between successive DCE phases.

Concordance Correlation Coefficient (CCC). The Concordance Correlation Coefficient (CCC) measures the overall agreement between the predicted and reference intensity curves. It accounts for both their correlation and differences in their means and variances, and is defined as

$$M_{\mathrm{CCC}} = \frac{2\sigma_{\hat{c},c^*}}{\sigma_{\hat{c}}^2 + \sigma_{c^*}^2 + \left(\mu_{\hat{c}} - \mu_{c^*}\right)^2}, \tag{30}$$

where

$$\mu_{\hat{c}} = \frac{1}{K} \sum_{k=1}^{K} \hat{c}_k, \qquad \mu_{c^*} = \frac{1}{K} \sum_{k=1}^{K} c_k^* \tag{31}$$

are the mean predicted and reference intensity changes over the post-contrast phases,

$$\sigma_{\hat{c}}^2 = \frac{1}{K} \sum_{k=1}^{K} \left(\hat{c}_k - \mu_{\hat{c}}\right)^2, \qquad \sigma_{c^*}^2 = \frac{1}{K} \sum_{k=1}^{K} \left(c_k^* - \mu_{c^*}\right)^2 \tag{32}$$

are their variances, and

$$\sigma_{\hat{c},c^*} = \frac{1}{K} \sum_{k=1}^{K} \left(\hat{c}_k - \mu_{\hat{c}}\right)\left(c_k^* - \mu_{c^*}\right) \tag{33}$$

is their covariance. The CCC ranges from -1 to 1, with a value of 1 indicating perfect agreement. The pre-contrast point is excluded because it is identically fixed to zero for both curves and would otherwise artificially increase the measured agreement.

Kinetic Curve Root Mean Squared Error (rMSE). The kinetic curve Root Mean Squared Error (rMSE) measures the point-wise discrepancy between the predicted and reference lesion-specific regional intensity changes by

$$M_{\mathrm{rMSE}} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\hat{c}_k - c_k^*\right)^2}. \tag{34}$$

A lower rMSE indicates closer agreement between the two curves. Because the phase-wise errors are squared before averaging, the kinetic curve RMSE places greater emphasis on phases with large intensity errors. The pre-contrast point is excluded because $\hat{c}_0 = c_0^* = 0$ by construction, also as proved empirically in Appendix A.10. The metric is expressed in the same intensity unit as the original images.

Slope Mean Absolute Error (S-MAE). The Slope Mean Absolute Error (S-MAE) evaluates whether the predicted curve reproduces the temporal rate of change of the reference regional intensity curve. The reference and predicted slopes between consecutive phases are defined as

$$m_k^* = \frac{c_{k+1}^* - c_k^*}{t_{k+1} - t_k}, \qquad \hat{m}_k = \frac{\hat{c}_{k+1} - \hat{c}_k}{t_{k+1} - t_k}, \qquad k = 0, \ldots, K-1. \tag{35}$$

S-MAE is then calculated as

$$M_{\text{S-MAE}} = \frac{1}{K} \sum_{k=0}^{K-1} \left| \hat{m}_k - m_k^* \right|. \tag{36}$$

A lower S-MAE indicates better agreement in the temporal evolution of the two curves. This metric is sensitive to differences in the rates of initial enhancement, subsequent plateau formation, and washout. Its unit is the original image-intensity unit divided by the time unit used for t_k. Since the absolute error is calculated separately for each temporal interval, positive and negative slope errors do not cancel.

A.8 Multi-level Evaluations on Anatomy Consistency

Additional Whole-Image Visualizations. Figure 9 provides additional whole-image comparisons with all baselines on the internal and PROSTATEx test sets. Image-to-image methods generally preserve the input anatomy but produce inconsistent enhancement across independently generated phases, whereas video-based methods improve inter-phase continuity but may exhibit structural drift or inaccurate lesion kinetics. In contrast, KineWorld maintains patient-specific anatomy throughout the sequence while generating lesion-specific intensity curves that more closely follow the reference uptake and delayed-phase trends.

