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

A Vision-Language-Action Model for Adaptive Ultrasound-Guided Needle Insertion and Needle Tracking

Yuelin Zhang, Qingpeng Ding, Longxiang Tang, Chengyu Fang, Shing Shin Cheng*

Abstract— **Ultrasound (US)-guided needle insertion is a critical yet challenging procedure due to dynamic imaging conditions and difficulties in needle visualization. Many methods have been proposed for automated needle insertion, but they often rely on hand-crafted pipelines with modular controllers, whose performance degrades in challenging cases. In this paper, a Vision-Language-Action (VLA) model is proposed for adaptive and automated US-guided needle insertion and tracking on a robotic ultrasound (RUS) system. This framework provides a unified approach to needle tracking and needle insertion control, enabling real-time, dynamically adaptive adjustment of insertion based on the obtained needle position and environment awareness. To achieve real-time and end-to-end tracking, a Cross-Depth Fusion (CDF) tracking head is proposed, integrating shallow positional and deep semantic features from the large-scale vision backbone. To adapt the pretrained vision backbone for tracking tasks, a Tracking-Conditioning (TraCon) register is introduced for parameter-efficient feature conditioning. After needle tracking, an uncertainty-aware control policy and an asynchronous VLA pipeline are presented for adaptive needle insertion control, ensuring timely decision-making for improved safety and outcomes. Extensive experiments on both needle tracking and insertion show that our method consistently outperforms state-of-the-art trackers and manual operation, achieving higher tracking accuracy, improved insertion success rates, and reduced procedure time, highlighting promising directions for RUS-based intelligent intervention.**

I. INTRODUCTION

Ultrasound (US)-guided needle insertion is frequently used in a variety of percutaneous procedures, including minimally invasive interventions like tissue biopsy, tumor ablation, and regional anesthesia [1]. As a non-invasive, portable, safe, and cost-effective imaging technique [2], US offers real-time intraoperative visualization of both the needle and surrounding tissue, thus helping to reduce the risk of inadvertent injuries.

Some US needle trackers have been proposed to provide real-time needle position feedback. However, even with accurate needle position feedback, precise needle insertion still requires skillful operators, which can be challenging in resource-limited settings. Robotic ultrasound (RUS) has therefore emerged as a transformative technology for automated needle insertion, offering enhanced precision and consistency. Recent advancements have enabled RUS systems to perform complex tasks such as probe manipulation, optimal image acquisition, and needle steering, thereby reducing operator dependency and improving procedure outcomes [3].

There are many challenges during needle insertion, where procedure outcomes can be affected by occlusion, imaging artifacts, and intermittent needle invisibility [4]. Traditional hand-crafted automated needle-insertion pipelines [5], [3], [6] remain fragile when facing these challenges, which require not only **accurate needle tracking** for closed-loop control but also **generalizable context awareness** to proactively adapt to dynamic environments. This highlights the need for better generalizability and high-level reasoning beyond conventional feature engineering methods.

Building upon the developments of Large-Language Models (LLMs) [7] and Vision-Language Models (VLMs) [8], Vision-Language-Action (VLA) [9] models have gained increasing attention in both open-world tasks and medical fields, as they can both understand multimodal information and generate actionable outputs. Although VLA models demonstrate robust logical reasoning abilities and strong generalization to dynamic environments, research on VLA-based automated needle insertion and tracking remains limited due to their high computational and data requirements. Achieving satisfactory accuracy while maintaining efficiency in VLA models remains challenging.

In this work, a novel VLA framework for automated and adaptive US-guided needle insertion on a RUS system is proposed for the first time. Without reliance on external sensors or pre-registration, the proposed framework holds promise for safer, more standardized, and operator-independent clinical workflows than conventional visual servoing needle steering pipelines. To enable adaptive needle insertion, accurate real-time needle position feedback is indispensable. Instead of incorporating a separate tracker like traditional methods, a Cross-Depth Fusion (CDF) Tracking Head is introduced as an end-to-end tracking head to incorporate cross-layer features from the pretrained vision backbone. Compared to hand-crafted non-end-to-end pipelines with independent needle trackers based on particle filter [3] or one-stream pipelines with computationally expensive VLM-based object grounding [10], the proposed framework not only ensures end-to-end training but also maintains efficiency.

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)).

Yuelin Zhang and Qingpeng Ding are with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong.

Longxiang Tang is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong.

Chengyu Fang is with the Shenzhen International Graduate School, Tsinghua University, China.

Shing Shin Cheng is with the Department of Mechanical and Automation Engineering, T Stone Robotics Institute, Shun Hing Institute of Advanced Engineering, Multi-Scale Medical Robotics Center, and Institute of Medical Intelligence and XR, The Chinese University of Hong Kong, Hong Kong.* sscheng@cuhk.edu.hk

By further integrating the proposed learnable Tracking-Conditioning (TraCon) Register, the pretrained vision backbone is parameter-efficiently conditioned for robust tracking.

Based on the needle position feedback, the VLA then controls the needle to reach the target with the proposed uncertainty-aware control policy, which ensures procedural success and safety when facing uncertainty due to occlusion, imaging artifacts, or intermittent tip invisibility. Leveraging large pretrained models, it provides better generalizability to dynamic US environments than traditional methods. An asynchronous VLA pipeline is proposed to decouple visual analysis from action generation, which operates at different latencies, thereby achieving precise action generation while ensuring real-time tracking. Extensive experiments demonstrate the effectiveness of the proposed VLA-based tracking-control pipeline.

The contributions of our work are fourfold:

- To provide accurate real-time needle position feedback for insertion control, a Cross-Depth Fusion (CDF) Tracking Head is developed to integrate shallow positional and deep semantic features, better adapting deep vision backbone for tracking tasks.
- A Tracking-Conditioning (TraCon) Register is introduced as a lightweight learnable token for parameter-efficient transferable vision backbone adaptation, facilitating tracking-oriented conditioning.
- To achieve adaptive needle insertion, an uncertainty-aware control policy is implemented, ensuring procedural safety and success under imaging artifacts and procedural uncertainty.
- To accommodate different execution speeds of tracking head and LLM, an asynchronous VLA pipeline is proposed, simultaneously ensuring real-time needle tracking and precise insertion control.

