2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Breaking the Latency Barrier: Synergistic Perception and Control for High-Frequency 3D Ultrasound Servoing

Yizhao Qian, Yujie Zhu, Jiayuan Luo, Li Liu, Yixuan Yuan, Hongen Liao, Guochen Ning*

***Abstract*— Tracking moving anatomical targets with robotic ultrasound is particularly challenging when the target motion is both fast and large in scale, as the end-to-end latency of existing systems prevents the perception–control loop from closing fast enough. In this paper, we argue that overcoming this limitation calls for the joint design of perception and control, rather than optimizing each in isolation. We present a tightly-coupled framework with two main components: (1) a Decoupled Dual-Stream Perception Network that estimates 3D translational state from 2D ultrasound images at high frequency, and (2) a Single-Step Flow Policy that outputs an entire action sequence in one forward pass, removing the need for iterative rollouts used in conventional policies. Together, the two modules enable closed-loop control at over 60 Hz. In phantom experiments with complex 3D trajectories, the system achieves a mean tracking error below 6.5 mm and re-acquires the target after resultant displacements exceeding 170 mm. It tracks targets moving at speeds up to 102 mm/s with a terminal error under 1.7 mm. In-vivo trials on a human volunteer further confirm that the approach transfers to realistic clinical conditions. To our knowledge, this is the first RUSS framework to unify high-bandwidth dynamic tracking with large-scale repositioning within a single architecture, offering a concrete step toward autonomous ultrasound operation in the presence of patient motion.**

I. INTRODUCTION

Robotic Ultrasound Systems (RUSS) have attracted growing interest for their potential to assist in medical diagnostics and image-guided interventions [1], [2]. In clinical practice, a central requirement is to maintain a stable ultrasound view of the target anatomy while the patient moves—whether due to respiration, involuntary body shifts, or postural adjustments—so that the clinician can safely carry out tasks such as needle insertion or catheter drainage. As shown in Fig. 1, this amounts to continuously adjusting the robotic probe so that the live image stays aligned with a reference view previously selected by the operating physician. Despite its practical importance, achieving this alignment in real time remains difficult, largely because existing systems cannot process visual feedback and generate motor commands fast enough to keep pace with rapid physiological motion.

This work was supported in part by the National Key Research and Development Program of China (2025YFC2425700), National Natural Science Foundation of China (82472115, 62201315, U22A2051).

Yizhao Qian and Yixuan Yuan are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China.

Guochen Ning, Hongen Liao, Yujie Zhu are with School of Biomedical Engineering, Tsinghua University, Beijing, China.

Li Liu and Jiayuan Luo are with Great Bay University, Dongguan, China.

*Corresponding author: Guochen Ning (email: ningguochen@tsinghua.edu.cn).

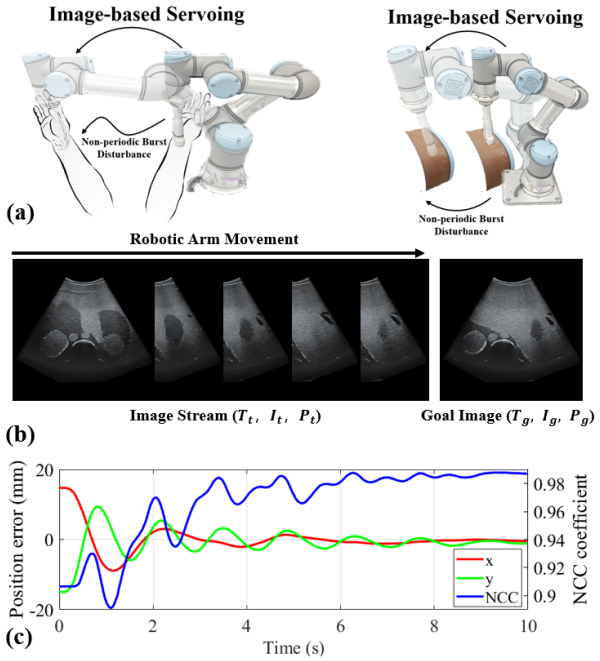

Fig. 1: Overview of the proposed high-frequency visual servoing. (a) Challenge: maintaining the ultrasound view under significant and unpredictable disturbances. (b) Control objective: aligning the live video stream with the target image using robotic manipulation. (c) Outcome: rapid reduction of positional errors (x, y) and maximization of image similarity, quantified by normalized cross-correlation (NCC).

Current approaches to this problem each face limitations when applied to continuous, high-frequency tracking. Event-driven strategies [3] react only after discrete triggers and are too slow for sustained motion compensation. Model-driven visual servoing methods [4], [5] rely on hand-crafted image features and typically require many update steps to converge. More recent learning-based methods, including Diffusion Policy [6], show promise but are constrained by their iterative denoising procedure, which limits control frequency to well below the **60 Hz frame rate of standard medical ultrasound hardware**. A control loop that cannot consume every incoming frame inevitably discards information, placing a hard ceiling on tracking performance. What is currently lacking is a framework in which both the perception and control modules are designed together to operate at the full sensor rate.

In this work we present such a framework, where perception and control are tightly coupled and jointly optimized for minimal end-to-end latency. The design rests on a straightforward observation: a fast control policy is of little use if the

perception module feeding it cannot keep up, and conversely a fast perception module is wasted if the downstream policy cannot act on its outputs in time. We realize this through two main components. First, a Decoupled Dual-Stream Perception Network separates in-plane geometric matching from out-of-plane semantic inference, enabling robust 3D state estimation at high frequency from 2D ultrasound images alone. Second, a Single-Step Flow Policy replaces the multi-step denoising of diffusion models with a single forward pass that produces a full predictive action sequence, removing the iterative latency that has limited prior generative approaches. This tightly-integrated perception-control loop is complemented by a sample-efficient Sim-to-Real transfer strategy that exploits the decoupled structure of the perception front-end for rapid domain adaptation. We validate the complete framework both on a dynamic phantom and through an in-vivo study on a human volunteer. The main contributions of this paper are:

- A RUSS framework that integrates a high-frequency Flow Policy with a co-designed perception front-end, achieving 62,Hz closed-loop tracking of dynamic anatomical targets.
- A dual-stream perception architecture that addresses the inherent ambiguity of out-of-plane motion estimation, enabling real-time 3D translational servoing from 2D images.
- A demonstration of sample-efficient Sim-to-Real transfer, where the system generalizes from simulation to a physical phantom using only 50 expert trajectories.