Cross-Cohort Validation (Duke and Internal). To evaluate robustness to cohort and acquisition-protocol shifts, we train KineWorld on one breast DCE-MRI cohort and directly evaluate it on the test set of the other, without fine-tuning or target-domain adaptation. As shown in Table 3 and Table 4, performance decreases under both cross-cohort settings, reflecting differences in patient composition, scanner characteristics, and temporal sampling. Nevertheless, the model trained on the larger and more heterogeneous internal cohort transfers more effectively to Duke than the reverse direction. This asymmetry is expected because the internal cohort contains both benign and malignant lesions with more diverse types and broader temporal coverage, whereas Duke contains only malignant cases and fewer post-contrast phases. The relatively moderate cross-cohort degradation, particularly for anatomical consistency, demonstrates that KineWorld learns transferable anatomy and enhancement dynamics rather than cohort-specific mappings.

Table 3: Cross-cohort validation of **anatomical consistency** between the Duke and internal breast DCE-MRI cohorts. No target-domain fine-tuning is performed.

Train → Test	PSNR↑	SSIM↑	LPIPS↓	rMSE↓
Duke → Duke	**26.06**	**0.7568**	**0.2245**	**164.5**
Internal → Duke	24.42	0.7136	0.2527	191.8
Internal → Internal	**24.61**	**0.7407**	**0.2453**	**178.0**
Duke → Internal	21.93	0.6564	0.3018	239.7

Table 4: Cross-cohort validation of **kinetic fidelity** between the Duke and internal breast DCE-MRI cohorts. No target-domain fine-tuning is performed.

Train → Test	CCC↑	rMSE↓	S-MAE↓
Duke → Duke	**0.8015**	**76.26**	**57.89**
Internal → Duke	0.7598	95.37	66.84
Internal → Internal	**0.7833**	**80.75**	**49.14**
Duke → Internal	0.6526	117.92	78.36

