# 3D MRI IMAGE PRETRAINING THROUGH CONTROLLABLE 2D SLICE NAVIGATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Self-supervised pretraining has become the mainstream approach for learning MRI representations from unlabeled scans. However, most existing objectives treat each scan as static volumes, which is inefficient use of the scans. We ask whether there exists an intrinsic self-supervision signal that is different from reconstructing the masked patches to assist improving representation quality? Through transforming the 3D volumes into controllable 2D dynamic sequences by rendering slices at continuous positions, orientations, and scales, we extract the action trajectories as the advantageous signal. We study this formulation with an action-conditioned pretraining objective, where a tokenizer encodes slice observations and a latent dynamics model predicts the evolution of latent features. Across representative anatomical and spatial downstream tasks, the proposed pretraining is evaluated against other strong 3D MRI self-supervision or supervision baselines, and ablation modules of ours. These results suggest that controllable MRI slice navigation provides a useful complementary pretraining interface for learning anatomical and spatial representations from large unlabeled MRI collections.

## 1 INTRODUCTION

Self-supervised pretraining has become a central strategy for learning 3D representations of Magnetic Resonance Imaging (MRI) from unlabeled scans (Zeng et al., 2024; Wald et al., 2025; Hatamizadeh et al., 2022). However, many existing objectives treat the MRI as collections of static observations, leaving relationships across viewing states underused(Wald et al., 2025; Zeng et al., 2024; Bai et al., 2019b). This leaves a useful source of supervision underexploited: MRI is volumetric in 3D, *while clinical inspection often proceeds 2D slices and changes of viewing plane* (Crowe et al., 2018; Nakashima et al., 2016). This motivates us to ask a focused research question: Can the controlled transitions in how a fixed MRI volume is viewed serve as a useful self-supervised signal for pretraining?

The transition signals are essentially inter-frame signals, which have been operationalized in two related ways. Video representation learning uses temporal context to predict latent content across frames, encouraging representations to capture dependencies beyond individual images. Action-conditioned world models go further by predicting how latent states evolve under actions, explicitly linking controls to changes in observations (Assran et al., 2025; Mur-Labadia et al., 2026; Maes et al., 2026; Hafner et al., 2025a;b). These approaches show how transitions can supervise representations through temporal prediction and action-conditioned dynamics.

In MRI, however, a transition between slices does not reflect physical changes in the anatomy: it arises from changing the viewing state of the same fixed volume. Figure 1 provides a motivating example: under controlled rotations of the test volumes, the representation learned with recorded viewing controls exhibits a substantially less relative Dice degradation than static baselines. This observation motivates us to ask how the known transition information should be incorporated into the pretraining objective.

Typically, scanning a 2D MRI slice requires a specific set of transition actions that defines its position, orientation and in-plane scaling factors. By applying a sequence of parameters, a 3D MRI volume can be transformed into a series of 2D image frames. These action parameters specify how the same anatomical structure is observed *from one viewing state to another*. Consequently, even an



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.

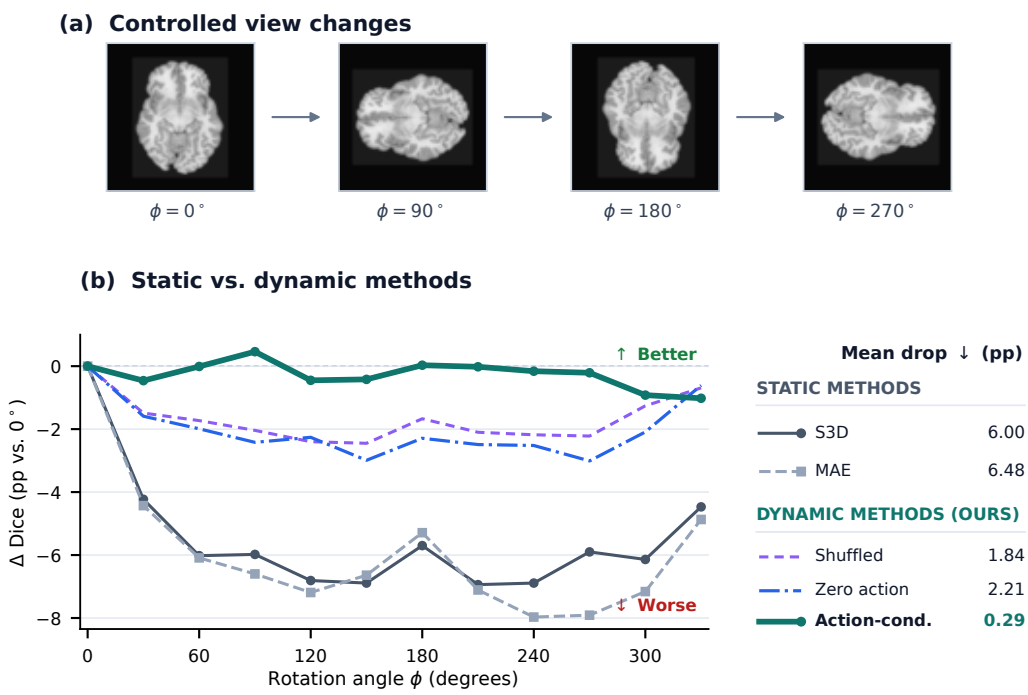


Figure 1: Rotation robustness under controlled view changes. (a) Illustrative in-plane rotations; the quantitative study rotates each 3D MRI volume about the $x$, $y$, and $z$ axes. (b) Change in tissue-segmentation Dice relative to $0^\circ$, caused by the rotated input, averaged across the three axes.

unlabeled MRI volume can yield informative frames, containing and offering insights about spatial and structural relationships (Zhang et al., 2024; Bai et al., 2019a;b).

To accurately model the relationship between action signals and observed frames and thereby improve representation quality, we formulate the following pretraining task: given a 2D MRI slice video and its recorded slice actions, the pretraining model learns to predict the current slice representation in the latent space. Action-conditioned latent denoising (Acuaviva et al., 2025) offers a useful starting point for learning from sequential observations, but its objective cannot be transferred to MRI slice sequences without clarifying the role of the actions. In many video settings, actions are used to model how a scene evolves over time. In our setting, the underlying anatomy remains fixed: an action changes only the slice plane, not the anatomy itself. Consequently, treating these controls as physical actions would misrepresent the transitions we want to learn from.

Inspired by clinical inspection across viewing planes, we reproduce the MRI rendering mechanism in the video game engine to produce the slice-action sequences. As illustrated in figure 2, we implement our proposed formulation with the re-rendering process and a two-stage action-conditioned pretraining pipeline. First, a MAE tokenizer is trained on rendered 2D MRI slices to produce spatial latent tokens. To capture the anatomy between voxels, we increase the patch size, yielding a rich spatial representation with fewer input patches for high-resolution slices. Second, with the tokenizer frozen, a causal latent denoising model receives historical slice tokens, actions, and a corrupted current latent, and learns to recover the clean latent of the current slice. Different with other action-conditioned methods, we remove the policy-oriented training objectives, focusing learning on the representation of controlled viewpoint changes for extracting the meaningful representation. At finetune stage, we compare tokenizer features with dynamics features, where the frozen dynamics model provides action-conditioned context from the slice video for dense downstream prediction. Thus, the action is used to add controllable view-transition information.