II. RELATED WORK

A. Robotic Ultrasound Systems: The System-Level Bottleneck for Dynamic Tracking

Recent advances in RUSS have shown success in automating quasi-static tasks like vascular screening [4], thyroid scanning [7], and standard plane localization [5], [8], [9]. The feasibility of maintaining stable probe contact is also well-established [10].

However, addressing patient and target motion, particularly high-frequency, unpredictable disturbances, remains a formidable challenge. One typical method, event-driven discrete compensation, employs a "Stop-Register-Resume" strategy [3], but its reported 336 ms registration latency makes it unsuitable for continuous clinical disturbances.

Another line of work pursues continuous tracking via model-driven visual servoing, achieving high control frequencies [11] (20 Hz) or sub-millimeter [5], [12] static accuracy. Yet, their system-level responsiveness is poor, with end-to-end convergence times on the order of seconds, even when using high-rate perception (60 Hz) [4], [5], [12]. This discrepancy proves a critical point: **component-level speed does not translate to system-level agility**, which indicates that a comprehensive and structured framework is needed to achieve low-latency dynamic responses.

Therefore, a critical gap exists for a RUSS framework comprehensively architected for high-bandwidth, unpredictable motion tracking. Recent surveys confirm that the lack of real-time [1] integrated perception and control [1], [2] is a key challenge in the field. Our work directly addresses this gap by proposing a framework where these subsystems are cohesively co-designed for a low-latency dynamic response.

B. Learning-based Control: The Quest for High-Frequency Policies and Robust Generalization

Learning-based methods, particularly imitation learning (IL), are effective for acquiring expert workflows in RUSS [13], [14]. The state-of-the-art is dominated by Diffusion Policies [15], but their reliance on an iterative denoising process for inference imposes a fundamental latency bottleneck. This limits their control frequency to 10-23 Hz [15], [16], a rate far below the 60 Hz update stream from the US probe, making real-time compensation of physiological motion impossible.

To overcome this, policies based on Flow Matching have emerged as a compelling alternative [17]. By enabling single-step inference, their recent work demonstrating speeds of 50 Hz—a nearly 7-fold improvement [18] over diffusion counterparts and highlighting their potential for high-frequency control [6].

However, a fast policy alone is insufficient. A critical second challenge is generalization against variations in US appearance. Existing frameworks are often too slow for dynamic tasks, relying on minute-long offline searches [19] or using perception modules that limit the system frame rate to a mere 3 fps [8]. This reveals a critical trade-off: existing methods sacrifice either real-time performance for generalization, or vice-versa.

Therefore, an effective framework must address both challenges in concert. To our knowledge, no prior work has presented a comprehensive framework where a high-frequency policy is cohesively co-designed with a fast, sample-efficient Sim-to-Real strategy to enable true, end-to-end dynamic tracking at over 60 Hz. This fusion of a high-bandwidth policy with a robust, low-latency generalization pipeline is the central methodological contribution of our paper.

C. The Perception Bottleneck for High-Frequency Servoing

The performance of any high-frequency control system is ultimately limited by the latency and accuracy of its perception front-end. Common RUSS perception pipelines, comprising segmentation, feature extraction, and matching, inherently accumulate latency and propagate errors [5], [20], rendering them unsuitable for real-time dynamic tracking.

This challenge is particularly acute in US due to a fundamental ambiguity: inferring out-of-plane (Z) motion from a 2D image sequence is a notoriously ill-posed problem [11], [21]. Existing systems often circumvent this with inefficient search strategies or are confined to 2D in-plane compensation only [22]. While end-to-end regression has been proposed [23], these methods have not been validated within a high-frequency dynamic tracking loop.

