DreamReg: Belief-Driven World Model for 2D–3D Ultrasound Registration

Luoyao Kang[1], Yuelin Zhang[1], Jiwei Shan[1], Haifan Gong[3], Qingpeng Ding[1], and Shing Shin Cheng[1,2]✉

[1] Department of Mechanical and Automation Engineering and T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Shun Hing Institute of Advanced Engineering and Multi-scale Medical Robotics Center, Hong Kong SAR, China
sscheng@cuhk.edu.hk
[3] Perelman School of Medicine, University of Pennsylvania, PA, USA

Abstract. Ultrasound (US) is widely used for surgical navigation, yet real-time registration between intraoperative 2D slices and preoperative 3D volumes remains challenging due to partial observability, speckle noise, and the action-dependent US acquisition. Existing methods are one-shot or short-horizon, making it hard for them to gather evidence over time or capture how surgeons adjust probe motion based on on-screen feedback. We propose DreamReg, a belief-driven world-model framework that formulates 2D–3D registration as belief updating over rigid transformations. DreamReg maintains a latent belief state that summarizes past observations and poses information, and continuously refines the transformation through learned dynamics as new slices arrive. During training, DreamReg is exposed to probe-motion trajectories that mimic clinical scanning behavior and learns to update its belief by conditioning pose refinement on the current US observation. During inference, DreamReg refines registration via internal imagination: it rolls out the learned world model to simulate candidate probe motions and their predicted observations, and integrates these imagined outcomes to converge to an accurate rigid transformation. Experiments on CAMUS and μ-RegPro datasets demonstrate improved robustness and competitive registration accuracy for real-time guidance compared with state-of-the-art methods. Project page: https://github.com/kangluoyao/DreamReg.

Keywords: 2D–3D Registration · World Model · Ultrasound Image · Surgical Navigation

1 Introduction

Ultrasound (US) is widely used for surgical navigation due to its real-time and safe imaging capabilities [9, 24, 26]. In many workflows, a preoperative 3D US volume provides global context, while intraoperative guidance relies on streaming 2D slices [5, 8, 15], making real-time 2D–3D registration essential for navigation. However, registration is fundamentally challenging: 2D slices offer partial,

speckle-corrupted observations, and subsequent views depend on probe motion, leading to strong action-observation coupling under partial observability [18].

Learning-based approaches for ultrasound registration can be broadly categorized into two paradigms. The first learns 2D/3D features and recovers rigid transformations via geometric fitting [6], leveraging CNN descriptors to establish correspondences before solving for 6-DoF pose [3, 21]. The second paradigm performs end-to-end pose regression [10, 17, 27, 33] models, but typically treat slices as passively observed and operate in a one-shot or short-horizon manner. In practice, clinicians actively adjust probe pose based on visual feedback to disambiguate anatomy [1, 23, 30]. Ignoring this coupling often leads to brittle performance under sparse views. Recent advances in world model provides a framework for decision-making under partial observability by learning latent dynamics that couple actions and observations [11–13]. It demonstrated strong performance in robotics and control by enabling internal simulation and imagination-based planning without exhaustive real-world interaction [20, 31]. In medical imaging, however, world-model formulations remain largely unexplored [22, 32], and existing registration approaches rarely model how probe motion influences future observations through learned latent dynamics.

Motivated by this gap, we propose **DreamReg**, a belief-driven world-model framework that reformulates 2D–3D US registration as sequential belief updating over rigid transformations. Rather than predicting pose from an isolated slice, DreamReg maintains a latent belief state representing the current alignment hypothesis under partial observability [11, 12]. This belief evolves through learned action-conditioned dynamics, coupling probe motion and observation transitions in latent space. During training, monitored probe maneuvers supervise belief transitions, yielding a data-driven prior over clinically plausible scanning dynamics. During inference, DreamReg performs imagination-based rollouts: the learned world model simulates candidate probe motions and predicted observations to iteratively refine belief, without dense correspondences or heuristic exploration. By belief updating and action-conditioned modeling, DreamReg departs from conventional one-shot regression and improves robustness in challenging intraoperative scenarios.