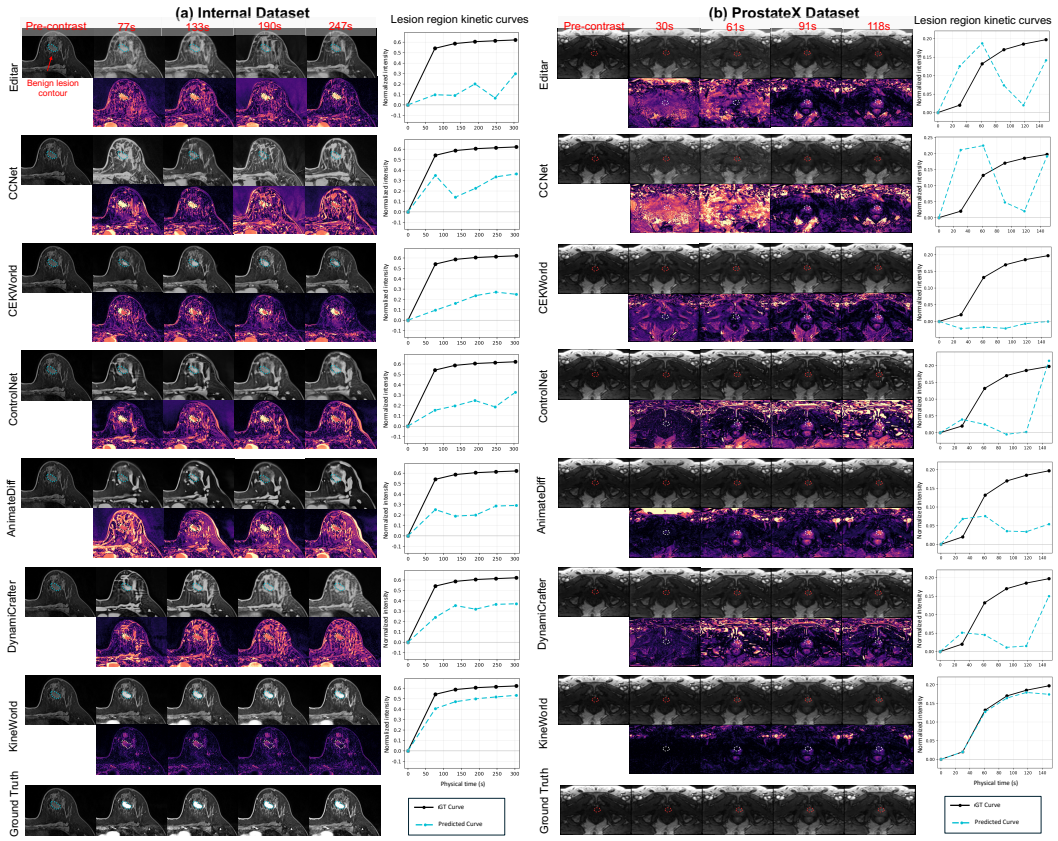

Figure 9: Additional whole-image comparisons on the **(a)** internal and **(b)** PROSTATEx test sets. Generated post-contrast phases and the corresponding lesion-specific intensity curves are shown for all methods. Compared with the baselines, KINEWORLD better preserves anatomical consistency throughout the sequence while reproducing reference faithful lesion-enhancement trajectories.

Lesion-level anatomical consistency To assess more fine-grained lesion-level anatomical fidelity, we additionally compute PSNR, SSIM, LPIPS, and rMSE within lesion-centered regions of interest. The results are presented in Table 5. For each lesion, the same ROI is extracted from the generated and ground-truth images at every post-contrast phase. The metrics are first averaged across phases and lesions for each subject and subsequently averaged across subjects. For the Duke and Internal datasets, regions of interest (ROIs) are derived directly from radiologist-annotated lesion masks. Since the ProstateX dataset provides only lesion-center coordinates, we define its ROIs as circular regions with a 6 mm radius centered at the provided coordinates.

Pathology stratified analysis on kinetics fidelity We further stratify the internal test lesions by pathology label. As shown in Table 6, all methods perform better on benign lesions, whose enhancement trajectories are generally smoother and more persistent, than on malignant lesions with more rapid and heterogeneous uptake and washout patterns. Nevertheless, KINEWORLD consistently achieves the highest CCC and the lowest rMSE and S-MAE in both subgroups, demonstrating robust kinetic modeling across distinct lesion phenotypes.

A.9 READER STUDY

Brief introduction on the kinetic Curve subtypes. We define six lesion-specific kinetic subtypes according to two complementary characteristics of the time–intensity curve (TIC), namely the initial contrast uptake rate and the delayed enhancement pattern. As illustrated in Fig. 10, Types I–III exhibit slow-to-moderate initial uptake followed by persistent or plateau enhancement, whereas Types IV–VI exhibit rapid uptake followed by persistent enhancement, a plateau, or washout, respectively. This fine-grained taxonomy extends the conventional persistent–plateau–washout classification by

Table 5: Analysis of **lesion-region anatomical consistency** on three datasets. Metrics are computed within lesion-centered regions of interest. **Bold** and underlined values indicate the best and second-best results, respectively.

Method	Duke Dataset				Internal Dataset				ProstateX Dataset			
	PSNR↑	SSIM↑	LPIPS↓	rMSE↓	PSNR↑	SSIM↑	LPIPS↓	rMSE↓	PSNR↑	SSIM↑	LPIPS↓	rMSE↓
Image-to-Image-Based Methods												
ControlNet Zhang et al. (2023)	17.05	0.4218	0.3224	388.6	16.47	0.3985	0.3431	402.8	19.38	0.6125	0.3129	298.7
T2I Mou et al. (2024)	14.62	0.3387	0.3715	417.5	14.88	0.3301	0.3894	435.3	16.74	0.5716	0.3417	284.1
CCNet Osuala et al. (2024)	18.74	0.3579	0.3862	365.8	18.21	0.3257	0.3716	411.2	20.06	0.5448	0.3325	225.9
EditAR Mu et al. (2025)	17.92	0.3741	0.4027	349.6	17.70	0.3469	0.4192	390.4	21.15	0.5966	0.3542	212.3
CEKWorld Kong et al. (2026)	18.51	0.4877	0.3183	306.4	18.11	0.5362	0.3297	326.8	21.62	0.6817	0.3055	114.8
Video-Generation-Based Methods												
LVDM He et al. (2022)	17.66	0.4385	0.3561	337.8	16.82	0.4144	0.3675	371.3	20.74	0.6258	0.3198	152.4
AnimateDiff Guo et al. (2024)	18.42	0.4517	0.3428	314.2	17.51	0.4827	0.3508	340.6	21.08	0.6469	0.3272	167.1
DynamiCrafter Xing et al. (2024a)	18.69	0.4723	0.3106	309.7	17.94	0.5171	0.3214	310.5	21.55	0.6702	0.2918	121.7
KINEWORLD	**23.12***	**0.6845***	**0.2471***	**212.6***	**21.64***	**0.6628***	**0.2762***	**231.5***	**24.27***	**0.7896***	**0.2387***	**67.84***