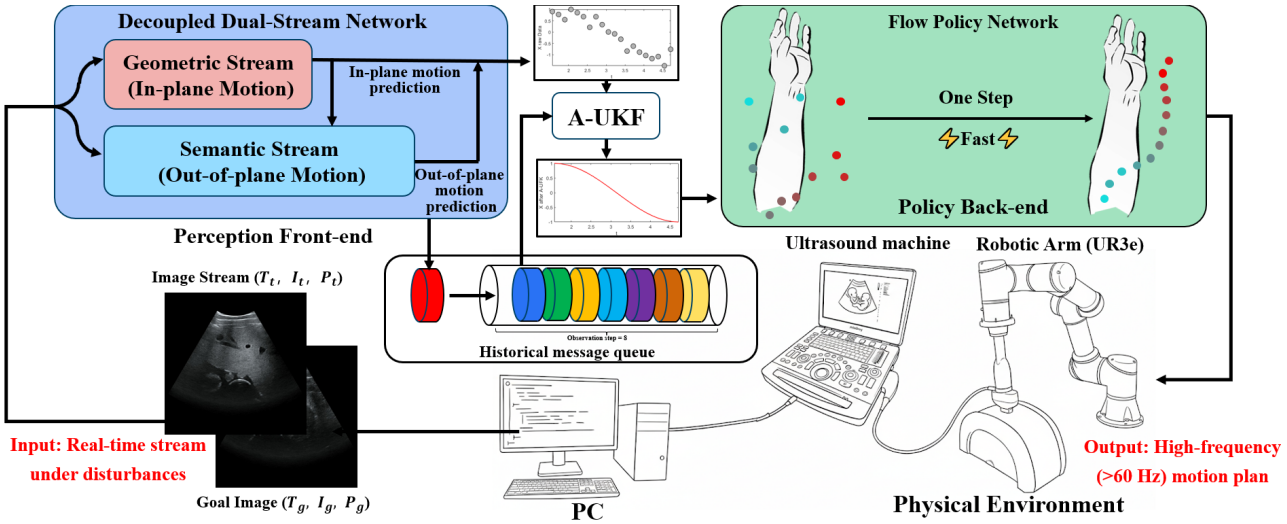

Fig. 2: Overview of our proposed high-frequency US servoing framework. The system takes a live image stream and a goal image as input. The Vision Front-end, composed of a Decoupled Dual-Stream Network and an Adaptive-UKF, estimates the 3D translational error. This state information is fed to the Flow Policy Network, which generates a short-horizon motion plan executed by the robotic arm in the Physical Environment.

Therefore, a perception module for this task must be low-latency and architected to resolve the out-of-plane ambiguity from image data directly. We address this by proposing a novel, decoupled dual-stream architecture that estimates the full 3D translational state at high frequency. This perception front-end is co-designed with our high-speed policy, forming the cornerstone of our cohesive framework.

III. METHODOLOGY

A. Problem Formulation

We formulate the dynamic visual servoing task as learning a policy, π_θ, that maps a temporal sequence of visual observations to a trajectory of future actions. The objective is to minimize the 3D translational error $\mathbf{e}_t$ between the current ultrasound frame I_t and a goal image I_g. In our setting, I_g is a reference frame pre-selected by the operator to define the target anatomy of interest; the system's role is to keep the probe aligned with this region even as the patient moves.

At each time step t, a perception front-end, ϕ, estimates this error (detailed in Sec. III-C):

$$\mathbf{e}_t = [dx_t, dy_t, dz_t]^T = \phi(I_t, I_g) \tag{1}$$

where dx_t, dz_t denote errors within the imaging plane (the XOZ plane of the probe), and dy_t is the error along the elevational (out-of-plane) direction. To capture temporal dynamics of the target, we define the system state, $\mathbf{s}_t$ as a sliding window of the k most recent errors ($k = 8$ in our work):

$$\mathbf{s}_t = (\mathbf{e}_t, \mathbf{e}_{t-1}, \ldots, \mathbf{e}_{t-k+1}) \in \mathbb{R}^{3\times k} \tag{2}$$

Rather than producing a single reactive command as in [24], our policy outputs a short-horizon motion plan of H future actions ($H = 8$), where each action $\mathbf{a}_{t+i} \in \mathbb{R}^3$ is a desired translational velocity command $[v_x, v_y, v_z]^T$:

$$\mathbf{A}_t = (\mathbf{a}_t, \mathbf{a}_{t+1}, \ldots, \mathbf{a}_{t+H-1}) \tag{3}$$

The core task is to learn the deterministic policy π_θ that maps the state history to this action sequence:

$$\mathbf{a}_t, \ldots, \mathbf{a}_{t+H-1} = \pi_\theta(\mathbf{s}_t) \tag{4}$$

At execution time, we adopt a receding-horizon strategy: only the first $h = 4$ actions of each predicted sequence are executed before the policy is re-queried. The choices of k and h reflect a practical trade-off — a larger observation window k improves prediction accuracy but limits the fastest trackable motion, while a larger execution horizon h reduces re-planning overhead at the cost of responsiveness to abrupt changes.

B. Framework Overview

Our approach to this predictive control problem is a framework for high-frequency dynamic visual servoing, illustrated in Fig.2. The central design principle is to minimize end-to-end latency by tightly coupling perception and control within a single pipeline. Rather than developing the two modules independently, we co-designed them so that the policy's fast inference does not idle waiting for perception updates, and conversely, high-rate state estimates from the perception front-end are consumed without unnecessary buffering. Concretely, the framework integrates a high-frequency perception front-end (Sec.III-C) with a single-step predictive policy (Sec. III-D). The resulting perception-to-action loop runs at over 60 Hz, which is sufficient for compensating the dynamic disturbances encountered in our target application.

C. High-Frequency Temporal Perception Front-End

The core task of our perception front-end is to robustly estimate the target's 3D translational motion from a 2D US stream. This requires balancing two conflicting objectives: **(1) Generalization** for performance across diverse subjects, and **(2) Real-time Performance** for tracking high-frequency motion at over 60 Hz. We address these requirements through

a structured visual observer paired with a predictive temporal filter.

1) Decoupled Architecture for Generalizable 3D Motion Features: To encourage generalization, we introduce physically-motivated structure into the network (Fig. 3). Specifically, we decouple the estimation of in-plane and out-of-plane motion, since these two components arise from different visual cues in ultrasound image changes.

This architecture consists of two specialized, parallel streams. The **Geometric Stream** first estimates in-plane displacement ($\mathbf{d_xz} = [d_x, d_z]$) by performing dense matching on low-level geometric feature maps $\phi_g(\cdot)$ via a cost volume:

$$C(u, v, \mathbf{d}) = \langle \phi_g(I_g)_{u,v}, \phi_g(I_t)_{u+d_x, v+d_z} \rangle. \tag{5}$$

This reliance on geometric correspondence makes it inherently robust to appearance shifts. Crucially, the estimated in-plane displacement $\mathbf{d}$ is then used to warp the feature maps for the second stream. The **Semantic Stream** analyzes these warped high-level semantic features, $\phi_s(\cdot)$, to infer the more ambiguous out-of-plane motion (d_y). It is trained to interpret changes in anatomical morphology as translational displacement.

2) Predictive State Estimation for Real-Time Performance: To meet the throughput requirement, we apply computational optimizations including feature caching and pre-computation for the static goal image I_g. The raw 3D displacement estimates $\mathbf{d}_t = [d_x, d_y, d_z]^T$, from the vision network are noisy, particularly along the elevational axis. We therefore pass them through a temporal filter to produce smoothed state estimates for the downstream policy.