Contributions. (1) Real-time 2D–3D ultrasound registration is reformulated as a belief-driven sequential decision problem under partial observability. (2) An action-conditioned world model is proposed to explicitly capture the coupling between probe motion and observation in a latent space for robust pose refinement. (3) A belief rollout mechanism is introduced to enable model-based imagination during inference. (4) Comprehensive evaluations are conducted on CAMUS [16] and μ-RegPro [2], where consistent improvements in both geometric accuracy and image similarity are demonstrated over the state-of-the-art baselines.

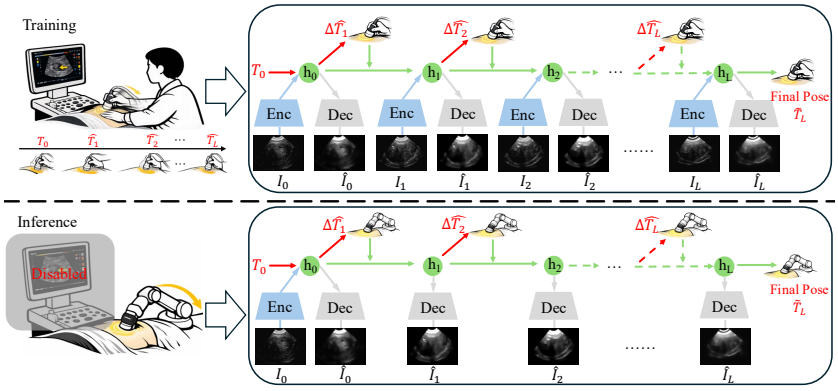

Fig. 1. Belief-driven world-model registration in training and inference. *Training (top):* Mimicking a clinician with screen feedback, the model iteratively refines the pose; *Inference (bottom):* Without observations, the model performs closed-loop refinement by imagination.

2 Methodology

2.1 Problem Formulation

Given a preoperative 3D US volume $V \in \mathbb{R}^{1\times D\times H\times W}$ and an intraoperative target 2D US slice $I_g \in \mathbb{R}^{1\times H\times W}$, the objective is to estimate a 6-DoF rigid US probe pose $T^\star \in \mathbb{R}^6$ such that rendering a slice from the volume at $T^\star$ matches I_g. A differentiable physics-based slice renderer is given by $I(T_t) = \mathcal{R}(V, T_t) \in \mathbb{R}^{1\times H\times W}$ with a noise-perturbed expert pose T_t and a sampling operator R. Instead of directly regressing $T^\star$ in a one-shot manner, we formulate registration as an iterative decision process, as shown in Fig.1: starting from an initial pose T_0, at each step t the agent predicts an incremental motion $\Delta\hat{T}_t \in \mathbb{R}^6$ and updates the probe-pose estimate as $\tilde{T}_{t+1} = \tilde{T}_t + \Delta\hat{T}_t$. Here, $\tilde{T}_t$ denotes the agent's current estimate of the probe pose at step t.

2.2 Belief-Driven World Model for Registration

Model overview. As illustrated in Fig. 1, training (top) performs an observe rollout where rendered slices are encoded to update the belief state via posterior inference, whereas inference (bottom) performs a closed-loop imagination rollout using only the learned prior without new observations. Our model, `DreamReg`, follows a Dreamer-style [12, 13] latent world model for iterative pose refinement. At each step t, the model maintains a latent belief state composed of: (i) a *stochastic latent variable* z_t, which represents the latent encoding of the current slice conditioned on pose and belief (see the prior/posterior blocks in Fig. 2). z_t captures a stochastic representation of the latent observation state under partial observability. And (ii) a *belief update* h_t, which aggregates temporal information