* indicates statistical significance ($p < 0.05$) over the strongest competing baseline under a two-sided paired Student's t-test on per-subject scores.

Table 6: Pathology-stratified analysis of **lesion-level kinetic fidelity** on the internal dataset. Benign and malignant lesions account for 37% and 63% of the eligible annotated lesions, respectively. **Bold** and underlined values indicate the best and second-best results within each subgroup.

Method	Benign Lesions (37%)			Malignant Lesions (63%)		
	CCC↑	rMSE↓	S-MAE↓	CCC↑	rMSE↓	S-MAE↓
Image-to-Image-Based Methods						
ControlNet Zhang et al. (2023)	0.2675	232.4	140.2	0.1721	320.7	205.3
T2I Mou et al. (2024)	0.1398	286.7	175.9	0.0739	381.2	251.4
CCNet Osuala et al. (2024)	0.3192	214.9	124.7	0.2167	294.6	193.1
EditAR Mu et al. (2025)	0.2026	263.8	160.5	0.1250	350.5	230.6
CEKWorld Kong et al. (2026)	0.4250	191.6	110.8	0.2215	271.4	175.2
Video-Generation-Based Methods						
LVDM He et al. (2022)	0.5850	138.1	82.0	0.3677	214.9	138.2
AnimateDiff Guo et al. (2024)	0.7060	107.5	73.50	0.4066	203.99	117.9
DynamiCrafter Xing et al. (2024a)	0.6905	116.8	76.04	0.5517	159.8	109.7
KINEWORLD	**0.8465***	**62.90***	**35.40***	**0.7462***	**91.23***	**57.21***

* indicates statistical significance ($p < 0.05$) over the strongest competing baseline under a two-sided paired Student's t-test on per-subject scores.

additionally distinguishing the initial wash-in behavior, enabling a more sensitive assessment of whether generated DCE-MRI series preserve clinically relevant lesion kinetics.

Persistent enhancement is generally associated with lower suspicion, whereas plateau and, particularly, washout patterns are increasingly associated with malignancy. Rapid initial uptake further increases suspicion by indicating strong early contrast inflow Kuhl et al. (1999); Wang et al. (2009). The resulting clinical tendencies are summarized in Table 7. These categories represent relative kinetic suspicion rather than definitive diagnoses and should be interpreted together with lesion morphology and other clinical findings.

Report-to-label Protocol. For each target lesion, the corresponding report entry is identified using laterality, anatomical location, and lesion index. Two radiologists, with 4 and 5 years of relevant imaging experience and no involvement in the reader study, independently extract the lesion-specific kinetic descriptor and assign it to one of six prespecified curve categories, following the descriptions in Table 7. The annotators are blinded to all generated images, method identities, and reader-study assessments. All disagreements are resolved by a senior radiologist with 10 years of experience. The resulting annotations serve as report-derived reference labels.

Reader study protocol interface. Figure 11 illustrates the interface of our reader study protocol. For each case, the interface presents all acquired DCE phases from a fixed lesion-containing slice, together with the overlaid ROI contours and corresponding lesion-specific intensity curves. Without access to the generation method or report-derived label, the reader assigns each ROI to the best-matching subtype among the six predefined kinetic patterns using a drop-down menu.