II. RELATED WORK

A. Ultrasound Needle Tracking and Robotic Ultrasound

Convolutional neural networks (CNNs) and transformer-based deep learning approaches [11], [12] have become widely adopted for tracking applications [13], [14], including ultrasound (US) needle tracking [15], [16], [17]. Given that the needle tip can be obscured by noise and artifacts, subsequent works employ segmentation of the needle shaft as a precursor to tip localization [18]. Another strategy is to incorporate temporal needle motion information into tracking. In [19], a motion prediction mechanism is combined with a visual tracker to enable motion cues beyond vision information. However, after acquiring needle position, the existing needle trackers failed to be integrated with control models to achieve intelligent intervention with needle position feedback.

To improve patient outcomes, the robotic ultrasound (RUS) systems are increasingly being integrated with artificial intelligence to enable autonomous procedures. Khadem et al. [6] introduced an autonomous US-guided intervention system that leverages nonlinear model predictive control (MPC) for accurate and adaptive needle targeting in soft tissue. By integrating particle filter-based needle tracking with closed-loop visual servoing, a real-time 3D US-guided robotic needle steering system is proposed in [3] for autonomous flexible needle insertion. With further integration with segmentation-based state estimation, [5] proposed a US needle insertion system addressing noisy tissue features. However, these modular feature-engineering pipelines with stacked separate modules remain brittle in dynamic environments, as they typically rely heavily on the accuracy of the physics-based model and require numerous hyperparameters for complex optimization, not to mention the impractical assumptions of persistent visibility and quasi-static status. It thus motivates a VLA-based controller that unifies perception and action for context-aware needle insertion control.

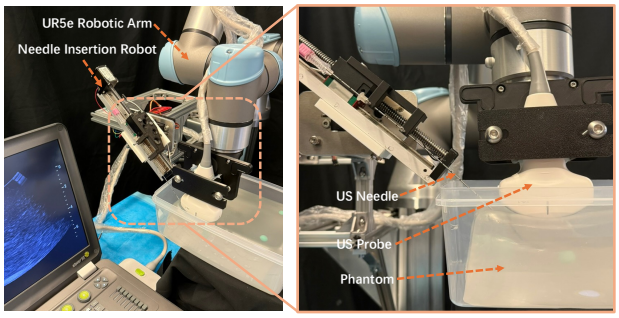

Fig. 1: The proposed robotic ultrasound (RUS) platform.

B. VLA Models in Medical Applications

Based on LLMs and VLMs, Vision-Language-Action (VLA) models have rapidly evolved, enabling systems to perceive, interpret, and interact with complex environments through multimodal understanding and reasoning [9], [20]. In the medical domain, VLA models have shown great promises. CapsDT [21] is proposed as a VLA model for endoscopy capsule robot manipulation with diffusion transformer [22], demonstrating state-of-the-art performance in complex gastrointestinal tasks. More recently, EndoVLA [10] introduces a dual-phase VLA framework designed for robotic endoscopy, enabling robust autonomous tracking of abnormal regions through end-to-end instruction imitating and reinforced with task-specific rewards. Although it achieves end-to-end object tracking and action generation within VLA, its use of a coupled structure for both tasks significantly compromises efficiency.

Due to the scarcity of RUS action datasets and the inherently high dynamics, noise, and indistinct target features of US images compared to other modalities, such VLA models for US procedures have yet to be proposed, despite their potential effectiveness.

III. SYSTEM DESIGN

A. Robotic Ultrasound Platform

A robotic ultrasound (RUS) platform was developed for automated needle insertion, which includes a needle insertion robot and a UR5e robotic arm, as shown in Fig. 1. The