We evaluate this formulation as a pretraining method, focusing on downstream probes that directly reflect its intended benefits. Brain-region segmentation and tissue segmentation test anatomical structure, while dense coordinate-field prediction tests whether the representation preserves 3D

spatial organization. To isolate the role of action-conditioned dynamics, we compare tokenizer-only features, zero-action dynamics, shuffled-action variants, and correctly action-conditioned dynamics. For segmentation, we further compare against static-volume and fixed-axis slice baselines (Wald et al., 2025; Hatamizadeh et al., 2022). Together, these experiments provide evidence that the proposed controllable slice dynamics provides a useful pretraining signal for MRI pretraining.

Our contributions are threefold:

- We define a controllable 2D slice-navigation formulation for 3D MRI pretraining, turning each unlabeled volume into many slice sequences with recorded controls over slice position, orientation, and scale.
- We develop an action-conditioned pretraining pipeline that combines slice tokenization with latent dynamics, enabling dense downstream prediction from tokenizer-plus-dynamics features.
- We provide targeted evidence that action-conditioned slice dynamics is a useful signal for MRI pretraining, using anatomical segmentation, dense coordinate prediction, action-alignment diagnostics, and comparisons with static-volume and fixed-axis baselines.

## 2 Related Work

**Self-supervised and spatially aware medical representation learning.** Self-supervised pretraining originated as a response to the annotation bottleneck in medical imaging, and it has since become a standard route for learning transferable representations from unlabeled scans. (Zeng et al., 2024; Wald et al., 2025; Hatamizadeh et al., 2022) Recent reviews and foundation-model studies summarize a broad family of objectives, including reconstruction, contrastive learning, masked modeling, distillation, and hybrid pretraining across modalities (Zeng et al., 2024; Zhang & Metaxas, 2024). In MRI and other volumetric settings, large-scale efforts such as LVM-Med, 3D masked autoencoding, and 3DINO further show that unlabeled medical volumes can support strong and transferable representations at scale (Xu et al., 2025; Lang et al., 2023; M. H. Nguyen et al., 2023). A complementary line of work originated from the observation that medical images often come with structured geometry: anatomy-oriented imaging planes and intrinsic spatial information can be exploited to pretrain representations that encode relative plane relations or physical crop location (Zhang et al., 2024; Bai et al., 2019a;b). However, these methods still treat geometry mainly as a static target or a spatial proxy. In contrast, we ask whether the geometry of MRI observation itself can be turned into a dense pretraining signal, so that a 3D volume becomes a collection of aligned observation transitions rather than isolated static views.

**World models and medical world modeling.** World models originated in model-based reinforcement learning, where latent dynamics are learned from action-observation trajectories rather than from independent frames, and recent predictive representation learning has reinforced the value of forecasting latent content instead of reconstructing pixels (Battaglia et al., 2013; Zhou et al., 2021; Hafner et al., 2025a). Recent medical work has begun to explore related formulations: CheXWorld studies image world modeling for radiographs, Medical World Model simulates disease evolution under clinical treatment decisions, and Cardiac Copilot uses a world model for echocardiographic probe guidance (Yue et al., 2025; Yang et al., 2025; Jiang et al., 2024). These studies establish that world-model formulations can be useful in medicine, but their action spaces are tied to clinical intervention, disease progression, or acquisition policy. Our setting is narrower and more data-centric: we treat the slice position, orientation, and scale as retrospective observation controls over a fixed MRI volume, and we study whether action-conditioned latent dynamics can serve as a complementary self-supervised objective for learning reusable anatomical and spatial representations.

## 3 Methods

### 3.1 Overall

Our pretraining framework consists of data preparation, tokenizer pretraining, and latent dynamics pretraining. Figure 3 illustrates the novel preparation pipeline. We load each 3D MRI volume with

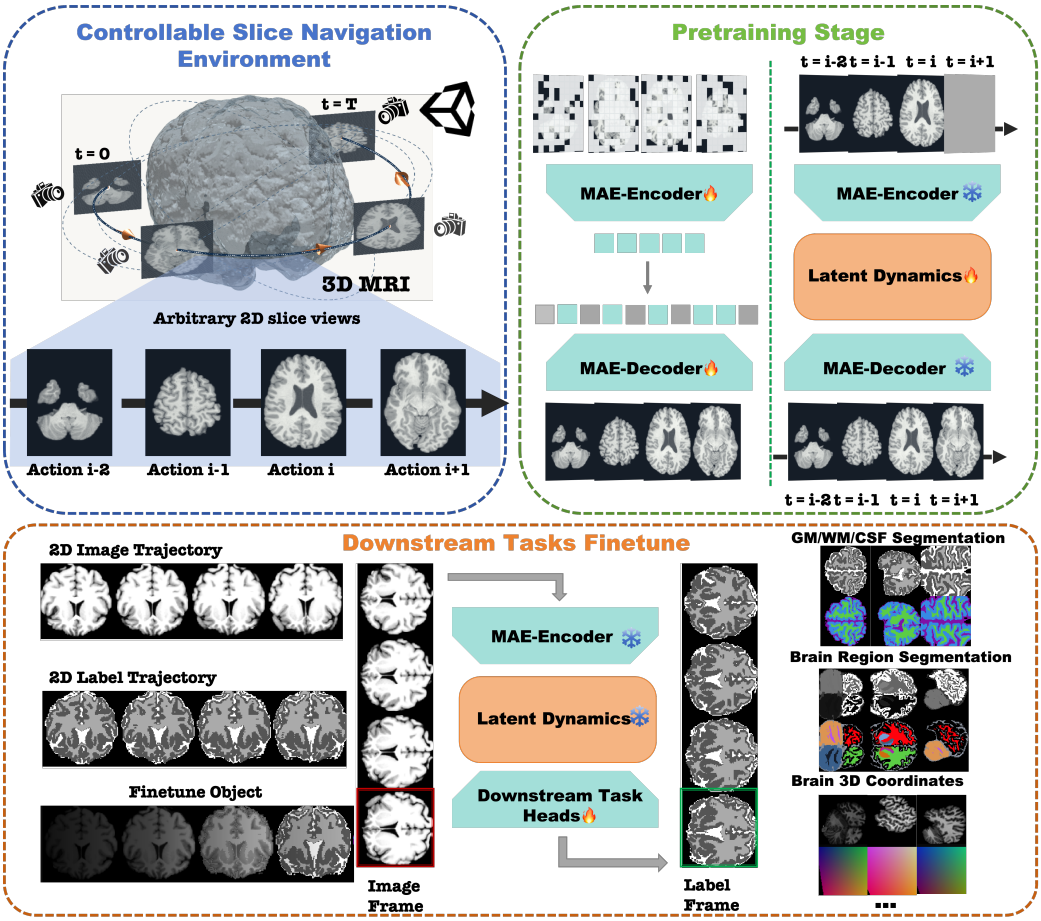


Figure 2: Overview of controllable 2D slice navigation for 3D MRI pretraining. A fixed 3D MRI volume is re-rendered in Unity into arbitrary 2D slice videos, where each slice frame is generated by a recorded action. The resulting slice-action sequences support tokenizer pretraining, action-conditioned latent dynamics pretraining, and downstream anatomical and spatial probes.

SimpleITK (Lowekamp et al., 2013) and render it in Unity using the open-source UnityVolumeRendering (Lavik, 2019). The resulting slice-action sequences are used to train a tokenizer and an action-conditioned latent dynamics model (Figure 2). During downstream transfer, both pretrained models are frozen while task-specific heads are trained.

