MRI Contrast Enhancement Kinetics World Model

Jindi Kong[1] Yuting He[1] Cong Xia[2] Rongjun Ge[3] Shuo Li[1*]
[1]Case Western Reserve University, Cleveland, OH, USA,
[2]Jiangsu Cancer Hospital, Nanjing, Jiangsu, China, [3]Southeast University, Nanjing, Jiangsu, China

Abstract

*Clinical MRI contrast acquisition suffers from inefficient information yield, which presents as a mismatch between the risky and costly acquisition protocol and the fixed and sparse acquisition sequence. Applying world models to simulate the contrast enhancement kinetics in the human body enables continuous contrast-free dynamics. However, the low temporal resolution in MRI acquisition restricts the training of world models, leading to a sparsely sampled dataset. Directly training a generative model to capture the kinetics leads to two limitations: (a) Due to the absence of data on missing time, the model tends to overfit to irrelevant features, leading to content distortion. (b) Due to the lack of continuous temporal supervision, the model fails to learn the continuous kinetics law over time, causing temporal discontinuities. For the first time, we propose **MRI** **C**ontrast **E**nhancement **K**inetics **World** model (MRI CEKWorld) with **S**patio**T**emporal **C**onsistency **L**earning (STCL). For (a), guided by the spatial law that patient-level structures remain consistent during enhancement, we propose Latent Alignment Learning (LAL) that constructs a patient-specific template to constrain contents to align with this template. For (b), guided by the temporal law that the kinetics follow a consistent smooth trend, we propose Latent Difference Learning (LDL) which extends the unobserved intervals by interpolation and constrains smooth variations in the latent space among interpolated sequences. Extensive experiments on two datasets show our MRI CEKWorld achieves better realistic contents and kinetics. Codes will be available at https://github.com/DD0922/MRI-Contrast-Enhancement-Kinetics-World-Model.*

1. Introduction

World models [8, 14, 33], which learn to simulate the dynamics of physical systems via deep neural representations [5, 44, 46, 51], offer a compelling direction for modeling MRI contrast enhancement kinetics. As illustrated in Fig. 1, such models can infer the pharmacokinetic evolution of contrast agents directly from an initial non-contrast MRI, thereby enabling the estimation of contrast-agent distribution at arbitrary time points and synthesizing the corresponding contrast-enhanced MRI images. Once realized, this capability yields two key advantages: **1)** *Contrast-free MR imaging paradigm.* By obviating the need for exogenous contrast administration, it mitigates injection-related risks [43] and reduces the economic [64] and procedural overhead incurred by additional contrast-enhanced acquisitions. **2)** *High temporal resolution modeling.* By producing continuous and temporally dense enhancement trajectories which are unconstrained by the sparse sampling of clinical protocols [24, 66], the model offers substantially higher information throughput and a more faithful reconstruction of underlying contrast-agent kinetics. These considerations naturally motivate a scientific question: "*Can we construct an MRI **C**ontrast **E**nhancement **K**inetics **World** model*

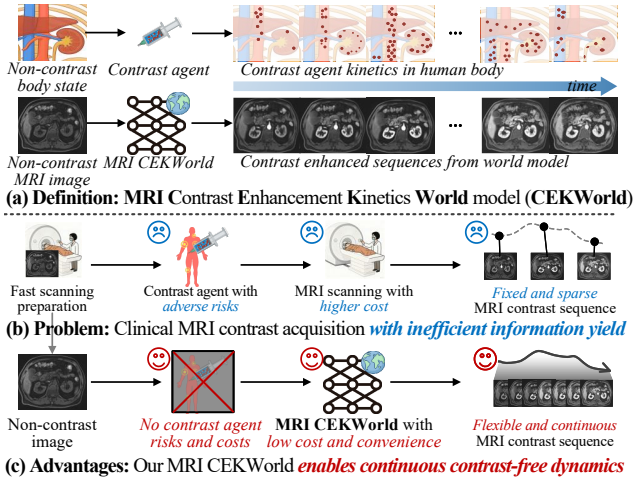

Figure 1. (a) **Task**: MRI CEKWorld generates contrast-enhanced sequences that conform to kinetics in the human body after contrast agent injection. (b) **Problem**: Clinical contrast MRI acquisition presents inefficient information yield with adverse risks and higher cost, but a fixed, sparse sequence. (c) **Advantages**: Our MRI CEKWorld enables continuous contrast-free dynamics with no contrast agent risks, low cost, and convenience.

*Corresponding author: shuo.li11@case.edu

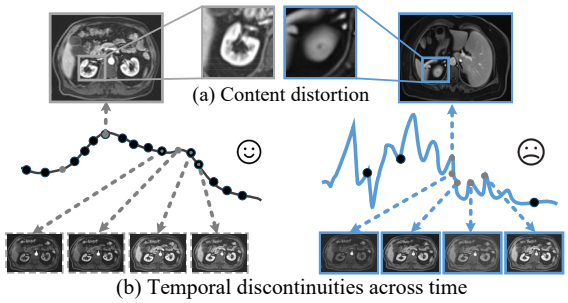

Figure 2. **Limitation**: MRI Acquisition-induced low temporal resolution in MRI CEKWorld leads to (a) Content distortion and (b) Temporal discontinuities across time.

(CEKWorld) that, relying solely on non-contrast MRI, faithfully reconstructs in vivo contrast-agent dynamics and produces high-fidelity temporal enhancement images?"

However, it is challenging to train the MRI CEKWorld owing to the low temporal resolution in MRI acquisition [10, 15, 48]. Different from the general domain where millisecond-level continuous videos can be obtained via continuous sampling for model training [2, 20, 27, 54], the sequence acquisition in the MRI CEKWorld is constrained by reconstruction duration [2, 20, 54] and patient respiratory cooperation [65]. Consequently, the acquired sequences are extremely sparse, with only second-level intervals [69], such a sparsely sampled dataset directly hinders the model's learning of contrast agent kinetic laws.

A practical compromise is to directly train a generative model on a sparsely sampled dataset in an attempt to capture the underlying contrast-agent kinetics. However, this strategy suffers from two fundamental limitations: (a) Content distortion in the spatial dimension. Owing to the absence of ground-truth frames at the missing time points, the model receives no supervision on the true anatomical state. Once overfitting occurs, it produces the distortions illustrated in Fig. 2(a), including structural deformation and organ misalignment. Although prior-based regularization [1, 21, 38, 53, 77] can encourage more realistic content at unsampled times, such priors still fail to preserve patient-specific anatomical details. (b) Discontinuity in the temporal dimension. Without continuously sampled data, the model is unable to learn the true kinetic law of the contrast agent, leading to mismatches with time conditions and temporal jumps between adjacent frames, as shown in Fig. 2(b). While post-hoc smoothing [13, 23, 26, 39] reduces visible discontinuities, pixel-space smoothing inevitably blurs fine details and deviates from the actual kinetics.

The contrast kinetics of MRI in the same patient follow an inherent spatiotemporal consistent law. *Spatially*, the anatomical structures such as organ contours, tissue boundaries, and their relative positions in the same patient remain consistent across time, unaffected by the dynamic changes of contrast agent kinetics. Guided by this spatial law, patient-level spatial consistency is enforced to constrain the model's learning, directing it to focus on relevant spatial features, effectively preserving content reality. *Temporally*, the enhanced sequences follow a consistent smooth evolutionary trend without abrupt jumps. Leveraging this temporal law directs the model to capture the inherent sequential dynamics of contrast agent metabolism in the latent space, ensuring the temporal smoothness of generated sequences and suppressing unnatural jumps.