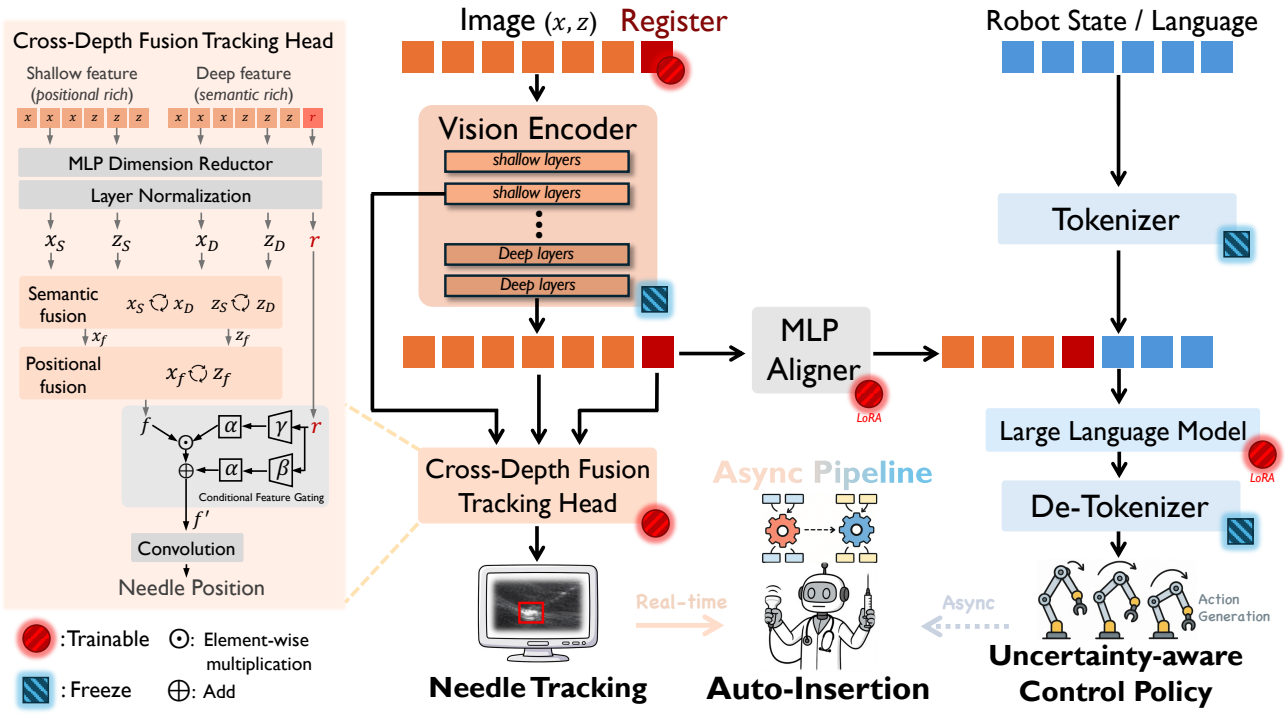

Fig. 2: Structure overview. The proposed VLA framework integrates the CDF tracking head and the TraCon register with a large-scale vision encoder, and decouples tracking from the proposed uncertainty-aware control through an asynchronous pipeline. This design enables accurate real-time needle tracking and precise, context-aware insertion control.

needle insertion robot consists of a linear manipulator for controlling the needle's axial displacement x_n (defined in the needle coordinate frame) and a servo for adjusting the insertion angle θ_n. The UR5e robotic arm is for manipulating the US probe in three translational degrees of freedom $\mathbf{x}_p = [x_p, y_p, z_p]$, expressed in the world coordinate frame.

B. Vision-Language-Action Model Overview

The proposed VLA is built based on a Qwen2.5-VL-3B model [8]. As shown in Fig. 2, the proposed VLA model adopts a pretrained vision encoder ϵ_V based on the Vision Transformer (ViT) [23], a pretrained LLM ϕ_L for interpreting input and generating actions, and a separate Cross-Depth Fusion (CDF) tracking head ϕ_T specifically designed for real-time needle tracking. In addition, the proposed Tracking-Conditioning (TraCon) register is appended to input image embeddings as a learnable, parameter-efficient, and task-oriented token, which is then applied in tracking head and LLM for task adaptation.

In object tracking [13], the network typically performs prediction based on the search map x and the template map z, where z encodes the target object's appearance from the initial frame, while x represents the region in subsequent frames where the target is likely to be located. The model identifies the object in x that most closely matches z and predicts a bounding box representing the target's location.

The framework receives ultrasound imaging observation $O_t = [x_t, z_t]$ and the language instruction I, predicts needle position $P = \phi_T(\epsilon_V(O_t))$, and then predicts action $\mathcal{A}$ via an asynchronous inference pipeline to achieve automated adaptive needle insertion.

C. Tracking-Conditioning Register

The proposed Tracking-Conditioning (TraCon) register is a lightweight learnable token for task-oriented parameter-efficient fine-tuning (PEFT). Most existing VLA models perform PEFT on the pretrained vision encoder by LoRA [24] to ensure that the pretrained model can be adapted to downstream tasks. However, such PEFT methods place high demands on the quantity and quality of labeled training data, while acquiring labeled datasets in medical scenarios is often difficult and expensive. Besides, in the proposed model, the output of the vision encoder is utilized by both the tracking head and the LLM. Directly fine-tuning the vision encoder would inevitably introduce performance degradation due to inconsistent training objectives.

Instead of applying PEFT to the whole vision encoder, the proposed TraCon register $R \in \mathbb{R}^{B \times L_r \times C}$ ($L_r = 4$ in this paper) is appended to the vision observation O_t as a lightweight trainable token (0.5 M parameters), externally conditioning the frozen vision backbone for object tracking and aligning it with both the tracking head and the LLM. After being encoded by the vision encoder ϵ_V, the encoded TraCon register $\epsilon_V(R)$ is subsequently fed into both the CDF tracking head ϕ_T and LLM ϕ_L. In CDF head, $\epsilon_V(R)$ is fused with vision embeddings via the proposed Conditional Feature Gating. In LLM, $\epsilon_V(R)$ is inserted into the input sequence to facilitate cross-modal knowledge transfer.

D. Cross-Depth Fusion Tracking Head

The Cross-Depth Fusion (CDF) tracking head ϕ_T is proposed as a dedicated module for needle tracking,

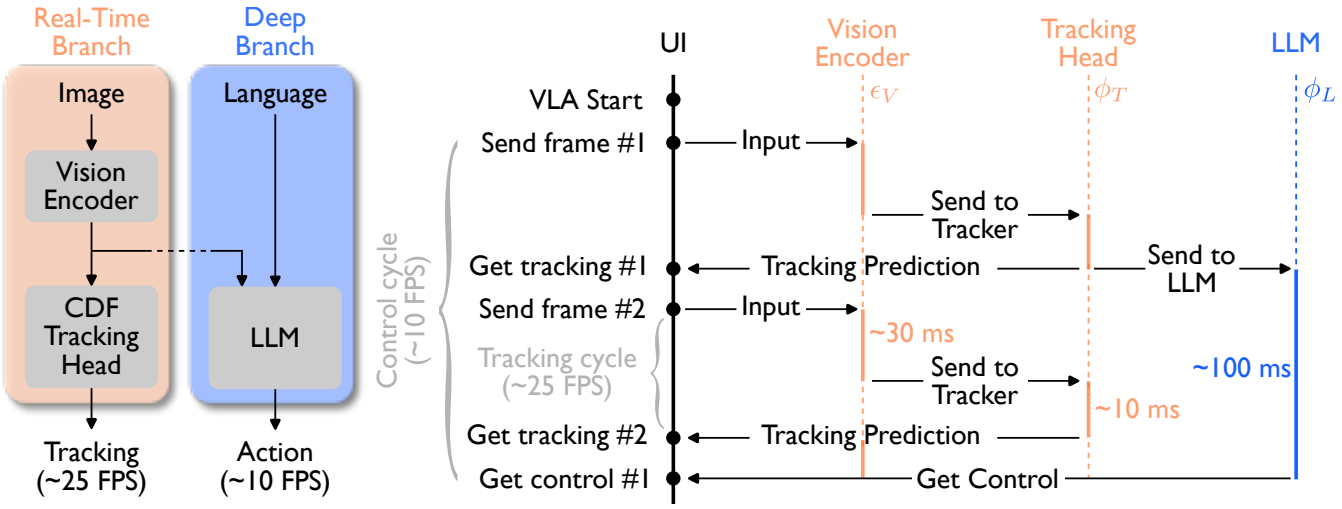

Fig. 3: Asynchronous VLA pipeline (left) and its sequence diagram (right). The runtime of the tracking cycle and control cycle is measured while they are running simultaneously on a single NVIDIA A800 GPU. The lengths in the sequence diagram are for illustration only.