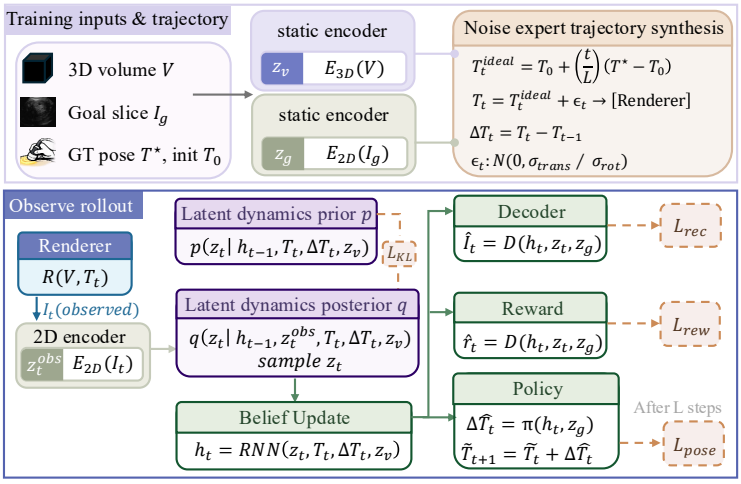

Fig. 2. Training-time rollout. A noise-perturbed expert trajectory is rendered to produce observations. The posterior infers latent states conditioned on observations and is regularized by the prior via the KL divergence. The belief state is updated recurrently and used for slice reconstruction, reward prediction, and pose refinement.

through recurrent updates (Fig. 1, green nodes). The latent transition is modeled by a learned prior $p_\theta(z_t \mid \cdot)$, while a posterior $q_\phi(z_t \mid \cdot)$ incorporates the rendered observation during training. The sampled z_t updates the belief state h_t, and conditions slice reconstruction, reward prediction, and policy learning (Fig. 2).

Encoders for static context and goal. As shown in Fig. 2, the 3D volume V and target slice I_g are encoded by 3D and 2D CNNs, yielding embeddings z_v and z_g.

Latent dynamics with prior/posterior. At training time, the current slice is rendered as $I_t = \mathcal{R}(V, T_t)$ along a known pose trajectory. The slice I_t is encoded into an observation embedding $z_t^{\text{obs}} = P_s(E_{2D}(I_t))$. The world model defines a Gaussian prior and posterior over z_t. The prior predicts the next latent state without observing the current slice, while the posterior incorporates the rendered observation during training (see purple blocks in Fig. 2):

$$p_\theta(z_t \mid h_{t-1}, \Delta T_t, T_t, z_v) = \mathcal{N}\big(\mu_t^p, \mathrm{diag}(\sigma_t^{p2})\big), \tag{1}$$

$$q_\phi(z_t \mid h_{t-1}, \Delta T_t, T_t, z_v, z_t^{\text{obs}}) = \mathcal{N}\big(\mu_t^q, \mathrm{diag}(\sigma_t^{q2})\big), \tag{2}$$

where $(\mu, \log\sigma^2)$ are produced by two MLPs. Sampling is performed using the reparameterization trick: $z_t = \mu_t^q + \sigma_t^q \odot \epsilon, \epsilon \sim \mathcal{N}(0, I)$.

Belief update. The belief state is updated by a recurrent module applied to a projected input:

$$u_t = W_u[\, z_t;\ \Delta T_t;\ T_t;\ z_v\,], \quad h_t = \mathrm{RNN}(u_t, h_{t-1}), \tag{3}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes concatenation and RNN is a recurrent transformer block [4]. Importantly, we do *not* feed the goal embedding z_g into the latent dynamics

update to prevent shortcut learning, where the model directly learns the final pose from the goal embedding. Instead, z_g only functions as a *task condition* for decoding and reward. This belief update accumulates information across time steps.

Imagination decoder and reward model. Given (h_t, z_t, z_g), we decode an imagined slice $\hat{I}_t$ and predict a scalar reward vector $\hat{r}_t$ (Fig. 2, Decoder and Reward head):

$$\hat{I}_t = D_\psi([h_t; z_t; z_g]), \qquad \hat{r}_t = R_\psi([h_t; z_t; z_g]), \tag{4}$$

where D_ψ reconstructs the slice from the latent belief, and R_ψ outputs two values measuring translational and rotational progress. The decoder provides reconstruction supervision for world-model learning, while the reward head supplies dense progress signals for action supervision.

