EndoVLA: Dual-Phase Vision-Language-Action for Precise Autonomous Tracking in Endoscopy

Chi Kit Ng[1*] **Long Bai**[1,2*] **Guankun Wang**[1*] **Yupeng Wang**[1] **Huxin Gao**[1]
Kun Yuan[1,2] **Chenhan Jin**[1] **Tieyong Zeng**[1] **Hongliang Ren**[1†]
[1] The Chinese University of Hong Kong [2] Technical University of Munich

Abstract: In endoscopic procedures, autonomous tracking of abnormal regions and following circumferential cutting markers can significantly reduce the cognitive burden on endoscopists. However, conventional model-based pipelines are fragile — each component (e.g., detection, motion planning) requires manual tuning and struggles to incorporate high-level endoscopic intent, leading to poor generalization across diverse scenes. Vision-Language-Action (VLA) models, which integrate visual perception, language grounding, and motion planning within an end-to-end framework, offer a promising alternative by semantically adapting to surgeon prompts without manual recalibration. Despite their potential, applying VLA models to robotic endoscopy presents unique challenges due to the complex and dynamic anatomical environments of the gastrointestinal (GI) tract. To address this, we introduce EndoVLA, designed specifically for continuum robots in GI interventions. Given endoscopic images and surgeon-issued tracking prompts, EndoVLA performs three core tasks: (1) polyp tracking, (2) delineation and following of abnormal mucosal regions, and (3) adherence to circular markers during circumferential cutting. To tackle data scarcity and domain shifts, we propose a dual-phase strategy comprising supervised fine-tuning on our EndoVLA-Motion dataset and reinforcement fine-tuning with task-aware rewards. Our approach significantly improves tracking performance in endoscopy and enables zero-shot generalization in diverse scenes and complex sequential tasks.

Keywords: Vision–Language Action, Continuum Robots, Autonomous Endoscopic Tracking, Reinforcement Learning

1 Introduction

Endoscopic procedures is the gold standard for robotic-assisted minimally invasive surgery that rely on flexible robotic endoscopes to navigate the long and complex GI tract and to locate and treat lesions [1, 2, 3, 4]. A critical capability during these interventions is tracking of endoluminal targets—such as polyps, lesion boundaries, or circumferential cutting markers—to guide precise manipulation. However, due to the non-intuitive nature of endoscope manipulation, surgeons must undergo extensive and systematic training to develop the required skills [5]. The success of endoscopic surgery highly depends on the surgeon's clinical experience, which can be influenced by limitations of preoperative and intraoperative imaging, restricted visual fields, and subjective biases, etc. These challenges can impair the accuracy of endoscopy planning and execution. Additionally, physical constraints such as hand tremors and fatigue may exacerbate procedural difficulty [6, 7].

Autonomous tracking capabilities for following abnormal regions and circumferential cutting markers could significantly reduce the cognitive burden on endoscopists [8, 9]. However, the GI environment is highly dynamic and unstructured, characterized by continuous tissue deformation and spec-

*Equal contribution.
†Corresponding author: hlren@ee.cuhk.edu.hk

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

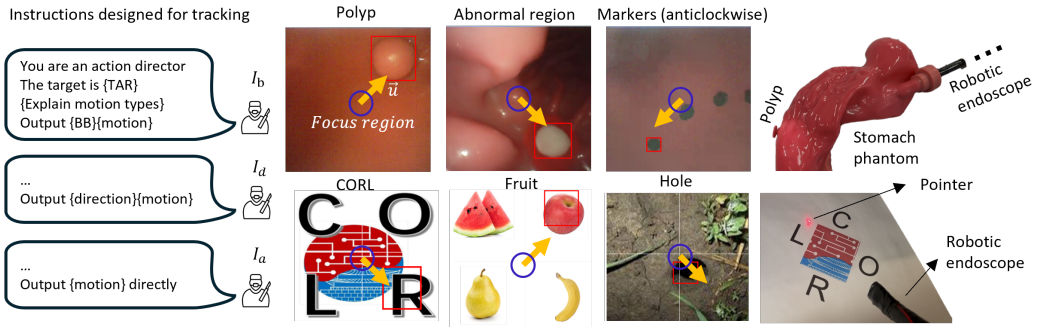

Figure 1: EndoVLA enables robust autonomous tracking in endoscopic procedures, demonstrating effective zero-shot generalization capabilities across general scenes and sequential tracking tasks.

ular reflections that cause rapidly changing appearances. These factors render conventional vision-based tracking pipelines brittle, as they typically decompose the task into distinct perception, motion planning, and control modules—each requiring manual tuning and lacking the ability to generalize across diverse anatomical scenes and incorporate high-level surgical intent [10, 11]. To tackle these constraints, learning-based strategies such as reinforcement learning have been proposed to address specific endoscopic tasks, aiming to improve precision and outcomes [12, 13, 14, 15]. However, these approaches often require intensive data, detailed reward engineering, and extensive parameter tuning. They also lack generalizability across different endoscopic scenarios, making it difficult to address domain gaps in safety-critical in vivo tasks, thereby limiting their autonomy [16, 17].

Multimodal Large Language Models (MLLMs) have gained significant attention for the ability to effectively process textual instructions and visual inputs [18, 19, 20, 21, 22]. These models have demonstrated substantial potential in vision-language-action (VLA) paradigms for robotic planning, control, and execution [16, 23, 24, 25, 26, 27]. MLLMs can directly interpret the surgeon's instructions through natural language (e.g., tracking the polyp), dynamically adjust task objectives as needed, and require minimal training or tuning. Despite these advantages, directly applying VLA frameworks to robotic endoscopy remains challenging due to the anatomical context varies significantly across patients and regions of the GI tract [17, 28]. While MLLMs are trained on general-purpose datasets, they still exhibit the semantic gap when applied to the complex and dynamic in vivo scenarios of endoscopic environments. Moreover, their open-ended nature can conflict with the deterministic and reliable behavior required in endoscopy [20]. Consequently, a customized solution is essential to ensure safe, accurate performance in this safety-critical in vivo environment [29].

To address these challenges, we propose EndoVLA, a dual-phase VLA framework designed specifically for autonomous tracking with continuum robotic endoscopes. Given endoscopic images and surgeon-issued tracking prompts, EndoVLA performs three core tracking tasks essential for autonomous assistance: polyp tracking, delineation and following of abnormal mucosal regions, and adherence to predefined circular markers during circumferential cutting. To achieve robust performance and generalization, our approach incorporates a novel dual-phase fine-tuning (DFT) strategy that combines supervised fine-tuning (SFT) with reinforcement fine-tuning (RFT) using task-aware verifiable rewards. Specifically, our contributions are as follows: (1) We introduce EndoVLA, an end-to-end VLA model tailored for the continuum endoscopic robot. We further employ the DFT approach, combining SFT and RFT for robust performance and generalization. (2) We construct the EndoVLA-Motion dataset for vision-language-kinematics modeling of endoscopic robots. (3) Extensive real-world experiments demonstrate that EndoVLA achieves state-of-the-art performance across three endoscopic tracking tasks while exhibiting remarkable zero-shot generalization to tracking tasks in general scenes and more challenging sequential tasks.