which is shown in Fig. 2. In CDF head, multi-depth features from the ViT vision encoder ϵ_V are simultaneously incorporated to learn shallow positional information and deep semantic information, since the shallow layer outputs in ViT predominantly encode positional information due to their proximity to the input embeddings and limited receptive field, whereas deeper layer outputs capture more semantic information as a result of accumulated global context and hierarchical feature abstraction [23].

Specifically, shallow features and deep features (together with $\epsilon_V(R)$) will first go through a MLP dimension reductor (to reduce backbone dimension 2048 to internal dimension $C = 512$) and a layer normalization [25] to get x_S, z_S, x_D, z_D, and r, respectively, where x_S and z_S are shallow outputs of search and template maps, x_D and z_D are deep outputs. To incorporate deep semantic features into tracking prediction without compromising the shallow positional features, a cross-depth semantic fusion is first performed separately to (x_S, x_D) and (z_S, z_D) to get x_f and z_f. The core of semantic fusion can be formulated as:

$$SemFus = (\text{softmax}(\frac{Q_{ch}(K_{ch})^{\mathrm{T}}}{\sqrt{L}})V_{ch})^{\mathrm{T}} \in \mathbb{R}^{B\times L\times C}, \quad (1)$$

where $Q_{ch} = \text{Linear}(q)^{\mathrm{T}}$, $K_{ch} = \text{Linear}(k)^{\mathrm{T}}$, and $V_{ch} = \text{Linear}(v)^{\mathrm{T}}$ are channel sequences in the dimension of $B \times C \times L$. The full semantic fusion block employs a standard transformer-style architecture [11], which includes a residual shortcut, layer normalization [25], and a MLP applied after the attention module.

With the cross-depth semantically enhanced maps x_f and z_f, the next stage is the positional correlation. The positional correlation correlates x_f with z_f by modeling positional attention to precisely localize target z_f on search map x_f:

$$PosCor = \text{softmax}(\frac{Q(K)^{\mathrm{T}}}{\sqrt{C}})V \in \mathbb{R}^{B\times L\times C}, \quad (2)$$

where $Q = \text{Linear}(q)$, $K = \text{Linear}(k)$, and $V = \text{Linear}(v)$ are all in $B \times L \times C$. Similar to semantic fusion, the remaining parts of positional correlation also follow the standard transformer block design [11].

Finally, before predicting the bounding box by a convolution layer ($stride = 1$, $kernel = 1 \times 1$), the fused map f is integrated with register r using the proposed Conditional Feature Gating. This operation conditions the primary vision feature map f on auxiliary register embeddings that capture task-specific priors, enabling dynamic modulation without altering the spatial structure of f. It is formulated by:

$$f' = f \odot (\alpha \cdot \gamma(r)) + \alpha \cdot \beta(r), \quad (3)$$

where $\gamma, \beta \in \mathbb{R}^{B\times C}$ are linear transformations, and α is a learnable scalar gate. Compared with the existing FiLM (Feature-wise Linear Modulation) [26], a learnable scalar α is added to adaptively control the modulation strengths. This mechanism enhances feature adaptability by conditioning on register-derived priors, promoting task-specific refinements in visual tracking.

With this regressive CDF tracking head, the proposed VLA model possesses several intrinsic advantages on tracking. Existing VLA models with object tracking like EndoVLA [10] typically couples action generation and tracking together to the LLM, where tracking is accomplished in an auto-regressive manner with LLM-based object grounding. This results in inefficiency, with tracking and control sharing a low-speed pipeline running at 2 FPS, which falls significantly short of real-time requirements. In contrast, by leveraging the dedicated CDF tracking head with decoupled action and tracking pipelines, the proposed VLA achieves significantly higher tracking **efficiency** ($\sim$ 25 FPS) with **fewer parameters**. Furthermore, compared to hand-crafted models that combine separate needle trackers and insertion controllers [6], [3], [5], the CDF head and action generator (LLM)

share the same pretrained vision backbone, enabling higher consistency between tracking and control.

E. Asynchronous VLA Pipeline

After obtaining the needle position, the next stage is to generate action with LLM for insertion control. If tracking waits in a blocking manner for the LLM to complete before processing the next frame, like the existing one-stream methods [10], real-time tracking then cannot be guaranteed. To address this issue, an asynchronous VLA pipeline is proposed, as shown in Fig. 3.

The proposed pipeline includes a real-time branch for needle tracking ($\sim$ 25 FPS) and a deep branch for action generation ($\sim$ 10 FPS). After the vision encoder ϵ_V finishes encoding one frame, the tracking will then start. Once the needle tracking prediction P is obtained by ϕ_T, it will be sent to both user interface (UI) for display and LLM ϕ_L for action generation. While the LLM generates actions, the vision encoding and tracking prediction for the next frame start immediately. Once the LLM completes action prediction, it will receive the latest available vision embedding without waiting to initiate the next round of action generation.

F. Uncertainty-Aware Control Policy

In US-guided needle insertion, the tip visibility is dynamic and fragile. When the tip fades, before the change is detected and reacted upon, the needle will lose tracking and a steady insertion speed can translate into millimeters of unseen advancement, leading to **discontinuities** in tip visualization and **clinical danger** near critical structures. Instead of relying on the operator to subjectively determine when to pause the needle advancement, a VLA-based uncertainty-aware control policy is proposed, following the control law by skilled operators: *see what you're doing*, and act more slowly when uncertainty increases (uncertainty is defined as the inverse of tip visibility) [27], [28].