Policy for pose refinement. At the inference phase, real slices are no longer needed while the pose is purely refined by imagination (Fig. 1, bottom). The policy predicts pose increments from the current belief and goal embedding: $\hat{\Delta T}_t = \pi_\omega([h_t; z_g])$. The belief is then updated using the learned prior:

$$z_t \sim p_\theta(z_t \mid h_{t-1}, \hat{\Delta T}_t, T_t, z_v), \quad h_t = \mathrm{RNN}(W_u[z_t; \hat{\Delta T}_t; T_t; z_v], h_{t-1}). \tag{5}$$

Importantly, the prior models how the latent state evolves *given* an action, whereas the policy predicts *which* action should be taken. The two components therefore play complementary roles: the prior learns latent transition dynamics, while the policy performs action selection from the belief state. Without the policy head, the model would not be able to generate pose updates during inference, since the prior requires an action input but does not predict one.

2.3 Training via Observe Rollout

Training follows the observe rollout pipeline shown in Fig. 2, which uses supervised pose labels $T^\star$ for each pair (V, I_g). We synthesize a length-L trajectory from T_0 to $T^\star$ and inject noise to emulate off-trajectory states.

Trajectory synthesis. Let L be the number of world-model steps. We define an "ideal" linear path $T_t^{\text{ideal}} = T_0 + \frac{t}{L}(T^\star - T_0)$ for $t = 1 \dots L$, and then add Gaussian noise $\eta_t \sim \mathcal{N}(0, \sigma^2 I)$ to obtain T_t. Actions are defined consistently as differences between consecutive noisy poses $\Delta T_t = T_t - T_{t-1}$.

Observe rollout. Given $(V, I_g, \{T_t\}_{t=1}^L, \{\Delta T_t\}_{t=1}^L)$, we run an observe rollout: at each step, we render $I_t = \mathcal{R}(V, T_t)$, infer the posterior $q_\phi(z_t|\cdot)$, update belief h_t, decode $\hat{I}_t$, predict $\hat{r}_t$, and accumulate the KL divergence.

World-model learning. We train DreamReg with a sum of rollout-averaged losses: $\mathcal{L} = \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{rew}} + \mathcal{L}_{\text{pose}}$. Latent dynamics are regularized by minimizing the KL divergence between the posterior and the prior, while the decoder reconstructs the rendered slices. Formally, we define

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_t[D_{\text{KL}}(q_\phi(z_t|\cdot) \| p_\theta(z_t|\cdot))], \quad \mathcal{L}_{\text{rec}} = \mathbb{E}_t\Big[\|\hat{I}_t - I_t\|_1 + \mathcal{L}_{\text{SSIM}}(\hat{I}_t, I_t)\Big], \tag{6}$$

Table 1. State-of-the-art comparison on CAMUS and μ-RegPro (mean±std). Arrows indicate whether lower (↓) or higher (↑) is better.

CAMUS							
Method	DisErr↓	I-NCC↑	SSIM↑	TransErr↓	RotErr↓	P-NCC↑	FPS↑
CUReg [17]	9.77±0.03	79.72±0.15	38.48±0.65	7.60±0.03	7.75±0.02	53.10±0.75	37
EUReg [27]	8.40±0.33	86.69±1.54	44.33±2.20	5.59±0.32	7.98±0.07	64.55±1.19	**189**
FVRNet [10]	9.38±1.14	82.23±5.01	40.08±4.15	6.72±1.14	8.10±0.22	58.49±7.58	55
DreamReg	**7.16**±0.13	**90.45**±0.14	**51.87**±0.35	**4.77**±0.07	**7.08**±0.15	**72.12**±1.62	40
μ-RegPro							
Method	DisErr↓	I-NCC↑	SSIM↑	TransErr↓	RotErr↓	P-NCC↑	FPS↑
CUReg [17]	12.43±0.48	55.19±6.18	22.27±2.60	8.53±0.72	9.30±0.03	22.37±15.38	36
EUReg [27]	11.89±0.06	63.03±0.57	27.61±0.43	7.74±0.12	9.29±0.01	36.26±2.10	**202**
FVRNet [10]	11.83±0.66	58.99±6.63	24.67±3.97	7.71±0.95	9.31±0.02	34.85±10.90	51
DreamReg	**11.06**±0.24	**66.68**±2.24	**30.99**±1.95	**6.86**±0.36	**9.19**±0.02	**44.39**±3.34	56

