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

NeuroVLA: Surgical Scenario-Aware Learning of Debulking Skills in Endoscopic Robotic Neurosurgery via Vision-Language-Action Model

Zhiwei Fang[1*], Chi Kit Ng[1*], Huxin Gao[1], Tao Zhang[1], Zhiqing Tang[1,2], Tat Ming Danny Chan[3], Hongbin Liu[4], Renzhi Wang[5], and Hongliang Ren[1†]

***Abstract*— Robotic surgical systems have attracted widespread attention due to their accuracy and efficiency during operations. Recent studies have shown that the development of Vision-Language-Action (VLA) models offers greater potential to enable autonomous task completion in complex environments. However, the application of VLA models in surgical robotics is often limited by insufficient data on surgical environments and robot kinematics. As a result, models trained with limited data often lack a comprehensive understanding of the surgical scene and the robot's behavior. In this paper, we propose NeuroVLA, a VLA model designed for the debulking task in neurosurgical robotic scenarios. We collected a dataset using a flexible parallel continuum robot in phantom-based debulking experiments. We formulate skill objectives in the debulking task as skill instructions in NeuroVLA. We develop a Vision-Language-Model-backboned scenario understanding within NeuroVLA to help the robot understand both the surgical debulking scenario and the robot itself through skill-based instruction. After training on 90 debulking episodes, NeuroVLA can infer corresponding actions from image observations, language instructions, and robot states for the four sequential skills of the debulking task. We evaluate NeuroVLA on the four skills defined in the debulking task: align, grasp, transfer, and release. Our approach reduces pixel distance error by at least 55 % and achieves mean pixel distances of 29.10 and 21.55 pixels in align and transfer skills, respectively. The success rates for grasp and release skills are 88.89 % and 100 %, respectively.**

I. INTRODUCTION

Neurosurgery encompasses a wide range of procedures, many of which are severe and technically challenging to treat [1]. The use of surgical robots can enhance the precision and efficiency of neurosurgical procedures [2]. Price et al. developed a teleoperated neurosurgical robot with two working channels and one camera channel, and validated the system across multiple neurosurgical procedures [3]. Despite these achievements, achieving high levels of autonomy in robotic neurosurgery remains challenging due to the complexities of dynamic, deformable surgical environments [4].

A critical obstacle in achieving higher levels of autonomy lies in the robot's ability to interpret and interact with unstructured and deformable tissues [5]. Traditional model-based control methods often struggle to adapt to such uncertainties, leading to suboptimal performance in real-world scenarios. For instance, kinematic and dynamic models fail to accurately account for soft-tissue deformation and tool–tissue interaction forces, resulting in failed procedures or incomplete tasks [6].

The recent advances of Vision-Language Models (VLMs) and Vision-Language-Action (VLA) Models offer a promising pathway toward more adaptive and generalizable robotic control [7]. These models leverage large-scale pre-training on diverse multimodal data (images, text, and actions) to acquire rich representations of scenes and tasks. In surgical contexts, VLMs have shown potential in interpreting anatomical structures, instrument states, and surgical instructions, thereby facilitating higher-level decision-making [8].

Despite these advances, training VLA models for surgical robots faces unique challenges. Foundation models are typically trained on publicly available video or image datasets, which, although diverse and extensive, generally lack knowledge specific to surgical robotic operations [9]. Collecting surgical datasets is expensive and time-consuming. Gathering demonstration data for surgical robots typically requires highly proficient operators. Controlled environments and carefully designed experimental setups are necessary to collect data from surgical robots. Privacy concerns further limit access to real surgical recordings [10].