For the first time, we propose the **S**patio**T**emporal **C**onsistency Learning (STCL), which utilizes the inherent consistency spatiotemporal law of contrast agent kinetics to enable the MRI CEKWorld, achieving realistic predictions and smooth simulations under the training from the sparsely sampled dataset. It has two innovations:

Latent Alignment Learning (LAL) for realistic content automatically constructs an explicit patient-specific template by leveraging region-specific responses to encode spatial consistent relationships and constraining the generated content at each time point to align with this template spatially. First, a patient-specific template is calculated by computing covariance matrices between features at each time point in the latent space, which represents the time-invariant spatial anatomical structure during enhancement process. It then normalizes and aggregates these features to form an explicit patient-level template. Subsequently, the equidistance constraint aligns the statistical features at each time point with this template, ensuring all time points adhere to the unified statistical rules of the template and thereby maintaining content consistency.

Latent Difference Learning (LDL) for temporal continuity interpolates in the unobserved intervals in the latent space and constrains smooth variations between consecutive points for semantic continuity. First, it uniformly inserts intermediate virtual time points between the original sparse acquisition sequence to construct a dense sequence, filling temporal gaps. Second, it calculates the variations of adjacent time points by computing the discrete second-order central differences of time points in the dense sequence, and constrains the variation to zero, which constrains the temporal changes to suppress abrupt jumps, and ensures the smoothness of temporal evolution.

Our contributions are summarized as follows: 1) For the first time, we propose MRI CEKWorld, which simulates the contrast agent kinetics in the human body and facilitates continuous contrast-free dynamics. 2) We propose STCL, which enforces content reality and temporal continuity under acquisition-induced low temporal resolution through spatiotemporal consistent physiological law. 3) Our LAL constructs an explicit patient-level template for each generation alignment, maintaining content consistency and reality. 4) Our LDL extends unobserved intervals and con-

strains variations between consecutive points, thereby ensuring the smoothness of temporal evolution.

2. Related Work

Virtual MRI Contrast Enhancement is an alternative to the use of contrast agents [5, 28, 31, 36], designed to emulate the visibility of specific tissues and bodily fluids. Compared to traditional contrast agents [16], it has three significant advantages: i.) Safe: it can address safety concerns regarding possible contrast agent deposition [43], ii.) Comfortable: it can mitigate patient discomfort while scanning [47], iii.) Economical: it can help cut down on human resources, hardware costs, and overall expense [64].

Recent virtual MRI contrast enhancement methods are divided into static generation and dynamic sequence generation. Static generation methods [4, 5, 22, 45] focus on synthesizing the single-phase contrast-enhanced images at a single time point, T1-weighted contrast-enhanced images from multiple non-contrast sequences such as T1-weighted, T2-weighted and the Apparent Diffusion Coefficient map. Such methods prioritize accurately simulating the final enhancement patterns of tumors or lesions to improve the accuracy of classification-based diagnosis. Dynamic sequence generation methods [49, 58, 62] aims to synthesize the time sequence, including multi-phase contrast-enhanced images at several time points. However, due to the physical acquisition limit, virtual MRI contrast agent remains confined to the scope of image-to-image mapping[34, 51, 57], failing to simulate contrast agent kinetics.

World Models are able to understand the world and predict the future [8, 14, 33]. Based on its great functionality and promise, the world model has been used in many aspects such as autonomous driving [3, 8, 12, 40, 71], video generation [27, 37, 40] and medical[59, 70, 74]. Existing world models are either continuous action-based [15, 15, 48, 61, 72], relying on a continuously available external control signal to guide and correct state transitions, or observation-driven [11, 52], learning the dynamics from densely sampled video sequences where the observations themselves serve as a continuous surrogate. However, continuous interaction and dense observations are costly, or even infeasible in MRI contrast agent kinetics world model, limiting the practicality of these approaches.

Spatiotemporal Consistency has gradually attracted growing attention which can be categorized into two families: Slow feature analysis [25, 68, 76] assumed that changes between adjacent frames in natural videos are slow and smooth, extracting temporally consistent representations by minimizing temporal derivatives; Contrastive learning [7, 9, 17–19, 55, 67, 73] is enabled to learn features insensitive to spatiotemporal perturbations through positive-negative sample contrastive constraints. While both paradigms successfully capture temporal stability in dense video data, they rely on continuous frame supervision and aim at representation robustness rather than generation. Given low temporal resolution training data, the former relies on continuous sampling and thus fails to estimate temporal variation trends, while the latter lacks sufficient samples for comparison, resulting in underfitting of the learned spatiotemporal consistent features.

3. Method

As shown in Fig.3, our spatiotemporal consistency learning implements MRI contrast agent kinetics world model (formulated in Sec.3.1) via constraining spatial information at each time points to the patient-level template to preserve content reality (LAL, see Sec.3.2) and constraining the latent representation in the dense interpolated sequence to be smooth (LDL, see Sec.3.3).

3.1. Formulation

The MRI contrast agent kinetics world model is formulated as an image time series modeling, which predicts the contrast enhanced MRI image $\mathcal{I}(t)$ at arbitrary time t based on a non-contrast image $\mathcal{I}_{p,0}$.

Dataset Due to low temporal resolution in MRI acquisition, the dataset is temporally sparsely sampled. For each patient p, we denote the image-time pairs as $\mathcal{D}_p = \{(\mathcal{I}_{p,i}, t_{p,i})\}_{i=0}^{T_p}$ where $\mathcal{I}_{p,i}$ represents the image acquired at time $t_{p,i}$, T_p represents the total of the acquisition time points. The complete dataset is defined as $\mathcal{D} = \{\mathcal{D}_p \mid p = 1, 2, \ldots, P\}$ where P is the patient number.

Training The model aims to learn a mapping from the initial non-contrast image $\mathcal{I}_{p,0}$ and a continuous time variable t to the corresponding contrast-enhanced image $\mathcal{I}_p(t)$. As shown in Fig.3 (a), t, $\mathcal{I}_{p,0}$ and $\mathcal{I}_p(t)$ are encoded separately. The groundtruth encoder E_{gt}, same as the encoder in VAE [30, 60] encodes the contrast-enhanced image $\mathcal{I}_p(t)_{\text{gt}}$. The time condition encoder E_t uses CLIP [56] processes the time variable which means the duration after contrast agent injection, converting temporal text information into high-dimensional features that guide the model to generate time-specific enhancement features. The image condition encoder E_{img}, encodes the non-contrast image $\mathcal{I}_{p,0}$ through zero-convolution [75] and adds to the layers in latent diffusion model [60], acting as a hint to guide the prediction.