### 3.2 Controllable MRI slice trajectories

Different with current methods which treat MRI as the fixed 3D volumes or 2D slices, Our pretraining framework is built on the slice-action sequences like videos. Let $V_i \in \mathbb{R}^{H\times W\times D}$ be the MRI volume of subject $i$. At trajectory time $t$, a renderer maps the viewing control $a_{i,t}$ to a 2D slice:

$$x_{i,t} = \mathcal{R}(V_i, a_{i,t}; \theta), \qquad a_{i,t} = (p_x, p_y, p_z, r_{\text{pitch}}, r_{\text{yaw}}, r_{\text{roll}}, s_x, s_z)_{i,t} \in [-1, 1]^8. \tag{1}$$

Here, $a_{i,t}$ specifies slice position, orientation, and in-plane scale. We use "action" to denote this viewing control. Because the underlying volume remains fixed, these controls specify changes in observation rather than changes in anatomy. They therefore provide an explicit geometric cue for relating views of the same anatomy. A trajectory and the pretraining dataset are

$$\tau_i = \{(x_{i,t}, a_{i,t})\}_{t=0}^{T_i}, \qquad \mathcal{D}_{\text{pre}} = \{\tau_i\}_{i=1}^{N}. \tag{2}$$

Each $a_{i,t}$ is aligned with the slice it renders, providing action-observation supervision without manual labels. For a downstream target volume $Y_i$, we replay the same controls to obtain aligned targets:

$$y_{i,t} = \mathcal{R}(Y_i, a_{i,t}; \theta). \tag{3}$$

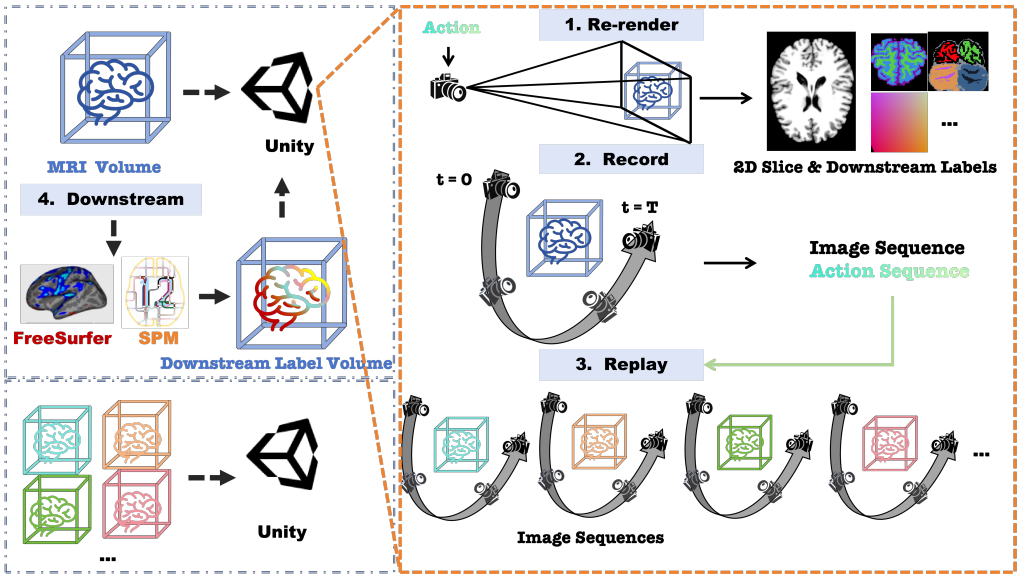


Figure 3: Data preparation pipeline for 2D slice-action and downstream supervision. (1) Re-render: each T1 MRI volume is loaded with SimpleITK and re-rendered in Unity into a 2D slice. (2) Record: the sequence is recorded in Unity. (3) Replay: the recorded sequence is replayed on MRI volumes and downstream tasks label volumes.

### 3.3 World-model pretraining objective

We use a two-stage pipeline: a tokenizer learns slice-level representations to provide a latent space, then a causal dynamics model trained to denoise the current slice latent using trajectory history and viewing controls for learning the dynamics in the latent space.

For slice $x_{i,t}$, let $P(x_{i,t}) \in \mathbb{R}^{N_p \times d_p}$ be its patchified representation. At each encoder call, we randomly sample a binary patch mask $m_{i,t} \in \{0,1\}^{N_p}$, where $m_{i,t,j} = 1$ indicates a masked patch. The tokenizer encoder $E_\psi$ and decoder $G_\psi$ produce

$$z^{\text{tok}}_{i,t} = E_\psi(P(x_{i,t}), m_{i,t}), \qquad \hat{P}_{i,t} = G_\psi(z^{\text{tok}}_{i,t}, m_{i,t}). \tag{4}$$

We train the tokenizer using masked reconstruction and a perceptual term:

$$\mathcal{L}_{\text{mae}} = \mathbb{E}_{x,m}\left[\frac{\sum_{j=1}^{N_p} m_j \|\hat{P}_j - P(x)_j\|_2^2}{\sum_{j=1}^{N_p} m_j}\right],$$
$$\mathcal{L}_{\text{token}} = \mathcal{L}_{\text{mae}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}}\big(\text{Unpatchify}(\hat{P}_{i,t}), x_{i,t}\big). \tag{5}$$

With the masked autoencoder, we could provide a semantic and compressed latent space for training the dynamics model. After tokenizer pretraining, we freeze $E_\psi$ and pack its output into the dynamics latent:

$$z_{i,t} = \text{Pack}\big(z^{\text{tok}}_{i,t}\big) = \text{Pack}\big(E_\psi(P(x_{i,t}), m_{i,t})\big), \qquad z_{i,t} \in \mathbb{R}^{S \times d_z}. \tag{6}$$

Here, $\text{Pack}(\cdot)$ is the spatial packing operation. The dynamics target is the frozen-tokenizer latent of the current slice.

To explicitly capture the anatomical representation constructed by the action-slice relationship, the dynamics target is the frozen-tokenizer latent of the current slice. We use $t$ for trajectory time and $d_t$ for the dynamics denoising step associated with frame $t$. The denoising step is independent of the tokenizer mask. For signal level $\sigma_{d_t} \in [0,1]$, the corrupted latent is

$$\tilde{z}^{(d_t)}_{i,t} = (1-\sigma_{d_t})\epsilon^{(d_t)}_{i,t} + \sigma_{d_t} z_{i,t}, \qquad \epsilon^{(d_t)}_{i,t} \sim \mathcal{N}(0, I). \tag{7}$$

Thus, $\sigma_{d_t} = 0$ gives pure noise and $\sigma_{d_t} = 1$ gives the clean latent. The shortcut embedding $\Delta_{d_t}$ specifies the denoising step size, not a temporal offset.

The causal dynamics model $F_\phi$ uses preceding views and the current viewing control and corrupted latent to predict the clean latent of the current slice:

$$\hat{z}_{i,t}^{(d_t)} = F_\phi\Big(\underbrace{\{\tilde{z}_{i,s}^{(d_s)}, a_{i,s}, \sigma_{d_s}, \Delta_{d_s}\}_{s<t}}_{\text{history}}; \underbrace{\tilde{z}_{i,t}^{(d_t)}, a_{i,t}, \sigma_{d_t}, \Delta_{d_t}}_{\text{current denoising slot}}\Big). \tag{8}$$

Its empirical objective is

$$\mathcal{L}_{\text{emp}} = \mathbb{E}_{\tau,t,d}\left[w(\sigma_{d_t})\|\hat{z}_{i,t}^{(d_t)} - z_{i,t}\|_2^2\right]. \tag{9}$$

### 3.4 Transfer to downstream tasks

For a $K$-frame clip ending at trajectory time $t$,

$$c_{i,t} = \{(x_{i,s}, a_{i,s})\}_{s=t-K+1}^{t}, \qquad h_{i,t} = \rho(c_{i,t}), \qquad \hat{y}_{i,t} = g_\omega(h_{i,t}), \tag{10}$$

where $\rho$ is the frozen tokenizer-dynamics backbone and $g_\omega$ is a lightweight task-specific head. The target $y_{i,t}$ is rendered from $Y_i$ using the same viewing control as in Eq. 3.