After the CDF head ϕ_T predicts the needle tip position P, LLM ϕ_L generates action $\mathcal{A} = \phi_L(\epsilon_V(O_t), I, P)$, where $\epsilon_V(O_t)$ is the encoded vision embeddings. P is given by the bounding box $[x1, y1, x2, y2]$ of the needle tip, represented by its top-left $(x1, y1)$ and bottom-right $(x2, y2)$ coordinates in pixels. The language instruction I depicts the basic information and provides high-level command, given by

> "*You are an ultrasound expert. The target position is [*`TARGET`*]. The insertion technique is [*`TECH`*]. Control the insertion based on visibility feedback: when visibility decreases, insert with a slower speed. Stop insertion upon reaching the target.*"

[TARGET] is the position of insertion target given by its center pixel coordinates. [TECH] is the insertion technique (in-plane-static or in-plane-moving [19]). The predicted action $\mathcal{A}$ is given by $\mathcal{A} = [\theta_n, v_n, \mathbf{v}_p]$, where θ_n is the needle insertion angle, v_n is the insertion speed, $\mathbf{v}_p = [v_{p,x}, v_{p,y}, v_{p,z}]$ is the probe moving speed. The needle and probe positions are then given by $x_n = v_n \Delta t + x_{n,0}$ and $\mathbf{x}_p = \mathbf{v}_p \cdot \Delta t + \mathbf{x}_{p,0}$, where $x_{n,0}$ and $\mathbf{x}_{p,0}$ are their initial positions. When the needle reaches the target, the model generates a [STOP] token to terminate the procedure.

Leveraging the pretrained LLM, the proposed VLA-based control policy provides high-level reasoning based on global contextual information, in contrast to the ungeneralizable segmentation priors [5] and explicit needle deflection predictors [6] used in traditional needle steering controllers. This language conditioning keeps the LLM architecture unchanged while encoding semantically rich control context. Moreover, unlike traditional methods that require hyperparameter-controlled decision making, the proposed VLA model directly learns the control policy introduced by the expert from the training dataset. The insertion can then be adaptively controlled by visual reasoning, thus avoiding lost tip tracking during low-visibility period and ensuring a lower safe velocity near complex tissue structures.

G. 2-Stage Training and Dataset Collection

The proposed VLA model is trained by 2 stages. Stage 1 is for pretraining on the tracking task, where only the CDF tracking head and TraCon register are trained and the remaining modules adopted from Qwen2.5-VL-3B [8] are frozen. A US needle tracking dataset $\mathcal{D}_1 = \{(o_k, p_k)\}_{k=1}^N$ was collected for stage 1, where o_k is the image, p_k is the needle position ground truth measured by an optical tracker ClaroNav MicronTracker 3 (with an acceptable RMSE of 0.189 mm in our setup). It was acquired using a Wisonic Clover 60 US machine and a Wisonic C5-1 convex transducer with an 18 gauge needle. $\mathcal{D}_1$ contains 41,075 frames from 239 videos (1920×1080) by 105 in-plane-static (IPS) trials and 134 in-plane-moving (IPM) trials [4], [19] at a velocity of 20 mm/s and 3 different insertion angles (30°, 45°, 60°). The needle and transducer (i.e. US probe) were manipulated by the aforementioned RUS system. Several different materials were used for needle insertion, including phantoms made from fresh pork and solidified agar, as well as simulators composed of silicone and artificial tumors. $\mathcal{D}_1$ is divided into training, validation, and testing sets in the ratio 7:1:2.

Stage 2 is for fine-tuning the VLA model for adaptive needle insertion. Only the LLM and MLP aligner are fine-tuned with LoRA [24] while the remaining modules including CDF head and TraCon register are frozen. The pretrained vision encoder is frozen during both stages. A US needle insertion dataset $\mathcal{D}_2 = \{(o_k, p_k, i_k, a_k)\}_{k=1}^N$ was collected using the same devices and setup as $\mathcal{D}_1$. In addition to image o_k and needle positions p_k, $\mathcal{D}_2$ also includes language instructions i_k and action ground truth a_k. a_k was collected through expert demonstrations, in which an experienced operator manually manipulated the needle to reach specified targets, following principles similar to the proposed uncertainty-based control policy. During these demonstrations, the needle velocity v_n, insertion angle θ_n, and probe velocity $\mathbf{v}_p$ were recorded in real time. The operator recorded a [STOP] sign upon reaching the target. $\mathcal{D}_2$ contains 3,852 frames from 18 videos by 9 IPS trials and 9 IPM trials. It is only used for training, not for testing.

IV. Experiments and Results

A. Implementation Details

The evaluation was carried out on needle tracking and needle insertion respectively [1]. All experiments are implemented using PyTorch on a server with two NVIDIA A800 GPUs (although often considered high-end, A800 is based on an old NVIDIA Ampere architecture that is slower than current flagships). For needle tracking evaluation, several state-of-the-art trackers are involved for comparison, including classic Siamese trackers [13], [29], [30], transformer-based trackers [31], [32], CNN-based trackers [33], and large-scale trackers [34]. All these models were trained on the same dataset (training set of $\mathcal{D}_1$) by the same strategy (350 epochs, batch size 48, AdamW optimizer) as ours. Scaling, blur, and shifting were adopted for augmentation. The learning rate was set to 1e-4 and dropped by 10 after 100 epochs.

For needle insertion evaluation, the model pretrained on $\mathcal{D}_1$ was further fine-tuned on dataset $\mathcal{D}_2$ for a single epoch with a batch size of 16 and an AdamW optimizer. The learning rate was initialized to 1e-4 and modulated by a cosine annealing scheduler.

B. Needle Tracking Evaluation and Ablation Studies

Models were trained and evaluated on the training set and testing set of $\mathcal{D}_1$. The result is reported as area under curve (AUC) [35] and precision (P) [36] in percentage, as well as average error (Err) and standard deviation (SD) in mm. As shown in Tab. I, the proposed tracking pipeline achieves the best performance in almost all comparisons against the other SOTA trackers. Our tracker achieves 10.7% and 16.0% improvement on Err and SD over the second best method, showing advancement on tracking accuracy and robustness. Furthermore, our framework is the only approach that achieves a SD of less than 2 on both the IPS and IPM tasks. The tracking demonstration in Fig. 4 further illustrates that the proposed method attains the most robust and accurate tracking under challenging conditions. This accurate tracking serves as the basis of successful automated insertion.

