ULTRAWORLD: LEARNING INTERACTIVE ULTRASOUND WORLD MODELS FROM UNTRACKED CLINICAL VIDEOS WITH ACOUSTIC SAMPLING MAP

Anonymous authors
Paper under double-blind review

ABSTRACT

World models can support autonomous ultrasound scanning by predicting probe-motion outcomes from local observations. Learning this action–observation relationship typically requires synchronized video–pose pairs, which are costly to collect at scale and generally absent from routine clinical recordings. World models must also account for ultrasound's cross-sectional sampling geometry to follow scanning actions reliably. We present UltraWorld, a self-distillation recipe that transfers clinical video priors into interactive ultrasound world models without real action annotations. Using clinical videos, we adapt a video foundation model into an ultrasound generator conditioned on reference images and masks. The resulting world model predicts from local observations and actions without masks or assets at inference. Our Acoustic Sampling Map (AsMap) encodes probe poses and imaging settings as pixel-wise 3D sampling positions, beam directions, and depths. Experiments show improved prediction fidelity and action following. Across nine simulated closed-loop local planning episodes, UltraWorld reduces mean final distance to the goal and orientation error by 29% and 38%, respectively, compared with visual servoing. Code: `https://github.com/sephohi/UltraWorld`.

1 INTRODUCTION

World models have shown promise for autonomous ultrasound scanning, supporting probe guidance (Yue et al., 2025) and policy learning from predicted observations (Fan et al., 2026). Predicting the outcomes of candidate actions from local observations supports planning and policy evaluation or improvement through simulated interactions (Hafner et al., 2020; Huang et al., 2026; Guo et al., 2026). This is particularly useful in ultrasound, where sonographers face the demanding task of relating 2D sections to 3D anatomy to guide probe motion (Jiang et al., 2025a). Two challenges limit this capability. First, acquiring synchronized ultrasound video–pose data is costly, limiting data scale and clinical diversity. Second, scanning operations alter which anatomical cross-sections are imaged and how they are sampled, making reliable action-controlled prediction difficult in ultrasound.

Recent world models transfer pretrained video generators for data efficiency but still require observation–action pairs to learn target-domain control (Huang et al., 2026). In ultrasound, collecting these pairs requires expert operators, calibrated tracking, synchronized recording, and participant recruitment, making large-scale acquisition labor-intensive. Existing clinical videos, meanwhile, usually lack probe-motion records. Simulation offers controlled interactions, but full-path ray tracing may omit tissue scattering and attenuation (Duelmer et al., 2025), while learning-based simulation may require specially acquired and registered CT–ultrasound pairs (Ao et al., 2026). This motivates a more scalable route to learning reliable action responses from untracked clinical videos.

Our key insight is that visual and temporal priors learned from clinical videos can be transferred into an interactive world model using independently constructed action supervision (Figure 1(a)). Clinical videos capture anatomical appearance and temporal changes (Mishra et al., 2026); their plane selection and scanning sequences also implicitly reflect clinical judgment and expertise. We can learn these priors in a controllable video generator, then construct action supervision using 3D anatomical assets and programmable scanning trajectories. Self-distillation provides a route for this

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.

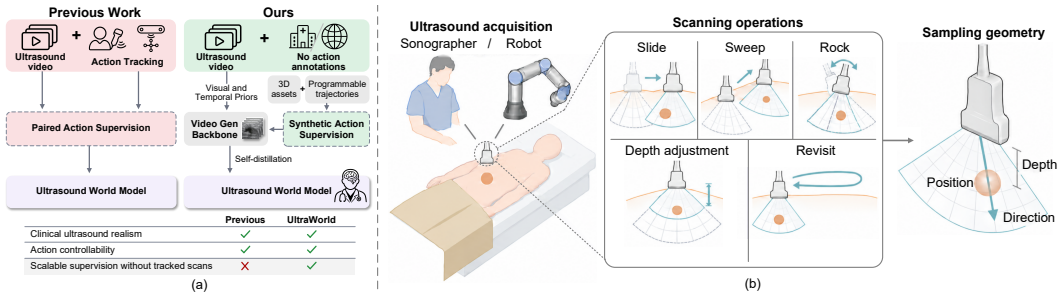

Figure 1: **Learning ultrasound world models from untracked clinical videos. (a)** Clinical videos provide appearance priors, while 3D anatomy provides motion supervision. Together, they yield synthetic action–video pairs for self-distillation. **(b)** Complex scanning operations change which anatomical cross-sections are imaged and how they are sampled. AsMap encodes the resulting sampling positions, beam directions, and depths to support reliable action control.

transfer, using generated outputs as supervision to learn new conditioning capabilities (Cai et al., 2025) or representations (Bahmani et al., 2026). Using generated action–video pairs, we can transfer this prior to prediction conditioned on local observations and scanning operations, without tracked clinical recordings.

Learning from this supervision also requires relating local 2D observations to probe motion in 3D: scanning operations change both the anatomical cross-section and its sampling geometry (Figure 1(b)). Pose-only inputs omit imaging-scale changes, while camera encodings describe viewing rays or projective relationships (Li et al., 2025). Ultrasound pixels instead correspond to depth-localized samples along acoustic beams (Szabo, 2014). Reliable action conditioning therefore needs to make these sampling locations explicit under probe motion and imaging adjustments.

We present UltraWorld, a recipe for learning interactive ultrasound world models from untracked clinical videos. We adapt a video foundation model into an ultrasound generator conditioned on a reference image and anatomical mask sequences. For action supervision, we generate action–video pairs using mask sequences sampled from 3D assets along programmable trajectories. Through self-distillation, we transfer the learned ultrasound priors into an observation- and action-conditioned world model initialized from the generator's weights, requiring neither masks nor 3D assets at inference. To link 3D probe motion to changes in 2D observations, we introduce the Acoustic Sampling Map (AsMap), encoding probe poses and imaging settings as pixel-wise 3D sampling positions, beam directions, and depths. Experiments show improved prediction fidelity and action following. We further evaluate consistency from unseen clinical frames and explore closed-loop local planning for robotic ultrasound scanning in simulation.

Our main contributions are as follows: (1) A **self-distillation recipe for learning ultrasound world models from untracked clinical videos**, combining clinical video priors with programmable supervision from 3D anatomy to learn action-conditioned prediction without real video–pose pairs. (2) The **Acoustic Sampling Map (AsMap)**, which grounds probe motion and imaging adjustments in pixel-wise 3D sampling positions, beam directions, and depths to support reliable action control in cross-sectional ultrasound imaging. (3) **Experimental validation and a robotic planning application**, demonstrating improved prediction fidelity and action following, evaluating consistency from unseen clinical frames, and exploring closed-loop local planning for robotic ultrasound scanning in simulation.

2 Related Works

Robotic Ultrasound. Robotic ultrasound guidance has been explored through visual servoing (Abolmaesumi et al., 2002; Chatelain et al., 2017; Ma et al., 2025), reinforcement learning (Li et al., 2021), and demonstration-based reward learning (Jiang et al., 2024b). Integrated systems further combine learned navigation, probe-orientation optimization, and force feedback for autonomous thyroid scanning (Su et al., 2024). More recently, ultrasound world models have demonstrated the

value of predicting acquisition outcomes for improving probe guidance and supporting policy learning. Cardiac Copilot and EchoWorld learn cardiac world models for echocardiography probe guidance (Jiang et al., 2024a; Yue et al., 2025), while Fan et al. (2026) use a diffusion world model to provide rewards for goal-conditioned policy training and demonstrate closed-loop goal-plane navigation. Despite these benefits, both world-model approaches rely on synchronized ultrasound–pose recordings to learn acquisition dynamics, leaving routine untracked clinical videos underexploited.

Ultrasound Synthesis. Ultrasound synthesis encompasses anatomical simulation from CT/MR-derived anatomy and learned generation (Kutter et al., 2009; Salehi et al., 2015). SonoGym (Ao et al., 2026) supports physics-based and generative ultrasound simulation from CT-derived anatomy for robotic learning. Echo from Noise (Stojanovski et al., 2023) uses semantic-conditioned diffusion to generate training images for cardiac segmentation, while BUSGen (Yu et al., 2026a) learns large-scale breast ultrasound priors to synthesize task-specific data for downstream analysis. Video diffusion models generate cardiac cycles from semantic maps (Van Phi et al., 2024) or arbitrary single frames (Zhang et al., 2026). These studies demonstrate the value of anatomical control and synthetic data, but visual realism alone does not establish faithful responses to prescribed scanning actions. Connecting clinical video priors with acquisition control requires consistency between the commanded trajectory and the evolving anatomy.

Geometric Conditioning. Geometric conditioning encodes pose and imaging geometry within visual models. Ultrasound world models encode probe motion using learned embeddings or multi-scale trigonometric features (Yue et al., 2025; Fan et al., 2026), leaving pixel-wise sampling relationships implicit. In camera-controlled video generation, CameraCtrl II (He et al., 2025) uses pixel-aligned Plücker embeddings with input-level feature addition, while VD3D (Bahmani et al., 2025) injects Plücker camera embeddings into video diffusion transformers through a ControlNet-like conditioning mechanism. GTA (Miyato et al., 2024) incorporates relative geometric transformations into attention, while PRoPE (Li et al., 2025) models relative camera-frustum relationships through both intrinsics and extrinsics. These camera-centric representations encode optical viewing rays or projective relationships. Adapting such conditioning to ultrasound requires representing depth-resolved acoustic sampling under probe motion and imaging configuration.

3 PRELIMINARIES

Controllable video generation. Following controllable video models (Jiang et al., 2025b), we denote a reference condition by $\mathbf{r}$ and temporally aligned controls by $\mathbf{C}_{1:T} = \{\mathbf{c}_t\}_{t=1}^{T}$. A conditional video generator models $p_\theta(\mathbf{X}_{1:T} \mid \mathbf{r}, \mathbf{C}_{1:T})$, where $\mathbf{X}_{1:T} = \{\mathbf{x}_t\}_{t=1}^{T}$ denotes the generated video.