## 4 Experiments

Our experiments evaluate whether controllable MRI slice navigation provides a useful pretraining signal for anatomical and spatial representation learning. The experiments are organized around four questions. First, does the pretrained representation improve anatomical dense prediction compared with baselines and other methods? Second, how much slice-video context is needed, and what trade-off does context length create between accuracy, inference time, and memory cost? Third, does the dynamics model actually use the recorded slice actions during pretraining? Fourth, does the learned representation preserve 3D spatial structure beyond local slice appearance?

### 4.1 Experimental setup

We use 184 T1-weighted MRI volumes from the HCP2 dataset (Van Essen et al., 2012). All experiments use subject-level splits with a fixed random seed: 148 subjects for training, 18 for validation, and 18 for testing. The test subjects are held out from both self-supervised pretraining and downstream training. Volumes are loaded with a SimpleITK-based NIfTI pipeline (Lowekamp et al., 2013); the Unity rendering and target replay pipeline is summarized in Figure 3.

Unless otherwise specified, all the experiments have been repeated at least three times, rendered slice frames are resized to $256 \times 256$ and paired with 8-dimensional slice states encoding position, orientation, and in-plane scale. The tokenizer is pretrained on grayscale slice videos with patch size 16. The dynamics model is trained on strided slice-action sequences with frame stride 4. For downstream transfer, we train lightweight prediction heads on frozen pretrained features, sweep clip lengths $K \in \{2, 4, 8, 16\}$, and select models by validation performance.

We evaluate three dense probes. Brain-region segmentation uses FreeSurfer-derived parcellation labels replayed with the same slice states, and evaluates whether the representation captures anatomical region structure (Fischl, 2012). GM/WM/CSF segmentation uses SPM-derived tissue labels and evaluates tissue-level anatomical information (Ashburner, 2009). Dense 3D coordinate-field prediction uses Unity-recorded coordinates for valid slice pixels and evaluates whether the representation preserves spatial organization in the underlying 3D volume. Segmentation performance is reported with Dice and mIoU, while coordinate prediction is reported with valid-mask MAE and per-axis MAE. Task-specific losses, valid-pixel handling, and metric definitions are provided in the supplementary material.

We compare tokenizer-only pretraining, zero-action dynamics, and action-conditioned dynamics. For segmentation, we additionally include static-volume baselines: supervised UNETR (Hatamizadeh et al., 2022) and S3D-style 3D self-supervised pretraining (Wald et al., 2025). These baselines test whether controllable slice-trajectory pretraining provides value beyond conventional 3D volume learning and fixed-axis 2D slice segmentation. For the converged comparison with 3D baselines, all methods use the same subject split and full-volume protocol. Our slice model traverses the z axis

Table 1: Comparison with 3D baselines after convergence under a common full-volume evaluation protocol. SC denotes a skip-connection decoder. All methods use the same subject split; every voxel contributes once to evaluation.

| | | GM/WM/CSF | | Brain region | |
|---|---|---|---|---|---|
| Method | Decoder | Dice ↑ | mIoU ↑ | Dice ↑ | mIoU ↑ |
| UNETR (Hatamizadeh et al., 2022) | SC | 0.8804 | 0.7935 | 0.6768 | 0.5284 |
| S3D (Wald et al., 2025) | SC | 0.8964 | 0.8185 | 0.7635 | **0.6297** |
| Ours | normal | 0.8552 | 0.7552 | 0.7380 | 0.5999 |
| Ours | SC | **0.9073** | **0.8942** | **0.8016** | 0.6105 |

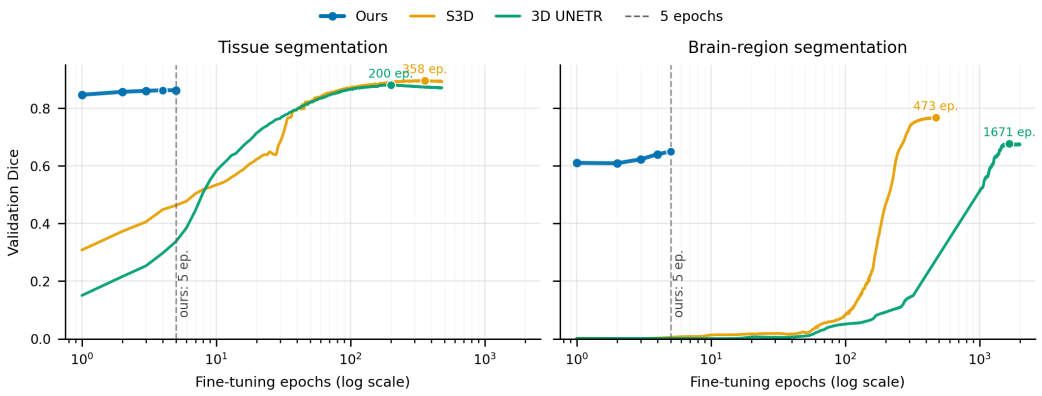


Figure 4: Fine-tuning efficiency on tissue and brain-region segmentation. Our action-conditioned world-model representation reaches strong downstream performance within 5 epochs, whereas static 3D baselines require hundreds to thousands of fine-tuning epochs. For S3D, only supervised fine-tuning epochs are counted; its masked 3D pretraining epochs are not included. Curves are parsed from the downloaded training logs.

once so that, as for the volume baselines, every voxel contributes exactly once to evaluation. The three-seed experiment fixes the split, pretrained checkpoints, training schedule, and hyperparameters while varying only downstream sample order and minibatch composition.

### 4.2 Main downstream results

Table 1 compares segmentation performance with the supervised static-volume segmentation method 3D UNETR (Hatamizadeh et al., 2022) and the self-supervised and fine-tuned static-volume segmentation method S3D (Wald et al., 2025). With the matched skip-connection decoder, our method achieves the highest Dice on GM/WM/CSF (0.9073) and brain-region segmentation (0.8016), as well as the highest GM/WM/CSF mIoU (0.8942). S3D remains strongest on brain-region mIoU (0.6297 versus 0.6105). The improvement from our normal decoder to the skip-connection decoder further indicates that the original lightweight decoder underused the pretrained representation. These results support the main hypothesis that correct actions provide useful transition supervision beyond static tokenizer pretraining.

As shown in Figure 4, our downstream adaptation trains lightweight heads for only 5 epochs with frozen pretrained modules, whereas the static 3D baselines are optimized for substantially longer schedules, with brain-region UNETR trained for 2000 epochs and S3D fine-tuned for 480 epochs. The key observation is that controllable slice dynamics already yields measurable gains under a short adaptation budget, suggesting that action-conditioned pretraining provides a useful signal for efficient downstream adaptation. The visualized results are shown in Figure 6.