Spatial, temporal and diffusion losses are utilized to regularize the generations. The former two will be introduced respectively in Sec.3.2 and Sec.3.3. The diffusion loss $\mathcal{L}_{\text{Diffusion}} = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_\theta\|^2\right]$ is used to constrain the accuracy of noise prediction. Here, t represents the denoising timestep. Let $\mathcal{M}_\theta$ denote the MRI contrast enhancement world model parameterized by θ.

$$\hat{\mathcal{I}}_p(t) = \mathcal{M}_\theta(\mathcal{I}_{p,0}, t), \quad t \in \mathbb{R}^+ \tag{1}$$

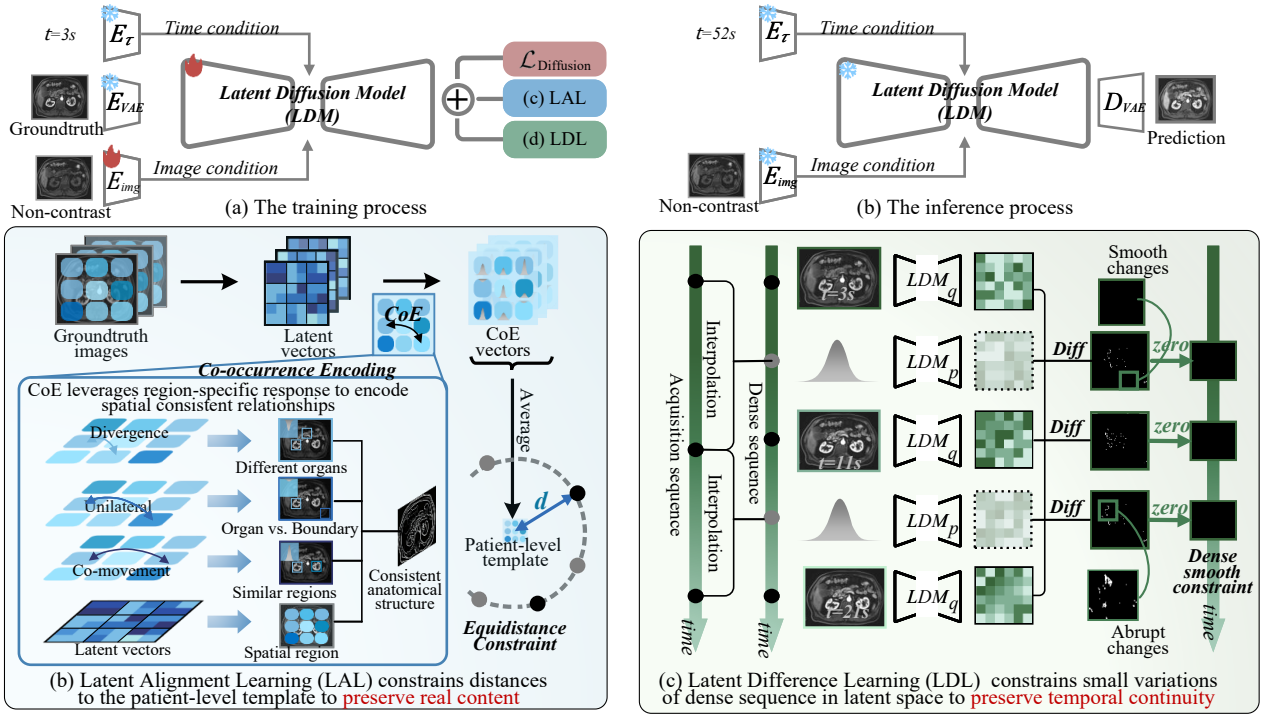

Figure 3. Overview framework of the MRI CEKWorld. (a) and (b) shows the training and inference processes. (c) LAL captures region-wise co-occurrence relationships and enforces anatomical consistency by aligning to a patient-level template. (d) LDL constructs a dense time series in the latent space and imposes a second-order difference (denoted as Diff) on adjacent moments for smooth evolution (p and q denote the inference).

Inference After training, the optimized model $\mathcal{M}_{\theta^*}$ takes the non-contrast image $\mathcal{I}_{p,0}$ and time t as input to predict the contrast enhanced image: $\hat{\mathcal{I}}_p(t) = \mathcal{M}_{\theta^*}(\mathcal{I}_{p,0}, t)$. In Fig.3 (b), the prediction is decoded by $\mathcal{D}_{img}$ after the U-Net to convert the latent variable into the pixel space.

3.2. Latent Alignment Learning *for real contents*

As shown in Fig.3 (c), on the basis of anatomical consistency, leveraging the differences in the response patterns of contrast agent signals across various regions, we encode the fluctuation relationship between these regions as a numerical statistical template for anatomical regions, and use this template to constrain the generated results to comply with this fluctuation relationship, suppressing the distortion.

Co-occurrence Encoding is implemented by a covariance matrix, which computes the spatial co-occurrence patterns of anatomical structures. Co-movement within regions corresponds to similar regions, while unilateral divergence between regions corresponds to boundary separation. This characterizes the consistent spatial content of the patient. Latent representation $\hat{x}_0$ is extracted by leveraging the reverse process of diffusion models. It uses the model-predicted noise $\boldsymbol{\epsilon}$ and the noisy samplex_t to provide a high-quality, structured latent space representation for subsequent statistical calculations and constraints. $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ represents the cumulative coefficient of α across t diffusion steps, x_t denotes the noisy sample after τ denoising steps: $\hat{x}_0 = \frac{x_\tau - \sqrt{1-\bar{\alpha}_\tau}\cdot\boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_\tau}}$. Then, in the latent space, we have the prediction series $\hat{x_0} = \{\hat{x_0}_t \in \mathbb{R}^{c\times h\times w}\}_{t=1}^{T}$, where c is the number of channels, h and w are the height and width of $\hat{x}_0$. We flatten each time point of $\hat{x_0}_t$ into $X_t \in \mathbb{R}^{c\times s}$ with $s = h\cdot w$ and center it along the spatial dimension $X_t^c = X_t - \frac{1}{s}\sum_{s_i=1}^{s} X_t$. This removes the spatial mean bias so that the covariance reflects the true distribution shape of the features. Covariance matrix for each time acquisition time point t is computed as $\Sigma_t = \frac{1}{S-1} X_t^c (X_t^c)^\top$ and then regularized with shrinkage and a small jitter to ensure positive-definiteness, so $\tilde{\Sigma}_t = (1-\gamma)\Sigma_t + \gamma I + \varepsilon I$ where γ controls the shrinkage strength, I is the identity matrix, and ε is a small jitter.

The patient-level template, obtained by computing the mean of covariance matrix of time points, represents a more stable patient-level spatial feature under the spatial law of invariant patient anatomical structures. We map each latent covariance Σ_t to an Euclidean vector via the log–Cholesky parameterization for numerical stability and positive–definiteness preservation for optimization in training [35]. Let $L_t = \mathrm{chol}(\Sigma_t)$; extract $\mathrm{lower}_t = \mathrm{vec}(\mathrm{tril}(L_t, -1))$ and $\mathrm{logdiag}_t = \log(\mathrm{diag}(L_t))$, and form $z_t = [\mathrm{lower}_t; \mathrm{logdiag}_t]$. Averaging gives the patient-level template vector $\bar{z} = \frac{1}{T}\sum_{t=1}^{T} z_t$, which we use sub-

sequently as the template representation.

Equidistance constraint constrains the z_t at different time points are consistent with the template by keeping the same distance from the template $\bar{z}$. It not only ensures statistical consistency but also allows reasonable dynamic changes to be preserved between time points, so that the generated sequence has a real spatial content under sparse supervision. The spatial loss from equidistance constraint of distance $d_t^2 = \left\| z_t - \bar{z} \right\|_2^2$ is defined as:

$$\mathcal{L}_{\text{Spatial}} = \frac{1}{P} \sum_{p=1}^{P} \frac{1}{T_p} \sum_{t=1}^{T_p} \left(d_{p,t}^2 \right). \tag{2}$$

where P represents the total patient number and T_p represents the total acquisition time points of the p-th patient.

3.3. Latent Difference Learning *for continuity*

As shown in Fig.3(d), LDL generates predictions in the latent space for the unobserved intermediate time points by interpolation, and imposes smoothness constraints on the interpolated dense sequence for semantic continuity to achieve temporal smooth transition.