where $\mathbb{E}_t$ denotes the average over the L time steps ($t = 1, \ldots, L$).

Progress reward and pose supervision. We supervise the reward head by aligning predicted actions with the direction to the ground-truth pose. The reward head provides supervision on directional progress. Let $\tilde{T}_t$ be the policy-updated pose and $d_t = T^\star - \tilde{T}_t$, we define:

$$r_t = [\cos(\hat{\Delta T}_t^{1:3}, d_t^{1:3}),\, \cos(\hat{\Delta T}_t^{4:6}, d_t^{4:6})], \quad \mathcal{L}_{\text{rew}} = \frac{1}{L}\sum_{t=1}^{L} \|\hat{r}_t - r_t\|_2^2 \tag{7}$$

Pose prediction is supervised by a Smooth$\mathcal{L}_1$ loss:

$$\mathcal{L}_{\text{pose}} = \text{SmoothL1}(\tilde{T}_L^{1:3}, T^{\star\,1:3}) + \text{SmoothL1}(\tilde{T}_L^{4:6}, T^{\star\,4:6}). \tag{8}$$

2.4 Inference via Imagination Rollout

During inference, we perform the imagination rollout shown in Fig. 1 (bottom). Given (V, I_g), we compute z_v and z_g once and initialize $T_0 = \mathbf{0}, h_0 = \mathbf{0}$. For L steps: (i) predict action $\hat{\Delta T}_t = \pi_\omega([h_t; z_g])$, (ii) update pose $\tilde{T}_{t+1} = \tilde{T}_t + \Delta T_t$ with angle wrapping to $(-\pi, \pi]$, (iii) sample $z_t \sim p_\theta(\cdot)$ and update belief via the prior. This closed-loop imagination refinement enables pose estimation without real-time image feedback. The final pose $\tilde{T}_L$ is returned.

3 Experiments and Results

3.1 Datasets and Implementation Details

We evaluate DreamReg on two public 3D ultrasound datasets: CAMUS [16] and the ultrasound subset of μ-RegPro [2]. For CAMUS, we follow the official split.

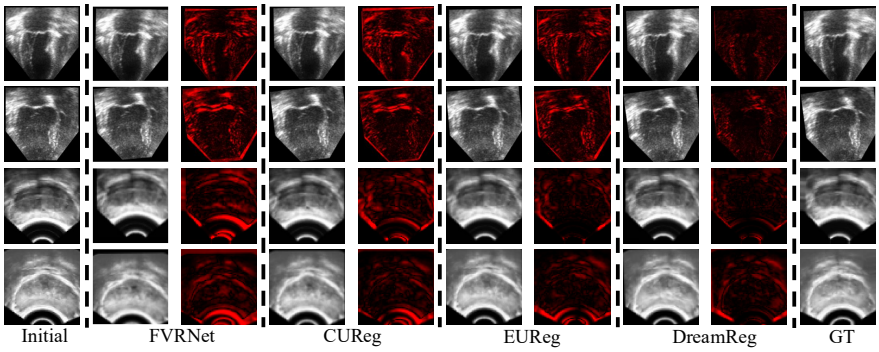

Fig. 3. Qualitative comparison on registration results of different methods. Top two rows: CAMUS; bottom two rows: μ-RegPro. Red overlays indicate intensity differences between the target slice and the resampled slice.