We adopt an **Adaptive Unscented Kalman Filter (A-UKF)** rather than a standard Extended Kalman Filter (EKF) for two reasons. First, the UKF propagates uncertainty through sigma points, avoiding the Jacobian linearization that the EKF requires; this is better suited to our setting where the mapping from image features to 3D pose is highly non-linear. Second, our A-UKF variant adaptively adjusts the process noise covariance Q based on the innovation sequence. In practice, this means the filter automatically places more trust in new measurements during rapid motion and tightens its estimates during stationary phases — a property that a fixed-Q EKF cannot provide without manual re-tuning. The filter state explicitly models the 3D position $\mathbf{p}_t$, velocity $\mathbf{v}_t$, and a sensor bias $\mathbf{b}_t$:

$$\mathbf{x}_t = [\mathbf{p}_t^T, \mathbf{v}_t^T, \mathbf{b}_t^T]^T \in \mathbb{R}^9 \tag{6}$$

By jointly estimating position, velocity, and bias, the A-UKF provides a temporally consistent state representation that accounts for systematic errors in the vision network, which matters for stable control at high tracking speeds.

D. Flow Matching for High-Speed Policy

Contemporary generative models such as Diffusion Policy [15], [16] require iterative denoising during inference, which introduces substantial latency and makes them impractical for high-frequency control tasks [6]. To address this, we adopt a policy based on **Flow Matching** [17], [18], which

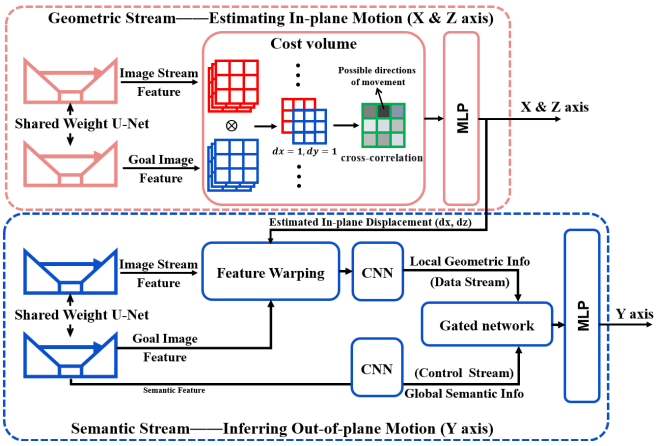

Fig. 3: The architecture of our Decoupled Dual-Stream Perception Network. The Geometric Stream uses a cost volume to estimate in-plane motion (X & Z axis) based on low-level feature. Concurrently, the Semantic Stream infers out-of-plane motion (Y axis) by interpreting higher-level feature.

can generate an entire action sequence in a single forward pass, keeping decision-making latency low within the control loop.

The policy learns to model the trajectory between a simple noise distribution p_0 (e.g., a standard Gaussian) and the distribution of expert actions p_1 by parameterizing a continuous, time-dependent vector field governed by an Ordinary Differential Equation (ODE). The vector field is conditioned on the state representation $\mathbf{s}_t$ provided by the perception front-end (Sec. III-C):

$$\frac{d\mathbf{x}_t}{dt} = v(\mathbf{x}_t, t|\mathbf{s}_t) \tag{7}$$

where the neural network $v(\cdot)$ approximates the conditional vector field. Once trained, the policy directly maps a sampled noise vector to a complete action sequence in one step, as illustrated in Fig. 4.

Because single-step inference keeps the policy latency well below the perception cycle, the temporal advantage of the 60 Hz front-end is preserved throughout the loop. This allows the system to react to fast and unpredictable patient motion during clinical procedures without introducing additional delay at the decision-making stage.

E. Sample-Efficient Sim-to-Real Transfer

To reduce the dependence on large-scale clinical data, we adopt a three-stage Sim-to-Real training pipeline that exploits the decoupled structure of the perception module described above.

*a) **Step 1: Vision Pre-training (Simulation)**:* The vision front-end is first trained on 20,000 simulated image pairs generated from CT volumes. We apply **domain randomization** over both visual properties (e.g., brightness, contrast) and US physics parameters (e.g., probe frequency, TGC curves), so that the learned features generalize across domain variations.

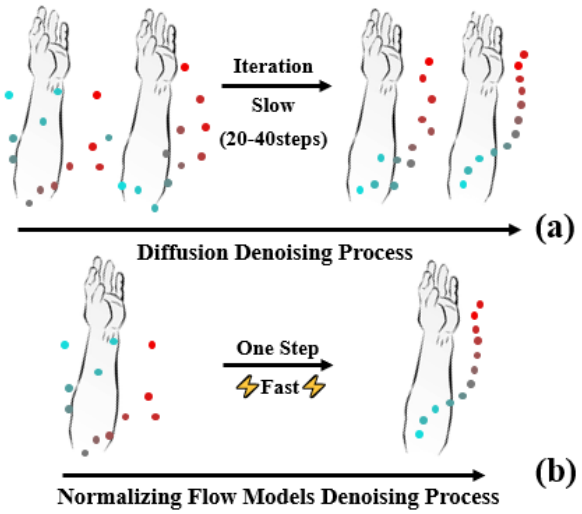

Fig. 4: Conceptual comparison of policy inference processes. (a) Diffusion Policies rely on iterative denoising, requiring multiple steps to produce an action. (b) Flow Policy generates the action sequence in a single forward pass, substantially reducing latency.

*b) **Step 2: End-to-End Pre-training (Simulation)**:* The full framework is then trained end-to-end on 1,000 simulated tracking trajectories, allowing the policy to learn the mapping from visual state representations to motion commands in a controlled setting.