Ultrasound acquisition parameters. Clinical ultrasound acquisition is determined jointly by probe pose and imaging configuration. We denote the acquisition state at time t as $\boldsymbol{\xi}_t = (\mathbf{T}_t, \boldsymbol{\kappa}_t)$, where $\mathbf{T}_t \in SE(3)$ transforms points from the probe coordinate system to the patient coordinate system, and $\boldsymbol{\kappa}_t$ denotes acquisition parameters such as scan geometry, physical sampling scale, imaging depth or field of view, and beam steering.

An ultrasound acquisition maps a physical point $\mathbf{p}$ to its image coordinate as

$$(u, v) = \pi_{\mathrm{US}}\left(\mathbf{T}_t \mathbf{p}; \boldsymbol{\kappa}_t\right), \tag{1}$$

where π_{US} denotes the acquisition-dependent mapping from spatial locations on the imaging plane to ultrasound image coordinates.

World models. A visual world model predicts the visual consequences of actions from an observed context (Zhu et al., 2025; Huang et al., 2026). Given $\mathbf{X}^c = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ and future actions $\mathbf{A}^f = \{\mathbf{a}_K, \ldots, \mathbf{a}_{K+H-1}\}$, it predicts $\widehat{\mathbf{X}}^f = \mathcal{W}_\theta(\mathbf{X}^c, \mathbf{A}^f)$, where $\widehat{\mathbf{X}}^f = \{\hat{\mathbf{x}}_{K+1}, \ldots, \hat{\mathbf{x}}_{K+H}\}$.

4 METHOD

In this section, we present **UltraWorld**, a recipe for learning interactive ultrasound world models from action-free clinical videos. Our key idea is to transfer the visual and temporal priors learned

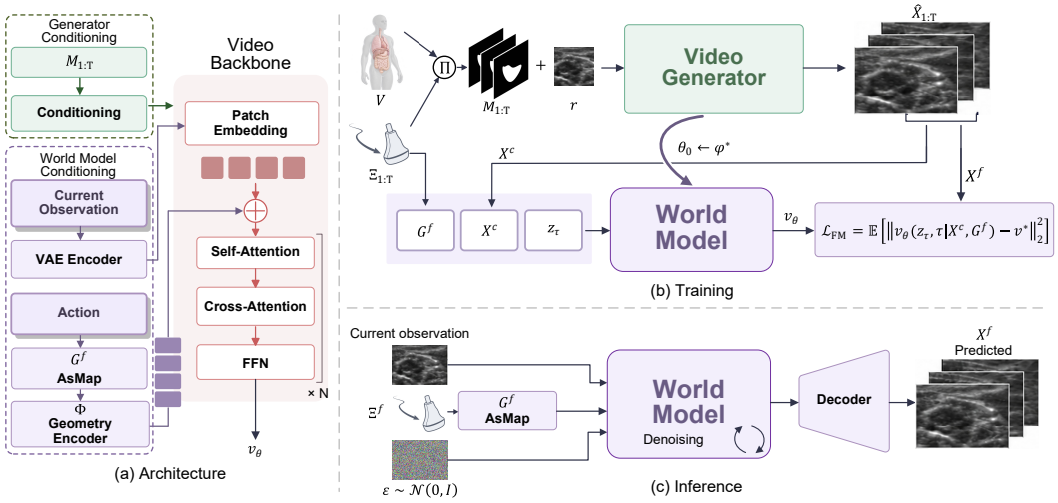

Figure 2: **Overall framework of UltraWorld.** (a) A shared video backbone supports mask-conditioned generation and action-conditioned world modeling through AsMap. (b) 3D anatomy and programmable trajectories provide synthetic action–video supervision, transferring the learned video prior to the world model through self-distillation and flow matching. (c) At inference, the world model predicts future ultrasound video from the current observation and prescribed acquisition trajectory.

from clinical videos into action-conditioned dynamics through anatomy-guided interaction supervision. We further introduce a pixel-aligned representation of ultrasound acquisition to explicitly encode the spatial effects of diverse clinical scanning operations. The overall approach is illustrated in Fig. 2.

4.1 From Action-Free Ultrasound Videos to Interactive Trajectories

We reformulate world-model learning as transferring priors from action-free clinical ultrasound videos to trajectories with explicit acquisition supervision. We first learn a controllable video prior, then construct the missing action supervision independently from anatomy-guided scanning trajectories.

Learning ultrasound priors from action-free videos. Given an action-free ultrasound sequence $\mathbf{X}_{1:T} = \{\mathbf{x}_t\}_{t=1}^T$, we associate it with a temporally aligned semantic anatomical condition $\mathbf{M}_{1:T} = \{\mathbf{m}_t\}_{t=1}^T$. Instantiating the structured video control $\mathbf{C}_{1:T}$ with $\mathbf{M}_{1:T}$, we learn a controllable ultrasound video generator

$$p_\phi\left(\mathbf{X}_{1:T} \mid \mathbf{r}, \mathbf{M}_{1:T}\right), \tag{2}$$

where $\mathbf{r}$ is a reference ultrasound image that anchors the appearance characteristics of the generated sequence. Crucially, this stage requires no corresponding acquisition states $\boldsymbol{\Xi}_{1:T}$, allowing ultrasound-specific visual and temporal priors to be learned directly from action-free clinical videos.

Anatomy-guided ultrasound scanning trajectories. The learned video prior captures realistic ultrasound appearance and temporal evolution, but action-free videos do not specify how observations should change under prescribed acquisition states. We obtain this missing interaction supervision from 3D anatomical volumes. Let $\mathcal{V}$ denote a semantic anatomical volume. We construct a scanning trajectory as a sequence of acquisition states

$$\boldsymbol{\Xi}_{1:T} = \{\boldsymbol{\xi}_t\}_{t=1}^T, \qquad \boldsymbol{\xi}_t = (\mathbf{T}_t, \boldsymbol{\kappa}_t), \tag{3}$$

where $\mathbf{T}_t$ denotes the transducer pose and $\boldsymbol{\kappa}_t$ specifies the ultrasound acquisition parameters. We programmatically sample trajectories covering common clinical scanning operations, including sweeps, slide, rock, revisits, and changes in imaging depth or field of view.

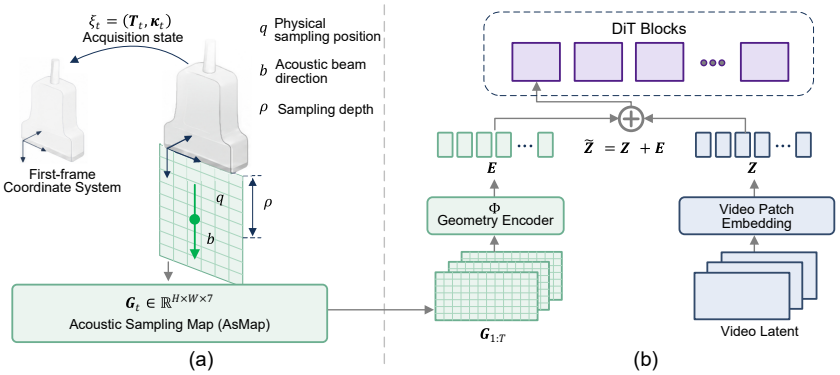

Figure 3: **Acoustic Sampling Map and geometry conditioning.** (a) Probe pose and imaging parameters determine each pixel's physical sampling position $\mathbf{q}$, beam direction $\mathbf{b}$, and sampling depth ρ. Positions and directions are expressed in the first-frame probe coordinate system, forming a seven-channel AsMap. (b) A geometry encoder Φ aligns the AsMap sequence with the video tokens, providing pixel-aligned conditioning before the first transformer block.

For each acquisition state, the corresponding anatomical observation is obtained by resampling the 3D volume under the specified acquisition configuration,

$$\mathbf{m}_t = \Pi\left(\mathcal{V}; \boldsymbol{\xi}_t\right), \tag{4}$$

where $\Pi(\cdot)$ denotes the volumetric sampling operator. Applying this operation along $\boldsymbol{\Xi}_{1:T}$ yields an anatomically consistent mask trajectory $\mathbf{M}_{1:T}$ whose acquisition states are known by construction. Consistent with prior ultrasound simulation, we use MR only to provide 3D anatomical geometry for $M_{1:T}$, while ultrasound appearance is inherited from the clinical-ultrasound prior p_ϕ (Kutter et al., 2009).

Interactive trajectory synthesis. We then use the anatomy-guided mask trajectory as the structured control of the learned ultrasound video prior:

$$\hat{\mathbf{X}}_{1:T} \sim p_\phi\left(\mathbf{X}_{1:T} \mid \mathbf{r}, \mathbf{M}_{1:T}\right). \tag{5}$$

Because $\mathbf{M}_{1:T}$ is generated from the prescribed acquisition trajectory $\boldsymbol{\Xi}_{1:T}$, the synthesized ultrasound sequence is paired by construction with its acquisition states. Repeating this process produces an interaction dataset

$$\hat{\mathcal{D}}_{\text{traj}} = \left\{\left(\hat{\mathbf{X}}_{1:T}, \boldsymbol{\Xi}_{1:T}\right)\right\}, \tag{6}$$

thereby converting the visual and temporal priors learned from action-free clinical videos into scalable action-labeled supervision for ultrasound world-model learning.

4.2 Acoustic Sampling Map

Clinical ultrasound acquisition depends jointly on probe motion and imaging configuration. For example, a sonographer may reduce the imaging depth to enlarge a lesion within the field of view while keeping the probe pose nearly unchanged. A 6-DoF pose alone cannot distinguish these acquisitions, although they correspond to different physical sampling geometries and therefore different image observations. To explicitly represent the spatial sampling induced by both probe motion and acquisition parameters, we introduce the *Acoustic Sampling Map* (AsMap), a dense, pixel-aligned representation of ultrasound acquisition geometry (Fig. 3).