Table 2: Effect of slice-video context length for action-conditioned dynamics. Accuracy is reported on the test set. Latency and memory are measured on a single NVIDIA A800 with batch size 1 and averaged across the three downstream probes.

| $L$ | Accuracy | | | Cost | |
|---|---|---|---|---|---|
| | Region Dice ↑ | Tissue Dice ↑ | Coord. MAE ↓ | Lat. (ms) | Mem. (GB) |
| 2 | 0.5927 | 0.8484 | 0.0100 | 16.87 | 0.380 |
| 4 | 0.5919 | 0.8493 | 0.0099 | 16.77 | 0.381 |
| 8 | 0.5899 | 0.8482 | 0.0093 | 16.65 | 0.382 |
| 16 | **0.6749** | **0.8542** | **0.0085** | 16.92 | 0.433 |

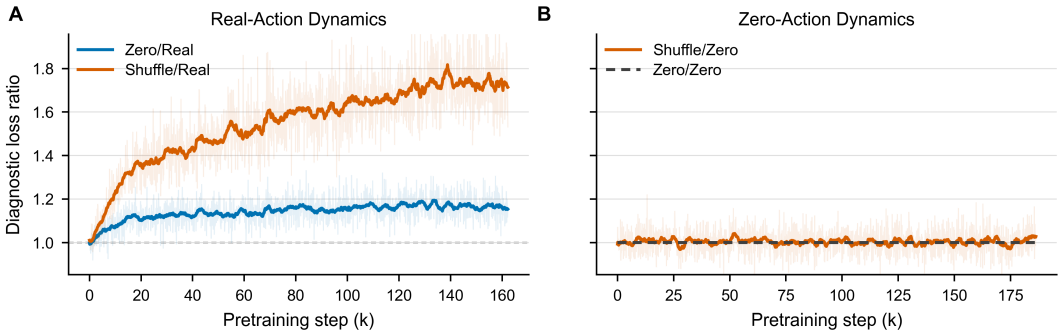


Figure 5: Action-use diagnostics during dynamics pretraining. In the Real-Action Dynamics model, we report diagnostic loss ratios under Zero and Shuffle actions relative to Real actions, denoted as Zero/Real and Shuffle/Real. In the Zero-Action Dynamics model, we use Zero actions as the denominator and report Shuffle/Zero and Zero/Zero. Curves show a centered moving average with translucent raw traces.

### 4.3 Effect of slice-video context length

As shown in Table 2, we change $L$, the number of context frames, to observe the performance of downstream tasks. The short context frames already yield competitive tissue segmentation and coordinate prediction performance, but the strongest overall results are obtained at $L = 16$. Increasing $L$ from 2 to 16 improves region Dice from 0.5927 to 0.6749, tissue Dice from 0.8484 to 0.8542, and coordinate MAE from 0.0100 to 0.0085. The improvement is especially pronounced, suggesting that longer slice trajectories provide additional anatomical and structural information for downstream tasks. Besides, the computational overhead is modest in this setting. Latency remains nearly unchanged across context lengths, ranging from 16.65 ms to 16.92 ms on a single A800, while memory increases from 0.380 GB at $L = 2$ to 0.433 GB at $L = 16$.

### 4.4 Action-use diagnostics

We further ask whether the dynamics model actually uses the slice actions during the pretraining. To test this, we replace only the action used by the dynamics model every single training step, but calculate gradients only with the correct action. For the real-action model, we compare three inputs: the recorded actions, zero actions, and shuffled actions. If the model relies on correct action-observation alignment, replacing the recorded actions should increase the latent prediction loss. As a control, we apply the same test to a zero-action model, which is trained without informative actions.

Figure 5 reports the loss ratios during training. In the real-action model, zero or shuffled actions produce higher prediction loss than the recorded actions, showing that the model learns to use the slice actions. In the zero-action model, the ratios remain much closer to one, indicating weaker dependence on the action input. These results support that the benefit of our dynamics pretraining comes from correct dynamics.

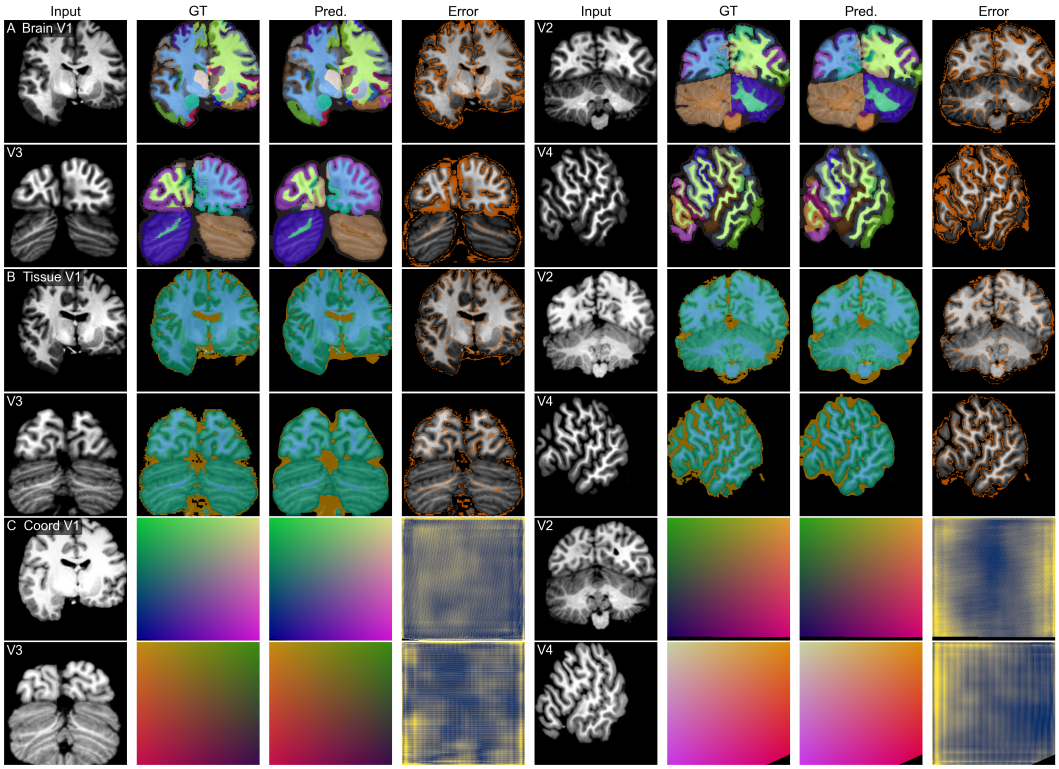


Figure 6: The visualized results of the brain region segmentation, the tissue segmentation, and the dense 3D coordinates understanding tasks.

## 5 Conclusion and Limitations

We studied whether controllable MRI slice navigation can serve as a self-supervised pretraining. By viewing a 3D MRI volume as an environment that can be observed through controllable slice states, each unlabeled scan can generate dense action-slice trajectories. We instantiated this idea with a world-model pretraining objective, where a tokenizer learns compact slice representations and an action-conditioned dynamics model learns latent transitions between anatomical views. Through downstream anatomical and spatial probes, together with action ablations, our results provide initial evidence that controllable slice dynamics is a useful signal for MRI representation learning.

This work has several limitations. First, our evaluation is intentionally focused: we use a limited set of dense downstream probes rather than a broad benchmark across MRI modalities, organs, and clinical tasks. Second, the current experiments are centered on HCP2 T1 MRI, so the generality of the formulation to other contrasts, scanners, pathologies, and body regions remains open. Third, our current environment uses slice observation controls rather than richer interactive or physical actions.