Kim et al. collected 16,000 trajectories from 34 gallbladders using a human-operated dVRK system and trained the Hierarchical Surgical Robot Transformer (SRT-H) on this dataset [11]. SRT-H achieved a 100 % success rate in eight ex vivo cholecystectomy procedures. However, the robot used by SRT-H was equipped with two wrist cameras in addition to the main endoscopic camera, providing extra views not available in conventional laparoscopic configurations. This setup is not compatible with standard laparoscopic systems, limiting the approach's transferability to other surgical robots. Long et al. achieved sim-to-real transfer for surgical task automation by applying reinforcement learning in a surgical simulator [11]. They employed a four-module frame-

This work was supported in part by the Hong Kong Innovation and Technology Fund (ITF) under Grant MHP/185/24; the National Key Research and Development Program of China under Grant 2024YFE0216200; Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) STIC Grant SGCX20250526153900001, 1+1+1 CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund – Young Scholar Projects (Ref: YSP10), ITF MHP/185/24, Ministry of Science and Technology (MOST) of China Key Project 2025YFE0122500, Hong Kong Research Grants Council (RGC) Collaborative Research Fund (CRF C4026-21GF), General Research Fund (GRF 14216022, 14204524, 14203323, 14206125), InnoHK, 2022 Shenzhen Higher Education Institutions Stable Support Program (Grant No. 2023SC0073), Hainan Province Clinical Medical Center.

* These authors contribute equally to this work.

[1]Department of Electronic Engineering, the Chinese University of Hong Kong (CUHK), Hong Kong, China. [2]Shenzhen Loop Area Institute, Shenzhen, China. [3]Department of Surgery, the Chinese University of Hong Kong (CUHK), Hong Kong, China. [4]The Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong, China. [5]The School of Medicine, the Chinese University of Hong Kong-Shenzhen, Shenzhen, China

* Corresponding author: Hongliang Ren: hlren@ieee.org.