Ablation studies were performed on four variants, as shown in Tab. I. $\mathcal{V}_{T,1}$ and $\mathcal{V}_{T,2}$ investigate the TraCon register and the impact of its length L_r. The results in $\mathcal{V}_{T,1}$ indicate that using a longer register does not consistently improve performance and may even lead to undesired outcomes. In contrast, removing the register results in performance degradation across almost all metrics in $\mathcal{V}_{T,2}$. These findings demonstrate that the proposed TraCon register can effectively enhance tracking through model conditioning. $\mathcal{V}_{T,3}$ and $\mathcal{V}_{T,4}$ investigate the significance of fusing cross-depth features. They represent variants where cross-depth semantic fusion is removed and only shallow feature or deep feature is kept respectively. The results show that receiving only shallow or deep features leads to degraded performance. This indicates that taking into account both shallow positional features and deep semantic features is critical for tracking.

[1]See examples of needle tracking and insertion in supplementary video.

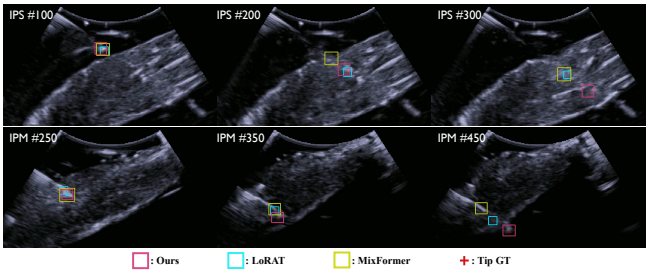

Fig. 4: Needle tracking demonstration in a tissue phantom. The proposed method outperforms two state-of-the-art trackers under dynamic environment and degraded visibility.

TABLE I: Evaluation and ablation study results of needle tracking in AUC (%), P (%), Err (mm), and SD (mm). The methods with the best and the second best performance are in red and cyan.

Method	In-plane-static (IPS)			In-plane-moving (IPM)			Mean		
	AUC↑	P↑	Err±SD↓	AUC↑	P↑	Err±SD↓	AUC↑	P↑	Err±SD↓
SiamRPN++ [13]	45.0	60.1	5.55±3.19	60.2	76.3	4.10±3.64	51.4	66.9	4.90±3.41
SiamCAR [29]	53.1	71.0	4.87±3.28	60.4	83.3	3.38±3.09	56.2	76.2	4.27±3.21
SiamBAN [30]	53.0	72.7	4.75±3.30	65.4	92.0	4.13±2.99	58.2	80.8	4.50±3.19
SwinTrack [31]	49.6	70.6	4.99±3.37	65.8	94.0	3.41±2.93	56.4	80.4	4.37±3.09
STMTrack [33]	52.1	74.1	4.96±3.20	64.3	92.5	3.17±3.03	57.2	81.9	4.16±2.98
MixFormer [32]	57.9	77.0	4.32±2.77	68.0	93.9	3.01±2.73	62.1	84.1	3.79±2.77
LoRAT [34]	59.7	79.0	3.82±2.55	65.1	93.7	2.73±2.68	62.0	85.2	3.37±2.62
Ours	60.3	84.0	3.45±1.92	67.9	95.5	2.29±1.98	63.5	88.9	3.01±2.20

Ablations	Mean		
	AUC↑	P↑	Err±SD↓
Baseline: default	63.5	88.9	3.01±2.20
$\mathcal{V}_{T,1}$: $L_r = 16$	63.7(+0.2)	88.5(-0.4)	3.10(+0.09)±2.19(-0.01)
$\mathcal{V}_{T,2}$: $L_r = 0$ (w/o TraCon R)	62.8(-0.7)	86.1(-2.8)	3.18(+0.17)±2.10(-0.10)
$\mathcal{V}_{T,3}$: w/o fusion, only shallow	62.1(-1.4)	86.4(-2.5)	3.49(+0.48)±2.62(+0.42)
$\mathcal{V}_{T,4}$: w/o fusion, only deep	61.7(-1.8)	87.0(-1.9)	3.32(+0.31)±2.95(+0.75)

C. Needle Insertion Evaluation and Ablation Studies

The model trained on $\mathcal{D}_1$ was further fine-tuned on dataset $\mathcal{D}_2$, where only LLM and MLP aligner were trained with LoRA. The comparison is performed between the proposed VLA-based RUS insertion and manual insertion regarding success rate (SUC) in percentage and procedure time (T) in seconds. For SUC, a successful attempt is defined as the needle tip reaching within 5 mm of the target point. Any deviation of the needle from the target or complete loss of needle visualization during the process is considered a failure. T denotes the average completion time across all successful trials. For VLA-based RUS insertion, a total of 40 attempts were performed (20 IPS and 20 IPM). For manual insertion, five experienced users each performed 4 IPS and 4 IPM insertions, also resulting in a total of 40 attempts (20 IPS and 20 IPM). Ablation studies on RUS insertion were conducted using the same protocol. For each target position, both RUS and manual insertion were conducted.

This study does not compare with conventional hand-crafted pipelines or existing VLA models for US needle insertion, as such methods are scarce and lack open-source implementations. Instead, we focus on comparisons with manual operation, which remains the mainstream approach in

clinical practice and directly reflects the practical benefits and translational potential for real-world adoption. As shown in Tab. II, the proposed VLA-based tracking-insertion pipeline achieves a significant 33.3% SUC improvement with less time consumption than manual insertion. For IPM insertion, the SUC improvement even reached 63.6% with an average time reduction of 7.1 s. Furthermore, with the proposed asynchronous pipeline, the average frame rates for tracking and action generation reached 25.1 FPS and 10.4 FPS, respectively. It can not only provide operators with real-time needle position feedback, but also ensure a safe and acceptable action generation frequency, which surpasses that of most existing VLA models [10], [37].