2 Related Work

Multimodal Large Language Models in Robotics. Recently, there has been substantial advancement in the development of multimodal large language models tailored for robotic applica-

tions [30, 31, 32, 33]. These approaches are trained on large-scale datasets, taking as input sequences of images from robots' cameras with task instructions in natural language, and predicting motion outputs [34]. VoxPoser [35] introduces an LLM-guided perception system capable of grounding daily commands into manipulation primitives using visual information. Robotics Transformer 2 [36] extended this concept with large-scale web and robotic data to support zero-shot generalization. Rather than relying on token-based action representations, Black et al. [37] adopt continuous action generation using flow matching, allowing precise, high-frequency control suitable for dexterous tasks. Although great progress has been achieved in robotics, these techniques often exhibit poor generalization to endoscopic domains, as MLLMs optimized for specific tasks tend to struggle when applied to more complex tasks in diverse endoscopic scenarios.

Robot Learning for Automatic Surgical Robotics. Endoscopy-assisted surgical robotics has increasingly adopted reinforcement learning (RL) and imitation learning (IL) to achieve the precision and adaptability necessary for delicate clinical procedures [38, 39]. Deep RL has been employed to acquire complex surgical skills such as suturing, tissue manipulation, and needle handling, achieving notable improvements in accuracy and autonomy across simulated and real-world environments [40]. To address the high data demands of RL, demonstration-guided RL integrates expert trajectories to accelerate convergence and reduce exploration in tasks like knot tying and percutaneous needle insertion [14]. In parallel, sequence-based IL from surgical video data enables the learning of instrument motion patterns and semantic understanding directly from intraoperative recordings, advancing applications such as laparoscope guidance and endoscopic navigation under limited supervision [41]. Moreover, large-scale video-based IL models trained on da Vinci robotic surgery datasets have demonstrated near-surgeon-level proficiency in foundational tasks, including needle manipulation and tissue retraction [42]. While prior techniques have shown promising performance, their application to continuum endoscopy remains unexplored, and challenges like fine-grained control and real-time perception for continuum endoscopic robots have yet to be addressed.

3 Robotic System and Dataset Collection

3.1 Endoscopic Robot System

To implement our proposed EndoVLA on an actual robotic system, we developed an endoscopic robot system for automatic control based on the Olympus endoscope. The commercial Olympus endoscope features two control knobs at its operating end, which regulate the 2-degree-of-freedom (2-DOFs) movement of the endoscopic end-effector. This 2-DOFs movement, as depicted in Figure 2 (a), allows the front segment of the robot to bend towards four perpendicular directions. By utilizing two motors to independently drive these knobs, the system achieves full coverage of movement within the imaging field, enabling real-time tracking of the target. The endoscopic system also provides a real-time video feed at 30 FPS. All experiments are conducted at a resolution of 400×400 pixels, corresponding to the maximum output resolution of the endoscopic camera. Based on the output from EndoVLA, motion is converted to angle under the assumption of linearity mapping between the bending angle of front segment of endoscope and motor rotation angle.

3.2 EndoVLA-Motion Dataset

Given the absence of multimodal data for continuum robots, we construct EndoVLA-Motion, a vision-language and kinematic dataset, to train our EndoVLA model. EndoVLA-Motion dataset comprises 6k image–action pairs across three tasks: polyp tracking, abnormal region localization, and circular-marker following. All data are acquired with the robotic endoscope system using two types of stomach phantoms embedding polyps, abnormal mucosal regions, and concentric arrays of black markers. Raw video streams are captured under teleoperated control and automatically annotated with bounding boxes via YOLOv5 [43], followed by manual curation to remove erroneous detections. Each bounding box is mapped to one of four discrete bending motions—upper-right, upper-left, lower-left, or lower-right—yielding the corresponding action label. As shown in Fig. 1, the circular shape of focus region (FR) is defined. The center of FR is image center. The threshold

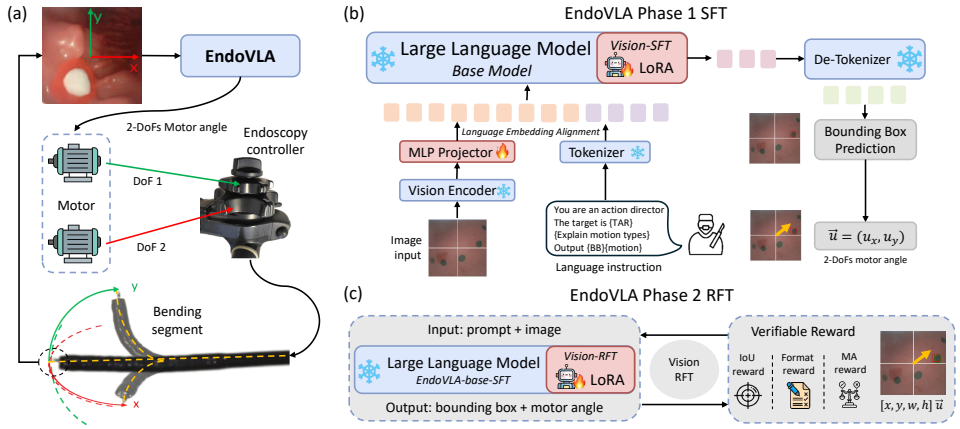

Figure 2: Overview of the setup of robotic endoscope and the DFT architecture of EndoVLA

of the desired observation region is set to be 28 pixels. If the distance between the center of the target and the center is less than the threshold, the motion type is modified to be "still". We split the EndoVLA dataset into training (80%) and evaluation (20%) partitions to ensure robust model assessment. The prompts consist of two crucial components: (1) a detailed visual scene description generated by the Qwen2-VL base model, and (2) an explicit formulation of the spatial relationship between the localized target and the corresponding optimal bending motion direction. This structured prompt design enables the model to establish clear correlations between visual inputs, spatial reasoning, and appropriate robotic motion commands.

4 Vision-Language-Action Instruction Tuning

4.1 Preliminaries

Our VLA model is built upon the Qwen2-VL backbone. We jointly process the current RGB observation $O_t \in \mathbb{R}^{H\times W\times 3}$ and the instruction prompt I to produce the next control command. At each timestep t, the current RGB frame O_t is mapped into a latent vector via a frozen visual encoder E_v, while the instruction I is tokenized and embedded by a fixed language encoder E_l. These two feature vectors are then combined and passed through the finetuned language model f_ϕ. The output tokens is given by $T_l = f_\phi\big(E_v(O_t),\, E_l(I)\big)$, where T_l is the decoded text.

4.2 EndoVLA Model

Architecture. Instead of full parameter training, we employ Low-Rank Adaptation (LoRA [44] to efficiently fine-tune the LLM while keeping most of the base model parameters frozen. For a given RGB image O_t and a language instruction I, the model predicts both a bounding box B_t of the target and an appropriate action A_t. This can be formulated as $(B_t, A_t) = f_\phi(O_t, I)$. The RGB image processing pipeline involves a fixed vision encoder that extracts visual features, followed by a trainable MLP projector that aligns these features with the language embedding space of the LLM. The language instruction is processed directly by the LLM's tokenizer. This multimodal fusion allows the model to understand both the visual context and the linguistic commands, enabling it to make accurate predictions for the tracking tasks. The bounding box prediction B_t is represented in the format $[x, y, w, h]$ where (x, y) denotes the top-left corner coordinates and (w, h) represents the width and height of the box. The action A_t corresponds to one of the predefined discrete actions.