Acoustic sampling geometry. For each image location (u, v), we define its probe-local sampling geometry as

$$\mathcal{S}_{\text{US}}(u, v; \kappa_t) = \left(\mathbf{q}_{t,u,v}, \mathbf{b}_{t,u,v}, \rho_{t,u,v}\right), \tag{7}$$

where $\mathbf{q}_{t,u,v} \in \mathbb{R}^3$ is the physical sampling position, $\mathbf{b}_{t,u,v} \in \mathbb{R}^3$ is the unit acoustic beam direction, and $\rho_{t,u,v}$ is the sampling depth.

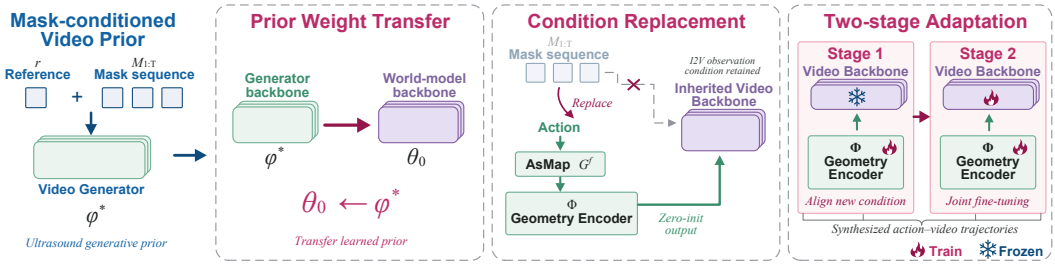

Figure 4: **Self-distillation for ultrasound world modeling.** The transformation consists of three steps: (1) **Prior weight transfer**: initialize the world model from the learned video generator; (2) **Condition replacement**: replace anatomical-mask conditioning with AsMap conditioning while retaining the observation input; (3) **Two-stage adaptation**: train the geometry encoder with a frozen video backbone, then jointly fine-tune both. Both stages use synthesized action–video trajectories.

To describe a trajectory consistently, we express positions and directions in the probe coordinate system of the first frame. With $T_t = (R_t, \mathbf{t}_t)$ denoting the probe-to-patient transformation,

$$\begin{bmatrix} \mathbf{q}^{(1)}_{t,u,v} \\ 1 \end{bmatrix} = T_1^{-1} T_t \begin{bmatrix} \mathbf{q}_{t,u,v} \\ 1 \end{bmatrix}, \qquad \mathbf{b}^{(1)}_{t,u,v} = R_1^\top R_t \mathbf{b}_{t,u,v}. \tag{8}$$

The AsMap at time t is then

$$\mathbf{G}_t(u,v) = \left[\mathbf{q}^{(1)}_{t,u,v}, \mathbf{b}^{(1)}_{t,u,v}, \rho_{t,u,v} \right] \in \mathbb{R}^7, \qquad \mathbf{G}_t \in \mathbb{R}^{H \times W \times 7}. \tag{9}$$

Unlike pose-only conditioning, AsMap explicitly changes with the physical sampling configuration even when the probe pose is fixed. The first-frame formulation also removes dependence on the arbitrary global patient coordinate frame; additional properties are discussed in Appendix A.

Encoding and injection. A lightweight geometry encoder Φ converts the AsMap sequence into features aligned with the video tokens,

$$\mathbf{E} = \Phi(\mathbf{G}_{1:T}), \qquad \widetilde{\mathbf{Z}} = \mathbf{Z} + \mathbf{E}, \tag{10}$$

where $\mathbf{Z}$ denotes the patch-embedded video features. Following the input-level conditioning strategy of CameraCtrl II (He et al., 2025), the AsMap features are added once before the first transformer block, providing frame- and pixel-aligned acquisition conditioning without modifying the internal transformer architecture.

4.3 Self-Distillation for Ultrasound World Modeling

The synthesized trajectories transfer geometry-controlled video generation into action-conditioned prediction of how ultrasound appearance changes with scanning operations. Following generative self-distillation (Cai et al., 2025; Bahmani et al., 2026), we transform the mask-conditioned ultrasound generator into an action-conditioned world model by replacing mask conditioning with AsMap while retaining the observation input (Fig. 4). We initialize the world-model backbone from the teacher, $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\phi}^*$, and train it on the synthesized trajectories to model $p_\theta(\mathbf{X}^f \mid \mathbf{X}^c, \mathbf{G}^f)$. For optimization, we first train the AsMap encoder Φ with the video backbone frozen and then jointly fine-tune both. Detailed training procedures are provided in Appendix B.

5 Experiments

5.1 Experimental Setup

Datasets. We use BUV (Lin et al., 2022) and CAMUS (Leclerc et al., 2019) to learn the mask-conditioned ultrasound video prior, and I-SPY1 (Newitt et al., 2016; Hylton et al., 2016) to construct anatomy-guided trajectories from volumetric breast MRI. CAMUS is used only for prior learning. **Training details.** We initialize the ultrasound video generator from Wan2.1-FunControl-1.3B-V1.1 (Wan Team et al., 2025) and train the generator and world model at 512×512 resolution.

Baselines. We compare AsMap with **Fourier Encoding** (Fan et al., 2026), **MLP Encoding** inspired by EchoWorld (Yue et al., 2025), and **PRoPE** (Li et al., 2025). All methods receive the same acquisition state $(\mathbf{T}_t, \boldsymbol{\kappa}_t)$.

5.2 Mask-Conditioned Ultrasound Video Generator

We first verify that the mask-conditioned generator preserves realistic ultrasound appearance while following the prescribed anatomical structure. As shown in Fig. 5, the generated sequences remain temporally coherent and follow the target anatomy across both breast ultrasound and echocardiography. Quantitative evaluation of generation fidelity and mask controllability is provided in Appendix E.1.

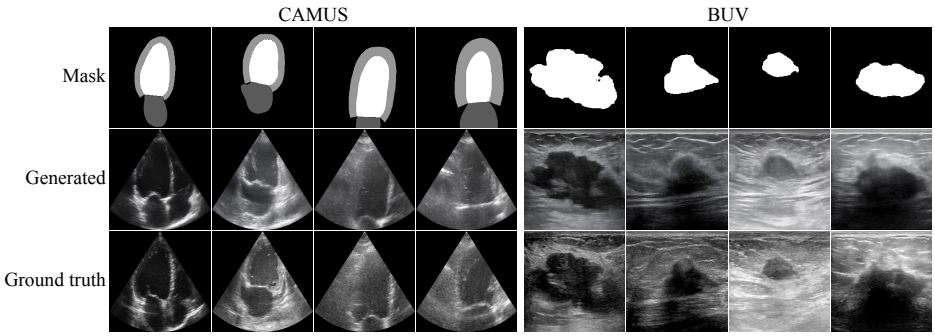

Figure 5: **Mask-conditioned ultrasound video generation.** Generated sequences follow the prescribed anatomy while preserving realistic ultrasound appearance.

5.3 World Model Quality Analysis

We evaluate whether UltraWorld predicts plausible changes in ultrasound texture and anatomical appearance in response to prescribed acquisition trajectories. Following recent discussions on world-model evaluation (Yu et al., 2026b), we consider both prediction fidelity and action-sensitive behavior. We first evaluate on paired synthesized trajectories with known reference futures and acquisition states. Counterfactual evaluations under alternative trajectories are provided in Appendix E.

Metrics. Following trajectory-conditioned world models (Zhu et al., 2025), we evaluate prediction fidelity using PSNR, LPIPS, and latent ℓ_2 distance, and action following using Δ-LPIPS (Li et al., 2026) and *Motion Error*. The latter measures the discrepancy between predicted and reference target-anatomy displacements. Let $\boldsymbol{\mu}_t$ and $\hat{\boldsymbol{\mu}}_t$ denote the reference and predicted target-anatomy centroids at frame t:

$$E_{\text{motion}} = \frac{1}{|\mathcal{P}|} \sum_{t \in \mathcal{P}} \left\| (\hat{\boldsymbol{\mu}}_{t+1} - \hat{\boldsymbol{\mu}}_t) - (\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t) \right\|_2 , \tag{11}$$

where $\mathcal{P}$ denotes the set of valid frame pairs.

Prediction fidelity and action following. Table 1 compares AsMap with representative global action encodings and projective geometric conditioning under the same acquisition states. AsMap achieves the best performance across both prediction-fidelity and action-following metrics, showing that explicitly representing pixel-wise ultrasound sampling geometry improves both visual prediction and responses to prescribed acquisition changes.

Figure 6 compares future ultrasound predictions from the same initial observation under four scanning operations. White arrows highlight where baseline predictions fail to follow the image changes induced by probe motion or imaging adjustments. Yellow insets reveal missing or unstable local features, including hypoechoic regions and surrounding echogenic tissue interfaces. UltraWorld more consistently predicts how these features evolve with the prescribed actions while maintaining plausible texture and local anatomical appearance across future frames. Additional examples are provided in Appendix E.2.

Table 1: **Comparison of different action-conditioning representations.** AsMap consistently improves both prediction fidelity and action following over global action encodings and projective geometric conditioning. Motion error is measured in pixels.