Intermediate points are inserted to generate a dense sequence in the latent space. The set of original sparse acquisition time points consists of real observed time values, defined as: $T_{\text{acq}} = \{t_{\text{acq},0}, t_{\text{acq},1}, \ldots, t_{\text{acq},N-1}\}$ where N is the number of original acquisition time points; $t_{\text{acq},i}$ denotes the absolute time of the i-th original observation and ordered by acquisition time. For each pair of adjacent original observations $(t_{\text{acq},i}, t_{\text{ac1},i+1})$ $(i \in [0, N-2])$, let K_i be the number of inserted intermediate points $(K_i \geq 1)$, on the time interval $\Delta t_i = t_{\text{acq},i+1} - t_{\text{acq},i}$, uniform interpolation to ensure intermediate points lie strictly between $t_{\text{acq},i}$ and $t_{\text{acq},i+1}$. The time value of the k-th intermediate point $(k \in [1, K_i])$ between the i-th pair of adjacent points is: $t_{\text{mid},i,k} = t_{\text{acq},i} + \frac{k}{K_i+1} \cdot \Delta t_i$. The dense sequence T_{dense} is the union of original observations and all intermediate points, sorted by time: $T_{\text{dense}} = T_{\text{acq}} \cup \left(\bigcup_{i=0}^{N-2} \{t_{\text{mid},i,1}, \ldots, t_{\text{mid},i,K_i}\} \right)$.

Dense prediction sequences in the latent space are constructed as a set of clean predictions at all dense time points, including both the outputs anchored at real acquisition times and those generated from noise at inserted intermediate times. For each observed acquisition point $t_{\text{acq},i}$, we recover its latent vector $\hat{x}_{0,\text{acq},i}$ from the corresponding noisy sample $x_{\tau,\text{acq},i}$ after τ denoising steps: $\hat{x}_{0,\text{acq},i} = \frac{x_{\tau,\text{acq},i} - \sqrt{1-\bar{\alpha}_{\tau_{\text{acq},i}}}\, \epsilon_\theta}{\sqrt{\bar{\alpha}_{\tau_{\text{acq},i}}}}$. For each inserted intermediate time point $t_{\text{mid},j}$, we sample a noise latent $x_{\tau,\text{mid},j} \sim \mathcal{N}(0, I)$ at the assigned timestep $\tau_{\text{dense},j}$, and obtain its latent prediction by applying denoising schedule of the latent diffusion model [75]: $\hat{x}_{0,\text{mid},j} =_\theta (x_{\tau,\text{mid},j}, \tau_{\text{dense},j})$, where p_θ denotes the standard inference denoising process that maps a noisy latent to its predicted latent vector [75]. Thus, the dense latent prediction sequence is finally assembled as

$$\hat{X}_{\text{dense}}[j] = \begin{cases} \hat{x}_{0,\text{acq},i}, & t_{\text{dense},j} = t_{\text{acq},i}, \\ \hat{x}_{0,\text{mid},j}, & t_{\text{dense},j} \text{ is intermediate.} \end{cases}$$

Dense smooth constraint limits the abrupt variance to zero through second-order difference center difference for smoothness over time. After de-duplication to remove possible duplicate time points of T_{dense}, we obtain an ordered time sequence $T_{\text{sort}} = \{t_0, t_1, \ldots, t_{T-1}\}$ $(T \leq M$, where t_k is the time value in seconds) and corresponding model outputs $\boldsymbol{y}_{\text{sort}} = \{y_{\text{sort}}^0, y_{\text{sort}}^1, \ldots, y_{\text{sort}}^{T-1}\}$ (where $y_{\text{sort}}^k \in \mathbb{R}^{1\times c\times h\times w}$ is the model output at t_k). For each point $k \in [1, T-2]$, the dense smooth constraint among discrete different time points through the center difference equation is derived in *Supplementary*, which is defined as:

$$\begin{aligned} \mathbf{D}_2^k = 2 \cdot \Bigg(& \frac{y_{\text{sort}}^{k-1}}{h_0^k \cdot (h_0^k + h_1^k)} - \frac{y_{\text{sort}}^k}{h_0^k \cdot h_1^k} \\ & + \frac{y_{\text{sort}}^{k+1}}{h_1^k \cdot (h_0^k + h_1^k)} \Bigg) \cdot w^k \end{aligned} \tag{3}$$

where $h_0^k = t_k - t_{k-1} + \delta$ and $h_1^k = t_{k+1} - t_k + \delta$ ($\delta = 10^{-6}$ to avoid division by zero) are adjacent time intervals; $w^k = \frac{1}{1+h_0^k+h_1^k}$ is the interval weight used for weaker penalty for larger intervals, adapting to varying temporal densities. The final loss is the average of these differences using the L1 norm for robustness to outliers, which is expressed as

$$\mathcal{L}_{\text{Temporal}} = \frac{1}{T-2} \sum_{k=1}^{T-2} \| \mathbf{D}_2^{(k)} \|_1. \tag{4}$$

4. Experiments

4.1. Comparison Study

4.1.1. Experiment Protocol

This section introduces the overview protocol in our study for complete and fair evaluations in our experiments.

Datasets Two Dynamic Contrast Enhancement - Magnetic Resonance Imaging (DCE-MRI) datasets are used: (1) Private Abdominal DCE-MRI Dataset (Abdominal DCE-MRI): This abdominal consists of 91 patients. There is one non-contrast image, and 15 contrast enhanced images within 300 seconds after contrast agent injection. Among these contrast enhanced images, 6 are in the arterial phase, 6 are in the venous phase, and 3 are in the delayed phase. (2) Public Duke Breast DCE-MRI Dataset (Breast DCE-MRI) [63]: This dataset contains 922 examination records of breast DCE-MRI. After the injection of contrast agent, contrast-enhanced data at 3 or 4 time points are acquired. Following [50], we crop the slices containing the lesion region and increase the width and height of the tumor bounding box to half the width and height of the full image. Both

Table 1. Quantitative results: Quantitative comparison of different methods on Abdominal and Breast DCE-MRI datasets show that our method achieves the best performance. Quantitative ablation results verify the effectiveness of our LAL and LDL. "Avg.SSIM" and "Avg.cSSIM" denote the average score over two datasets in spatial SSIM and temporal cSSIM metrics, respectively.

Method	Abdominal DCE-MRI					Breast DCE-MRI					Avg.	Avg.
	PSNR↑	SSIM↑	LPIPS↓	rMSE↓	cSSIM↑	PSNR↑	SSIM↑	LPIPS↓	rMSE↓	cSSIM↑	SSIM↑	cSSIM↑
CustomDiff[32]	17.73	0.4130	0.4646	295.7	0.3551	11.19	0.2463	0.4706	1094	0.3835	0.3296	0.3693
T2I[41]	16.89	0.4396	0.3100	124.1	0.3396	17.73	0.4130	0.2905	297.4	0.1347	0.4263	0.2372
CCNet[50]	**24.35**	0.5794	0.2735	43.44	0.7098	**21.47**	0.4043	0.3128	289.3	0.3155	0.4918	0.5127
EditAR[42]	22.65	0.5571	0.3314	49.76	0.7536	19.85	0.4170	0.3119	288.9	0.3886	0.4870	0.5711
ControlNet$_{baseline}$ [75]	23.61	0.7178	0.2719	43.08	0.8286	19.79	0.5196	0.2625	308.0	0.3370	0.6187	0.5828
+ *LAL*	23.92	0.7227	0.2666	40.29	0.8439	20.86	0.5442	0.2640	250.7	0.3879	0.6335	0.6159
+ *LDL*	24.05	0.7369	0.2623	40.25	0.8411	20.21	0.5391	0.2636	261.5	0.3392	0.6380	0.5901
MRI CEKWorld	24.06	**0.7419**	**0.2622**	**40.08**	**0.8451**	21.09	**0.5599**	**0.2620**	**243.5**	**0.3900**	**0.6509**	**0.6176**

datasets are resized to the 256×256, normalized to $[-1, 1]$ [60] and then stacked into 3 channels as the image input. Since the acquisition time of DCE-MRI sequences in both datasets is manually controlled, the acquisition of sequences is not strictly fixed at specific time points. Thus, the time points in the test set rarely appear in the training set.