Problem Formation. Given an instruction I that specifies the target name, the LLM must first transform this linguistic input into an object detection task on the RGB image to localize the target, and then decide whether to actuate the robot or hold position based on the localization result. Let $p_t = (u_t, v_t)^\top \in \mathbb{R}^2$ denote the pixel coordinates of the detected target center in the current image frame, and $p_f = (u_f, v_f)^\top \in \mathbb{R}^2$ denote the pixel coordinates of the desired focus region center.

Denote by $\boldsymbol{\theta} = [\theta_1, \theta_2]^\top \in \mathbb{R}^2$ the incremental rotations of the two actuation motors (one pair for bending towards right/left, one pair for bending towards up/down), and assume each bending angle α_i is linearly related by $\alpha_i = k\,\theta_i$ for some constant k. We define the (nonlinear) mapping from motor rotations to image coordinates of the target as $p_t = \mathcal{M}(\boldsymbol{\theta})$, $\mathcal{M} : \mathbb{R}^2 \to \mathbb{R}^2$. We define a discrete action set $\mathcal{A} = \{\text{upper-right, upper-left, lower-left, lower-right, stop}\}$, each non-stop action $a \in \mathcal{A}$ corresponding to a fixed motor increment $\Delta\boldsymbol{\theta}(a)$:

$$\begin{aligned} \Delta\boldsymbol{\theta}(\text{upper-right}) &= [+\,\delta\theta_1,\ +\delta\theta_2]^\top, \quad \Delta\boldsymbol{\theta}(\text{lower-left}) &= [-\,\delta\theta_1,\ -\delta\theta_2]^\top, \\ \Delta\boldsymbol{\theta}(\text{upper-left}) &= [-\,\delta\theta_1,\ +\delta\theta_2]^\top, \quad \Delta\boldsymbol{\theta}(\text{lower-right}) &= [+\,\delta\theta_1,\ -\delta\theta_2]^\top, \end{aligned} \tag{1}$$

and the special "stop" action is executed when $\|p_t - p_f\|_2 \le \epsilon$, indicating mission success. The ϵ is a fixed value, which is defined when the training dataset is built. We set $\epsilon = 18$ pixels. A vision–language action LLM parameterized by ϕ maps the current image and instruction I to a discrete action: $a_t = f_\phi\big(E_v(O_t),\ E_l(I)\big) \in \mathcal{A}$. When "stop" signal has not yet been issued, applying a_t for updating the motors: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \Delta\boldsymbol{\theta}(a_t)$, and the new target position is $p_{t+1} = \mathcal{M}\big(\boldsymbol{\theta}_{t+1}\big)$.

The control objective is to choose actions so that the pixel-distance $\|p_t - p_f\|_2$ is reduced below a threshold ϵ as quickly as possible, and then issue "stop":

$$a_t^* = \begin{cases} \arg\min_{a \in \mathcal{A}} \Big\| \mathcal{M}\big(\boldsymbol{\theta}_t + \Delta\boldsymbol{\theta}(a)\big) - p_f \Big\|_2, & \|p_t - p_f\|_2 > \epsilon, \\ \text{stop}, & \|p_t - p_f\|_2 \le \epsilon. \end{cases} \tag{2}$$

Input. EndoVLA processes two input components: an RGB frame O_t captured by the eye-in-hand endoscopic camera, and a prompt I that combines a textual description I_T with a motion guideline I_G. To generate I_T, we leverage the image captioning capabilities of the pretrained LLM to describe the visual scene. For I_G, we incorporate in-context learning principles by providing exemplars that demonstrate the reasoning process for selecting appropriate actions based on target localization [45]. Specifically, we include examples that show how to reason about the spatial relationship between the detected target and the desired focus point, followed by the corresponding action selection.

4.3 Dual-phase Fine-tuning (DFT)

Supervised Fine-tuning. We implement a LoRA adaptor [44] to align visual representations with the LLM enabling effective processing of endoscopic images with language instructions. The training objective minimizes prediction errors on ground-truth bounding boxes and action prediction.

Reinforcement Learning with Verifiable Rewards. To further enhance the model's performance beyond supervised learning, we employ Reinforcement Learning with Verifiable Rewards. This approach uses reward signals that can be objectively computed from the model's predictions rather than relying on subjective human preferences or black-box reward models.

The optimization is across different target–focus configurations (groups), and we employ Group Relative Policy Optimization (GRPO) [46]. Let $\mathcal{G}$ be the set of groups (i.e. different target positions or image contexts) and for each group $g \in \mathcal{G}$ let trajectories $\tau^g = (s_0, a_0, \dots)$ be collected under the old policy $\pi_{\phi_{\text{old}}}$. Define the group-specific importance ratio and advantage:

$$r_t^g(\phi) \;=\; \frac{\pi_\phi(a_t \mid s_t, g)}{\pi_{\phi_{\text{old}}}(a_t \mid s_t, g)}, \quad A_t^g \;=\; R_t^g - V_\psi(s_t, g), \tag{3}$$

where R_t^g is the cumulative reward (e.g. IoU, MA, Format) for group g, and $V_\psi(s_t, g)$ is a value function estimate. The GRPO surrogate objective with clipping and a KL penalty across groups is:

$$\begin{aligned} L^{\text{GRPO}}(\phi) = &\sum_{g \in \mathcal{G}} w_g\, \mathbb{E}_{t \sim \tau^g}\Big[\min\big(r_t^g(\phi)\, A_t^g,\ \text{clip}(r_t^g(\phi), 1-\epsilon, 1+\epsilon)\, A_t^g\big)\Big] \\ &- \alpha \sum_{g \in \mathcal{G}} w_g\, \mathbb{E}_{t \sim \tau^g}\Big[\text{KL}\big[\pi_{\phi_{\text{old}}}(\cdot \mid s_t, g) \,\big\|\, \pi_\phi(\cdot \mid s_t, g)\big]\Big]. \end{aligned} \tag{4}$$

where w_g are nonnegative group-weighting coefficients, ϵ is the clipping threshold, and α scales the KL penalty. The policy parameters are updated by $\phi^* = \arg\max_\phi \; L^{\mathrm{GRPO}}(\phi)$, ensuring that each group's policy improves relative to its own baseline while maintaining stability across updates.

In our framework, verifiable rewards provide quantitative feedback for the model's predictions based on direct measurement of prediction quality. We designed three complementary reward functions: 1. **IoU Reward**: This reward measures the Intersection over Union between the predicted bounding box and the ground truth bounding box:

$$R_{IoU}(B_{\text{pred}}, B_{\text{gt}}) = \frac{\text{Area}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{Area}(B_{\text{pred}} \cup B_{\text{gt}})} \tag{5}$$

where $B_{\text{pred}} = [x_{\text{pred}}, y_{\text{pred}}, w_{\text{pred}}, h_{\text{pred}}]$ is the predicted bounding box and $B_{\text{gt}} = [x_{\text{gt}}, y_{\text{gt}}, w_{\text{gt}}, h_{\text{gt}}]$ is the ground truth bounding box.