Future work can extend this direction in several ways. A first step is to scale pretraining to larger and more diverse MRI collections, including multi-contrast and multi-site data. A second direction is to study richer controllable action spaces beyond geometric slice motion alone. Finally, downstream evaluation should be expanded to broader anatomical and clinical tasks, including low-label and cross-domain transfer settings. We hope this paper is an initial step toward using controllable 2D slice actions as powerful signals for 3D MRI image pretraining.

### AI use statement

We used OpenAI's Codex generative AI assistant to discuss manuscript organization and framing, revise language. It was not used to collect or generate MRI data, annotate scans, write or execute experimental code, run model training or evaluation, compute primary experimental metrics, or create or alter figures; no AI-generated measurements or images are reported. All AI-assisted text was

reviewed and edited by the authors. The authors take responsibility for the final manuscript and its claims.

### ETHICS STATEMENT

This work is a secondary analysis of previously collected HCP2 and MSD Task02 Heart MRI data. We did not recruit or contact participants, collect new data, or perform interventions. The data were used under the access and use conditions specified by their providers; consent and ethics procedures for the original data collections are those documented by the respective providers.

### REPRODUCIBILITY STATEMENT

The Methods and Appendix describe the subject splits, MRI rendering and target-replay pipeline, model architecture and objectives, training settings, downstream protocols, and evaluation metrics. The Appendix also reports the additional rotation, cross-dataset, cross-trajectory, and ablation experiments. HCP2 and MSD Task02 Heart data are obtained from their respective providers subject to the providers' access and use conditions.

## REFERENCES


Pablo Acuaviva, Aram Davtyan, Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Alexandre Alahi, and Paolo Favaro. From generation to generalization: Emergent few-shot learning in video diffusion models. *arXiv preprint arXiv:2506.07280*, 2025.

John Ashburner. Computational anatomy with the spm software. *Magnetic resonance imaging*, 27 (8):1163–1174, 2009.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 541–549, Cham, 2019a. Springer International Publishing. ISBN 978-3-030-32245-8.

Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International conference on medical image computing and computer-assisted intervention*, pp. 541–549. Springer, 2019b.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110(45):18327–18332, 2013.

Emily M Crowe, Iain D Gilchrist, and Christopher Kent. New approaches to the analysis of eye movement behaviour across expertise while viewing brain mris. *Cognitive research: principles and implications*, 3(1):12, 2018.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pp. 1–7, 2025a.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models, 2025b. URL https://arxiv.org/abs/2509.24527.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758, 2022. doi: 10.1109/WACV51458.2022.00181.

Haojun Jiang, Zhenguo Sun, Ning Jia, Meng Li, Yu Sun, Shaqi Luo, Shiji Song, and Gao Huang. Cardiac Copilot: Automatic Probe Guidance for Echocardiography with World Model . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15001. Springer Nature Switzerland, October 2024.

Daniel M Lang, Eli Schwartz, Cosmin I Bercea, Raja Giryes, and Julia A Schnabel. Multispectral 3d masked autoencoders for anomaly detection in non-contrast enhanced breast mri. In *MICCAI Workshop on Cancer Prevention through Early Detection*, pp. 55–67. Springer, 2023.

Matias Lavik. Unityvolumerendering: A volume renderer for unity3d. https://github.com/mlavik1/UnityVolumeRendering, 2019. Open-source software, MIT License, version 1.8.0.

Bradley C. Lowekamp, David T. Chen, Luis Ibanez, and Daniel Blezek. The design of simpleitk. *Frontiers in Neuroinformatics*, Volume 7 - 2013, 2013. ISSN 1662-5196. doi: 10.3389/fninf.2013.00045. URL https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2013.00045.

Duy M. H. Nguyen, Hoang Nguyen, Nghiem Diep, Tan Ngoc Pham, Tri Cao, Binh Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, Daniel Sonntag, and Mathias Niepert. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27922–27950. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/58cc11cda2a2679e8af5c6317aed0af8-Paper-Conference.pdf.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint*, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Ryoichi Nakashima, Yuya Komori, Eriko Maeda, Takeharu Yoshikawa, and Kazuhiko Yokosawa. Temporal characteristics of radiologists' and novices' lesion detection in viewing medical images presented rapidly and sequentially. *Frontiers in Psychology*, 7:1553, 2016.

D.C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T.E.J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S.W. Curtiss, S. Della Penna, D. Feinberg, M.F. Glasser, N. Harel, A.C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S.E. Petersen, F. Prior, B.L. Schlaggar, S.M. Smith, A.Z. Snyder, J. Xu, and E. Yacoub. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2012.02.018. URL https://www.sciencedirect.com/science/article/pii/S1053811912001954. Connectivity.

Tassilo Wald, Constantin Ulrich, Stanislav Lukyanenko, Andrei Goncharov, Alberto Paderno, Maximilian Miller, Leander Maerkisch, Paul Jaeger, and Klaus Maier-Hein. Revisiting mae pre-training for 3d medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5186–5196, 2025.

Tony Xu, Sepehr Hosseini, Chris Anderson, Anthony Rinaldi, Rahul G Krishnan, Anne L Martel, and Maged Goubran. A generalizable 3d framework and model for self-supervised learning in medical imaging. *npj Digital Medicine*, 8(1):639, 2025.

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 (ICCV)*, pp. 8319–8329, October 2025.

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 2025.

Xiangrui Zeng, Nibras Abdullah, and Putra Sumari. Self-supervised learning framework application for medical image analysis: a review and summary. *BioMedical Engineering OnLine*, 23(1):107, 2024.

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, 91:102996, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102996. URL https://www.sciencedirect.com/science/article/pii/S1361841523002566.

Tianwei Zhang, Dong Wei, Mengmeng Zhu, Shi Gu, and Yefeng Zheng. Self-supervised learning for medical image data with anatomy-oriented imaging planes. *Medical Image Analysis*, 94:103151, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103151. URL https://www.sciencedirect.com/science/article/pii/S1361841524000768.

S Kevin Zhou, Hoang Ngan Le, Khoa Luu, Hien V Nguyen, and Nicholas Ayache. Deep reinforcement learning in medical imaging: A literature review. *Medical image analysis*, 73:102193, 2021.

Table 3: Fixed subject-level split used in the HCP experiments.

| Split | Subjects | Percent | Use |
|---|---|---|---|
| Train | 148 | 80.4% | Pretraining and downstream-head training |
| Validation | 18 | 9.8% | Model selection |
| Test | 18 | 9.8% | Held-out evaluation |
| Total | 184 | 100.0% | |

Table 4: Eight-dimensional slice state/action.

| Index | Symbol | Group | Meaning |
|---|---|---|---|
| 1–3 | $(p_x, p_y, p_z)$ | Position | Slice-plane position in the renderer coordinate system. |
| 4–6 | $(r_{\text{pitch}}, r_{\text{yaw}}, r_{\text{roll}})$ | Orientation | Slice-plane orientation. |
| 7–8 | $(s_x, s_z)$ | In-plane scale | Scale along the two slice axes. |

# A Supplementary Details and Additional Experiments

This appendix provides implementation details for the data, rendering, training, and evaluation protocols, followed by additional experiments requested during review. Unless stated otherwise, experiments use the subject-level HCP split described below.

## A.1 Dataset, Rendering, and Targets

**Subject-level split.** We use 184 T1-weighted MRI volumes from the HCP2 source collection. Subjects are assigned once to train, validation, or test, so no subject contributes slices to multiple splits. The fixed split is used for both pretraining and downstream adaptation; test subjects are excluded from pretraining and model selection. The reviewer-requested analyses with a 36-subject HCP test set use a separate fixed split from the main-paper 148/18/18 split: 138 subjects are used for development and 36 are held out for testing; when validation is required, these 138 are split into 124 training and 14 validation subjects. Results from these protocols are reported separately and should not be compared as if they shared one split.