*c) **Step 3: Targeted Fine-tuning (Physical Phantom)**:* Finally, the pre-trained model is fine-tuned on only **50 expert trajectories** collected on a physical phantom. During this stage, we freeze most weights of the **geometric stream**, whose learned representations are largely domain-invariant, and concentrate updates on the domain-sensitive **semantic stream**. This selective fine-tuning bridges the sim-to-real gap with minimal real data while retaining the knowledge acquired in simulation.

IV. EXPERIMENTS

We conducted a series of experiments on both a dynamic phantom and a human volunteer to evaluate our system. The experiments are organized around four questions: (1) How accurately can the system converge to a target view? (2) How fast can it track a moving target? (3) How robust is it under complex trajectories? (4) How does it perform in an in-vivo scenario? In addition, we performed a comparative study against baseline methods to quantify the contributions of our framework's key components.

A. Experimental Setup

Our platform (Fig. 5) consists of a 6-DoF UR3e manipulator, a Mindray M8 US machine with a C5-1s convex probe, and a CIRS Model 057A abdominal biopsy phantom. The phantom contains clinically relevant structures such as liver and portal vein, providing a realistic test environment. Ground-truth probe positions were measured by an optical tracker (NDI Polaris) rigidly attached to the probe. All algorithms ran on a workstation equipped with an NVIDIA RTX 4080 GPU, running Ubuntu 20.04 and ROS Noetic.

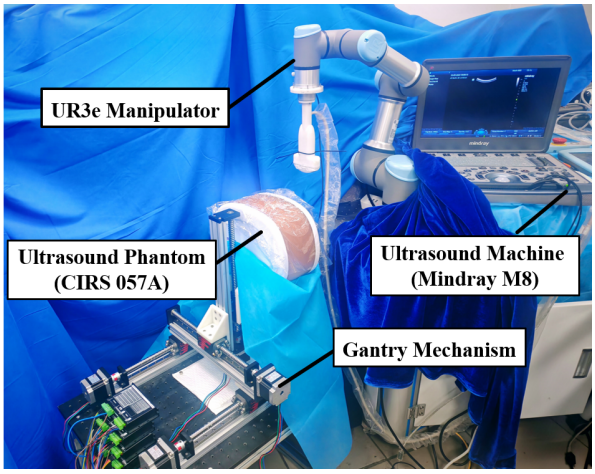

Fig. 5: Overview of the experimental setup, showing the UR3e manipulator, the CIRS phantom and the US system.

B. Baseline Performance Evaluation

We first evaluate the system in two scenarios: (1) static and quasi-static repositioning to a target view, and (2) continuous tracking of a target moving at high velocity. These tests quantify the system's accuracy, repeatability, and dynamic response.

1) Static and Repositioning Accuracy: We validated positioning accuracy through two tests: a local convergence test requiring recovery from minor manual displacements, and a large-scale repositioning test where the probe was displaced by over 170 mm (corresponding to simultaneous 100 mm offsets along X, Y, and Z, yielding a resultant magnitude of $\sqrt{3} \times 100 \approx 173$ mm).

As shown in Table I, the system converges to a terminal error of approximately 1.52 mm with image similarity above 0.92 (NCC) in both cases. The convergence dynamics for the large-scale test are plotted in Fig. 6a and b, showing a steady error decay from the initial 173 mm offset. These results confirm that the geometric stream can handle large spatial deviations while the semantic stream provides fine-grained alignment near the target.

TABLE I: Performance in Local Recovery and Global Repositioning Tasks

Experiment	**Movement Dist** (mm)	**Terminal Error** (mm)	**Terminal NCC**
Local Recovery	21.2	1.5148	0.9481
Global Repositioning	173.2	1.5219	0.9246

2) Dynamic Tracking Performance: To test end-to-end responsiveness, we designed a high-velocity tracking experiment in which significant latency in the perception-to-action loop would directly cause tracking failure.

As reported in Table II, the system tracks a target moving at over 100 mm/s with a mean error of approximately

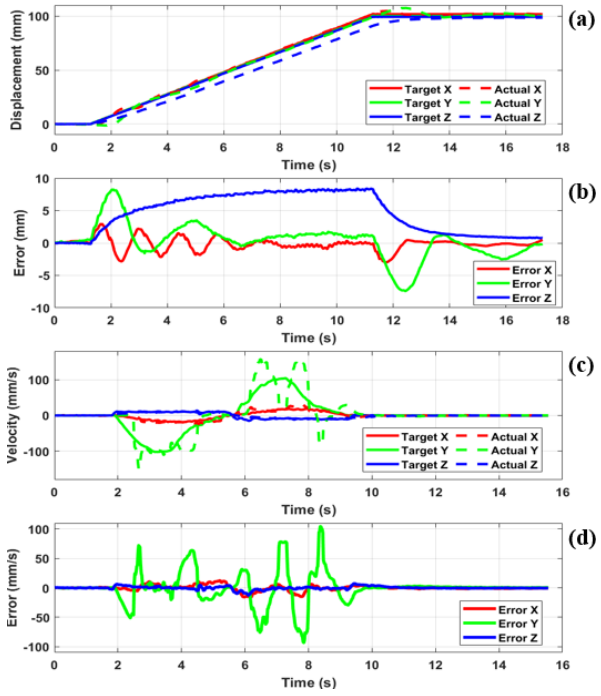

Fig. 6: System dynamic performance in large-scale repositioning and high-velocity tracking. **(a-b)** Convergence dynamics for the >170 mm repositioning test. **(c-d)** Agility during the high-velocity dynamic tracking test.

6.12 mm. The velocity profiles in Fig. 6c show that the robot's actual velocity closely follows the commanded profile. Fig. 6d further confirms that the velocity error remains bounded, indicating sufficient control bandwidth for high-speed operation.