2. **Motion Angle (MA) Reward**: This binary reward evaluates whether the predicted action matches the ground truth action:

$$R_{MA}(A_{\text{pred}}, A_{\text{gt}}) = \begin{cases} 1.0 & \text{if } A_{\text{pred}} = A_{\text{gt}} \\ 0.0 & \text{otherwise} \end{cases} \tag{6}$$

where actions are selected from a discrete action set $\mathcal{A}$ defined at Eq. 1.

3. **Format Reward**: This reward ensures the model's output adheres to the required format of $[x, y, w, h]a_t$, where the data type of discrete action a_t is character:

$$R_{Format}(output) = \begin{cases} 1.0 & \text{if output matches [digit, digit, digit, digit] character} \\ 0.0 & \text{otherwise} \end{cases} \tag{7}$$

5 Experiments

5.1 Implementation Details

Our experiments address two key questions: (1) Can EndoVLA accurately track targets in dynamic endoluminal environments using RGB images and textual prompts with limited training data? (2) Does explicit object localization improve motion prediction compared to direct motion commands?

EndoVLA is first SFT using Unsloth [47] and then RFT with VLM-R1 [48]. Our base model is Qwen2-VL-7B [49], with LoRA-based DFT on an NVIDIA A6000 GPU. The training of our model on the EndoVLA-motion dataset was performed for a single epoch. For the SFT phase, we employed the AdamW optimizer with an initial learning rate of 2e-4 and implemented a linear decay schedule, completing the training in approximately 3 hours with a batch size of 2. Subsequently, during the RFT phase, we maintained the AdamW optimizer but reduced the initial learning rate to 1e-5, utilizing a batch size of 4 and a group size of 4 (g=4 in Eq. 3) for 4 hours training.

We evaluate three endoscopic tasks (Figure 3): polyp tracking (PP), abnormal region tracking (AR), and circular cutting marker following (CC), with increasing localization difficulty. We test three instruction types (Figure 1): bounding box localization (I_b), directional localization (I_d), and direct action prediction (I_a). Additionally, we assess generalization in non-endoscopic scenes (Figure 4): hole tracking, sequential character tracking (CORL), and sequential fruit tracking.

5.2 Motion Prediction with Base Model

We evaluated the base Qwen2-VL model using instructions I_d and I_a on both raw images (Img_{raw}) and images with ground truth bounding boxes (Img_{gt}). Table 1 shows that localization instructions (I_d) improve performance, particularly for PP.

Table 1: **Motion Type Prediction of Base Model**

Tasks	PP/I_a	PP/I_d	AR/I_a	AR/I_d	CC/I_a	CC/I_d
Random	20.0	20.0	20.0	20.0	25.0	25.0
Img_{raw}	35.8	44.0	36.5	31.5	**26.5**	20.8
Img_{gt}	**38.4**	**50.2**	**42.1**	**45.1**	18.2	**24.7**
↑ (%)	7.3	14.1	15.3	43.2	-31.3	18.8

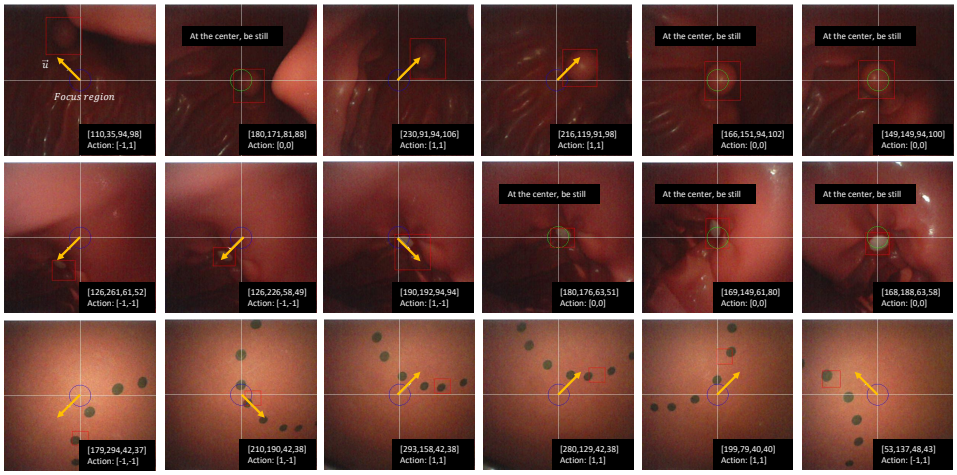

Figure 3: Example of successful rollouts on the real-world endoscopic tasks (PP, AR and CC) on a robotic endoscope setup and stomach phantoms.

Adding ground-truth boxes under I_d yielded improvements of +7.3% (PP), +43.2% (AR), and +18.8% (CC), confirming that pairing explicit localization with reliable bounding boxes boosts motion prediction.

Table 2: **IoU of bounding box and PR (%) of action prediction**

Model	PP (IoU, I_b)	PP (I_b)	PP (I_a)	AR (IoU, I_b)	AR (I_b)	AR (I_a)	CC (IoU, I_b)	CC (I_b)	CC (I_a)
SFT	**62.1**	**87.7**	**66.4**	45.7	**81.5**	**70**	**11.0**	57.1	26.8
RFT	58.0	73.5	57.6	**48.2**	61.3	34.9	2.8	**65.5**	**34.4**
SFT + RFT	**86.1**	**94.8**	**78.7**	**79.1**	**92.0**	**85**	**48.5**	**89.7**	**56**
↑ (%)	38.6	8.1	18.5	64.1	12.9	21.4	340.9	36.9	62.8

5.3 Ablation Study

Table 2 presents IoU for bounding box predictions and precision rate (PR) for motion predictions across different FT strategies. Single-phase approaches showed task-specific strengths: SFT performed well on PP (87.7% with I_b) but struggled with complex localization in CC (11.0% IoU), while RFT showed complementary strengths in CC but underperformed in PP.

Our DFT approach demonstrated superior performance across all metrics and tasks, with IoU improvements of 38.6% (PP), 64.1% (AR), and 340.9% (CC) compared to single-phase approaches. Motion prediction accuracy reached 94.8% (PP), 92.0% (AR), and 89.7% (CC) under I_b, representing improvements of 8.1%, 12.9%, and 36.9% respectively.