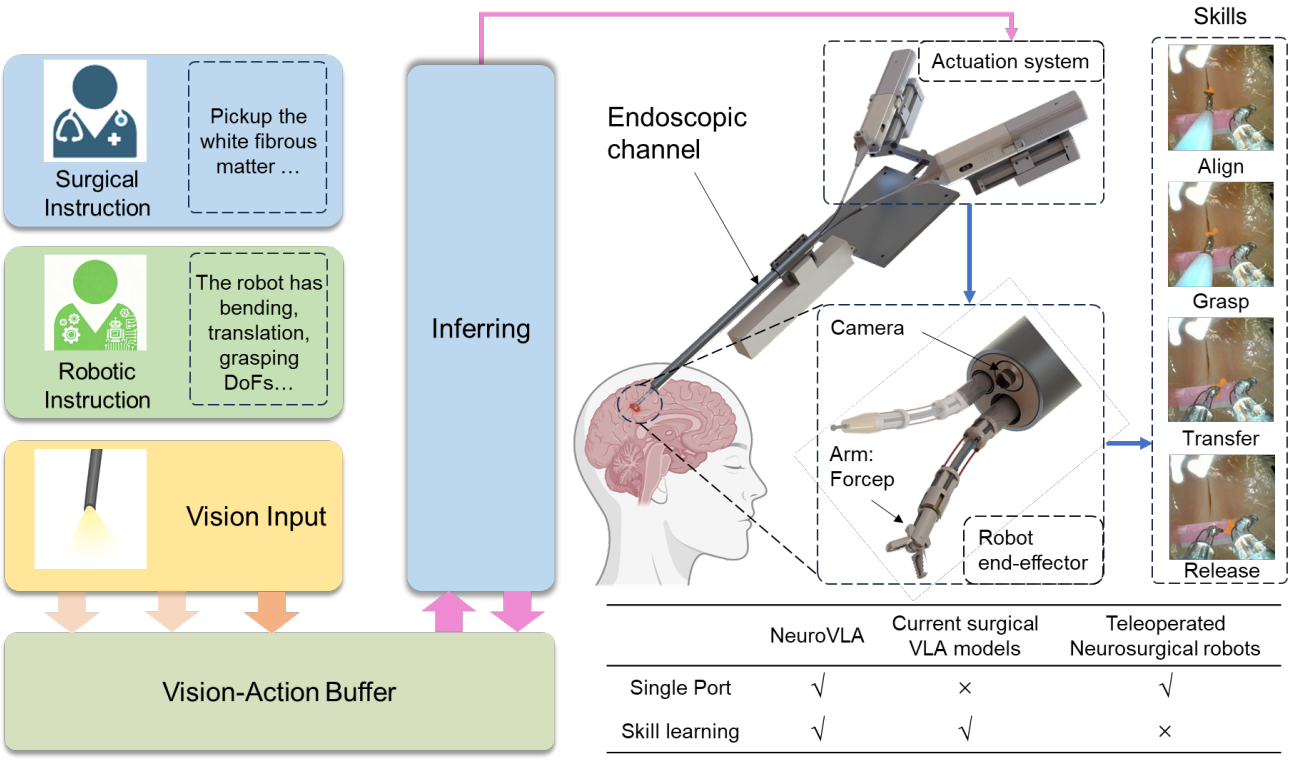

Fig. 1. The proposed NeuroVLA system framework. NeuroVLA takes instruction, vision, action, and robot state as model input and guides an endoscopic neurosurgical robot to perform skills in the debulking task.

work for visual parsing, perception, reinforcement learning, and visual servoing. The visual parsing module produces segmentation and depth estimation based on a vision foundation model. The capability for autonomous laparoscopic task execution was validated in both in–vivo and ex–vivo experiments. Although coordination among modules enables zero-shot transfer, inter-module communication increases the system's internal computational cost. In addition, the control performance of position-based visual servoing in dynamic scenarios remains to be investigated.

On the other hand, these methods have been demonstrated on tasks executed by multi-arm surgical robots in regions such as the gastrointestinal tract. However, to minimize invasiveness in neurosurgery, robots are commonly introduced via a single-port endoscopic approach. Compared with multi-port robotic procedures, the single-port approach reduces surgical injury but constrains the spatial deployment of surgical instruments and the camera [12]. The spacing between instrument channels and the camera channel typically does not exceed 5 *mm*, and the endoscope diameter is generally no greater than 10 *mm* [13]. This confined geometry results in fewer visual features in the camera field of view and makes it less intuitive to observe the manipulator's motion through the camera [14]. Therefore, research on automation in the field of neurosurgical robotics remains limited.

To address these challenges, we propose NeuroVLA, a first VLA model designed for the debulking task in neurosurgical procedures. Debulking refers to the removal of fibrous tumor tissue in neurosurgery. Typically, forceps grasp the tumor tissue and an aspiration tube removes it from the body. In this study, cotton pieces are used to simulate fibrous tumors, and a brain phantom serves as the experimental testbed. Fig. 1 illustrates the overall workflow of the proposed VLA and robotic system. Surgical and robotic instructions provide the model with the task specification and an understanding of the robotic system. The model achieves history-informed scene understanding from visual input and action-state information. We define four sequential skills for the debulking task and formulate skill objectives that are integrated into model inference. During each skill, to improve the understanding of the scenario in the debulking skills, the relevant representation texts of the skill - as an instruction are used to guide the model's action predictions. We employ a surgical robot based on a parallel continuum mechanism [15] with two working channels and one camera channel. In this study, we use a forceps tool in one working channel to collect 90 debulking episodes on a brain phantom. Across nine validation experiment groups, NeuroVLA achieves mean distances of 29.10 and 21.55 pixels for the align and transfer skills, respectively, and success rates of 88.89% and 100% for the grasp and release skills, respectively. The core contributions of this work include the following:

- An end-to-end scenario-aware VLA model that controls a parallel continuum surgical robot to perform debulking skills in robotic neurosurgery. A scenario understanding context as instruction provides scenario-aware represen-

tations that guide the model's action outputs.
- A VLA dataset collected for debulking tasks on a brain phantom. The dataset is suitable for training and evaluating models for debulking and similar continuum-robot tasks
- Quantitative Experiments demonstrate that NeuroVLA achieves mean pixel distances of 29.10 and 21.55 pixels in align and transfer skills, respectively. The success rates for grasp and release skills are 88.89 % and 100 %, respectively.

II. Related Works

Autonomous control of surgical robot. Autonomous control of surgical robots has progressed rapidly, advancing from basic task automation to performing complete surgical steps with minimal human input [16], [17], [18]. A landmark study by Shademan et al. demonstrated supervised autonomy for soft tissue surgery to perform automated suturing, achieving outcomes superior to expert surgeons in preclinical trials [19]. Zhong et al. presented an integrated planning and control framework by exploring the geometrical characteristics of surgical robots and validated task-level consistency and reliability in surgical automation [20]. Shin et al. implemented a vision-guided, learning-based model predictive controller on a surgical robot for autonomous soft-tissue manipulation, foreshadowing a paradigm in which the surgeon mainly provides high-level decisions [21].