Conditioning	Prediction fidelity			Action following	
	a PSNR↑	LPIPS↓	Latent ℓ_2 ↓	Δ-LPIPS↓	Motion err.↓
Fourier Enc. (Fan et al., 2026)	22.946	0.274	0.120	0.509	4.540
MLP Enc. (Yue et al., 2025)	23.031	0.262	0.121	0.486	4.471
PRoPE (Li et al., 2025)	22.634	0.275	0.127	0.504	4.457
AsMap (Ours)	**24.994**	**0.236**	**0.095**	**0.466**	**4.155**

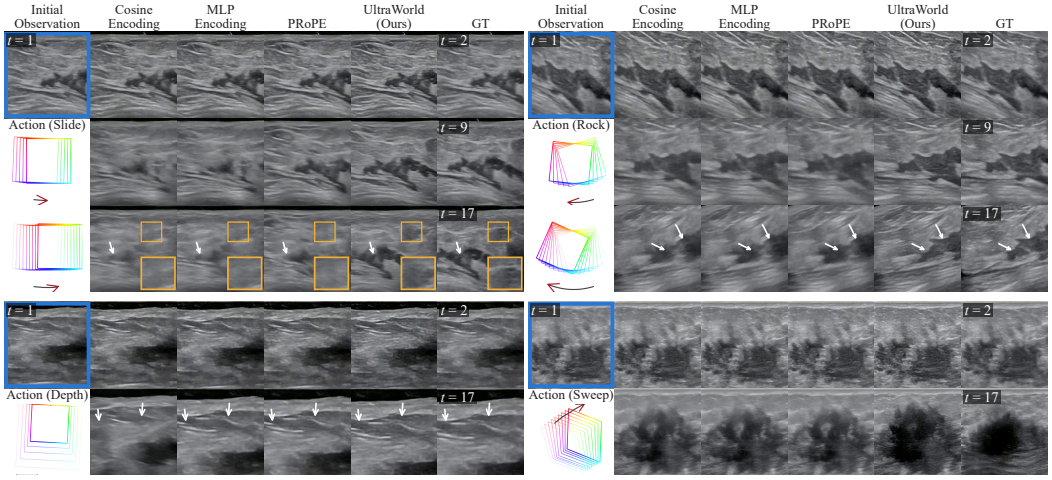

Figure 6: **Qualitative comparison across scanning operations.** UltraWorld follows prescribed operations more consistently. White arrows highlight differences in action following; yellow insets highlight missing or unstable hypoechoic regions and anatomical features in baseline predictions.

Table 2: **Ablation studies on action encoding and initialization.** "Matched" approximately matches the AsMap encoder parameter count and uses the same first-frame-relative coordinates. For each ablation, the remaining components follow the full model configuration.

Variant	Prediction fidelity			Action following	
	PSNR↑	LPIPS↓	Latent ℓ_2 ↓	Δ-LPIPS↓	Motion err.↓
Matched MLP conditioning	23.402	0.258	0.113	0.482	4.429
Wan2.1 initialization	24.640	**0.235**	0.099	0.472	4.414
Full model (Ours)	**24.994**	0.236	**0.095**	**0.466**	**4.155**

Ablation studies. A parameter-matched MLP using the same first-frame-relative coordinates remains worse than AsMap (Table 2), indicating that the gain is not explained by encoder capacity or coordinate convention. Initialization from Wan2.1 performs comparably, suggesting that self-distillation does not critically rely on generator-weight inheritance.

5.4 UltraWorld for Robotic Ultrasound Planning

Building on world-model-based robotic planning (Hansen et al., 2024; Zhu et al., 2025; Zhou et al., 2025), we evaluate UltraWorld for goal-directed local ultrasound planning in simulation. We use

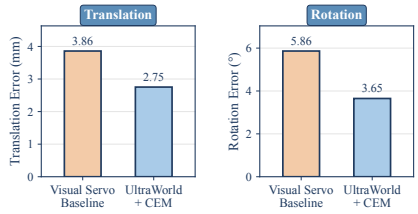

Figure 7: Closed-loop planning errors.

Table 3: **Generalization to real clinical ultrasound.**

Conditioning	BUSI		BUV	
	Zero-action↓	Round-trip↓	Zero-action↓	Round-trip↓
Fourier Enc.	0.110	0.231	0.148	0.276
MLP Enc.	0.172	0.238	0.244	0.253
PRoPE	0.159	0.227	0.226	0.317
AsMap (Ours)	**0.042**	**0.062**	**0.059**	**0.076**

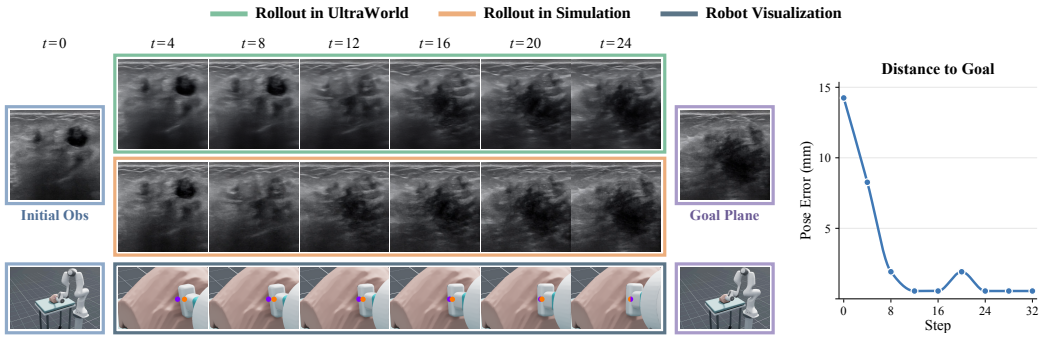

Figure 8: **Closed-loop robotic ultrasound planning with UltraWorld.** A representative episode showing imagined rollouts, executed observations, and the target view. Only the first optimized action is executed before replanning from the new observation.

model predictive control (MPC) (Mayne et al., 2000), optimizing candidate probe-action sequences with the cross-entropy method (CEM) (De Boer et al., 2005). We compare against an intensity-based visual-servoing baseline inspired by Nadeau & Krupa (2013); implementation details are provided in Appendix C. Across nine episodes, UltraWorld achieves lower mean terminal translation and rotation errors than visual servoing (Fig. 7). Figure 8 shows a representative closed-loop episode in which imagined rollouts guide the probe toward the target view.

5.5 Generalization to Real Clinical Ultrasound

We evaluate unseen real initial frames from patient-held-out BUV and independent BUSIdataset (Al-Dhabyani et al., 2020). Without paired futures, we use zero-action and round-trip LPIPS to assess stability and recovery consistency (Li et al., 2026). AsMap achieves the lowest errors on both datasets (Table 3); qualitative rollouts are provided in Appendix E.

6 Conclusion

We presented UltraWorld, a recipe for learning ultrasound world models from untracked clinical videos. It combines a mask-conditioned ultrasound video prior with anatomy-guided trajectory synthesis for self-distillation, and introduces the Acoustic Sampling Map (AsMap) to represent pixel-wise ultrasound sampling geometry. Experiments validate the fidelity and controllability of the video generator, and show that UltraWorld improves prediction fidelity and acquisition following over alternative conditioning representations. We further demonstrate closed-loop local scan planning with lower terminal translation and rotation errors than visual servoing, and improved consistency from unseen real clinical observations. Together, these results provide a practical route from untracked clinical videos to action-conditioned ultrasound world models. Future work will validate the approach with physical robotic systems.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring mandatory disclosure under the ICLR 2027 AI Policy, including developing theoretical or conceptual frameworks, formulating mathematical claims or proofs, proposing or refining hypotheses, designing research methodology or experiments, implementing the proposed methods, generating synthetic datasets, cleaning or reformatting datasets, conducting qualitative or thematic data analysis, or interpreting experimental results. Translation and proof-related assistance were not applicable to this work.

We used generative AI tools for tasks for which disclosure is recommended, including improving the readability and clarity of the manuscript, assisting with software code editing and debugging, brainstorming, and supporting literature search and summarization.

All AI-assisted text was reviewed and revised by the authors. AI-assisted code was manually inspected and validated through testing and experimental execution. Literature-related suggestions and summaries were checked against the original sources, and all references and factual claims included in the paper were independently verified by the authors. Generative AI outputs were not treated as authoritative sources.

We take responsibility for the final content of this work, including all text, claims, code, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available datasets and does not involve human subjects or personally identifiable information. We are not aware of any specific ethical concerns beyond those generally associated with the use of machine learning systems.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, model architecture, training procedure, hyperparameter settings, dataset preprocessing, and evaluation protocols in the main paper and appendices. Additional implementation details and experimental results are included in the supplementary material. The code and configuration files necessary to reproduce the main experiments will be released upon publication.

REFERENCES

Purang Abolmaesumi, Septimiu E. Salcudean, Wen-Hong Zhu, Mohammad Reza Sirouspour, and Simon P. DiMaio. Image-guided control of a robot for medical ultrasound. *IEEE Transactions on Robotics and Automation*, 18(1):11–23, 2002. doi: 10.1109/70.988970.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. doi: 10.1016/j.dib.2019.104863.

Yunke Ao, Masoud Moghani, Mayank Mittal, Manish Prajapat, Luohong Wu, Frederic Giraud, Fabio Carrillo, Andreas Krause, and Philipp Fürnstahl. Sonogym: High performance simulation for challenging surgical tasks with robotic ultrasound. *Advances in Neural Information Processing Systems*, 38, 2026.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. In *International Conference on Learning Representations*, volume 2025, pp. 66712–66737, 2025.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. In *International Conference on Learning Representations*, volume 2026, pp. 82850–82880, 2026.

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18434–18443. IEEE, 2025.

Pierre Chatelain, Alexandre Krupa, and Nassir Navab. Confidence-driven control of an ultrasound probe. *IEEE Transactions on Robotics*, 33(6):1410–1424, 2017. doi: 10.1109/TRO.2017.2723618.

Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

Felix Duelmer, Mohammad Farid Azampour, Magdalena Wysocki, and Nassir Navab. Ultraray: Introducing full-path ray tracing in physics-based ultrasound simulation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 653–662. Springer, 2025.

Siqi Fan, Mingcong Chen, Ran Liu, Zixuan Yang, Xiaoyu Fu, Xiaoqing Gao, Yunhui Liu, and Hongbin Liu. Action-conditioned world model for goal plane probe guidance in robotic ultrasound, 2026.

Yanjiang Guo, Lucy Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 6121–6138, 2026.

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

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13416–13426, 2025.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. In *International Conference on Learning Representations*, 2026.