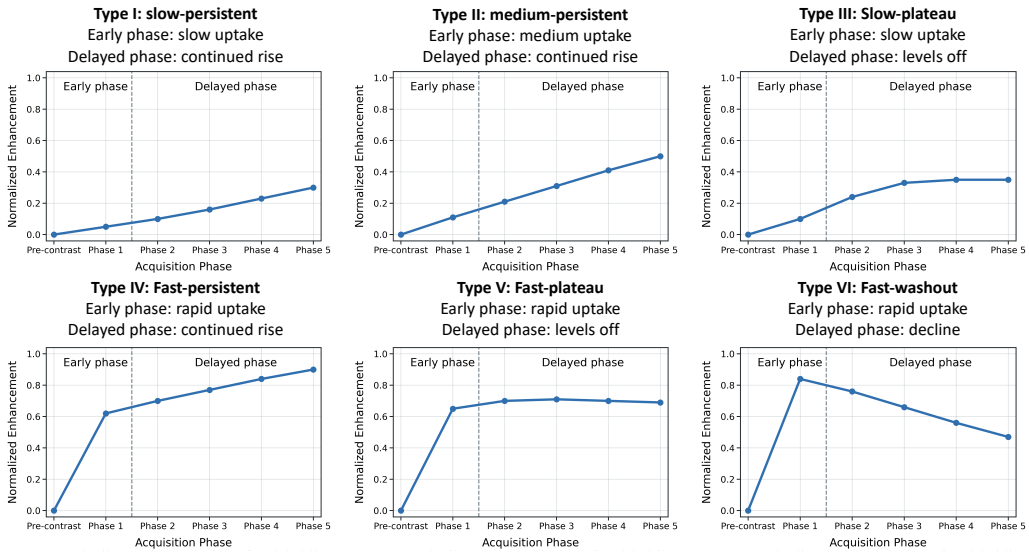

Figure 10: *Simulated* demo of the six lesion-specific TIC subtypes used in our study. Each subtype is characterized jointly by its initial contrast uptake rate and delayed enhancement pattern. The vertical dashed line separates the early and delayed phases. All curves are normalized and illustrate kinetic shape rather than absolute signal intensity.

Table 7: Definitions and relative clinical tendencies of the six predefined lesion-specific TIC subtypes. Suspicion levels indicate kinetic associations rather than standalone diagnoses.

Subtype	**Kinetic description**	**Relative malignancy**
Type I	Slow initial uptake followed by a continued delayed-phase rise	Strongly benign
Type II	Moderate initial uptake followed by a continued delayed-phase rise	Usually benign
Type III	Slow initial uptake followed by delayed-phase stabilization	Benign-leaning
Type IV	Rapid initial uptake followed by a continued delayed-phase rise	Borderline
Type V	Rapid initial uptake followed by a delayed-phase plateau	Usually malignant
Type VI	Rapid initial uptake followed by delayed-phase washout	Strongly malignant

A.10 Effectiveness of the Frozen VAE for DCE-MRI.

We validate the frozen SD1.5 VAE through direct MRI reconstruction. Evaluated on 300 cases, with 100 cases sampled from each of the Duke, Internal and PROSTATEx datasets, the VAE achieves an average SSIM of 0.9960 and an average rMSE of 4.75. At the lesion level, the lesion-specific kinetic curves achieve an average CCC of 0.9911 and an average rMSE of 10.11. These results confirm that the frozen VAE retains both anatomical and kinetic information in its latent representation. We therefore use it directly for latent encoding and image decoding without introducing a modality-specific adapter.

A.11 Hyperparameter Sensitivity Analysis.

We evaluate sensitivity to the local-window size L, Relay-Guard threshold τ_d, and horizon warm-up duration E_{warm} by varying one parameter at a time on the held-out validation split of each dataset, while keeping all other settings fixed. As shown in Fig. 12, both SSIM and CCC remain stable around the selected values $L = 2$, $\tau_d = 0.10$, and $E_{\text{warm}} \geq 1$ epoch. Increasing L from 1 to 2 improves performance by providing sufficient recent context, whereas larger windows offer little additional benefit due to temporal redundancy. For Relay-Guard, a moderate threshold performs consistently well, while $\tau_d = 0$ causes excessive resets and an overly large threshold permits trajectory