Vision-language-action Model. Recent years have seen rapid progress in applying end-to-end VLA models for embodied control. Google DeepMind's RT-2 [22] pioneered the paradigm of pretraining vision–language models for perception and understanding, followed by fine-tuning for robot control. OpenVLA [7] demonstrated strong performance with a relatively compact 7B backbone by coupling DINOv2/SigLIP visual encoders with LLAMA-2, and demonstrated that parameter-efficient fine-tuning enables fast task adaptation with lower computation resources. More recently, models such as π_0 [23] and G0 [24] have introduced new training and architectural strategies to further boost VLA capability. π_0 combines a SigLIP image encoder and a Gemma backbone and employs a diffusion-based flow matching mechanism so an action expert can output high-frequency action flows for dexterous manipulation. G0 presents comprehensive manipulation and navigation experiments with a dual-system design: a three-stage G0-VLA for instruction following and action execution, and a G0-VLM (finetuned from Qwen2.5-VL on Galaxea) that parses natural-language goals and scene context to propose subtask instructions. Despite impressive results of VLA models in household and industrial settings, these systems tend to underperform on endoscopic surgical robots due to limited exposure to surgical scenes and robot-specific control priors. Translating VLA to clinical workflows requires models that can parse anatomy, instruments, and safety constraints while issuing precise, low-latency actions. RoboNurse-VLA features language-conditioned, vision-guided assistance, locating and handing the correct instrument, and respecting sterile zones and no-fly regions [25]. EndoVLA focuses on autonomous, prompt-conditioned autonomous tracking in endoscopy [26]. VLA models in medical applications suggest the tight coupling between scene understanding and actuation, focusing on scenario-aware learning skills.

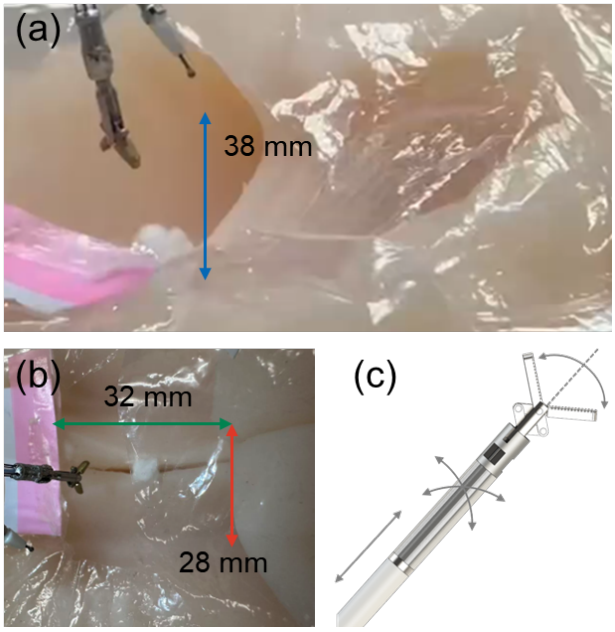

Fig. 2. Neurosurgical robot setup for the debulking task. (a) Lateral view of the experimental setup: the robot workspace has a vertical extent of approximately 38 *mm*. (b) Top view of the experimental setup: the robot workspace covers a horizontal area of approximately 28 *mm* × 32 *mm*. (c) The schematic of the robot end-effector. The forceps has two bending DoFs, one translation DoF, and one clamp DoF.

III. Neurosurgical Robot System

To accomplish the debulking task, we developed a neurosurgical robotic system based on the flexible parallel continuum robotic mechanism and equipped with a micro-RGB camera. The system provides two operational channels capable of accommodating instruments such as forceps and scissors. Fig. 2 illustrates the debulking experiment phantom and the schematic of the forceps end-effector. The debulking task is executed using four degrees-of-freedom (DoF) of a single forceps arm (pitch, yaw, translation, and clamp) as depicted in Fig. 2 (c). The end-effector is remotely connected to and actuated by servo motors (Maxon RE13 395114). The integrated camera captures images at up to 30 frames per second (FPS) with a resolution of 400 × 400 pixels and communicates with the control host via USB. All motors are controlled by a motion controller (Galil DMC 4183, Galil), which provides encoder values for motion state recording. Teleoperation and data collection were performed using a joystick.