Table 3: **Performance in real-world endoscopic tracking tasks.**

Model	PP (I_b, c)	PP (I_b, r)	PP (I_a, c)	PP (I_a, r)	AR (I_b, c)	AR (I_b, r)	AR (I_a, c)	AR (I_a, r)
SFT	70.0	33.3	50.0	30.0	83.0	40.0	57.0	20.0
RFT	37.0	37.0	43.0	0.0	0.0	0.0	0.0	0.0
SFT+RFT	**100.0**	**100.0**	**100.0**	**63.0**	**100.0**	**100.0**	**100.0**	**57.0**

(a) **SR (%) of PP and AR tracking tasks**

Model	CR (I_b)	SR (I_b)
SFT	50.0	0.0
RFT	45.0	0.0
SFT+RFT	**66.7**	**10.0**

(b) **CR (%) and SR (%) of CC task**

5.4 Real Robot Evaluation

For PP and AR, the maximum steps are 30 across 30 trials to position the targets within the FR. We measured SR using two metrics: moving closer (**c**) and moving within FR (**r**). For CC, models had 200 steps across 10 trials, with completion rate (CR) and SR measurements. The DFT approach

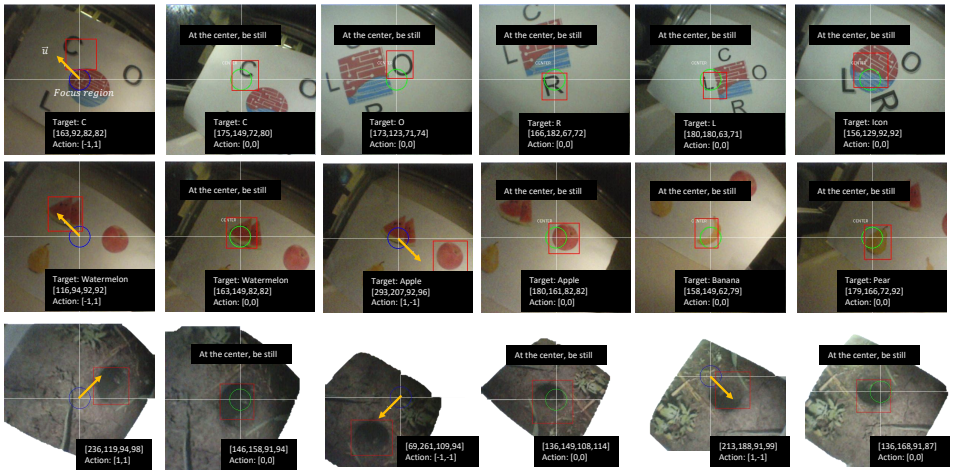

Figure 4: Example of successful rollouts on a real-world robotic setup of tasks: CORL character and icon sequential tracking, fruit sequence tracking, and hole tracking in the field.

achieved 100% success in moving toward targets in both PP and AR tasks, with significant improvements in precise positioning (63.0% for PP and 57.0% for AR with I_a), outperforming single-phase approaches by 110.0% and 185.0%, respectively. For the challenging CC task, only the DFT model achieved any success (10.0%) in completing the entire circle.

Table 4: **SR (%) in general scene tracking tasks**

Model	**CORL Character Sequence**						**Fruit Sequence**					**Hole**
	Full	C_{FR}	O_{FR}	R_{FR}	L_{FR}	$Icon_{FR}$	**Full**	W_{FR}	A_{FR}	P_{FR}	B_{FR}	
SFT	0.0	0.0	0.0	10.0	0.0	0.0	90.0	100.0	100.0	100.0	90.0	30.0
RFT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SFT+RFT	**50.0**	**100.0**	**70.0**	**60.0**	**60.0**	**80.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**90.0**

5.5 Generalization to General Scenes

We evaluated generalization to non-endoscopic tasks: CORL character and icon tracking, fruit sequence tracking, and hole tracking in the field. The maximum number of steps across 10 trials is 200, as measured by the SR. As shown in Table 4, the DFT approach demonstrated remarkable zero-shot performance, achieving a 50% success rate for the complete CORL sequence, 100% for fruit tracking, and 90% success in tracking hole in the field. In contrast, the RFT-only model generated excessively large bounding boxes and showed inconsistent motion predictions, while the SFT model struggled with visually different scenes (30% success in tracking hole).

These results confirm that the DFT enhances both endoscopic performance and zero-shot generalization by developing robust localization capabilities that transfer effectively across visual contexts.

6 Conclusion

In this work, we present EndoVLA, a novel Vision-Language-Action model tailored for autonomous tracking in endoscopic procedures using continuum robots. Addressing the critical challenges of limited endoscopic data and domain-specific generalization, we introduce a DFT strategy that combines SFT and RFT guided by verifiable, task-specific rewards. We construct the EndoVLA-Motion dataset, a vision-language-kinematics dataset, to facilitate training under realistic robotic settings. Our extensive experiments across both endoscopic and non-endoscopic tasks demonstrate that EndoVLA not only achieves state-of-the-art performance in precise target localization and motion prediction but also exhibits robust zero-shot generalization to novel environments and task sequences. This work establishes a foundation for more autonomous, precise robotic assistance in VLA for continuum robots.

Acknowledgments

This work was supported by the Hong Kong Research Grants Council Collaborative Research Fund under Grant CRF-C4026-21G, by the Hong Kong Research Grants Council Research Impact Fund under Grant RIF-R4020-22, Hong Kong RGC GRF (14204524, 14203323, 14216022), NSFC/RGC Joint Research Scheme N_CUHK420/22.

References

[1] H. Gao, X. Yang, X. Xiao, X. Zhu, T. Zhang, C. Hou, H. Liu, M. Q.-H. Meng, L. Sun, X. Zuo, et al. Transendoscopic flexible parallel continuum robotic mechanism for bimanual endoscopic submucosal dissection. *The International Journal of Robotics Research*, 43(3):281–304, 2024.

[2] J. T. Maple, B. K. A. Dayyeh, S. S. Chauhan, J. H. Hwang, S. Komanduri, M. Manfredi, V. Konda, F. M. Murad, U. D. Siddiqui, and S. Banerjee. Endoscopic submucosal dissection. *Gastrointestinal endoscopy*, 81(6):1311–1325, 2015.

[3] D.-H. Lee, B. Cheon, J. Kim, and D.-S. Kwon. easyendo robotic endoscopy system: Development and usability test in a randomized controlled trial with novices and physicians. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 17(1):1–14, 2021.

[4] K. Peng, Y. Chang, G. Lang, J. Xu, Y. Gao, J. Yin, and J. Zhao. Image segmentation network for laparoscopic surgery. *Biomimetic Intelligence and Robotics*, page 100236, 2025.

[5] X. Zhang, E. K. Ly, S. Nithyanand, R. J. Modayil, D. O. Khodorskiy, S. Neppala, S. Bhumi, M. DeMaria, J. L. Widmer, D. M. Friedel, et al. Learning curve for endoscopic submucosal dissection with an untutored, prevalence-based approach in the united states. *Clinical Gastroenterology and Hepatology*, 18(3):580–588, 2020.

[6] H. Lin, B. Li, X. Chu, Q. Dou, Y. Liu, and K. W. S. Au. End-to-end learning of deep visuomotor policy for needle picking. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8487–8494. IEEE, 2023.

[7] M. Hwang, D. Seita, B. Thananjeyan, J. Ichnowski, S. Paradis, D. Fer, T. Low, and K. Goldberg. Applying depth-sensing to automated surgical manipulation with a da vinci robot. In *2020 international symposium on medical robotics (ISMR)*, pages 22–29. IEEE, 2020.

[8] M. Islam, Y. Li, and H. Ren. Learning where to look while tracking instruments in robot-assisted surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 412–420. Springer, 2019.

[9] G. Cui, S. Su, H. Gao, K. Zhuo, K. Yang, and H. Wu. Soft objects grasping evaluation using a novel vcfn-yolov8 framework. *Biomimetic Intelligence and Robotics*, page 100232, 2025.

[10] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G. L. Gallia, R. H. Taylor, and G. D. Hager. Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data. *IEEE transactions on medical imaging*, 37(10):2185–2195, 2018.