Evaluation Metrics Both spatial and temporal metrics are utilized to validate the performance of virtual MRI contrast enhancement prediction in all experiments. *Spatially*, following the studies [6, 29, 50, 75], Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) and root Mean Squared Error (rMSE) are used to evaluate in all experiments. *Temporally*, continuous SSIM (cSSIM) is designed to quantify the structural consistency between adjacent frames in the temporal dimension. Assuming a time series contains N consecutive frames , corresponding to time points $t = 1, 2, ..., N$, denoted as $I_1, I_2, ..., I_N$ with adjacent frame pairs (I_t, I_{t+1}) (a total of $N - 1$ pairs), its formula is: $cSSIM = \frac{1}{N-1} \sum_{t=1}^{N-1} \text{SSIM}(I_t, I_{t+1})$ where $\text{SSIM}(I_t, I_{t+1})$ is the structural similarity of a single pair of adjacent frames. We then define a spatial average score Avg.SSIM by averaging the SSIM over both datasets, and a temporal average score Avg.cSSIM by averaging the cSSIM over both datasets. Both average metrics lie in [0,1], with larger values indicating better overall performance.

Implementation Details We adopt the ControlNet [75] as the model backbone and make U-Net [60], image encoder E_{img} trainable as well. Training and testing were implemented on an NVIDIA A100 GPU with 40GB of memory. For all of the hyper parameters in training process, total epoch is 14, batch size is 4, $\lambda_{\text{Spatial}} = 6.0$, $\lambda_{\text{Temporal}} = 1.0$, the $K_i = 2$ for the best performance in Abdominal DCE-MRI; $\lambda_{\text{Spatial}} = 4.0$, $\lambda_{\text{Temporal}} = 1.0$, the $K_i = 2$ obtains the best performance in Breast DCE-MRI. The whole pipeline is two-stage, first stage performs a diffusion warm-up to stabilize the initial latent space, then introduces spatial regularization to establish patient-specific content. The loss function is $\mathcal{L}_1 = \mathcal{L}_{\text{Diffusion}} + \lambda_{\text{Spatial}} \mathcal{L}_{\text{Spatial}}$ where λ_{Spatial} represents the hyperparameter for the strength of spatial regularization. The second stage switches to temporal regularization to leverage the aligned content and achieve smooth transitions. The loss of the second stage is $\mathcal{L}_2 = \mathcal{L}_{\text{Diffusion}} + \lambda_{\text{Temporal}} \mathcal{L}_{\text{Temporal}}$.

Comparisons Setting Recent controllable image generation frameworks [32, 41, 42, 75]. DCE-MRI contrast enhancement method [50], ContrastControlNet (CCNet) are compared. All the hyperparameter experiment settings are the same, the time inputs are formulated as "HH:MM:SS", where HH, MM, SS mean the hours, minutes and seconds of the time interval between the pre-contrast images and the generation. CCNet, T2I and ControlNet are all based on SD1.5 [60] following [42].

4.1.2. Comparison Analysis

Quantitative Results Analysis Quantitative results on both datasets show our method outperforms others in spatial fidelity and temporal smoothness (Tab. 1), leading in Avg.S and Avg.T. *Spatially*, our LAL module preserves anatomical consistency, achieving the best SSIM, LPIPS, and rMSE. While CCNet attains high PSNR, it fails to fully converge under the same training settings—producing over-smoothed predictions that lose structural detail, hence poor SSIM, LPIPS, and rMSE. Excluding this case, our method reaches 24.06 PSNR, 0.7419 SSIM, and 0.2622 LPIPS on the Abdominal DCE-MRI dataset, and leads all spatial metrics on the Breast dataset. The higher rMSE of the Breast dataset (intensity range 0–4000) stems from data distribution rather than generation performance. *Temporally*, our method achieves the highest cSSIM (0.8451 for Abdominal, 0.3900 for Breast DCE-MRI), preserving inter-frame structural coherence with smoother enhancement kinetics and fewer abrupt intensity changes.

Visualization Results Analysis As shown in Fig.4 and Fig.5, the visualization sequences of both datasets demonstrate that our method achieves high spatial reality and natural kinetics, both closely matching the ground-truth. CustomDiff and T2I generated images with severe deviations from the ground-truth, suffering from blurred organ contours and distorted dynamic enhancement gradients of con-

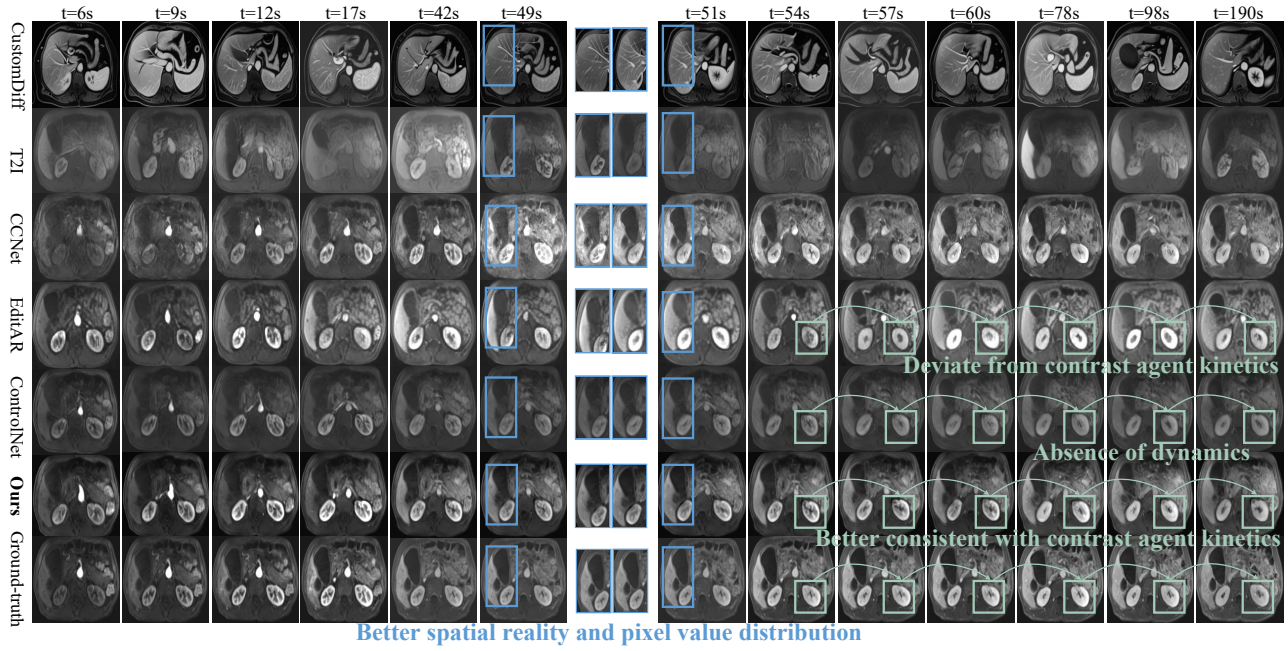

Figure 4. **Visualization results** in Abdominal DCE-MRI: The visualization results of our methods on different time points exhibit better spatial reality (zoom-in regions in blue boxes) and temporal continuities than comparisons (connected green boxes), whereas other methods show deviations from realistic kinetics or lack of dynamic consistency.

trast agents. CCNet failed to converge, leading to excessively smooth images, severe spatial structural distortion, and the appearance of color blocks, which is consistent with the previously analyzed characteristic of high PSNR. EditAR and ControlNet have normal spatial structures, but their kinetics both deviate from the normal pattern. More visualization results are in *Supplementary*.