**Controllable rendering.** For subject volume $V_i$, the renderer produces a grayscale slice observation

$$x_t = \mathcal{R}(V_i, a_t; \theta), \tag{11}$$

where $\mathcal{R}$ is the Unity slice renderer, $a_t$ is the state of the slicing plane at frame $t$, and $\theta$ denotes fixed rendering settings. Observations are resized to $256 \times 256$. The action is an observation control over a fixed volume, not a clinical intervention. Its eight independently normalized components are listed in Table 4; values are clipped to $[-1, 1]$ at the model input. The aligned trajectory is $\tau_i = \{(x_t, a_t)\}_{t=0}^{T}$.

Downstream targets are aligned by replaying the same action sequence on the target volume $Y_i$:

$$y_t = \mathcal{R}(Y_i, a_t; \theta). \tag{12}$$

Segmentation labels use nearest-neighbor sampling. Brain-region labels are derived from FreeSurfer `aparc+aseg` and remapped to 105 classes in the reported runs; background is label 0. GM/WM/CSF labels are derived from SPM12 tissue maps and form four classes including background. Coordinate targets are generated from the slice geometry and include a valid-pixel mask. The normalized source-volume coordinate at pixel $(u, v)$ is

$$c(u, v) = 2\,\text{dataCoord}(u, v) - 1 \in [-1, 1]^3, \tag{13}$$

where a pixel is valid when all components of $\text{dataCoord}(u, v)$ lie in $[0, 1]$.

## A.2 Model and Training Details

**Tokenizer.** For tokenizer pretraining, $m_t \in \{0, 1\}^{N_p}$ is a randomly sampled patch mask, where $m_{t,j} = 1$ marks a patch selected for reconstruction. The tokenizer encodes the masked patches

Table 5: Downstream targets and their sources.

| Task | Target source | Target |
|---|---|---|
| Brain-region segmentation | FreeSurfer `aparc+aseg`, resampled to the T1 grid | Multi-class labels |
| GM/WM/CSF segmentation | SPM12 tissue probability maps | Four discrete classes including background |
| Dense coordinate field | Replayed slice geometry | Three normalized coordinates and a valid mask |

Table 6: Tokenizer configuration for the grayscale MRI experiments.

| Setting | Value | Notes |
|---|---|---|
| Input resolution | $256 \times 256 \times 1$ | Grayscale slice |
| Patch size | $16 \times 16$ | 256 patches per frame |
| Model width / heads / depth | 384 / 6 / 10 | Transformer encoder and decoder |
| Latent slots / dimension | 32 / 64 | Tokenizer bottleneck |
| Random mask rate | 0.10–0.60 | Sampled for each training example |
| Perceptual-loss weight | 0.02 | AlexNet LPIPS |
| Optimizer | AdamW | Learning rate $2 \times 10^{-4}$; weight decay $10^{-2}$ |
| Training schedule | Up to 300,000 steps | Project training-entrypoint setting |

into tokens and reconstructs the selected patches using the masked reconstruction and perceptual objectives in the main text. The mask $m_t$ is unrelated to the dynamics denoising condition below. Table 6 summarizes the MRI tokenizer settings.

**Latent dynamics.** After tokenizer pretraining, its encoder is frozen. We denote the packed tokenizer representation of the rendered frame by $z_t = \mathrm{Pack}(E_\psi(x_t))$. The sequence index $t$ always denotes a rendered observation. Independently, $d$ denotes the dynamics denoising condition, with signal level $\sigma_d$ and shortcut step size $\Delta_d$. The corrupted current-frame latent is

$$\tilde{z}_t^{(d)} = (1 - \sigma_d)\epsilon_t + \sigma_d z_t, \qquad \epsilon_t \sim \mathcal{N}(0, I). \tag{14}$$

Thus $\sigma_d = 0$ gives pure noise and $\sigma_d = 1$ gives the clean latent. The causal dynamics model uses the preceding slice-action history and the corrupted target slot, including the action that generated the current observation, to recover the clean latent of that same frame:

$$\hat{z}_t = F_\phi\Big(\{(\tilde{z}_s^{(d_s)}, a_s, \sigma_{d_s}, \Delta_{d_s})\}_{s<t}; \tilde{z}_t^{(d)}, a_t, \sigma_d, \Delta_d\Big). \tag{15}$$

This is current-frame latent denoising conditioned on history and the current slice state, not an image-space reconstruction objective. The empirical loss and shortcut self-consistency term are

$$\mathcal{L}_{\mathrm{dyn}} = \mathbb{E}_{\tau,t,d}\big[w(\sigma_d)\|\hat{z}_t - z_t\|_2^2\big] + \lambda_{\mathrm{sc}}\mathcal{L}_{\mathrm{sc}}. \tag{16}$$

The self-consistency term is shared across dynamics variants; the variants differ in whether they receive recorded, zero, or shuffled actions.

### A.3 Evaluation Metrics

Segmentation metrics are computed over valid pixels; background label 0 and ignored labels are excluded from foreground class averaging. For foreground class $c$,

$$\mathrm{Dice}_c = \frac{2|\hat{Y}_c \cap Y_c|}{|\hat{Y}_c| + |Y_c|}, \qquad \mathrm{IoU}_c = \frac{|\hat{Y}_c \cap Y_c|}{|\hat{Y}_c \cup Y_c|}. \tag{17}$$

Reported Dice and mIoU are arithmetic means over foreground classes present in the prediction or target after valid-mask filtering. Coordinate error is the mean absolute error over valid pixels and all three coordinate channels:

$$\mathrm{MAE}_{\mathrm{coord}} = \frac{1}{3|\Omega|}\sum_{(u,v)\in\Omega}\|\hat{c}(u,v) - c(u,v)\|_1, \tag{18}$$

where $\Omega$ is the valid-pixel set.

Table 7: Latent-dynamics and downstream settings.

| Setting | Value | Notes |
|---|---|---|
| Dynamics input | Frozen tokenizer latents | Trained in latent, not pixel, space |
| Action dimension | 8 | Position, orientation, and in-plane scale |
| Dynamics frame stride | 4 | Temporal subsampling during data preparation |
| Dynamics schedule | Up to 200,000 steps | Project entrypoint setting |
| Downstream clip length | $K \in \{2, 4, 8, 16\}$ | Frames sampled at stride 4 |
| Downstream training | Task-specific head | Pretrained tokenizer and dynamics are frozen |
| Segmentation loss | Weighted cross-entropy and soft Dice | Selected by validation performance |
| Coordinate loss | Valid-mask $L_1$ | Evaluated only on valid target pixels |

Table 8: Full rotation study. Values are changes in GM/WM/CSF Dice (percentage points) from the unrotated reference.