[11] Z. Fu, Z. Jin, C. Zhang, Z. He, Z. Zha, C. Hu, T. Gan, Q. Yan, P. Wang, and X. Ye. The future of endoscopic navigation: a review of advanced endoscopic vision technology. *IEEE Access*, 9:41144–41167, 2021.

[12] Y. Long, W. Wei, T. Huang, Y. Wang, and Q. Dou. Human-in-the-loop embodied intelligence with interactive simulation environment for surgical robot learning. *IEEE Robotics and Automation Letters*, 8(8):4441–4448, 2023.

[13] J. Xu, B. Li, B. Lu, Y.-H. Liu, Q. Dou, and P.-A. Heng. Surrol: An open-source reinforcement learning centered and dvrk compatible platform for surgical robot learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1821–1828. IEEE, 2021.

[14] T. Huang, K. Chen, B. Li, Y.-H. Liu, and Q. Dou. Guided reinforcement learning with efficient exploration for task automation of surgical robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4640–4647. IEEE, 2023.

[15] Y. Tian, R. Hao, Y. Huang, D. Xie, C. P. L. Chan, J. Y. K. Chan, and H. Ren. Learning to perform low-contact autonomous nasotracheal intubation by recurrent action-confidence chunking with transformer. *arXiv preprint arXiv:2508.01808*, 2025.

[16] S. Li, J. Wang, R. Dai, W. Ma, W. Y. Ng, Y. Hu, and Z. Li. Robonurse-vla: Robotic scrub nurse system based on vision-language-action model. *arXiv preprint arXiv:2409.19590*, 2024.

[17] M. Moghani, L. Doorenbos, W. C.-H. Panitch, S. Huver, M. Azizian, K. Goldberg, and A. Garg. Sufia: language-guided augmented dexterity for robotic surgical assistants. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6969–6976. IEEE, 2024.

[18] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[20] G. Wang, L. Bai, J. Wang, K. Yuan, Z. Li, T. Jiang, X. He, J. Wu, Z. Chen, Z. Lei, et al. Endochat: Grounded multimodal large language model for endoscopic surgery. *arXiv preprint arXiv:2501.11347*, 2025.

[21] C. Zhang, J. Chen, J. Li, Y. Peng, and Z. Mao. Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics*, 3(4):100131, 2023.

[22] L. Bai, B. Ma, R. Wang, G. Wang, B. Cui, Z. Jiang, M. Islam, Z. Min, J. Lai, N. Navab, et al. Multimodal graph representation learning for robust surgical workflow recognition with adversarial feature disentanglement. *Information Fusion*, page 103290, 2025.

[23] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[24] J. Liu, M. Liu, Z. Wang, P. An, X. Li, K. Zhou, S. Yang, R. Zhang, Y. Guo, and S. Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024.

[25] H. Arai, K. Miwa, K. Sasaki, K. Watanabe, Y. Yamaguchi, S. Aoki, and I. Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1933–1943. IEEE, 2025.

[26] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[27] Z. Min, J. Lai, and H. Ren. Innovating robot-assisted surgery through large vision models. *Nature Reviews Electrical Engineering*, pages 1–14, 2025.

[28] R. Rudiman. Minimally invasive gastrointestinal surgery: from past to the future. *Annals of Medicine and Surgery*, 71:102922, 2021.

[29] K. Fan, Z. Chen, G. Ferrigno, and E. De Momi. Learn from safe experience: Safe reinforcement learning for task automation of surgical robot. *IEEE Transactions on Artificial Intelligence*, 5(7):3374–3383, 2024.

[30] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[31] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[32] Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, T. Zhang, H.-S. Fang, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis. *arXiv preprint arXiv:2312.08782*, 2023.

[33] X. He, M. Su, X. Jiang, L. Bai, J. Lai, and H. Ren. Capsdt: Diffusion-transformer for capsule robot manipulation. *arXiv preprint arXiv:2506.16263*, 2025.

[34] S. Schmidgall, J. W. Kim, A. Kuntz, A. E. Ghazi, and A. Krieger. General-purpose foundation models for increased autonomy in robot-assisted surgery. *Nature Machine Intelligence*, pages 1–9, 2024.

[35] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[36] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[37] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[38] J. Wang, W. Chen, X. Xiao, Y. Xu, C. Li, X. Jia, and M. Q.-H. Meng. A survey of the development of biomimetic intelligence and robotics. *Biomimetic Intelligence and Robotics*, 1:100001, 2021.

[39] T. Bin, H. Yan, N. Wang, M. N. Nikolić, J. Yao, and T. Zhang. A survey on the visual perception of humanoid robot. *Biomimetic Intelligence and Robotics*, 5(1):100197, 2025.

[40] C. Qian and H. Ren. Deep reinforcement learning in surgical robotics: enhancing the automation level. *Handbook of Robotic Surgery*, pages 89–102, 2025.

[41] J. Liu, A. Andres, Y. Jiang, X. Luo, W. Shu, and S. A. Tsaftaris. Surgical task automation using actor-critic frameworks and self-supervised imitation learning. *arXiv preprint arXiv:2409.02724*, 2024.

[42] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger. Surgical robot transformer (srt): Imitation learning for surgical tasks. *arXiv preprint arXiv:2407.12998*, 2024.

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[44] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[45] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[46] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[47] M. H. Daniel Han and U. team. Unsloth, 2023. URL http://github.com/unslothai/unsloth.

[48] H. Shen, P. Liu, J. Li, C. Fang, Y. Ma, J. Liao, Q. Shen, Z. Zhang, K. Zhao, Q. Zhang, R. Xu, and T. Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[49] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

7 Limitations

Dataset Size and Diversity. The EndoVLA-Motion dataset is relatively small (about 6k image-action pairs) and limited to only three phantom-based tasks. This restricted dataset may constrain the model's generalizability to a wider range of endoscopic scenarios. For example, we have not tested the model on real procedures with varied pathologies (such as active bleeding or mucosal irregularities) or on additional tasks that occur in clinical settings. The controlled, simplified nature of our training data (clean phantoms) may not capture the full complexity of in vivo conditions.

Environmental and Visual Variability. Our current evaluations use phantom models under ideal conditions. In contrast, real endoscopic environments often involve challenges like significant soft-tissue deformation, specular reflections, bleeding, smoke, mucus, or partial occlusions from surgical instruments. These factors can severely degrade visual tracking performance, yet they were not present in our experiments. As a result, the robustness of our model in such adverse visual conditions remains unverified.

Motion Control and Actuation Assumptions. The system uses a discrete set of motion commands (four bending directions and a stop) as a simplification of the continuum motion space. It also assumes a simple linear mapping between the predicted command and the actual motor rotation angle. These choices ignore complex factors such as nonlinear kinematics, actuation delays, or varying compliance in the endoscope mechanism. This coarse discretization and linear approximation may limit fine-grained manipulation precision and accuracy.

Inference Speed and Temporal Responsiveness. The full EndoVLA model runs at roughly 2 Hz, which is significantly lower than video frame rates (30 Hz). This latency means that fast-moving targets could outpace the controller, potentially causing delayed reactions or overshooting. Moreover, our model currently processes each frame independently without explicit temporal context. It cannot anticipate future motion or utilize history, which further hinders performance in rapidly changing scenes.

Observed Failure Modes. We observed several task-specific failure cases in our experiments.