4.2. Ablation Study and Model Analysis

Component Ablation The ablation studies in both datasets show the effectiveness of our proposed innovations. As shown in the downside part of Table.1, the LAL achieves 2.46% SSIM and 1.07 PSNR improvement in Breast DCE-MRI, which demonstrates the effectiveness of consistency owing to LAL. If using alone, significant promotion also represents its temporal smoothness compared with baseline, which achieves 1.25% SSIM in Abdominal DCE-MRI and 5.09% improvement in Breast DCE-MRI. When combining two innovations, the improvement results show that under the premise that LAL has formed better spatial structural consistency, further enhances both temporal dynamic smoothness and spatial structural consistency.

Contrast Agent Kinetics Time Curve As shown in Fig.6, the performance of the curves across the three key phases in clinical use demonstrates out method has a stronger capability in modeling the contrast agent kinetics. We performed equidistant sampling for the artery phase (1–15 s), vein phase (55–72 s), and delay phase (90–300 s) according to the common acquisition time. The sampling

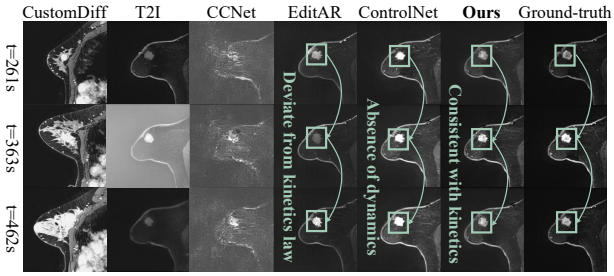

Figure 5. **Visualization results** in Breast DCE-MRI: The visualization results achieves temporal consistent (connected green boxes) with contrast agent kinetics, demonstrating superior fidelity in breast DCE-MRI sequence generation.

results of the mean gray value at each time point for the renal region of interest from each method were normalized to compare their smoothness and stability. In (a) artery phase, MRI CEKWorld exhibits a stable increase, which precisely matches the physiological process of rapid contrast agent filling in the artery phase. In (b) vein phase, ours shows a curve pattern of smooth transition, which reflects its accurate capture of the kinetics process of contrast agent in the vein phase. The curves increase by the accumulation of contrast agent within the interstitial space of the renal parenchyma then decrease owing to the washout phase. In contrast, competing methods such as CCNet and EditAR exhibit obvious abrupt fluctuations in their curves. In (c) delay phase, ours first maintain a stable signal level and then decays smoothly over time because this is fully consistent

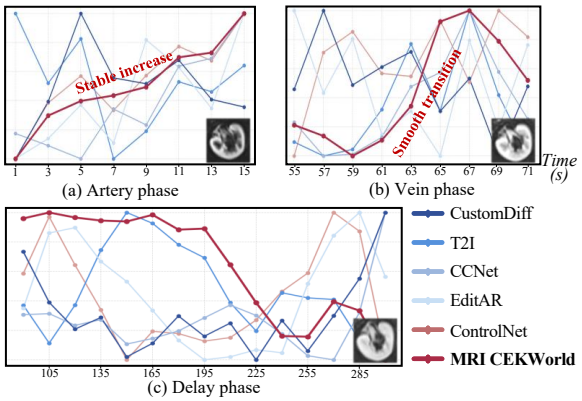

Figure 6. **Contrast agent kinetics time curve**: The curves in (a) artery phase, (b) vein phase and (c) delay phase are more stable and smooth than compared methods.

with the physiological mechanism of gradual contrast agent clearance in the delay phase, verifying its reliable modeling capability for long-term enhancement dynamics.

Hyper-parameter Ablation of λ_{Spatial} In Fig.7 (a), the transition of λ_{Spatial} shows a trend of first increasing and then decreasing in terms of SSIM and PSNR metrics. Since λ_{Spatial} determines the spatial regularization strength of LAL, when λ_{spatial} is small, the constraint of distance consistency is weak, and the generated results deviate from the template; when λ_{Spatial} is moderate which λ_{Spatial} is 6, the constraint strength achieves a balance between following the template and preserving feature diversity; when λ_{Spatial} is excessively large, this constraint rigidly enforces features to stay close to the template, suppressing the reasonable feature differences that should exist between time points.

Hyper-parameter Ablation of K_i As shown in Fig.7, the variation of K_i also shows a first increasing then decreasing trend in terms of cSSIM that measures continuity. With the increase of K_i, the newly added intermediate sampling points exactly fill the gaps in the sparse temporal sequence, which provides the model with more refined intermediate states in temporal evolution, enabling it to more accurately learn the continuous changes of contrast agent kinetic laws. However, when K_i exceeds 2, an excessive number of intermediate sampling points do not come from the real data distribution and carry noise that deviates from real patterns. This interferes with the model's learning of real temporal features, leading to a subsequent decrease.

Latent Representation at Continuous Time Points As shown in Fig.8, the distribution of continuous sequences generated by MRI CEKWorld is continuous and consistent, which demonstrates the spatial consistency and temporal continuity are preserved in the latent space. We visualize such a distribution by compressing the latent space vectors obtained from the latent space in ControlNet during the reverse process into a low-dimensional space via principal

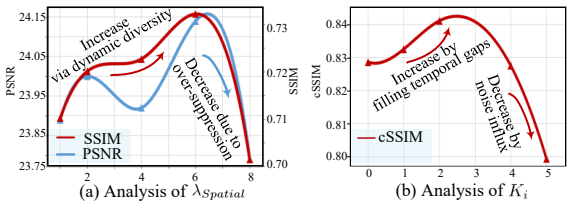

Figure 7. **Hyperparameter analysis**: Analysis of $\lambda_{Spatial}$ and K_i shows a trend of first increasing and then decreasing. (a) PSNR and SSIM increase by allowing more dynamic diversity, decrease due to over-suppression on the dynamic. (b) cSSIM increases by filling temporal gaps, and decreases due to noise influx.

component analysis, and using the corresponding time for coloring (the lighter and yellower the color, the larger the time value). In Fig.8 (a), the dots show a disperse state, indicating that their latent features have no obvious temporal pattern in the low-dimensional space and the feature distribution across different time points is chaotic. In contrast, our dots are distributed consistently and continuously, which suggests that the features at different time points have a strong clustering property, indicating that the consistency is stably preserved over time. Furthermore, as time progresses, the color sequentially changes from light to dark, reflecting that the features transition smoothly. It is worth noting that the outliers in the upper right corner of Fig.8 (b) correspond to the feature points at $= 0\,\text{s}$ and $= 1\,\text{s}$. Due to the limitation of the central difference in Eq.3, the constraints on these two points are neglected.