TABLE II: Dynamic Tracking Performance at High Velocity

Max Speed (mm/s)	**Avg. Error** (mm)	**Terminal Error** (mm)	**Terminal NCC**
102.47	6.124 ± 0.386	1.629	0.9548

C. Robustness on Complex 3D Trajectories

We evaluated the system on 11 complex 3D trajectories covering spirals, square waves, and random paths to examine its behavior under high curvature, abrupt acceleration, and stochastic disturbances.

As shown in Fig. 8 and Table III, the robot's trajectory closely follows the ground truth across all tested paths, with a mean tracking error below **6.4 mm** and an average **NCC>0.91**. These numbers indicate that both geometric accuracy and anatomical view stability are preserved during unpredictable motion. The result can be attributed to the high-frequency state updates from the perception front-end, which supply the predictive policy with timely input even under rapid direction changes.

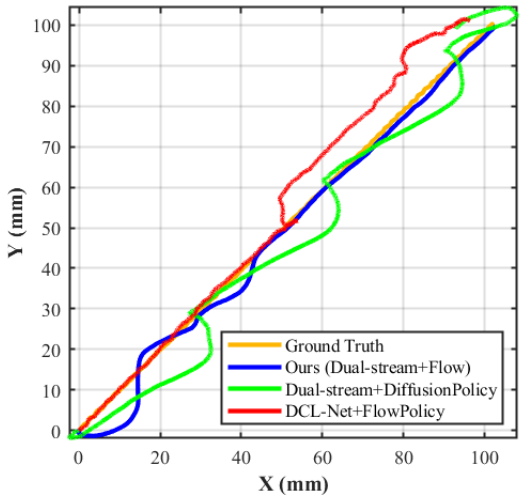

Fig. 7: XY-plane trajectory comparison from the dynamic tracking ablation study. Our full framework (blue) closely follows the ground truth (orange). While the Diffusion Policy (green) and DCL-Net module (red) lags significantly.

D. Comparative Analysis

To isolate the contribution of each module, we replaced our components with two strong alternatives and evaluated them on the high-velocity tracking task. For the policy we selected Diffusion Policy [15], a widely adopted imitation learning method; for the perception front-end we chose DCL-Net [25], a registration network designed for dynamic US.

Results are summarized in Table IV and Fig. 7. The Diffusion Policy variant suffered from high inference latency (>128 ms, corresponding to roughly **8 Hz**), which capped its tracking speed at **31 mm/s** and led to a repositioning error of **9.84 mm**. The DCL-Net variant, despite being computationally fast (14 ms per frame), failed to converge during dynamic tracking: its outputs were not sufficiently stable for the downstream policy, causing immediate instability.

Our framework achieves **16.2 ms** inference latency and a **62 Hz** control loop, supporting tracking speeds above **100 mm/s** with a final error below **1.6 mm**. These results suggest that neither fast perception nor fast control alone is sufficient; both modules must be designed in concert for reliable high-speed servoing.

TABLE IV: Ablation and Comparative on Dynamic Tracking

Framework Configuration	**Latency Metrics**		**Performance Metrics**	
	Time (ms)	**Freq. (Hz)**	**Max. Speed (mm/s)**	**Error (mm)**
Ours (Dual-stream + Flow)	≈ 16.2	≈ 62	**102.47**	**1.52**
Dual-stream + Diffusion	≈ 128.2	≈ 8	30.98	9.84
DCL-Net + Flow	≈ 13.7	73	**Failed to Converge**	6.508

E. Robustness to Out-of-Plane Rotational Disturbances

Our framework targets 3D translational servoing. To define its operational limits, we tested its stability under unmodeled Z-axis rotational disturbances introduced during a spiral

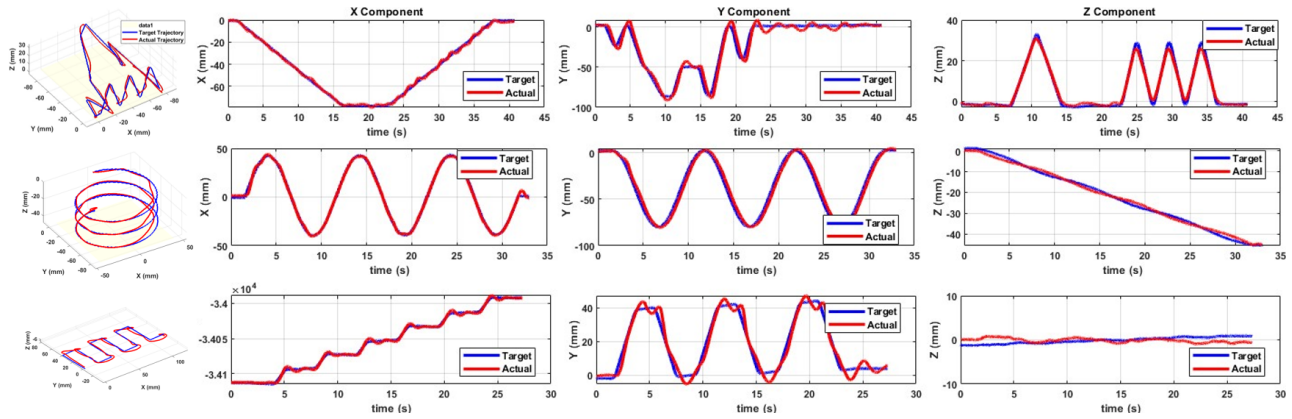

Fig. 8: High-fidelity tracking on three complex 3D trajectories (rows, top to bottom: random polyline, spiral, square wave). Each row shows the 3D path (left) and per-axis tracking (X,Y,Z).