- In the circular cutting task, for instance, the EndoVLA sometimes tracked the target ring in the wrong rotational direction (clockwise instead of counterclockwise).
- In zero-shot general scene tests, the model occasionally misidentified targets (for example, confusing similar letters or shapes) and lost track of the intended object. These errors high-

light challenges in disambiguating visual cues and following instructions when multiple plausible targets are present.

- In sequential tracking scenarios, our current approach lacks a mechanism for detecting or recovering when the next target falls outside the field of view after completing tracking of the current target. The system continues to issue motion commands without correction when tracking is lost or when model predictions become uncertain. We have not implemented confidence estimation or a re-localization strategy that would allow the system to recognize tracking failures and recover from them. This absence of uncertainty-aware control and recovery mechanisms means the system may fail silently in scenarios where tracking continuity is compromised.

Future Works. While EndoVLA has demonstrated strong performance under controlled conditions, its path toward clinical deployment requires several key advancements. First, we will integrate temporal and uncertainty-aware modeling to enhance safety and robustness. By extending the model to incorporate sequence-based architectures and real-time confidence estimation, the system can anticipate rapid target motion, detect out-of-distribution inputs, and trigger failsafe behaviors—such as pausing actuation or reverting to manual control—when uncertainty exceeds safe thresholds. We will also explore continuous control outputs, moving beyond the current discrete action set to predict continuous bending angles or velocities, and investigate model compression and hardware acceleration (e.g., distillation) to boost inference speed toward real-time rates ($\geq$30 Hz).

Second, we plan to broaden the training data and evaluation scenarios to bridge the gap between phantom studies and the complexity of in vivo endoscopy. This includes collecting and annotating diverse clinical videos that capture bleeding, smoke, soft-tissue deformation, and occlusions, as well as designing multi-step and multi-target tracking benchmarks. Complementing real data with simulated variations and domain-adaptation techniques will help address data scarcity. Finally, we will conduct human-in-the-loop user studies in wet-lab environments to assess endoscopists' trust cognitive load, and procedural efficiency, comparing EndoVLA's assistance against traditional methods, which will be essential to refine the system for safe and effective use in real endoscopic workflows.

Supplementary Materials

A Dataset Creation Pipeline

To construct the EndoVLA-Motion dataset, we developed an automated labeling pipeline using YOLOv5-based object detection, followed by human-in-the-loop curation. The dataset includes three task-specific tracking categories: (1) **PP**: polyp tracking, (2) **AR**: delineation and following of abnormal mucosal regions, and (3) **CC**: adherence to predefined circular markers during circumferential cutting.

A.1 Data Collection Setup

Figure 5a illustrates the physical setup used for data acquisition. Two different phantoms were employed to simulate gastrointestinal environments. Figure 5b shows the robotic setup, which includes a tendon-driven continuum endoscope manipulated by a motorized actuation unit, facilitating controlled and repeatable motions during data collection.

A.2 Automated Labeling for PP and AR Tasks

For both PP and AR tasks, frames were sampled from endoscopic videos (30 FPS). YOLOv5s, pre-trained on COCO, was employed to detect the target (e.g., polyp or abnormal region) in each selected frame. The highest-confidence detection bounding box $B = [x, y, w, h]$ was retained.

We defined a circular *Focus Region* (FR) centered in the image with a radius threshold $r = 20\sqrt{2}$ pixels (resolution is 400 $\times$ 400). The center of the bounding box was computed and compared to the FR center to determine motion labels:

- If $\|\text{center}(B) - \text{center(FR)}\| < r$, the action label is [0,0] (no movement).
- Otherwise, the image is partitioned into four quadrants (upper-right, upper-left, lower-left, lower-right), and the label of motor angle ($\mathcal{A}_{MA}$) increment vector is assigned accordingly: $(1, 1), (-1, 1), (-1, -1), (1, -1)$.

Each selected frame is saved in three variants: raw, with central cross lines, and with annotated bounding boxes. The results are represented as `[x, y, w, h]`$\mathcal{A}_{MA}$.

A.3 Automated Labeling for CC Task

The task CC involves tracking a sequence of markers arranged in a circular pattern. To determine the next movement step, we first detect all visible bounding boxes using YOLOv5 with low thresholds (`conf` $= 0.02$, `iou` $= 0.04$), extract their centroids, and fit a circle via least-squares optimization.

We then select the next target point in anti-clockwise order relative to the current center-most marker. The corresponding bounding box is retrieved and the label of motor angle ($\mathcal{A}_{MA}$) increment vector is assigned accordingly: $(1, 1), (-1, 1), (-1, -1), (1, -1)$.

A.4 Manual Curation and Finalization

Figure 6 highlights examples of labeling failures encountered during the automated annotation of the CC task. Specifically, it shows cases where the next target point along the anti-clockwise circular trajectory was incorrectly. The labeling accuracy was approximately 65% due to:

- High sensitivity to detection noise,
- Overfitting to hyperparameters such as detection thresholds,
- Errors in angle sorting during circular fitting.

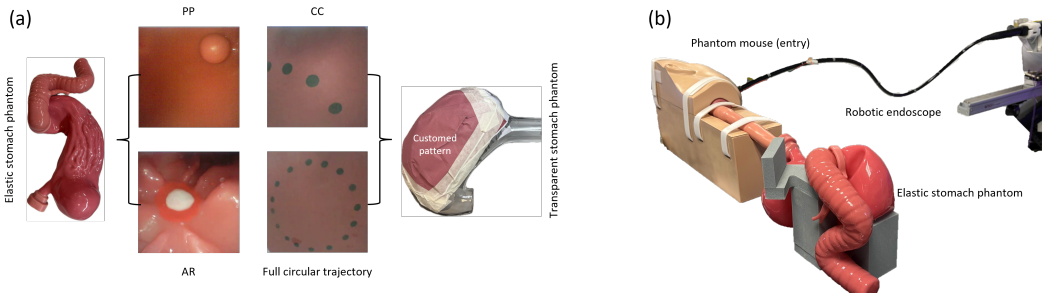

Figure 5: (a) The data is collected by two phantoms. (b) Robotic setup.

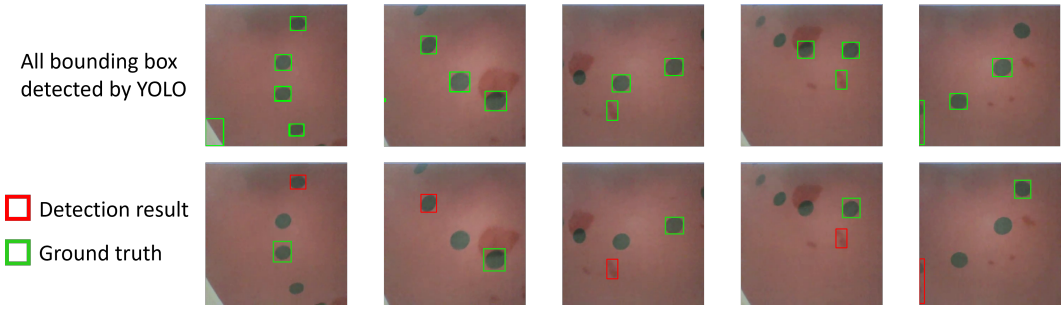

Figure 6: The examples for wrongly annotated the next tracking point in anti-clockwise direction.