Nola M. Hylton, Constantine A. Gatsonis, Mark A. Rosen, Constance D. Lehman, David C. Newitt, Savannah C. Partridge, Wanda K. Bernreuter, Etta D. Pisano, Elizabeth A. Morris, Paul T. Weatherall, Sandra M. Polin, Gillian M. Newstead, Helga S. Marques, Laura J. Esserman, and Mitchell D. Schnall. Neoadjuvant chemotherapy for breast cancer: Functional tumor volume by mr imaging predicts recurrence-free survival—results from the acrin 6657/calgb 150007 i-spy 1 trial. *Radiology*, 279(1):44–55, 2016. doi: 10.1148/radiol.2015150013.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 190–199. Springer, 2024a.

Haojun Jiang, Andrew Zhao, Qian Yang, Xiangjie Yan, Teng Wang, Yulin Wang, Ning Jia, Jiangshan Wang, Guokun Wu, Yang Yue, et al. Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system. *Nature Communications*, 16(1):7893, 2025a.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025b.

Zhongliang Jiang, Yuan Bi, Mingchuan Zhou, Ying Hu, Michael Burke, and Nassir Navab. Intelligent robotic sonographer: Mutual information-based disentangled reward learning from few demonstrations. *The International Journal of Robotics Research*, 43(7):981–1002, 2024b. doi: 10.1177/02783649231223547.

Oliver Kutter, Ramtin Shams, and Nassir Navab. Visualization and gpu-accelerated simulation of medical ultrasound from ct images. *Computer methods and programs in biomedicine*, 94(3): 250–266, 2009.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Ostvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan Dhooge, Lasse Lovstakken, and Olivier Bernard. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Medical Imaging*, 38(9):2198–2210, 2019. doi: 10.1109/TMI.2019.2900516.

Keyu Li, Jian Wang, Yangxin Xu, Hao Qin, Dongsheng Liu, Li Liu, and Max Q.-H. Meng. Autonomous navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8302–8308, 2021. doi: 10.1109/ICRA48506.2021.9561295.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0540.

Yaxuan Li, Zhongyi Zhou, Yefei Chen, Yaokai Xue, and Yichen Zhu. dworldeval: Scalable robotic policy evaluation via discrete diffusion world model. *arXiv preprint arXiv:2604.22152*, 2026.

Zhi Lin, Junhao Lin, Lei Zhu, Huazhu Fu, Jing Qin, and Liansheng Wang. A new dataset and a baseline model for breast lesion detection in ultrasound videos. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 614–623. Springer, 2022. doi: 10.1007/978-3-031-16437-8_59.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xihan Ma, Mingjie Zeng, Jeffrey C. Hill, Beatrice Hoffmann, Ziming Zhang, and Haichong K. Zhang. Guiding the last centimeter: Novel anatomy-aware probe servoing for standardized imaging plane navigation in robotic lung ultrasound. *IEEE Transactions on Automation Science and Engineering*, 22:6569–6580, 2025. doi: 10.1109/TASE.2024.3448241.

David Q Mayne, James B Rawlings, Christopher V Rao, and Pierre OM Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000.

Adrien Meyer, Aditya Murali, Farahdiba Zarin, Didier Mutter, and Nicolas Padoy. Ultrasam: a foundation model for ultrasound using large open-access segmentation datasets. *International Journal of Computer Assisted Radiology and Surgery*, 21(1):93–102, 2026.

Divyanshu Mishra, Mohammadreza Salehi, Pramit Saha, Olga Patey, Aris Papageorghiou, Yuki Asano, and Alison Noble. Self-supervised learning of echocardiographic video representations via online cluster distillation. *Advances in Neural Information Processing Systems*, 38:61229–61254, 2026.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. GTA: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.10375.

Caroline Nadeau and Alexandre Krupa. Intensity-based ultrasound visual servoing: Modeling and validation with 2-d and 3-d probes. *IEEE Transactions on Robotics*, 29(4):1003–1015, 2013. doi: 10.1109/TRO.2013.2256690.

David Newitt, Nola Hylton, and I-SPY 1 Network and ACRIN 6657 Trial Team. Multi-center breast dce-mri data and segmentations from patients in the i-spy 1/acrin 6657 trials, 2016.

Mehrdad Salehi, Seyed-Ahmad Ahmadi, Raphael Prevost, Nassir Navab, and Wolfgang Wein. Patient-specific 3d ultrasound simulation based on convolutional ray-tracing and appearance optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 510–518. Springer, 2015.

David Stojanovski, Uxio Hermida, Pablo Lamata, Arian Beqiri, and Alberto Gomez. Echo from noise: Synthetic ultrasound image generation using diffusion models for real image segmentation. In *Simplifying Medical Ultrasound*, volume 14337 of *Lecture Notes in Computer Science*, pp. 34–43. Springer, 2023. doi: 10.1007/978-3-031-44521-7_4.

Kang Su, Jingwei Liu, Xiaoqi Ren, Yingxiang Huo, Guanglong Du, Wei Zhao, Xueqian Wang, Bin Liang, Di Li, and Peter Xiaoping Liu. A fully autonomous robotic ultrasound system for thyroid scanning. *Nature Communications*, 15:4004, 2024. doi: 10.1038/s41467-024-48421-y.

Thomas L. Szabo. *Diagnostic Ultrasound Imaging: Inside Out*. Academic Press, 2 edition, 2014. ISBN 9780123964878. doi: 10.1016/C2011-0-07261-7.

Nguyen Van Phi, Tran Minh Duc, Pham Huy Hieu, and Tran Quoc Long. Echocardiography video synthesis from end diastolic semantic map via diffusion model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13461–13465. IEEE, 2024.

Wan Team, Ang Wang, Baole Ai, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haojun Yu, Youcheng Li, Nan Zhang, Zihan Niu, Xuantong Gong, Yanwen Luo, Haotian Ye, Siyu He, Quanlin Wu, Wangyan Qin, Mengyuan Zhou, Jie Han, Jia Tao, Ziwei Zhao, Di Dai, Di He, Dong Wang, Binghui Tang, Ling Huo, James Zou, Qingli Zhu, Yong Wang, and Liwei Wang. A foundation generative model for breast ultrasound image analysis. *Nature Biomedical Engineering*, 2026a. doi: 10.1038/s41551-026-01639-1.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? a decision-making-centric position, 2026b.

Yang Yue, Yulin Wang, Haojun Jiang, Pan Liu, Shiji Song, and Gao Huang. Echoworld: Learning motion-aware world models for echocardiography probe guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25993–26003, 2025. doi: 10.1109/CVPR52734.2025.02421.

Jiansong Zhang, Xiaying Yang, Xiaoling Luo, and Linlin Shen. Echovdiff: Cardiac-cycle echocardiography video generation from arbitrary single frame. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9040–9050, 2026.

Gaoyue Zhou, Hengkai Pan, Yann Lecun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025. URL `https://proceedings.mlr.press/v267/zhou25t.html`.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9834–9844, 2025. doi: 10.1109/ICCV51701.2025.00917.

A PROPERTIES OF THE ACOUSTIC SAMPLING MAP

We provide additional analysis of the Acoustic Sampling Map (AsMap) introduced in Sec. 4.2. AsMap converts the acquisition state $\xi_t = (T_t, \kappa_t)$ into pixel-wise physical sampling geometry and expresses the resulting geometry in the coordinate system of the first observed frame. Below, we clarify several properties of this representation.

Sensitivity to imaging configuration. A probe pose alone does not uniquely determine an ultrasound observation, because the physical samples represented by the image also depend on the acquisition configuration κ_t. Consider two acquisition states

$$\xi_t = (T_t, \kappa_t), \qquad \xi'_t = (T_t, \kappa'_t), \tag{12}$$

that share the same probe pose but differ in their imaging configurations. Their corresponding AsMaps are

$$\mathbf{G}_t = \mathrm{AsMap}(T_t, \kappa_t; T_1), \qquad \mathbf{G}'_t = \mathrm{AsMap}(T_t, \kappa'_t; T_1). \tag{13}$$

If κ_t and κ'_t induce different physical sampling locations, beam directions, or sampling depths, then in general

$$\mathbf{G}_t \neq \mathbf{G}'_t. \tag{14}$$

For example, changing the imaging depth or field of view changes the physical locations associated with image pixels even if T_t remains fixed. AsMap therefore distinguishes acquisition changes that are not observable from a 6-DoF probe pose alone.

Invariance to the global patient coordinate frame. AsMap expresses acquisition geometry relative to the probe coordinate system of the first frame. This construction removes dependence on the particular global coordinate system used to describe the trajectory.

Consider an arbitrary rigid transformation $S \in SE(3)$ applied to the patient coordinate system. The probe poses then become

$$T'_t = ST_t, \qquad T'_1 = ST_1. \tag{15}$$

The relative transformation used to construct AsMap satisfies

$$(T'_1)^{-1}T'_t = (ST_1)^{-1}(ST_t) = T_1^{-1}S^{-1}ST_t = T_1^{-1}T_t. \tag{16}$$

Therefore, the first-frame-relative physical sampling positions and acoustic beam directions are unchanged under a rigid change of the global patient coordinate system. Consequently,

$$\mathbf{G}'_t = \mathbf{G}_t. \tag{17}$$

This property is useful because the conditioning describes how the current acquisition differs geometrically from the observed initial acquisition, rather than encoding an arbitrary external reference frame.

Independence from numerical pose parameterization. AsMap is constructed after the probe pose has been converted to a rigid transformation $T_t \in SE(3)$. Let $\boldsymbol{\eta}_t$ and $\boldsymbol{\eta}'_t$ denote two numerical pose representations, and let $f(\cdot)$ map either representation to $SE(3)$. If

$$f(\boldsymbol{\eta}_t) = f(\boldsymbol{\eta}'_t) = T_t, \tag{18}$$

then both representations produce the same relative transformation $T_1^{-1}T_t$ and therefore the same AsMap.

Thus, after conversion to the physical acquisition geometry, the conditioning does not depend on whether the same rigid pose was originally represented, for example, by Euler angles, quaternions, or another equivalent parameterization. This property concerns the numerical representation of an identical physical pose and should not be interpreted as invariance to different probe poses or imaging configurations.