The dataset contains 1000 cardiac US volumes. We uniformly sample 4 slices per case and resample each 3D volume to $(32, 192, 192)$. For μ-RegPro, we use a case-level split with a 7:1:2 ratio for training/validation/testing, which contains 73 prostate volumes in total. We sample 8 slices per case and resample volumes to $(64, 64, 64)$. During training, the initial 6-DoF pose perturbations are randomly sampled within ± 10 (mm for translation and degrees for rotation) to simulate realistic misalignment, consistent with prior frame-to-volume registration settings [17, 27]. Following EUReg [27], training uses random sampling on-the-fly, while for evaluation, we pre-sample a fixed test set during preprocessing to ensure reproducibility (400 slices for CAMUS and 120 for μ-RegPro).

All experiments are conducted on a single NVIDIA A800 GPU with a batch size of 64 for 100 training epochs. Optimization is performed using AdamW [19] with a two-stage learning rate schedule. The initial learning rate is set to 1×10^{-4} for CAMUS and 1×10^{-5} for μ-RegPro. The number of world-model rollout steps L is set to 7 for CAMUS and 5 for μ-RegPro. For evaluation, we report seven metrics covering geometric accuracy, image alignment, parameter consistency, and efficiency [17, 27]. Specifically, we use DisErr (mm), TransErr (mm), and RotErr (°) for pose accuracy; I-NCC (%) and SSIM (%) for image-level alignment [25, 28]; P-NCC (%) for parameter consistency; and frames per second (FPS) for runtime efficiency.

3.2 Comparison with the State-of-the-art Methods

We compare DreamReg with the SOTA 2D–3D registration baselines, including CUReg [17], EUReg [27], and FVRNet [10], on CAMUS [16] and μ-RegPro [2].
Overall performance across datasets. Across both CAMUS and μ-RegPro (Table 1), DreamReg consistently achieves the best overall performance in both geometric accuracy and image similarity metrics, demonstrating the robustness of belief-driven latent modeling over one-shot regression approaches. On both

Table 2. Ablation study on CAMUS (mean±std).

Method	DisErr↓	I-NCC↑	SSIM↑	TransErr↓	RotErr↓	P-NCC↑
only pose	7.35±0.04	90.12±0.23	50.81±0.96	4.84±0.07	7.28±0.06	70.85±0.34
w/o rew	7.34±0.03	89.82±0.29	49.93±0.54	4.87±0.05	7.19±0.07	71.33±0.67
w/o recon	7.33±0.12	89.84±0.37	50.12±0.73	4.89±0.08	7.16±0.17	71.06±1.58
DreamReg	**7.16±0.13**	**90.45±0.14**	**51.87±0.35**	**4.77±0.07**	**7.08±0.15**	**72.12±1.62**

datasets, DreamReg yields the lowest distance and translation errors while simultaneously achieving the highest I-NCC, SSIM, and P-NCC scores. Notably, the improvements are particularly pronounced in similarity-based metrics, indicating enhanced structural alignment rather than merely reduced pose discrepancies. Compared with the strongest baseline (EUReg), DreamReg consistently increases structural similarity across datasets, suggesting that iterative belief updates and latent representation modeling lead to more anatomically coherent registrations. The consistent gains across multiple datasets indicate that explicitly modeling the interaction between action and observation produces more stable and clinically meaningful alignment.

Analysis of runtime and practical considerations. In addition to accuracy, we report inference speed (FPS) in Table 1. Although EUReg achieves higher throughput (189–202 FPS), DreamReg operates at 40–56 FPS, exceeding real-time requirements (FPS > 30). The additional computational cost arises from iterative belief updates and latent dynamics modeling. While this reduces raw throughput, it yields improved geometric and structural accuracy. Given that 40+ FPS already ensures smooth visual feedback in ultrasound-guided procedures [7, 14, 29], trading a moderate amount of inference latencies for improved registration reliability is both acceptable and clinically meaningful.