To ensure data quality, we manually reviewed all annotated frames and removed erroneous samples, yielding a high-confidence vision-language-kinematics dataset. The final curated dataset comprises 6k samples.

B EndoVLA-Motion Dataset

B.1 Prompt Design for Directional Action Labeling

We designed structured instruction prompts to enable the Vision-Language model to infer directional motion commands based on visual inputs. The prompts explicitly describe the target type and provide a consistent spatial-to-action mapping, allowing the model to learn task-aware control behaviors.

Instruction Format. Each prompt is composed of:

- A visual grounding prefix (`<image>`) to indicate the presence of image input;
- A task-specific target item's name (e.g., "sphere polyp" or "abnormal region"). The target is depicted by requesting qwen2-VL with prompt–What you see in the image?;
- A spatial-to-motion mapping that converts relative target location into one of five discrete MA action vectors:
  - Upper right $\rightarrow$ [1, 1]
  - Upper left $\rightarrow$ [-1, 1]
  - Lower left $\rightarrow$ [-1, -1]
  - Lower right $\rightarrow$ [1, -1]
  - Center $\rightarrow$ [0, 0]

Instruction Examples. We used three variants of prompts for each task:

(1) Direct Action Output:

- $I_a(PP)$: "<image> I want you to act as a direction leader. Please find the target that is a sphere polyp in the image. If the target is at lower left, output [-1,-1]; upper right, [1,1]; lower right, [1,-1]; upper left, [-1,1]; at center, [0,0]. Output action only."

(2) Bounding box localization format:

- $I_b(PP)$: "<image> I want you to act as a expert for object detection and direction leader. Please find the target that is a sphere polyp in the image. To determine the bounding box of target, define its left lower start point as (x,y), the width is w, the height is h. If the target is at lower left, output [-1,-1]; upper right, [1,1]; lower right, [1,-1]; upper left, [-1,1]; at center, [0,0]. Output bounding box and action."

(3) Directional format:

- $I_d(PP)$: "<image> I want you to act as a expert for direction leader. Please find the target that is a sphere polyp in the image. If the target is at lower left, output "lower left[-1,-1]"; upper right, "upper right[1,1]"; lower right, "lower right[1,-1]"; upper left, "upper left[-1,1]"; at center, "center[0,0]". Output direction and action."

B.2 Dataset Annotation and Statistics

Table 5: **Type and number of the motion annotation**

MA	[1,1]	[-1,1]	[-1,-1]	[1,-1]	[0,0]	All
TS (PP)	373	69	270	216	125	1053
TS (AR)	67	236	493	202	132	1130
TS (CC)	747	458	804	824	N/A	2833
ES (PP)	113	21	61	48	36	279
ES (AR)	16	62	107	40	41	266
ES (CC)	191	131	201	195	N/A	718
All Number	1507	977	1936	1525	334	6279

The total number of annotation sample N_t is 6279. Robotic endoscope enable four bending motion types: upper right, upper left, lower left, and lower right. The dataset is split into 80% training set (TS) and 20% evaluation set (ES) and still when object is within the FR. Table 5 summarizes the distribution. We trained the model with two prompt type: I_a and I_b. The total number of samples is $2 \times N_t$ = 12558.

To evaluate the grounding performance of different training paradigms, we present comparisons in Figure 8. Each sub-image displays the predicted bounding box and action for a given task and training setting.

C Grounding Capability across SFT, RFT, and DFT

Compare with the training scene and evaluation scene.

The targets in our evaluation phases are not overly "too visible". Training images are bright and clear, but our real phantom experiments present significant visual challenges. As shown in Fig. 7, the system handles motion blur, dim lighting, and complex tissue-like textures in the scene. These conditions make the targets far less salient than in the training data, yet EndoVLA adapts and contin-

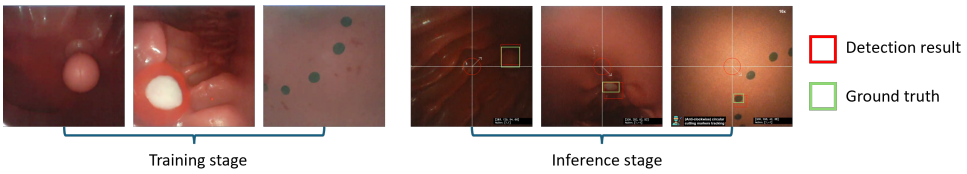

Figure 7: Comparison between training and evaluation images. Real experiments include conditions like dim lighting, motion blur and small target.

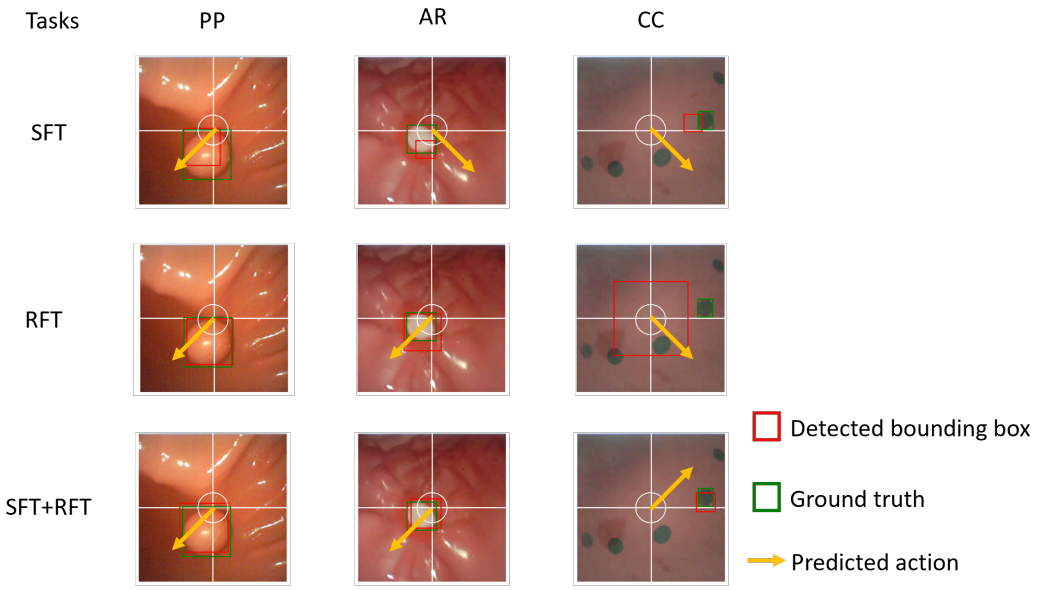

Figure 8: Qualitative results of bounding box prediction (green: GT, red: prediction) and motion output (yellow arrows) across the three tasks: PP, AR and CC. Top to bottom (fine-tuning method): SFT, RFT, SFT+RFT.

ues to track effectively. This demonstrates the system's robustness to reduced-visibility conditions in practical settings.

Observation.

- **SFT** generally predicts well-formed bounding boxes, especially in simpler visual scenes (e.g., polyp tracking). However, action prediction can be suboptimal.
- **RFT** fails to localize objects accurately as the bounding box is too large.
- **SFT+RFT** synergizes both strengths, yielding correct target localization and precise directional action across all tasks (PP, AR, and CC).