Metric interpretation. When the acquisition system is calibrated in physical units, $\mathbf{q}_{t,u,v}$ represents a physical sampling location, $\rho_{t,u,v}$ represents a physical sampling depth, and $\mathbf{b}_{t,u,v}$ is a unit acoustic beam direction. The rigid first-frame transformation preserves these physical quantities. Therefore, AsMap retains the physical magnitude of probe translations and changes in sampling scale rather than representing them only through an arbitrary normalized action vector.

This metric interpretation assumes metric calibration of both the probe pose and the mapping from image coordinates to physical sampling geometry. Without such calibration, AsMap still defines a spatially aligned acquisition representation, but its position and depth channels should not be interpreted as absolute physical distances.

Relation to camera-ray conditioning. Camera-conditioned video models commonly associate image pixels with optical viewing rays or encode relative projective geometry. In these representations, the ray primarily specifies how a pixel observes the scene from a given camera configuration. Ultrasound acquisition differs because a B-mode pixel corresponds to a depth-resolved acoustic sample rather than only to a viewing direction.

AsMap makes this distinction explicit by representing

$$(\mathbf{q}_{t,u,v}, \mathbf{b}_{t,u,v}, \rho_{t,u,v}) \tag{19}$$

for every image location. The beam direction $\mathbf{b}_{t,u,v}$ describes the local acoustic propagation direction, while $\mathbf{q}_{t,u,v}$ and $\rho_{t,u,v}$ specify the corresponding depth-resolved physical sample. This allows changes in probe pose and imaging configuration to be converted into a common pixel-aligned sampling representation before being provided to the video model.

Spatial alignment with visual tokens. Global pose encodings associate one acquisition vector with an entire frame. In contrast, AsMap retains the correspondence between acquisition geometry and image location. After spatial and temporal alignment by the AsMap encoder Φ, each geometry feature is associated with the video tokens representing the corresponding region and acquisition time.

This distinction is especially relevant when an acquisition change induces spatially non-uniform changes in the observed anatomy. Rather than requiring the video backbone to infer the mapping from a global pose vector to pixel-level image changes, AsMap exposes this correspondence explicitly in the conditioning signal.

Scope of the representation. AsMap represents ultrasound *sampling geometry*. In particular, the representation does not explicitly encode tissue-dependent scattering, attenuation, speckle formation, or other acoustic interactions responsible for ultrasound appearance. In UltraWorld, these visual and temporal characteristics are learned from clinical ultrasound videos by the generative prior, while AsMap specifies how the prescribed acquisition trajectory changes the spatial sampling of the underlying anatomy.

This separation is central to our formulation: the learned video prior provides ultrasound appearance and temporal dynamics, whereas AsMap provides the acquisition geometry required to transfer these priors to action-conditioned prediction.

B UltraWorld Implementation Details

Algorithms 1 and 2 summarize the training recipe and action-conditioned inference of UltraWorld. We denote the video transformer parameters by θ and the AsMap encoder parameters by ψ. The pretrained VAE and image/text encoders remain fixed. For brevity, $\mathcal{C}_1$ groups the inherited first-frame latent, image features, and text conditioning. The AsMap encoder Φ_ψ aligns sampling maps with video tokens, and its output is added once before the first transformer block.

Anatomical masks are used only to train the video prior and synthesize the supervision videos. During self-distillation, the mask-control channels remain zero, while the initial-frame conditioning is retained. $\mathrm{MSE}_{\mathrm{future}}$ denotes the mean squared error over future latent frames, excluding the observed initial latent frame.

Algorithm 1 UltraWorld Training Recipe

Require: Pretrained video generator p_{ϕ_0}; untracked clinical videos $\mathcal{D}_{\mathrm{US}}$; segmentation model $\mathcal{S}$; semantic anatomical volumes $\mathbb{V}$
Ensure: World-model parameters θ and AsMap encoder parameters ψ

1. Learn an ultrasound video prior

Construct video–mask pairs $(\mathbf{X}_{1:T}, \mathcal{S}(\mathbf{X}_{1:T}))$ from $\mathcal{D}_{\mathrm{US}}$
Associate each pair with a reference ultrasound image $\mathbf{r}$
Adapt p_{ϕ_0} using LoRA and conditional flow matching to obtain $p_{\phi^\star}(\mathbf{X} \mid \mathbf{r}, \mathbf{M})$

2. Synthesize acquisition-supervised trajectories

$\widehat{\mathcal{D}}_{\mathrm{traj}} \leftarrow \emptyset$
for each trajectory to synthesize **do**
    Sample a volume $\mathcal{V} \in \mathbb{V}$ and a programmed acquisition trajectory $\mathbf{\Xi}_{1:T} = \{(\mathbf{T}_t, \boldsymbol{\kappa}_t)\}_{t=1}^{T}$
    Obtain anatomical masks by volumetric resampling: $\mathbf{m}_t \leftarrow \Pi(\mathcal{V}; \mathbf{T}_t, \boldsymbol{\kappa}_t)$
    Select a clinical reference image $\mathbf{r}$
    Generate $\widehat{\mathbf{X}}_{1:T} \sim p_{\phi^\star}(\mathbf{X}_{1:T} \mid \mathbf{r}, \mathbf{M}_{1:T})$
    Store $(\widehat{\mathbf{X}}_{1:T}, \mathbf{\Xi}_{1:T})$ and associated conditioning metadata in $\widehat{\mathcal{D}}_{\mathrm{traj}}$
end for

3. Self-distill into an action-conditioned world model

Initialize the shared video-model weights θ from $\phi^\star$
Initialize Φ_ψ with a zero-output projection
for training stage $s \in \{1, 2\}$ **do**
    Set trainable parameters: $\Omega_1 = \{\psi\}$ and $\Omega_2 = \{\theta, \psi\}$
    for each optimization step **do**
        Sample a synthesized video, its acquisition trajectory, and conditioning metadata
        Encode the clean video latent $\mathbf{z}_0 \leftarrow \mathcal{E}(\widehat{\mathbf{X}}_{1:T})$
        Construct $\mathcal{C}_1$ from the initial frame and associated text condition
        Build first-frame-relative sampling maps: $\mathbf{G}_{1:T} \leftarrow \mathrm{AsMap}(\mathbf{\Xi}_{1:T}; \mathbf{T}_1)$
        Encode token-aligned geometry features: $\mathbf{E} \leftarrow \Phi_\psi(\mathbf{G}_{1:T})$
        Randomly drop $\mathbf{E}$ according to the stage-specific conditioning-dropout probability
        Sample a shifted flow time τ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
        $\mathbf{z}_\tau \leftarrow (1 - \tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}$
        Predict $\widehat{\boldsymbol{v}} \leftarrow \boldsymbol{v}_\theta(\mathbf{z}_\tau, \tau \mid \mathcal{C}_1, \mathbf{E})$
        $\mathcal{L} \leftarrow w(\tau)\ \mathrm{MSE}_{\mathrm{future}}\big(\widehat{\boldsymbol{v}}, \boldsymbol{\epsilon} - \mathbf{z}_0\big)$
        Update Ω_s with AdamW using $\mathcal{L}$
    end for
end for
return (θ, ψ)

C Robotic Ultrasound Planning Details

Algorithm 3 summarizes our closed-loop planning implementation. The planner interacts with an ultrasound scanning environment that provides the current observation and supports feasible probe motion. At each planning step, the planner accesses the current observation, the goal image, and the motion information required for planning, but not the target pose or future observations. UltraWorld remains frozen throughout planning.

Each candidate is predicted jointly from the current observation. Only the selected prefix is executed, after which the planner obtains a new observation from the environment and replans. The procedure runs for a fixed budget of eight MPC iterations without early termination or additional MR resampling.

Algorithm 2 Action-Conditioned Inference with UltraWorld

Require: Initial ultrasound frame $\mathbf{x}_1$; prescribed acquisition trajectory $\boldsymbol{\Xi}_{1:T}$; text condition t; trained parameters (θ, ψ); descending flow schedule $1 = \tau_S > \cdots > \tau_0 = 0$
Ensure: Predicted future ultrasound frames $\widehat{\mathbf{X}}_{2:T}$
Construct initial-frame and text conditioning $\mathcal{C}_1 \leftarrow \mathrm{I2VCondition}(\mathbf{x}_1, t)$
$\mathbf{G}_{1:T} \leftarrow \mathrm{AsMap}(\boldsymbol{\Xi}_{1:T}; \mathbf{T}_1)$
$\mathbf{E} \leftarrow \Phi_\psi(\mathbf{G}_{1:T})$
Initialize the complete video latent $\mathbf{z}_{\tau_S} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $i = S, \ldots, 1$ **do**
    $\boldsymbol{v} \leftarrow \boldsymbol{v}_\theta(\mathbf{z}_{\tau_i}, \tau_i \mid \mathcal{C}_1, \mathbf{E})$
    $\mathbf{z}_{\tau_{i-1}} \leftarrow \mathbf{z}_{\tau_i} + (\tau_{i-1} - \tau_i)\boldsymbol{v}$
end for
Decode $\widehat{\mathbf{X}}_{1:T} \leftarrow \mathcal{D}_{\mathrm{VAE}}(\mathbf{z}_{\tau_0})$
return $\widehat{\mathbf{X}}_{2:T}$