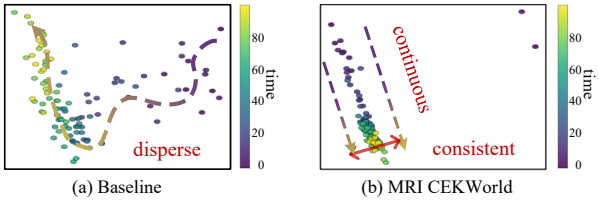

Figure 8. **Distribution in latent space**: The distribution of feature points in the latent space demonstrates generations of continuous time points adhere to temporal continuity and spatial consistency.

5. Conclusion

In this paper, we introduced MRI CEKWorld, the first contrast enhancement kinetics world model designed to simulate contrast agent kinetics in human body for the inefficient information yield in clinical MRI acquisition. Exploiting the inherent spatiotemporal consistency of contrast enhancement, we devised a spatiotemporal consistency learning under a sparsely sampled dataset which includes the latent alignment and difference learning. Despite its strong performance, we will extend to other contrast-enhanced imaging modalities, such as computed tomography, aiming for a unified contrast kinetics world model.

Acknowledgments This work was supported in part by the National Institutes of Health (NIH) under Grants R01HL173186, R01HL177813, and by the National Science Foundation (NSF) under Grant No. 2306545.

References

[1] Somayeh Akbari, Mahdi Tabassian, Joao Pedrosa, Sandro Queiros, Konstantina Papangelopoulou, and Jan D'Hooge. Beas-net: A shape-prior-based deep convolutional neural network for robust left ventricular segmentation in 2-d echocardiography. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 71(11):1565–1576, 2024. 2

[2] Gregorio Andria, Filippo Attivissimo, Giuseppe Cavone, and Anna Maria Lucia Lanzolla. Acquisition times in magnetic resonance imaging: optimization in clinical use. *IEEE Transactions on Instrumentation and Measurement*, 58(9): 3140–3148, 2009. 2

[3] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025. 3

[4] Chao Chen, Catalina Raymond, William Speier, Xinyu Jin, Timothy F Cloughesy, Dieter Enzmann, Benjamin M Ellingson, and Corey W Arnold. Synthesizing mr image contrast enhancement using 3d high-resolution convnets. *IEEE Transactions on Biomedical Engineering*, 70(2):401–412, 2022. 3

[5] Ka-Hei Cheng, Wen Li, Francis Kar-Ho Lee, Tian Li, and Jing Cai. Pixelwise gradient model with gan for virtual contrast enhancement in mri imaging. *Cancers*, 16(5):999, 2024. 1, 3

[6] Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1472–1480, 2024. 6

[7] Ishan Rajendrakumar Dave, Mamshad Nayeem Rizve, Chen Chen, and Mubarak Shah. Timebalance: Temporally-invariant and temporally-distinctive video representations for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2341–2352, 2023. 3

[8] 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*, 2024. 1, 3

[9] Shuangrui Ding, Rui Qian, and Hongkai Xiong. Dual contrastive learning for spatio-temporal representation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5649–5658, 2022. 3

[10] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024. 2

[11] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. *arXiv preprint arXiv:2503.18938*, 2025. 3

[12] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024. 3

[13] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7548–7558, 2024. 2

[14] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018. 1, 3

[15] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. 2, 3

[16] Carmel Hayes, Anwar R Padhani, and Martin O Leach. Assessing changes in tumour vascular function using dynamic contrast-enhanced magnetic resonance imaging. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 15(2):154–163, 2002. 3

[17] Yuting He and Shuo Li. Vector contrastive learning for pixel-wise pretraining in medical vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19827–19837, 2025. 3

[18] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9538–9547, 2023.

[19] Yuting He, Boyu Wang, Rongjun Ge, Yang Chen, Guanyu Yang, and Shuo Li. Homeomorphism prior for false positive and negative problem in medical image dense contrastive representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[20] Kieren Grant Hollingsworth. Reducing acquisition time in clinical mri by data undersampling and compressed sensing reconstruction. *Physics in Medicine & Biology*, 60(21): R297, 2015. 2

[21] Huimin Huang, Han Zheng, Lanfen Lin, Ming Cai, Hongjie Hu, Qiaowei Zhang, Qingqing Chen, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, et al. Medical image segmentation with deep atlas prior. *IEEE Transactions on Medical Imaging*, 40(12):3519–3530, 2021. 2

[22] Hongyan Huang, Junyang Mo, Zhiguang Ding, Xuehua Peng, Ruihao Liu, Danping Zhuang, Yuzhong Zhang, Genwen Hu, Bingsheng Huang, and Yingwei Qiu. Deep learning to simulate contrast-enhanced mri for evaluating suspected prostate cancer. *Radiology*, 314(1):e240238, 2025. 3

[23] Zhilin Huang, Yijie Yu, Ling Yang, Chujun Qin, Bing Zheng, Xiawu Zheng, Zikun Zhou, Yaowei Wang, and Wenming

Yang. Motion-aware latent diffusion models for video frame interpolation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 1043–1052, New York, NY, USA, 2024. Association for Computing Machinery. 2

[24] 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. 1

[25] Dinesh Jayaraman and Kristen Grauman. Slow and steady feature analysis: higher order temporal coherence in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3852–3861, 2016. 3

[26] Seungwoo Jeong, Wonjun Ko, Ahmad Wisnu Mulyadi, and Heung-Il Suk. Deep efficient continuous manifold learning for time series modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):171–184, 2023. 2

[27] 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. *arXiv preprint arXiv:2411.02385*, 2024. 2, 3

[28] Eunjin Kim, Hwan-Ho Cho, Junmo Kwon, Young-Tack Oh, Eun Sook Ko, and Hyunjin Park. Tumor-attentive segmentation-guided gan for synthesizing breast contrast-enhanced mri without contrast agents. *IEEE journal of translational engineering in health and medicine*, 11:32–43, 2022. 3

[29] JungEun Kim, Hangyul Yoon, Geondo Park, Kyungsu Kim, and Eunho Yang. Data-efficient unsupervised interpolation without any intermediate frame for 4d medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11353–11364, 2024. 6

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Jens Kleesiek, Jan Nikolas Morshuis, Fabian Isensee, Katerina Deike-Hofmann, Daniel Paech, Philipp Kickingereder, Ullrich Köthe, Carsten Rother, Michael Forsting, Wolfgang Wick, et al. Can virtual contrast enhancement in brain mri replace gadolinium?: a feasibility study. *Investigative radiology*, 54(10):653–660, 2019. 3

[32] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 6

[33] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 1, 3

[34] Shangxuan Li, Baoer Liu, Feilin Deng, Yikai Xu, and Wu Zhou. Image synthesis of hepatobiliary phase using contrast-enhanced mri and diffusion model. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 3

[35] Zhenhua Lin. Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019. 4

[36] Thomas Lindner, Hanna Debus, and Jens Fiehler. Virtual non-contrast enhanced magnetic resonance imaging (vnc-mri). *Magnetic Resonance Imaging*, 81:67–74, 2021. 3

[37] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 3

[38] Yuhui Ma, Jiang Liu, Yonghuai Liu, Huazhu Fu, Yan Hu, Jun Cheng, Hong Qi, Yufei Wu, Jiong Zhang, and Yitian Zhao. Structure and illumination constrained gan for medical image enhancement. *IEEE Transactions on Medical Imaging*, 40(12):3955–3967, 2021. 2