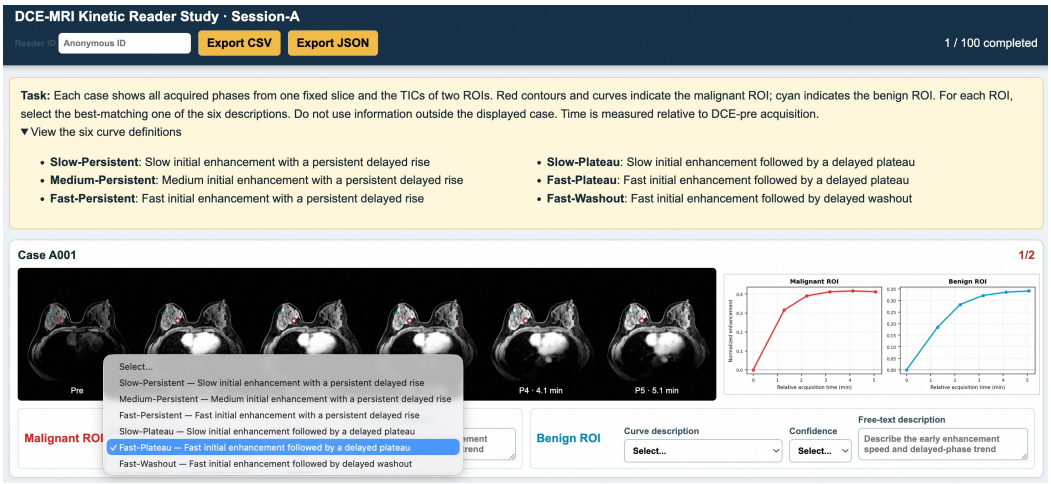

Figure 11: Reader-study interface. All DCE phases from a fixed slice are shown alongside the ROI contours and corresponding TICs. For each ROI, the reader selects one of the six predefined kinetic subtypes from the drop-down menu.

drift. Similarly, removing horizon warm-up degrades both metrics, whereas $E_{\text{warm}} \in [1, 3]$ epochs yield a stable plateau. Considering both performance and training efficiency, we set $E_{\text{warm}} = 1$. The consistent trends across Duke and PROSTATEx demonstrate that KINEWORLD is robust over broad hyperparameter ranges and does not rely on narrowly tuned values.

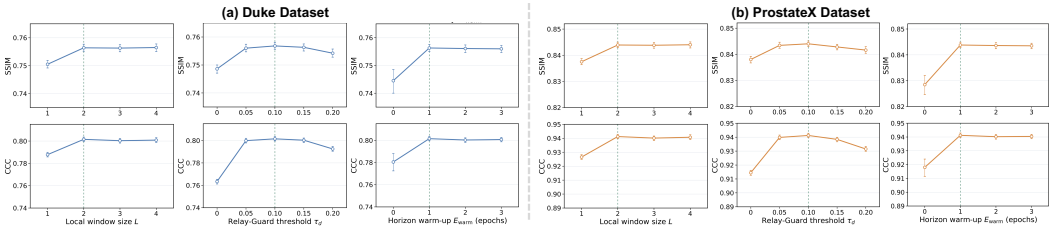

Figure 12: Hyperparameter sensitivity on the held-out validation subsets of the **(a)** Duke and **(b)** PROSTATEx datasets. We vary the local-window size L, Relay-Guard threshold τ_d, and horizon warm-up duration E_{warm} independently, and report SSIM and CCC. The green dashed line in each sub figure indicates the selected configurations. Performance remains stable within broad neighborhoods of the selected values.

A.12 RELATED WORK

DCE-MRI Pharmacokinetics. DCE-MRI kinetics are conventionally characterized using compartmental models, such as the Tofts and extended Tofts models, or lesion-specific time–intensity curves Tofts (1997); Tofts et al. (1999); Ingrisch & Sourbron (2013); Kuhl et al. (1999). These approaches extract physiologically meaningful descriptors from acquired dynamic sequences, but do not synthesize missing enhancement trajectories.

DCE-MRI Virtual Contrast Enhancement. Deep virtual contrast enhancement predicts post-contrast images from non-contrast acquisitions. Recent methods synthesize selected early and late phases from multimodal inputs Ramanarayanan et al. (2024) or condition generation on acquisition time Osuala et al. (2024). However, they primarily learn direct mappings from a pre-contrast image to individual target phases, without explicitly propagating a kinetic state through time. MRI CEKWorld Kong et al. (2026) is the closest work, casting continuous contrast synthesis as world

modeling and introducing latent alignment and difference constraints to preserve patient anatomy and temporal smoothness under sparse observations. Nevertheless, each phase is generated directly from the pre-contrast image and acquisition time, rather than autoregressively evolving a patient-specific kinetic state. In contrast, KINEWORLD explicitly evolves a patient-specific latent state by integrating pharmacokinetics-inspired velocities, thereby modeling contrast influx, efflux, and historical state dependence as components of the transition dynamics rather than solely as spatial or smoothness constraints.