Visualization results. Fig. 3 presents qualitative comparisons of different registration methods. The initial poses exhibit large structural misalignment. DreamReg produces resampled slices that are visually closer to the ground truth, with reduced error responses in the difference maps, indicating more accurate and stable pose estimation.

3.3 Ablation Study

We conduct ablations on CAMUS to evaluate the contribution of each supervision signal (Table 2). "only pose" trains the policy using pose regression alone; "w/o rew" removes the direction-based progress supervision; "w/o recon" removes slice reconstruction supervision. Removing the progress-alignment signal slightly degrades all metrics, indicating that reward-based progress supervision improves action quality and stabilizes belief updates. Removing reconstruction mainly decreases appearance consistency (SSIM and PNCC) and also increases geometric errors, suggesting that slice reconstruction regularizes latent dynamics. The pose-only variant performs worst overall, confirming that direct pose regression without belief-driven imagination is insufficient under partial observability.

4 Conclusion

We proposed DreamReg, a belief-driven world-model framework for 2D–3D ultrasound registration. By modeling action–observation coupling, DreamReg performs iterative belief refinement instead of one-shot pose regression. Experiments on CAMUS and μ-RegPro demonstrate consistent improvements in geometric and image similarity metrics under partial observability. The evaluation is conducted on 2D slices resampled from 3D volumes and does not model real tissue deformation, which may limit generalization to fully deformable settings. Although sequential rollout introduces additional computation, the achieved frame rate satisfies real-time clinical requirements. Future work will improve efficiency and extend the framework to deformable and closed-loop robotic ultrasound systems.

Acknowledgments. Research reported in this work was supported in part by Research Grants Council (RGC) of Hong Kong (CUHK 14217822, CUHK 14207823, CUHK 14211425, T45-401/22-N, and AoE/E-407/24-N) and in part by Innovation and Technology Commission of Hong Kong (MHP/096/22, ITS/235/22, ITS/224/23, ITS/225/23, and Multi-scale Medical Robotics Center (InnoHK initiative)).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bahner, D.P., Blickendorf, J.M., Bockbrader, M., Adkins, E., Vira, A., Boulger, C., Panchal, A.R.: Language of transducer manipulation: codifying terms for effective teaching. Journal of Ultrasound in Medicine **35**(1), 183–188 (2016)
2. Baum, Z.M.C., Saeed, S.U., Min, Z., Hu, Y., Barratt, D.C.: MR to ultrasound registration for prostate challenge (2023)
3. Brandstätter, S., Seeböck, P., Fürböck, C., Pochepnia, S., Prosch, H., Langs, G.: Rigid single-slice-in-volume registration via rotation-equivariant 2d/3d feature matching. In: International Workshop on Biomedical Image Registration. pp. 280–294. Springer (2024)
4. Bulatov, A., Kuratov, Y., Burtsev, M.: Recurrent memory transformer. Advances in Neural Information Processing Systems **35**, 11079–11091 (2022)
5. Ferrante, E., Paragios, N.: Slice-to-volume medical image registration: A survey. Medical image analysis **39**, 101–123 (2017)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
7. Giangrossi, C., et al.: Requirements and hardware limitations of high-frame-rate 3-d ultrasound imaging systems. Applied Sciences **12**(13), 6562 (2022)
8. Gillies, D.J., Gardi, L., De Silva, T., Zhao, S.r., Fenster, A.: Real-time registration of 3d to 2d ultrasound images for image-guided prostate biopsy. Medical physics **44**(9), 4708–4723 (2017)
9. Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Computers in biology and medicine **155**, 106389 (2023)