Two examples are shown in Fig. 5, where the insertion speed is dynamically adjusted by the proposed VLA framework to ensure consistent needle visualization and improved outcomes. In the IPM case, needle insertion is slowed during the initial stage due to increased uncertainty. The speed gradually recovers in the middle stage, and as the needle approaches the target, insertion slows again to account for ambiguity caused by tissue occlusion. This dynamic adaptation allows the needle to respond appropriately at any position, thereby improving the success rate.

Ablation studies were performed on four variants. $\mathcal{V}_{I,1}$ adopts the same structure as $\mathcal{V}_{T,1}$ in Tab. I. The performance degradation demonstrates that poor tracking leads to a higher rate of insertion failures. Without the TraCon register, $\mathcal{V}_{I,2}$ results in more insertion failures and increased time consumption. It further demonstrates the significance of PEFT enabled by the TraCon register, highlighting its substantial improvement of PEFT results even in scenarios with extremely limited training data (only 18 videos in $\mathcal{D}_2$). $\mathcal{V}_{I,3}$ shows a significant SUC decrease when the asynchronous pipeline is removed, due to the inability to respond promptly to the latest frames. Since the entire pipeline operates synchronously, tracking and action generation must share an inference speed of 13.1 FPS, which falls short of real-time requirements and results in an inevitable discrepancy between the estimated and actual needle positions. $\mathcal{V}_{I,4}$ removes the uncertainty-aware control policy, inserting the needle at a constant velocity of 10 $\mathrm{mm/s}$ and stopping upon reaching the target. The observed decrease in SUC indicates that, when uncertainty increases, continuing insertion at a fixed speed without intervention can lead to tracking loss and insertion failure. This highlights the importance of controlling insertion with environment awareness.

D. Discussion

The improvement can be attributed to two factors: first, the *efficient adaptation of the large-scale pretrained vision backbone*; second, the *LLM-enabled high-level reasoning*.

In contrast to existing large-scale vision trackers like LoRAT [34] that directly use the deepest layer outputs of large backbones, the proposed CDF tracking head in this work efficiently fuses and utilizes multi-level features from different layers of the deep vision encoder, avoiding positional information loss like LoRAT. It can be observed in

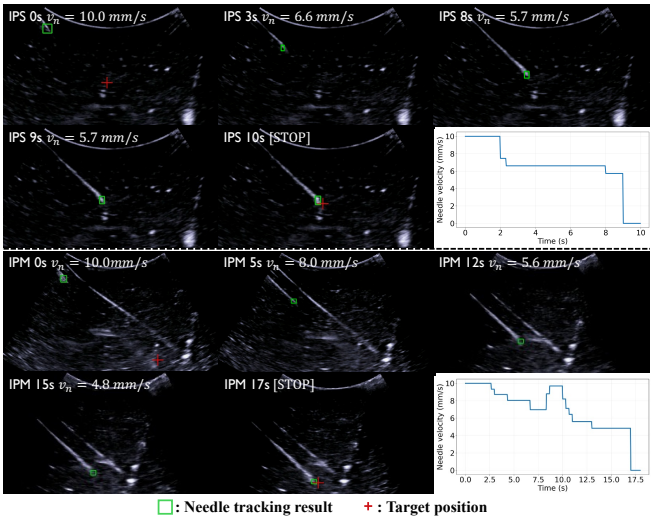

Fig. 5: Two examples of adaptive needle insertion with uncertainty-aware control. The plot of v_n and t is given. The target position is specified by the operator. The proposed framework proactively responds to environmental uncertainty by adjusting the speed when the needle is occluded by tissue to ensure tracking and safety.

TABLE II: Evaluation and ablation study results of needle insertion in SUC (%), T (s).

Method	In-plane-static (IPS)		In-plane-moving (IPM)		Mean	
	$\mathrm{SUC}_{\uparrow}$ (%)	T (s)	$\mathrm{SUC}_{\uparrow}$ (%)	T (s)	$\mathrm{SUC}_{\uparrow}$ (%)	T (s)
Ours	**70.0**	12.1	**90.0**	23.9	**80.0**	17.3
Manual	65.0	17.1	55.0	31.0	60.0	23.2

Ablations	Mean	
	$\mathrm{SUC}_{\uparrow}$ (%)	T (s)
Baseline: default	80.0	17.3
$\mathcal{V}_{I,1}$: $L_r = 16$	75.0(-5.0)	16.1(-1.2)
$\mathcal{V}_{I,2}$: $L_r = 0$ (w/o TraCon R)	72.5(-7.5)	24.5(+7.2)
$\mathcal{V}_{I,3}$: w/o *async* pipeline	67.5(-12.5)	15.0(-2.3)
$\mathcal{V}_{I,4}$: w/o *uncertainty* control	70.0(-10.0)	8.7(-8.6)

Tab. I that our method outperforms LoRAT on most metrics, benefiting from the CDF head's cross-depth fusion. Besides, unlike LoRAT, this work does not fine-tune the vision encoder for tracking, as fine-tuning the vision encoder shared by both the tracking head and the LLM could introduce feature imbalance between them, which would negatively affect LLM-based action generation. Furthermore, ablation studies on the TraCon register show that the vision encoder can be **externally conditioned** instead of internally fine-tuned, even when utilizing a minimal number of trainable parameters. Such a lightweight PEFT approach is crucial for preventing model overfitting in US tasks, given that US datasets usually consist of relatively few samples and insufficient diversity. These improvements ensure *accurate needle position tracking* even in dynamic environments.

Based on accurate needle position feedback, adaptive needle insertion control can be achieved. The results in Tab. II demonstrate the effectiveness of the proposed framework. The LLM is trained on large-scale open-world data,

inherently capturing both semantic information and physical commonsense. Compared to the subjective and user-dependent nature of manual insertion, the proposed method leverages the LLM's *generalizable environment-aware capability* to enable adaptive control. When imaging is affected by occlusion, artifacts, or intermittent needle visibility, the proposed control policy can make context-aware decisions to ensure continuous visualization of the needle.