[39] Jean-Laurent Mallet. Discrete smooth interpolation. *ACM Transactions on Graphics (TOG)*, 8(2):121–144, 1989. 2

[40] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*, 2023. 3

[41] 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*, pages 4296–4304, 2024. 6

[42] Jiteng Mu, Nuno Vasconcelos, and Xiaolong Wang. Editar: Unified conditional generation with autoregressive models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7899–7909, 2025. 6

[43] Gustav Müller-Franzes, Luisa Huck, Soroosh Tayebi Arasteh, Firas Khader, Tianyu Han, Volkmar Schulz, Ebba Dethlefsen, Jakob Nikolas Kather, Sven Nebelung, Teresa Nolte, et al. Using machine learning to reduce the need for contrast agents in breast mri through synthetic images. *Radiology*, 307(3):e222211, 2023. 1, 3

[44] Gustav Müller-Franzes, Luisa Huck, Maike Bode, Sven Nebelung, Christiane Kuhl, Daniel Truhn, and Teresa Lemainque. Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast mri. *European Radiology Experimental*, 8(1):53, 2024. 1

[45] Gowtham Murugesan, F Yu Fang, Michael Achilleos, John DeBevits, Sahil Nalawade, Chandan Ganesh, Ben Wagner, Ananth J Madhuranthakam, and Joseph A Maldjian. Synthesizing contrast-enhanced mr images from noncontrast mr images using deep learning. *American Journal of Neuroradiology*, 45(3):312–319, 2024. 3

[46] Gowtham Murugesan, Fang F. Yu, Michael Achilleos, John DeBevits, Sahil Nalawade, Chandan Ganesh, Ben Wagner, Ananth J Madhuranthakam, and Joseph A. Maldjian. Synthesizing contrast-enhanced mr images from noncontrast mr images using deep learning. *American Journal of Neuroradiology*, 45(3):312–319, 2024. 1

[47] Keisuke Nitta, Koji Matsumoto, Hajime Yokota, Taisuke Murata, Yoshitada Masuda, and Takashi Uno. Relationship between patient-friendly audiovisual systems and mri contrast agent to adverse reactions. *Journal of Magnetic Resonance Imaging*, 59(6):2013–2020, 2024. 3

[48] Masashi Okada and Tadahiro Taniguchi. Dreamingv2: Reinforcement learning with discrete world models without reconstruction. In *2022 IEEE/RSJ International Conference*

on Intelligent Robots and Systems (IROS), pages 985–991. IEEE, 2022. 2, 3

[49] Richard Osuala, Smriti Joshi, Apostolia Tsirikoglou, Lidia Garrucho, Walter HL Pinaya, Oliver Diaz, and Karim Lekadir. Pre-to post-contrast breast mri synthesis for enhanced tumour segmentation. In *Medical Imaging 2024: Image Processing*, pages 226–237. SPIE, 2024. 3

[50] 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*, pages 713–723. Springer, 2024. 5, 6

[51] Richard Osuala, Smriti Joshi, Apostolia Tsirikoglou, Lidia Garrucho, Walter HL Pinaya, Daniel M Lang, Julia A Schnabel, Oliver Diaz, and Karim Lekadir. Simulating dynamic tumor contrast enhancement in breast mri using conditional generative adversarial networks. *Journal of Medical Imaging*, 12(S2):S22014–S22014, 2025. 1, 3

[52] Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. *Advances in neural information processing systems*, 35:23178–23191, 2022. 3

[53] Vu Minh Hieu Phan, Zhibin Liao, Johan W Verjans, and Minh-Son To. Structure-preserving synthesis: Maskgan for unpaired mr-ct translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–65. Springer, 2023. 2

[54] Donald B Plewes and Walter Kucharczyk. Physics of mri: a primer. *Journal of magnetic resonance imaging*, 35(5): 1038–1054, 2012. 2

[55] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021. 3

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[57] Sriprabha Ramanarayanan, Arunima Sarkar, Matcha Naga Gayathri, Keerthi Ram, Mohanasankar Sivaprakasam, 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*, pages 5174–5183, 2024. 3

[58] Sriprabha Ramanarayanan, Kishore Kumar, Keerthi Ram, Harsh Agarwal, Ramesh Venkatesan, Mohanasankar Sivaprakasam, et al. Dcetriformer: A hybrid attention transformer for dce-mri synthesis in prostate imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1062–1071, 2025. 3

[59] 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*, pages 680–690. Springer, 2025. 3

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 6

[61] Tankred Saanum, Peter Dayan, and Eric Schulz. Simplifying latent dynamics with softly state-invariant world models. *Advances in Neural Information Processing Systems*, 37:38355–38382, 2024. 3

[62] S Sadhana, Sriprabha Ramanarayanan, Arunima Sarkar, Keerthi Ram, Matcha Naga Gayathri, Suresh Joel, Harsh Agarwal, Ramesh Venkatesan, and Mohanasankar Sivaprakasam. Dce-former: A transformer-based model with mutual information and frequency-based loss functions for early and late response prediction in prostate dce-mri. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 3

[63] 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, 2021. 5

[64] Prasad R Shankar, Kushal Parikh, and Matthew S Davenport. Financial implications of revised acr guidelines for estimated glomerular filtration rate testing before contrast-enhanced mri. *Journal of the American College of Radiology*, 15(2):250–257, 2018. 1, 3

[65] Travis B Smith. Mri artifacts and correction strategies. *Imaging in Medicine*, 2(4):445, 2010. 2

[66] Gustav J Strijkers, Willem J M Mulder, Geralda A F van Tilborg, and Klaas Nicolay. Mri contrast agents: current status and future perspectives. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*, 7(3):291–305, 2007. 1

[67] Jiabin Tang, Lianghao Xia, Jie Hu, and Chao Huang. Spatiotemporal meta contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2412–2421, 2023. 3

[68] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. 3

[69] Takahiro Yamada, Takayuki Masui, Masako Sasaki, Motoyuki Katayama, Yuji Iwadate, Naoyuki Takei, and Mitsuharu Miyoshi. Time resolved dce-mri of the kidneys: Evaluation of the renal vasculatures and tumors using f-disco with and without compressed sensing in normal and wide-bore 3t systems. *Medicine*, 101(31):e29971, 2022. 2

[70] Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, et al. Medical world model: Generative simulation of tumor evolution for treatment planning. *arXiv preprint arXiv:2506.02327*, 2025. 3

[71] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Ur-

tasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 3

[72] Zhuoran Yang, Xi Guo, Chenjing Ding, Chiyu Wang, Wei Wu, and Yanyong Zhang. Instadrive: Instance-aware driving world models for realistic and consistent video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25410–25420, 2025. 3

[73] Liangzhe Yuan, Rui Qian, Yin Cui, Boqing Gong, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, and Ting Liu. Contextualized spatio-temporal contrastive learning with self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13977–13986, 2022. 3

[74] 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 Computer Vision and Pattern Recognition Conference*, pages 25993–26003, 2025. 3

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3, 5, 6

[76] Zhang Zhang and Dacheng Tao. Slow feature analysis for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):436–450, 2012. 3

[77] Clement Zotti, Zhiming Luo, Alain Lalande, and Pierre-Marc Jodoin. Convolutional neural network with shape prior applied to cardiac mri segmentation. *IEEE journal of biomedical and health informatics*, 23(3):1119–1128, 2018. 2