To simulate the debulking task, we use the ventricle region of a brain phantom as the surgical area and represent fibrous tumor tissue with cotton pieces. The spatial dimensions of this setup are shown in Fig. 2 (a) and (b). The horizontal workspace of the robot in an area measures 32 *mm* in

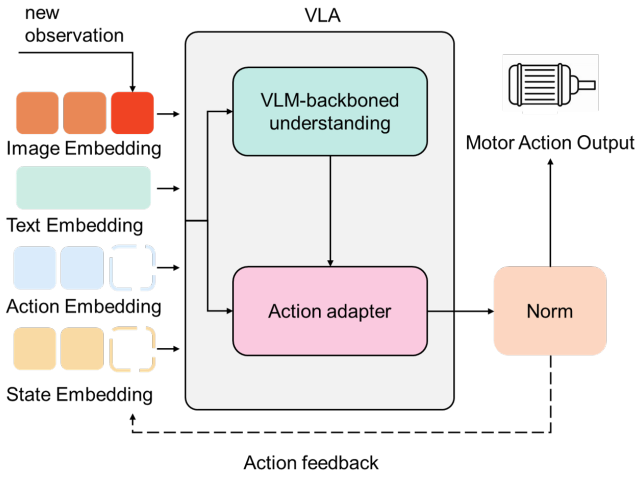

Fig. 3. Inference framework of the NeuroVLA model.

length and 28 *mm* in width. The maximum vertical distance between the phantom area and the initial forceps position is 38 *mm*. Cotton pieces are placed in one of three possible regions: left, center, or right. A transparent plastic film with an opening serves as the entry for the robot. In the actual debulking experiment, tumor tissue is placed adjacent to the aspiration tube. A red plate was used to represent the target location; successful placement of a cotton piece onto the red plate validates the robot's ability to transfer the object to a prescribed location and to enable the aspiration tube to perform suction.

We collected motor motion data and image data at 20 Hz, together with initial textual prompts, to construct the AutoDebulk dataset. The AutoDebulk dataset comprises a total of 90 debulking episodes across three phantom regions, the original data duration exceeds 6 hours. After data refinement (removal of empty motion data), the refined dataset contains 290,000 image-motion pairs. In each episode, the robot started from the initial position, reached the target to perform a grasping action, transferred the cotton piece to the target place, released the forceps, and returned to the initial position. Action–vision pairs were temporally aligned during collection using thread synchronization, ensuring a strict interval of 50 *ms*. To accomplish the debulking task, the operator observes the position of the robot and the cotton piece via the endoscopic camera image displayed on the computer screen. During the episode recording, camera images are stored locally in real time with timestamps. While the image-saving thread is running, the computer reads the real-time motor encoder values from the motor control board and stores them in temporary memory. After each episode ends, the operational data in memory is then temporally aligned with the image data and exported as the RLDS dataset [27].

IV. Method

A. Problem Formation

We consider sequential visuomotor control for neurosurgical debulking as conditional sequence prediction. Each episode is represented as

$$E = \{(o_t, s_t, a_t, I, \mathbf{y}_t)\}_{t=1}^{T},$$

where $o_t \in \mathbb{R}^{H\times W\times 3}$ is the RGB observation, $s_t \in \mathbb{R}^4$ is the robot state, $a_t \in \mathbb{R}^4$ is the action, $I = (I_{\text{surg}}, I_{\text{robo}})$ denotes the surgical and robotic instructions, and $\mathbf{y}_t$ is the discretized action-text sequence. When a phase is completed, the target sequence terminates with the literal token `Finished`.

For notation simplicity, we denote

$$s_t = [p_t, y_t, r_t, c_t], \quad a_t = [\Delta p_t, \Delta y_t, \Delta r_t, \Delta c_t],$$

corresponding to pitch, yaw, translation, and clamp, respectively.