| Angle | S3D | Tokenizer | Zero-action | Shuffled-action | Action-conditioned |
|---|---|---|---|---|---|
| $0^\circ$ | 0 | 0 | 0 | 0 | 0 |
| $30^\circ$ | $-4.23$ | $-4.43$ | $-1.59$ | $-1.49$ | $-0.46$ |
| $60^\circ$ | $-6.02$ | $-6.09$ | $-1.99$ | $-1.73$ | $-0.01$ |
| $90^\circ$ | $-5.98$ | $-6.60$ | $-2.42$ | $-2.04$ | $+0.46$ |
| $120^\circ$ | $-6.81$ | $-7.19$ | $-2.26$ | $-2.40$ | $-0.45$ |
| $150^\circ$ | $-6.89$ | $-6.64$ | $-2.99$ | $-2.45$ | $-0.42$ |
| $180^\circ$ | $-5.70$ | $-5.29$ | $-2.29$ | $-1.67$ | $+0.03$ |
| $210^\circ$ | $-6.94$ | $-7.11$ | $-2.49$ | $-2.10$ | $-0.02$ |
| $240^\circ$ | $-6.89$ | $-7.97$ | $-2.52$ | $-2.18$ | $-0.16$ |
| $270^\circ$ | $-5.90$ | $-7.91$ | $-3.01$ | $-2.22$ | $-0.21$ |
| $300^\circ$ | $-6.14$ | $-7.16$ | $-2.09$ | $-1.26$ | $-0.92$ |
| $330^\circ$ | $-4.47$ | $-4.87$ | $-0.61$ | $-0.69$ | $-1.02$ |
| Mean drop | 6.00 | 6.48 | 2.21 | 1.84 | **0.29** |

## A.4 Additional Reviewer-Requested Experiments

### A.4.1 Rotation robustness

We rotate each of the 36 held-out HCP volumes about each of the three axes, render the transformed volume, and evaluate without retraining. Entries in Table 8 are GM/WM/CSF Dice changes in percentage points relative to $0^\circ$, averaged over axes. Mean drop is the negative mean change over the 11 non-zero angles. Dynamics improve robustness over the tokenizer; among the dynamics variants, action-conditioned dynamics have the smallest mean drop (0.29 points, versus 1.84 for shuffled actions and 2.21 for zero actions). This supports an advantage under this controlled orientation shift, not a uniform in-domain downstream gain.

A separate rotation comparison on 36 held-out subjects evaluates the action-conditioned model against S3D without retraining. Its mean Dice drop over non-zero rotations is 0.002825 for our model (95% CI: [0.001500, 0.004177]) and 0.059975 for S3D (95% CI: [0.058680, 0.061203]); the paired difference is $-0.057150$ ($p = 2.91 \times 10^{-11}$). This comparison measures rotation sensitivity only: S3D has higher absolute Dice at every angle (0.9039 versus 0.7810 at $0^\circ$), so it does not show superior absolute segmentation accuracy.

### A.4.2 Cross-dataset transfer

We transfer HCP-pretrained representations to left-atrial segmentation on MSD Task02 Heart, which differs in anatomy, acquisition domain, and dataset. All methods use the same 14/3/3 split and evaluation setting. Our method exceeds UNETR by 0.1242 Dice but remains 0.0826 below S3D. This is initial evidence of cross-dataset transfer, not evidence of broad clinical generality; the held-out test set contains only three volumes.

Table 9: Transfer to MSD Task02 Heart. Metric: left-atrial Dice.

| Method | Dice ↑ | Train/val/test |
|---|---|---|
| UNETR | 0.5510 | 14/3/3 |
| S3D | 0.7578 | 14/3/3 |
| Ours | 0.6752 | 14/3/3 |

Table 10: Fine-tuned subject-macro Dice across trajectory designs.

| Pretraining → downstream | Brain-region Dice ↑ | GM/WM/CSF Dice ↑ |
|---|---|---|
| Tr1 → Tr3 | 0.7380 | 0.8485 |
| Tr2 → Tr3 | 0.6957 | 0.8305 |
| Tr1 → Tr2 | 0.3006 | 0.8449 |
| Tr2 → Tr2 | 0.3045 | 0.8450 |

### A.4.3 Cross-trajectory transfer

Tr1 is the manually defined reference trajectory, Tr2 an independently generated random trajectory in the same position–orientation–scale space, and Tr3 deterministic native-resolution axial sampling. $\mathrm{TrA} \rightarrow \mathrm{TrB}$ denotes pretraining on TrA followed by downstream adaptation and evaluation on TrB. Both pretraining trajectories transfer to Tr3, which is not used for pretraining. On downstream Tr2, tissue Dice differs by only 0.0001 between Tr1 and Tr2 pretraining. The lower Tr2-to-Tr3 scores show that trajectory choice can still affect transfer quality; the Tr1-to-Tr2 and Tr2-to-Tr2 region results were low (0.3006 and 0.3045 Dice), so transfer is task-dependent.

### A.4.4 Action-injection ablation

The eight normalized action components are zero-padded to a shared 16-dimensional interface and projected by a two-layer MLP (hidden dimension $2D$) into an independent $D$-dimensional action token. The token interacts with visual latents through block-causal attention. We compare this design with injecting pose through the positional embedding. The independent token performs better on these five measures. The pose-encoding model was evaluated on the same axial test protocol, but the action-token reference was not retrained as a matched seed-level control; the result is descriptive and does not establish that geometric positional encodings are generally inferior.

### A.4.5 Automated trajectory generation

The method requires aligned action–observation pairs, not manual trajectory recording. The reference route defines one trajectory once and replays it over volumes; the two alternatives generate trajectories automatically, either by interpolating random valid-space keyframes or by sampling and resampling planes directly without Unity. Each route produces 500 frames at $256 \times 256$ from a $182 \times 182 \times 218$ volume. Measurements used an Intel i9-10940X CPU, 64 GiB RAM, and an RTX 3090; Unity 2022.3.62f3c1 was used where applicable. Random-keyframe generation has similar replay cost to the reference route; direct resampling avoids Unity but takes longer. At inference, a volume is sampled once along a predefined or automatically generated trajectory.

Table 11: Action-injection mechanisms under the same downstream evaluation protocol.

| | Brain region | | GM/WM/CSF | | Coordinate |
|---|---|---|---|---|---|
| Injection | Dice ↑ | mIoU ↑ | Dice ↑ | mIoU ↑ | MAE ↓ |
| Independent action token | **0.7380** | **0.5999** | **0.8552** | **0.7501** | **0.0308** |
| Pose-conditioned position | 0.7171 | 0.5747 | 0.8450 | 0.7459 | 0.0461 |

Table 12: Trajectory-generation cost per volume (mean $\pm$ standard deviation).

| Route | Manual input | CPU | Peak memory | Time |
|---|---|---|---|---|
| Reference definition + batch replay | Once, $\sim$3 min | 3.69% | 709.8 MiB | $3.500 \pm 0.126$ s |
| Random keyframes + interpolation | None | 3.64% | 707.5 MiB | $3.528 \pm 0.086$ s |
| Direct plane sampling/resampling | None | 3.55% | 694.1 MiB | $13.173 \pm 0.187$ s |

Table 13: Downstream performance with the original and Swin-UNETR tokenizers. The confidence interval is over three Swin-UNETR runs.

| Tokenizer setting | GM/WM/CSF Dice ↑ | Brain-region Dice ↑ |
|---|---|---|
| Original, five epochs | 0.8552 | 0.5736 |
| Swin-UNETR, five epochs | 0.7299 | 0.5964 |
| Swin-UNETR, three runs | $0.7256 \pm 0.0037$ | $0.6069 \pm 0.0092$ |
| 95% confidence interval | [0.7165, 0.7348] | [0.5842, 0.6296] |

### A.4.6 Alternative tokenizer backbone

We replaced the original tokenizer with Swin-UNETR while keeping the data and downstream protocol fixed. The five-epoch comparison and three-run Swin-UNETR summary are reported in Table 13. The alternative backbone is trainable and gives stable results, with a different trade-off between tissue-level and fine-grained region segmentation. This supports architectural portability but does not imply that tokenizer choice is inconsequential.