Last but not least, the asynchronous VLA pipeline enables non-interfering synergy between tracking and control, ensuring real-time tracking and action generation. This contributes to a higher insertion success rate, as shown in Tab. II ($\mathcal{V}_{I,3}$). As such, the asynchronous pipeline is an indispensable component of the proposed VLA insertion-tracking model.

V. CONCLUSIONS

In this paper, a VLA framework is proposed for adaptive US-guided needle insertion and tracking. Extensive experiments demonstrate that our framework achieves state-of-the-art tracking accuracy and significantly improves insertion success rates compared to manual operation. These results highlight the potential of VLA models for enhancing safety and efficiency in RUS needle insertion. One limitation is that the tracking speed only barely meets real-time requirements, indicating room for further efficiency improvements. In future work, multi-degree-of-freedom probe manipulation will be developed to proactively enhance needle visibility. A dataset with a larger cohort will be collected in the future.

REFERENCES

[1] S. Liu *et al.*, "Deep learning in medical ultrasound analysis: a review," *Engineering*, vol. 5, no. 2, pp. 261–275, 2019.

[2] R. R. Richardson, MD and R. R. Richardson, "Imaging modalities: Advantages and disadvantages," *Atlas of Acquired Cardiovascular Disease Imaging in Children*, pp. 1–4, 2017.

[3] P. Chatelain *et al.*, "3d ultrasound-guided robotic steering of a flexible needle via visual servoing," in *2015 IEEE international conference on robotics and automation (ICRA)*, pp. 2250–2255, IEEE, 2015.

[4] A. Kimbowa *et al.*, "Advancements in needle visualization enhancement and localization methods in ultrasound: a literature review," *Artificial Intelligence Surgery*, vol. 4, no. 3, pp. 149–169, 2024.

[5] G. Lapouge *et al.*, "Towards 3d ultrasound guided needle steering robust to uncertainties, noise, and tissue heterogeneity," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 4, pp. 1166–1177, 2020.

[6] M. Khadem *et al.*, "Ultrasound-guided model predictive control of needle steering in biological tissue," *Journal of Medical Robotics Research*, vol. 1, no. 01, p. 1640007, 2016.

[7] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[8] S. Bai *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[9] Y. Ma *et al.*, "A survey on vision-language-action models for embodied ai," *arXiv preprint arXiv:2405.14093*, 2024.

[10] C. K. Ng *et al.*, "Endovla: Dual-phase vision-language-action model for autonomous tracking in endoscopy," *arXiv preprint arXiv:2505.15206*, 2025.

[11] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[12] Y. Zhang *et al.*, "A unified framework for microscopy defocus deblur with multi-pyramid transformer and contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11125–11136, 2024.

[13] B. Li *et al.*, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4282–4291, 2019.

[14] Y. Zhang *et al.*, "Motion-guided dual-camera tracker for endoscope tracking and motion analysis in a mechanical gastric simulator," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 01–07, IEEE, 2025.

[15] C. Mwikirize, J. L. Nosher, and I. Hacihaliloglu, "Learning needle tip localization from digital subtraction in 2d ultrasound," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 1017–1026, 2019.

[16] Y. Zhang *et al.*, "Mambaxctrack: Mamba-based tracker with ssm cross-correlation and motion prompt for ultrasound needle tracking," *IEEE Robotics and Automation Letters*, vol. 10, no. 5, pp. 5130–5137, 2025.

[17] Y. Zhang *et al.*, "Mrtrack: Register mamba for needle tracking with rapid reciprocating motion during ultrasound-guided aspiration biopsy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 407–417, Springer, 2025.

[18] A. M. Wijata, B. Pyciński, and J. Nalepa, "A needle in a (medical) haystack: Detecting a biopsy needle in ultrasound images using vision transformers," in *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 3017–3023, IEEE, 2024.

[19] W. Yan *et al.*, "Learning-based needle tip tracking in 2d ultrasound by fusing visual tracking and motion prediction," *Medical Image Analysis*, vol. 88, p. 102847, 2023.

[20] S. Wang *et al.*, "Trackvla: Embodied visual tracking in the wild," *arXiv preprint arXiv:2505.23189*, 2025.

[21] X. He *et al.*, "Capsdt: Diffusion-transformer for capsule robot manipulation," *arXiv preprint arXiv:2506.16263*, 2025.

[22] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

[23] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[25] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[26] E. Perez *et al.*, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[27] Y. Sato, K. Matsueda, and Y. Inaba, "Basic techniques and technical tips for ultrasound-guided needle puncture," *Interventional Radiology*, vol. 9, no. 3, pp. 80–85, 2024.

[28] F. Casanova, P. R. Carney, and M. Sarntinoranont, "Effect of needle insertion speed on tissue injury, stress, and backflow distribution for convection-enhanced delivery in the rat brain," *PLoS One*, vol. 9, no. 4, p. e94919, 2014.

[29] D. Guo *et al.*, "Siamcar: Siamese fully convolutional classification and regression for visual tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6269–6277, 2020.

[30] Z. Chen *et al.*, "Siamban: Target-aware tracking with siamese box adaptive network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 5158–5173, 2022.

[31] L. Lin *et al.*, "Swintrack: A simple and strong baseline for transformer tracking," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16743–16754, 2022.

[32] Y. Cui *et al.*, "Mixformer: End-to-end tracking with iterative mixed attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13608–13618, 2022.

[33] Z. Fu *et al.*, "Stmtrack: Template-free visual tracking with space-time memory networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13774–13783, 2021.

[34] L. Lin *et al.*, "Tracking meets lora: Faster training, larger model, stronger performance," in *European Conference on Computer Vision*, pp. 300–318, Springer, 2024.

[35] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2411–2418, 2013.

[36] M. Muller *et al.*, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 300–317, 2018.

[37] S. Li *et al.*, "Robonurse-vla: Robotic scrub nurse system based on vision-language-action model," *arXiv preprint arXiv:2409.19590*, 2024.