Algorithm 3 Closed-Loop Ultrasound Planning with UltraWorld

Require: Frozen UltraWorld $\mathcal{W}$; scan environment $\mathcal{E}$; initial state i; goal image $\mathbf{x}_g$
Ensure: Executed observation–action sequence $\mathcal{R}$
$\mathbf{x} \leftarrow \mathrm{Observe}(\mathcal{E}, i)$; $\mathcal{R} \leftarrow \emptyset$
for MPC iteration $t = 1, \ldots, 8$ **do**
    Initialize four categorical action distributions uniformly over five motion levels
    Construct feasible four-block action sequences under the environment motion constraints
    $\mathcal{B} \leftarrow \emptyset$ ▷ Evaluated candidates
    for CEM iteration $\ell = 1, \ldots, 3$ **do**
        Sample up to 16 unevaluated feasible plans using the current action distributions
        Include the stationary plan in the first iteration
        for each sampled plan $\mathbf{U}$ **do**
            Repeat each block action for four steps to obtain a 16-step motion trajectory
            Construct AsMap conditions $\mathbf{G}^{\mathbf{U}}$ relative to the current probe pose
            Predict $\widehat{\mathbf{X}}^{\mathbf{U}}_{0:16} \leftarrow \mathcal{W}(\mathbf{x}, \mathbf{G}^{\mathbf{U}})$ using a shared sampling seed
            $J(\mathbf{U}) \leftarrow 0.7\, d_{\mathrm{LPIPS}}(\hat{\mathbf{x}}^{\mathbf{U}}_4, \mathbf{x}_g) + 0.3\, d_{\mathrm{LPIPS}}(\hat{\mathbf{x}}^{\mathbf{U}}_8, \mathbf{x}_g)$
            Add the plan and its cost to $\mathcal{B}$
        end for
        Select the four lowest-cost plans from all candidates in $\mathcal{B}$
        Update action distributions using smoothed elite frequencies and uniform exploration
    end for
    Select the lowest-cost evaluated plan $\mathbf{U}^\star$ from $\mathcal{B}$
    Execute only its first four-step block in $\mathcal{E}$
    Update i to the reached state and obtain $\mathbf{x} \leftarrow \mathrm{Observe}(\mathcal{E}, i)$
    Append the executed actions and observations to $\mathcal{R}$
end for
return $\mathcal{R}$

D Additional Experimental Details

D.1 Data Preparation and Evaluation Sets

Data sources and splits. BUV (Lin et al., 2022) and CAMUS (Leclerc et al., 2019) provide untracked clinical ultrasound videos for learning the mask-conditioned ultrasound video prior, covering breast ultrasound and echocardiography, respectively. The videos are paired with frame-wise anatomical masks. I-SPY1 (Newitt et al., 2016; Hylton et al., 2016) provides volumetric breast DCE-MRI and lesion annotations used to construct anatomy-guided acquisition trajectories. All datasets are split at the patient level, and derived clips remain in the partition of their source patient.

Synthetic trajectory preparation. Programmed acquisition trajectories determine the anatomical cross-sections resampled from each I-SPY1 volume. The resulting mask trajectories are rendered

Table 4: **Architectures of token-level action encoders.**

Encoder	Architecture	Parameters
Fourier Encoding	Fourier features + MLP	1.11M
MLP Encoding	MLP	0.79M
MLP Encoding (Matched)	Matched-size MLP	53.49M
PRoPE	30-layer Linear	70.82M
AsMap (Ours)	Convolution + ResBlock	53.48M

into ultrasound videos by the mask-conditioned generator, yielding paired video–acquisition trajectories for world-model training and evaluation. World-model clips contain 17 frames at 512×512 resolution. Forward and backward subsequences use temporal strides of one or two and are supplemented with palindromic and stationary trajectories. RGB intensities are normalized to $[-1, 1]$ before VAE encoding. Video latents, first-frame image features, text embeddings, and synchronized acquisition states are precomputed for training.

Real-clinical evaluation sets. Real-initial-frame experiments use the patient-held-out BUV split and the independent BUSI dataset. Each clinical image serves as the initial observation, and the world model predicts a subsequent sequence under a prescribed acquisition trajectory. No paired clinical future observations are available, so these experiments evaluate observation stability and recovery consistency rather than direct prediction against a ground-truth future.

D.2 Baseline and Ablation Settings

Action-conditioning baselines. All conditioning variants use the same video backbone and receive the same acquisition state, initial-frame observation condition, and text condition. Fourier Enc. (Fan et al., 2026) and MLP Enc. (Yue et al., 2025) convert frame-level acquisition information into global embeddings that are broadcast across spatial tokens and added to the input video tokens. AsMap instead produces spatially aligned token features. PRoPE (Li et al., 2025) applies projective geometric conditioning inside self-attention and is therefore not converted to the same additive token-conditioning interface. Table 4 summarizes the action encoders used in our comparisons. Parameter counts include only the additional conditioning encoder and exclude the shared video backbone.

Controlled ablations. The matched MLP uses the same first-frame-relative coordinate convention as AsMap and approximately matches its encoder parameter count. This ablation therefore controls for encoder capacity and reference-frame convention without imposing an identical temporal representation.

For the initialization ablation, we retain AsMap and exactly the same synthesized action–video trajectories while replacing initialization from the mask-conditioned ultrasound generator with initialization from the corresponding Wan2.1 base checkpoint. This isolates the effect of prior-weight inheritance while keeping the generator-produced supervision fixed.

Training and inference configuration. The mask-conditioned ultrasound generator is initialized from Wan2.1-FunControl-1.3B-V1.1 (Wan Team et al., 2025). All models are trained on a single NVIDIA H200 GPU using 512×512 videos with 17 frames per clip and a batch size of 8. We use AdamW (Loshchilov & Hutter, 2019) for optimization.

The mask-conditioned generator is trained for 2K steps with a learning rate of 1×10^{-4}. The ultrasound world model is trained for 10K steps with a learning rate of 2×10^{-6}. As described in Sec. 4.3, world-model adaptation first trains the acquisition encoder with the video backbone frozen and then jointly fine-tunes the acquisition encoder and video backbone.

At inference, all methods use 25 Euler sampling steps with a timestep-shift factor of 5. Text and acquisition guidance scales are both set to 1.

D.3 Evaluation Metrics and Protocols

Prediction fidelity. For world-model fidelity evaluation, the conditioning frame is excluded from all reported metrics, and videos decoded from clean latents are used as references. We report PSNR, LPIPS, and latent ℓ_2 distance. LPIPS uses an AlexNet backbone, and latent ℓ_2 is implemented as mean-squared error in latent space.

Segmentation-based generator evaluation. We apply a frozen UltraSam segmenter (Meyer et al., 2026) to generated and reference ultrasound frames using identical box prompts derived from connected lesion components in the conditioning masks. Mean IoU between segmentations of generated and reference frames measures anatomical consistency, while mean Dice between generated segmentations and conditioning masks measures mask controllability. Frames with no lesion in the conditioning mask are excluded. No empty predictions occurred in this evaluation.

Action following. Following Li et al. (2026), we compute Δ-LPIPS over a four-frame interval. For reference and predicted videos, respectively,

$$\mathbf{D}_t = \mathbf{x}_{t+4} - \mathbf{x}_t, \qquad \widehat{\mathbf{D}}_t = \hat{\mathbf{x}}_{t+4} - \hat{\mathbf{x}}_t. \tag{20}$$

Each difference image is independently RMS-normalized:

$$\Delta\text{-LPIPS} = \frac{1}{|\mathcal{P}_4|} \sum_{t \in \mathcal{P}_4} d_{\text{LPIPS}} \left(\mathcal{N}(\widehat{\mathbf{D}}_t), \mathcal{N}(\mathbf{D}_t) \right), \qquad \mathcal{N}(\mathbf{D}) = \frac{\mathbf{D}}{\sqrt{\text{mean}(\mathbf{D}^2) + \epsilon}}, \tag{21}$$

where $\epsilon = 10^{-8}$. The set $\mathcal{P}_4$ contains valid four-frame pairs whose reference difference exceeds the static RMS threshold. Scores are first averaged over valid pairs within each video and then across videos; videos with no valid pairs are excluded.

Motion Error, defined in Eq. 11, compares adjacent-frame target-anatomy displacements. Reference and predicted lesion centroids are obtained using UltraSAM with identical annotation-derived box prompts. The annotation masks are used only to construct prompts and do not provide the reference centroid locations. The set $\mathcal{P}$ in Eq. 11 contains adjacent-frame pairs for which both reference and predicted centroids exist at both time points. Motion errors are averaged over valid pairs within each video and then over valid videos.

Real-clinical consistency. Real-initial-frame experiments are evaluated separately on BUSI and BUV. Zero-action LPIPS measures the average perceptual distance between the real input observation and predicted future frames while the acquisition remains stationary.

Round-trip LPIPS uses a 17-frame single-axis forward-and-return trajectory, $0 \rightarrow 8 \rightarrow 0$, and compares the real input observation with the final predicted frame after returning to the starting acquisition state. Both metrics use the real input frame, rather than a generated conditioning frame, as the reference. Results are averaged over cases within each dataset. These metrics evaluate observation stability and return consistency without requiring paired clinical futures; round-trip consistency alone does not establish action-following accuracy.

Robotic planning evaluation. We report terminal translation error and terminal rotation error. Let $(\mathbf{p}_f, \mathbf{R}_f)$ and $(\mathbf{p}_g, \mathbf{R}_g)$ denote the final executed and target probe poses. The terminal errors are

$$\begin{aligned} e_{\text{pos}} &= \|\mathbf{p}_f - \mathbf{p}_g\|_2, \\ e_{\text{rot}} &= \frac{180}{\pi} \left\| \text{rotvec}(\mathbf{R}_g^\top \mathbf{R}_f) \right\|_2, \end{aligned} \tag{22}$$

reported in millimeters and degrees, respectively. Both metrics are evaluated at the end of the fixed execution budget. The target pose is used only for evaluation; candidate selection during planning uses predicted-image LPIPS and never accesses the target pose.