Controllable Image Generation. Conditional diffusion models enable pretrained generative models to incorporate additional semantic or spatial controls. Custom Diffusion Kumari et al. (2023) adapts selected cross-attention parameters, while ControlNet Zhang et al. (2023) and T2I-Adapter Mou et al. (2024) inject external conditions into pretrained diffusion backbones. CCNet Osuala et al. (2024) specializes this paradigm for DCE-MRI by conditioning on acquisition time and imaging metadata. Although these methods provide strong image-level control, independently generating each phase does not enforce the state transitions/rollout underlying contrast uptake, peak enhancement, and washout.

World Models and Video Generation. World models learn compact state representations and transition dynamics for predicting future observations Ha & Schmidhuber (2018); LeCun (2022); Ding et al. (2026a). Their applications span autonomous driving Yang et al. (2023); Bar et al. (2025), video-based physical modeling and robot learning Mendonca et al. (2023); Kang et al. (2025), and medical problems Robertshaw et al. (2025); Yang et al. (2025); Yue et al. (2025a); Yang et al. (2026b); Ding et al. (2026b); Yue et al. (2025b). Many established formulations learn action-conditioned transitions Hafner et al. (2020); Okada & Taniguchi (2022); Saanum et al. (2024), while video-centric approaches infer actions or dynamics from densely sampled observations Mendonca et al. (2023); Gao et al. (2025); Pan et al. (2022). These settings assume either available control signals or sufficiently dense visual observations, neither of which is readily available in clinical DCE-MRI.

Sharing the goal of predicting future visual states, video diffusion models Xing et al. (2024b) learn generic priors over spatiotemporal evolution. AnimateDiff Guo et al. (2024) introduces reusable motion modules, DynamiCrafter Xing et al. (2024a) animates conditioning images using pretrained video priors, and LVDM He et al. (2022) performs video generation in a compressed spatiotemporal space. DCE-MRI differs from natural video because anatomy should remain stationary, while tissue-specific signal changes follow pharmacokinetic dynamics at sparse and irregular acquisition times. KINEWORLD therefore replaces generic motion modeling with an anatomy-anchored, PK-inspired latent state rollout.

Retrieval-Augmented Long-Horizon Generation. Long-range dependencies have been modeled using recurrent memory Dai et al. (2019), compressed historical states Rae et al. (2019), and retrieval-augmented context Khandelwal et al. (2020); Borgeaud et al. (2022); Wu et al. (2022); Xiao et al. (2026); Yang et al. (2026a). LongLive-RAG Hu et al. (2026) extends retrieval augmentation to autoregressive video generation by storing generated latents as searchable history. KineBank follows this general principle but specializes retrieval to DCE-MRI kinetics. In KA-Retrieval, each key jointly represents enhancement state and velocity, while kinetics-aware MMR balances relevance and diversity to prevent the retrieved context from being dominated by redundant states.

Stable Autoregressive Rollout. Autoregressive models trained only on ground-truth histories encounter a different state distribution at inference, causing errors to accumulate over long rollouts. Scheduled sampling Bengio et al. (2015), DAgger Ross et al. (2011), and Professor Forcing Lamb et al. (2016) mitigate this mismatch by exposing the model to its own predictions or aligning teacher-forced and free-running dynamics. Related in spirit, Relay-OPD Xu et al. (2026) briefly hands failed on-policy trajectories to a teacher for correction before returning control to the student. In DCE-MRI synthesis, rollout errors can additionally contaminate the kinetic memory used by subsequent transitions. Sharing similar spirit, Relay-guarded training therefore optimize KINEWORLD on self-generated histories while restoring a ground-truth state only upon excessive deviation, limiting error propagation and memory contamination while preserving predominantly on-policy exposure.