TABLE III: Performance on Complex 3D Trajectories

Trajectory Type	Total Time (s)	Position Metrics (mm)					Speed Metrics (mm/s)	Image Metrics	
		Error X	Error Y	Error Z	Avg. Tracking Error	Terminal Error	Speed Error	Avg. NCC	Terminal NCC
Spiral-like Trajectories									
Spiral	38.895	1.712 ± 0.046	3.252 ± 0.098	1.217 ± 0.039	4.212 ± 0.088	1.582	3.759 ± 0.166	0.9194 ± 0.0018	0.9592
Elliptical Spiral	32.042	4.800 ± 0.206	3.232 ± 0.140	0.912 ± 0.030	6.313 ± 0.222	1.139	6.078 ± 0.269	0.9181 ± 0.0022	0.9602
Square Wave Trajectory									
Square Wave	30.310	1.392 ± 0.055	2.836 ± 0.105	1.536 ± 0.048	3.906 ± 0.094	2.959	6.514 ± 0.254	0.9565 ± 0.0010	0.9655
Random Polyline Trajectories									
Random 1	42.234	0.900 ± 0.028	3.740 ± 0.131	1.081 ± 0.038	4.337 ± 0.120	1.686	4.896 ± 0.207	0.9196 ± 0.0017	0.9485
Random 2	41.294	0.972 ± 0.025	2.366 ± 0.100	1.103 ± 0.037	3.038 ± 0.097	2.460	2.915 ± 0.129	0.9432 ± 0.0018	0.9572
Random 3	38.177	1.114 ± 0.046	3.901 ± 0.140	0.768 ± 0.033	4.350 ± 0.138	1.584	5.390 ± 0.204	0.9432 ± 0.0016	0.9664
Random 4	39.015	1.349 ± 0.043	5.716 ± 0.215	1.241 ± 0.044	6.357 ± 0.203	1.684	9.669 ± 0.351	0.9263 ± 0.0018	0.9609
Random 5	83.185	1.189 ± 0.027	4.239 ± 0.102	0.756 ± 0.016	4.701 ± 0.096	1.216	5.981 ± 0.185	0.9398 ± 0.0010	0.9555
Random 6	52.897	3.509 ± 0.065	4.050 ± 0.121	1.249 ± 0.035	6.051 ± 0.105	5.052	7.528 ± 0.233	0.9204 ± 0.0013	0.9412
Random 7	55.311	2.066 ± 0.049	3.948 ± 0.109	1.064 ± 0.029	5.019 ± 0.097	2.719	6.489 ± 0.201	0.9174 ± 0.0012	0.9373
Random 8	50.819	2.500 ± 0.063	3.132 ± 0.099	1.166 ± 0.031	4.772 ± 0.081	4.710	5.868 ± 0.193	0.9315 ± 0.0010	0.9403

tracking experiment, with offsets ranging from 0° to 25°. Rotations around the X and Y axes, which degrade US image quality, are outside the scope of this work [26].

As shown in Table V, the system remains stable up to **15°**, maintaining high image similarity and low positional error. Beyond this point performance degrades quickly, with instability at 20° and tracking failure at 25°. This is expected: the perception network is trained to interpret visual shearing as *translational* motion, so large rotational offsets introduce a perceptual ambiguity that the controller cannot compensate for.

TABLE V: Stability under Out-of-Plane Rotational Error

Rotational Offset	Positional Error	Avg. NCC	Status
0°	4.212 ± 0.088	0.9181 ± 0.0018	Stable
5°	4.659 ± 0.093	0.9161 ± 0.0015	Stable
10°	5.096 ± 0.103	0.9477 ± 0.0014	Stable
15°	8.170 ± 0.179	0.9370 ± 0.0007	Stable
20°	9.916 ± 0.183	0.9153 ± 0.0013	**Unstable**
25°	N/A	0.6962 ± 0.0096	**Failed**

F. In-vivo Validation on Human Volunteers

We conducted an in-vivo study on a human volunteer's forearm to examine clinical applicability. The system autonomously tracked over **20 cm** of motion during a **27-second** scan. Despite non-rigid tissue deformation and physiological motion, it reached a terminal **NCC of 0.946**, confirming that the framework generalizes from phantom to in-vivo conditions. Throughout the experiment, a force sensor with predefined force and velocity limits ensured contact safety. Further safety redundancy, such as real-time tissue-contact monitoring and emergency retraction, would be required before eventual clinical deployment.

V. Conclusion

In this work, we presented a framework for dynamic robotic ultrasound tracking that achieves closed-loop control above 60 Hz. By tightly coupling a high-frequency, decoupled perception front-end with a single-step Flow Matching policy, the system can compensate for target motion at speeds exceeding 100 mm/s and transfer from simulation to real hardware with limited real-world data. These results were validated on both a dynamic phantom and a human volunteer. While the current implementation handles 3D translational motion, future work will extend to full 6-DoF pose control, explore applicability to other planar imaging modalities such

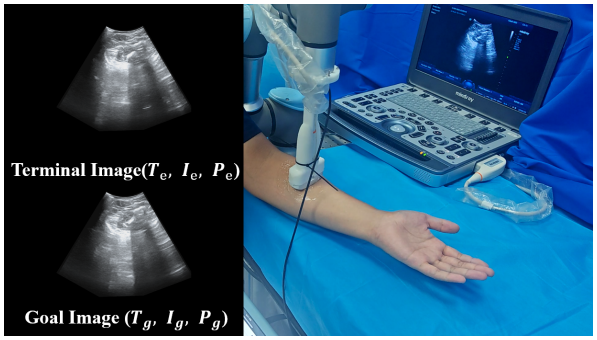

Fig. 9: In-vivo validation of the proposed framework on a human volunteer.

as photoacoustic and OCT, and address the safety requirements necessary for broader in-vivo deployment.

VI. Acknowledgment

The authors utilized Gemini to assist with language editing to ensure clarity. All scientific and technical content and logical reasoning remain the sole work of the authors.

References

[1] Z. Jiang, S. E. Salcudean, and N. Navab, "Robotic ultrasound imaging: State-of-the-art and future perspectives," *Medical image analysis*, vol. 89, p. 102878, 2023.