10. Guo, H., et al.: End-to-end ultrasound frame to volume registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 56–65. Springer (2021)
11. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 **2**(3), 440 (2018)
12. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. In: International Conference on Learning Representations
13. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. arXiv preprint arXiv:2010.02193 (2020)
14. Hidalgo, E.M., et al.: Evaluating the impacts of network latency, haptics, and ergonomics in a haptically-enabled robot for teleoperated echocardiography. Computers in Biology and Medicine **195**, 110450 (2025)
15. Hu, Y., Ahmed, H.U., Taylor, Z., Allen, C., Emberton, M., Hawkes, D., Barratt, D.: Mr to ultrasound registration for image-guided prostate interventions. Medical image analysis **16**(3), 687–703 (2012)
16. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019)
17. Lei, L., Zhou, J., Pei, J., Zhao, B., Jin, Y., Teoh, Y.C.J., Qin, J., Heng, P.A.: Epicardium prompt-guided real-time cardiac ultrasound frame-to-volume registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 618–628. Springer (2024)
18. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T.: Deep learning in medical ultrasound analysis: a review. Engineering **5**(2), 261–275 (2019)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Lu, G., Jia, B., Li, P., Chen, Y., Wang, Z., Tang, Y., Huang, S.: Gwm: Towards scalable gaussian world models for robotic manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9263–9274 (2025)
21. Markova, V., Ronchetti, M., Wein, W., Zettinig, O., Prevost, R.: Global multimodal 2d/3d registration via local descriptors learning. In: International conference on medical image computing and computer-assisted intervention. pp. 269–279. Springer (2022)
22. Mu, L., Huang, Z., Gu, Y., Qin, S., Zhang, S., Zhang, X.: Ehrworld: A patient-centric medical world model for long-horizon clinical trajectories. arXiv preprint arXiv:2602.03569 (2026)
23. Mulder, T.A., van de Velde, T., Dokter, E., Boekestijn, B., Olgers, T.J., Bauer, M.P., Hierck, B.P.: Unravelling the skillset of point-of-care ultrasound: a systematic review. The Ultrasound Journal **15**(1), 19 (2023)
24. Pavone, M., Seeliger, B., Teodorico, E., Goglia, M., Taliento, C., Bizzarri, N., Lecointre, L., Akladios, C., Forgione, A., Scambia, G., et al.: Ultrasound-guided robotic surgical procedures: a systematic review. Surgical endoscopy **38**(5), 2359–2370 (2024)
25. Rao, Y.R., Prathapani, N., Nagabhooshanam, E.: Application of normalized cross correlation to image registration. International Journal of Research in Engineering and Technology **3**(5), 12–16 (2014)
26. Smit, J.N., et al.: Ultrasound-based navigation for open liver surgery using active liver tracking. International journal of computer assisted radiology and surgery **17**(10), 1765–1773 (2022)

27. Wang, H., Wang, Y.: Eureg: End-to-end framework for efficient 2d-3d ultrasound registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 175–185. Springer (2025)
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
29. Wei, C.W., et al.: Real-time integrated photoacoustic and ultrasound (paus) imaging system to guide interventional procedures: ex vivo study. IEEE transactions on ultrasonics, ferroelectrics, and frequency control **62**(2), 319–328 (2015)
30. Weld, A., Dixon, L., Dyck, M., Anichini, G., Ranne, A., Camp, S., Giannarou, S.: Identifying visible tissue in intraoperative ultrasound: a method and application. International Journal of Computer Assisted Radiology and Surgery **20**(10), 2107–2117 (2025)
31. Wu, P., Escontrela, A., Hafner, D., Abbeel, P., Goldberg, K.: Daydreamer: World models for physical robot learning. In: Conference on robot learning. pp. 2226–2240. PMLR (2023)
32. Yang, Y., Wang, Z.Y., Liu, Q., Sun, S., Wang, K., Chellappa, R., Zhou, Z., Yuille, A., Zhu, L., Zhang, Y.D., et al.: Medical world model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8319–8329 (2025)
33. Yeung, P.H., Aliasi, M., Papageorghiou, A.T., Haak, M., Xie, W., Namburete, A.I.: Learning to map 2d ultrasound images into 3d space with minimal human annotation. Medical Image Analysis **70**, 101998 (2021)