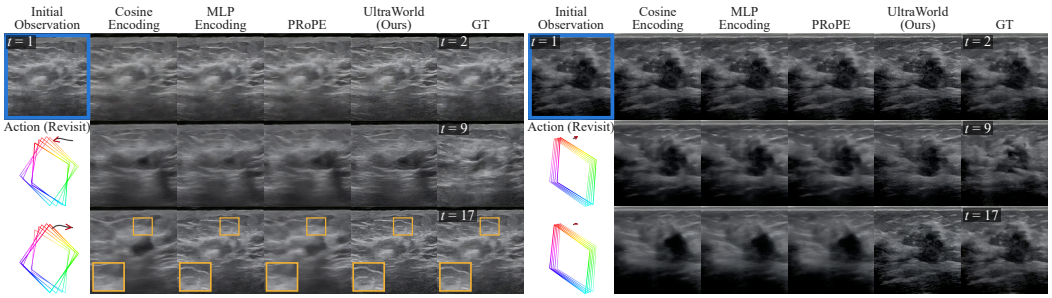

Figure 9: **Qualitative comparison under revisiting.** UltraWorld and alternative conditioning methods predict future ultrasound frames from the same initial observation and prescribed trajectory. Comparisons with reference futures highlight differences in action following and the continuity of local tissue appearance.

E Additional Results and Limitations

E.1 Mask-Conditioned Generator Evaluation

Generation fidelity. We evaluate the mask-conditioned ultrasound generator using both image-level and task-aware metrics. Generated videos achieve an SSIM of 0.4265 and an LPIPS of 0.4071. Following prior work evaluating synthetic ultrasound through downstream analysis (Yu et al., 2026a; Stojanovski et al., 2023), we additionally apply a pretrained ultrasound segmenter to generated and ground-truth frames. The resulting segmentations achieve a mean IoU of 0.8736, indicating that the generated videos preserve the relevant anatomical structures.

Mask controllability. We further evaluate whether the generated anatomy follows the prescribed mask trajectories. Comparing lesion regions segmented from generated frames with the conditioning masks yields a Dice score of 0.9260, compared with 0.9538 between ground-truth segmentations and the same masks. Frames without lesions in the conditioning masks are excluded. Metric and segmentation protocols are detailed in Appendix D.3.

E.2 Action-Conditioned Appearance Dynamics

A local 2D ultrasound observation can be compatible with different underlying 3D anatomies, leaving future sections only partially determined. Our focus is therefore on learning how scanning operations change ultrasound and anatomical appearance. Synthetic action–video pairs provides explicit supervision for these changes, while self-distillation transfers the resulting action–observation relationships into a model that predicts from local observations and actions.

Figure 9 extends Figure 6 to revisiting. UltraWorld shows more consistent responses to the prescribed operations while retaining plausible hypoechoic regions and surrounding tissue appearance. These results illustrate the transfer from generation under explicit anatomical control to predicting motion-induced visual changes while maintaining coherent local structure.

E.3 Additional world-model rollouts for robotic planning

Figure 10 provides additional closed-loop robotic planning examples, showing UltraWorld rollouts together with the corresponding simulator executions and probe-pose errors to the goal over the planning process. These examples complement the representative episode shown in Fig. 8 of the main paper.

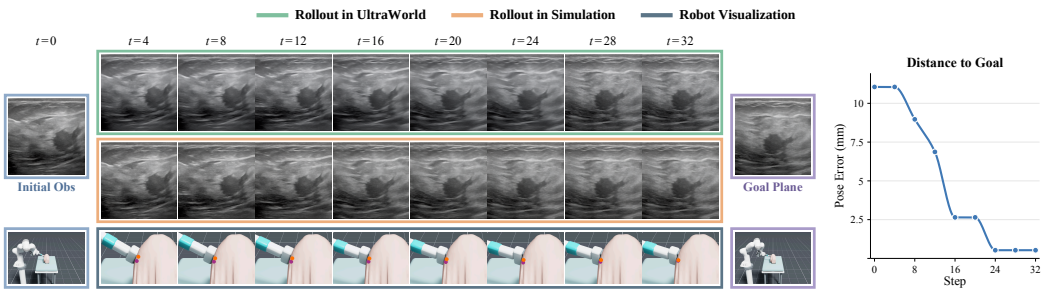

Figure 10: **Additional qualitative rollouts for robotic planning.** Additional closed-loop examples showing UltraWorld rollouts and the corresponding simulator executions, together with the probe-pose error to the goal during planning.

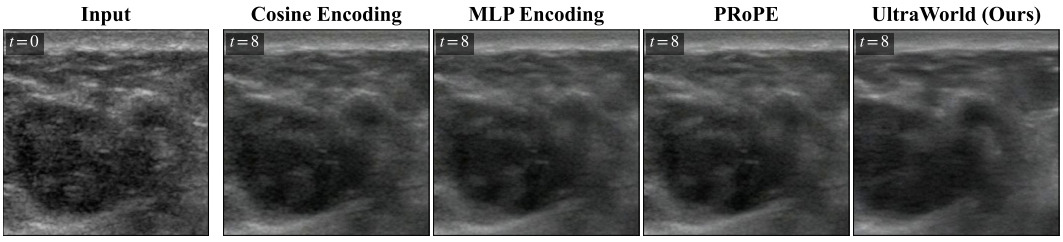

Figure 11: **Action-conditioned rollouts from real clinical ultrasound observations.** Each method is initialized from the same real ultrasound image and follows a prescribed acquisition trajectory. The intermediate predictions visualize how the predicted anatomy evolves in response to the prescribed action.

E.4 Qualitative Generalization to Real Clinical Ultrasound

To complement the quantitative real-clinical evaluation in Table 3, we visualize action-conditioned predictions initialized from unseen clinical ultrasound images in Fig. 11. The qualitative examples illustrate how the predicted anatomy evolves in response to prescribed acquisition actions, while the round-trip results in Table 3 evaluate whether the prediction returns close to the initial observation after a forward-and-return trajectory. Because paired future observations are unavailable, these experiments assess action-sensitive behavior and recovery consistency rather than direct future-prediction fidelity.

E.5 Counterfactual Acquisition Trajectories

We additionally evaluate action sensitivity by applying alternative prescribed acquisition trajectories to the same initial observation, as illustrated in Fig. 12. These counterfactual examples isolate the effect of the conditioning trajectory: the visual context is fixed, while the requested probe motion or imaging configuration is changed. As shown in Fig. 12, different prescribed trajectories lead to different predicted anatomical evolutions, highlighting whether each action conditioning method can produce distinct and plausible motion-dependent responses.

E.6 Limitations.

UltraWorld still has several limitations. First, the world model is trained primarily with generator-synthesized action–video trajectories rather than real ultrasound sequences with synchronized probe tracking, so errors or biases in the generative prior may propagate to the distilled world model. Second, AsMap represents acquisition geometry but does not explicitly model tissue-dependent acoustic effects such as scattering, attenuation, and speckle formation, and its metric interpretation assumes calibrated probe poses and imaging geometry. Third, evaluation on real clinical ultrasound is limited to stability and round-trip consistency because paired future observations with known

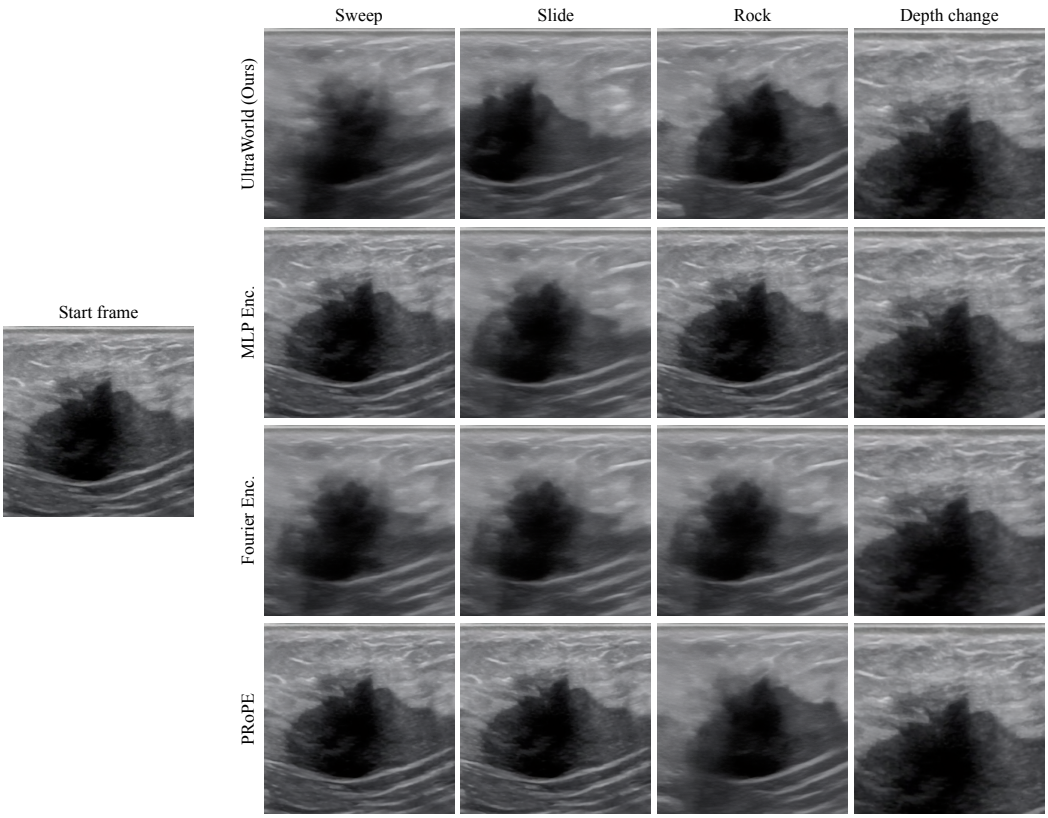

Figure 12: **Counterfactual predictions under alternative acquisition trajectories.** Starting from the same ultrasound observation, we prescribe sweep, slide, rock, and depth-change trajectories and compare the resulting predictions across action conditioning methods.

actions are unavailable, while closed-loop robotic planning is currently validated only in simulation and on a limited number of episodes. Future work will explore using a small amount of tracked clinical data for calibration and validation, while preserving the scalability of learning from large-scale untracked clinical videos, as well as extending to broader anatomical and device domains and physical robotic systems.