[2] K. Munir, A. F. Al-Battal, A. Al-Sheghri, H. Becher, M. Noga, and K. Punithakumar, "A survey of autonomous robotic ultrasound scanning systems," *IEEE Access*, 2025.

[3] Z. Jiang, N. Danis, Y. Bi, M. Zhou, M. Kroenke, T. Wendler, and N. Navab, "Precise repositioning of robotic ultrasound: Improving registration-based motion compensation using ultrasound confidence optimization," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[4] Z. Jiang, Z. Li, M. Grimm, M. Zhou, M. Esposito, W. Wein, W. Stechele, T. Wendler, and N. Navab, "Autonomous robotic screening of tubular structures based only on real-time ultrasound imaging feedback," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 7, pp. 7064–7075, 2022.

[5] X. Ma, M. Zeng, J. C. Hill, B. Hoffmann, Z. Zhang, and H. 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*, 2024.

[6] Q. Rouxel, A. Ferrari, S. Ivaldi, and J.-B. Mouret, "Flow matching imitation learning for multi-support manipulation," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2024, pp. 528–535.

[7] J. Zhou, H. Tian, W. Wang, *et al.*, "Fully automated thyroid ultrasound screening utilizing multi-modality image and anatomical prior," *Biomedical Signal Processing and Control*, vol. 87, p. 105430, 2024.

[8] Z. Jiang, Y. Bi, M. Zhou, Y. Hu, M. Burke, and N. Navab, "Intelligent robotic sonographer: Mutual information-based disentangled reward learning from few demonstrations," *The International Journal of Robotics Research*, vol. 43, no. 7, pp. 981–1002, 2024.

[9] Y. Huang, W. Xiao, C. Wang, H. Liu, R. Huang, and Z. Sun, "Towards fully autonomous ultrasound scanning robot with imitation learning based on clinical protocols," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3671–3678, 2021.

[10] S. Ipsen, D. Wulff, I. Kuhlemann, A. Schweikard, and F. Ernst, "Towards automated ultrasound imaging—robotic image acquisition in liver and prostate for long-term motion monitoring," *Physics in Medicine & Biology*, vol. 66, no. 9, p. 094002, 2021.

[11] T. Chen, X. Zhao, Y. Zhang, G. Zheng, L. Hou, Q. Ling, B. Tao, and Z. Yin, "Ultrasound-guided robotic autonomous operation based on real-time deformation tracking and prediction," *IEEE Transactions on Industrial Informatics*, 2024.

[12] J. Tan, J. Li, Y. Li, B. Li, Y. Leng, Y. Rong, and C. Fu, "Autonomous trajectory planning for ultrasound-guided real-time tracking of suspicious breast tumor targets," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 3, pp. 2478–2493, 2023.

[13] G. Ning, H. Liang, X. Zhang, and H. Liao, "Autonomous robotic ultrasound vascular imaging system with decoupled control strategy for external-vision-free environments," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 11, pp. 3166–3177, 2023.

[14] G. Ning, H. Liang, X. Zhang, and H. Liao, "Inverse-reinforcement-learning-based robotic ultrasound active compliance control in uncertain environments," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 2, pp. 1686–1696, 2024.

[15] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[16] H. Wang, Y. Long, Y. Chen, H.-C. Yip, M. Scheppach, P. W.-Y. Chiu, Y. Yam, H. M.-L. Meng, and Q. Dou, "Learning dissection trajectories from expert surgical videos via imitation learning with equivariant diffusion," *Medical Image Analysis*, p. 103599, 2025.

[17] Y. Fang, X. Zhang, H. Cheng, X. Zang, R. Song, and J. Zhao, "Flow policy: Generalizable visuomotor policy learning via flow matching," *IEEE/ASME Transactions on Mechatronics*, 2025.

[18] Q. Zhang, Z. Liu, H. Fan, G. Liu, B. Zeng, and S. Liu, "Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, 2025, pp. 14 754–14 762.

[19] Y. Long, A. Lin, D. H. C. Kwok, L. Zhang, Z. Yang, K. Shi, L. Song, J. Fu, H. Lin, W. Wei, *et al.*, "Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery," *Science Robotics*, vol. 10, no. 104, p. eadt3093, 2025.

[20] A. Tyagi, A. Tyagi, M. Kaur, R. Aggarwal, K. D. Soni, J. Sivaswamy, and A. Trikha, "Nerve block target localization and needle guidance for autonomous robotic ultrasound guided regional anesthesia," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 5867–5872.

[21] D. Dall'Alba, L. Busellato, T. R. Savarimuthu, Z. Cheng, and I. Iturrate, "Imitation learning of compression pattern in robotic assisted ultrasound examination using kernelized movement primitives," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[22] X. Liu, C. He, M. Wu, A. Ping, A. Zavodni, N. Matsuura, and E. Diller, "Transformer-based robotic ultrasound 3d tracking for capsule robot in gi tract," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–8, 2025.

[23] E. Zakeri, A. Spilkin, H. Elmekki, A. Zanuttini, L. Kadem, J. Bentahar, W.-F. Xie, and P. Pibarot, "Robust deep feature ultrasound image-based visual servoing: focus on cardiac examination," *IEEE/ASME Transactions on Mechatronics*, 2025.

[24] H. Yoon and S.-W. Kim, "Efficient and robust fabrication of soft sensors via injection with auxiliary suction in multilayered microchannels with a liquid metal alloy," *IEEE Sensors Journal*, vol. 25, no. 13, pp. 23 948–23 957, 2025.

[25] H. Guo, S. Xu, B. Wood, and P. Yan, "Sensorless freehand 3d ultrasound reconstruction via deep contextual learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 463–472.

[26] Y. Qian, Y. Zhang, M. Q.-H. Meng, and L. Liu, "Autonomous in-plane normal positioning in robotic ultrasound scanning," in *2024 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2024, pp. 342–347.