An episode starts with an surgical language instruction I_{surg}, an robotic language instruction I_{robo}, and an initial image observation o_*. The surgical language instruction describes the experiment scenario and the goals of the debulking task. The robotic language instruction describes the robot configuration and the type of end-effector in the robot arm. Combined with the initial image observation from the camera, the agent is capable of relating the image observation and the surgical robotic task, thus enhancing the task accomplishment.

B. Assumptions

a) First-order integrator: We assume a simple additive transition in state space:

$$s_{t+1} \approx s_t + a_t,$$

which is reasonable for the low-level, small-step joint deltas with high-accuracy servo motors.

b) Short-horizon temporal grounding: Two past frames and explicit numeric history $(S_{t-2:t-1}, A_{t-2:t-1})$ suffice to resolve local temporal ambiguities during fine dexterous motions.

c) Scenario-aware skill decomposition: The debulking task is decomposed into four skill phases (*Align*, *Grasp*, *Transfer*, *Release*); a *scenario understanding instruction context* conditions the action decoder, enabling skill-specific patterns.

C. NeuroVLA Policy and Inference

Let f_θ be a multimodal VLA model with image, text, state, and action embeddings. At step t, the model consumes

$$X_t = \big(\text{Enc}_{\text{img}}(O_{t-2:t}),\ \text{Enc}_{\text{text}}(I),\ \text{Enc}_{\text{num}}(S_{t-2:t-1}, A_{t-2:t-1})\big).$$

The *scenario understanding stage* receives the current skill token and produces a short scene description l_t. After scenario understanding, the NeuroVLA f_θ conditioned on (X_t, l_t), autoregressively generated a token sequence $\hat{y}_t$ representing $\hat{a}_t$ (discretized) and may emit `Finished` to terminate:

$$\hat{y}_t = \arg\max_{y} \prod_{i} f_\theta(y_{t,i} \mid y_{t,<i}, X_t, l_t).$$

We then de-tokenize $\hat{y}_t$ to obtain $\hat{a}_t \in \mathbb{R}^4$ (0.1-resolution). The decoder both *acts* and *decides when to stop* by emitting `Finished`.

We operationalize each phase with explicit inputs, objectives, and termination rules to the next skill. Let $\mathbf{p}_t \in \mathbb{R}^2$ be the image-plane pixel of the forceps tip detected at time t [28], g_t be the visual proxy of the forceps opening state, and $\mathbf{c}_t \in \mathbb{R}^2$ be the centroid of the cotton piece mask. Let $R_{\text{plate}} \subset \mathbb{R}^2$ denote the rectangular region (red plate) and define the point-to-rectangle distance as:

$$\text{dist}(\mathbf{p}_t, R_{\text{plate}}) = \begin{cases} 0, & \mathbf{p}_t \in R_{\text{plate}}, \\ \min_{\mathbf{r} \in \partial R_{\text{plate}}} \|\mathbf{p}_t - \mathbf{r}\|_2, & \text{otherwise.} \end{cases}$$

The scenario understanding stage is implemented as a deterministic phase-to-text mapping. Given the current phase token $k_t \in \{\texttt{Align}, \texttt{Grasp}, \texttt{Transfer}, \texttt{Release}\}$, a short phase-specific text description l_t is generated from a fixed template library. This text is concatenated with the instruction context and used to condition the policy decoder.

- **Align (tip-to-target approach and alignment).** *Inputs:* o_t, s_t, forceps tip pixel $\mathbf{p}_t$, cotton piece centroid $\mathbf{c}_t$.
  Objective: Align the forceps tip with the cotton piece with contact.
  Progress signal: pixel distance
  $$\text{PD}_t^{\text{align}} = \|\mathbf{p}_t - \mathbf{c}_t\|_2.$$
  Termination (Both):
  1) *Proximity*: $\text{PD}_t^{\text{align}} \le \varepsilon_{\text{align}}$.
  2) *Contact cue*: an action following the tip's approach causes centroid changes $\|\mathbf{c}_t - \mathbf{c}_{t-1}\|_2 \ge \varepsilon_{\text{move}}$.

  Guards: keep w_t unchanged, bound action displacement $\|\Delta \mathbf{a}_t\|_2 \le \alpha_{\max}$.
  Hysteresis: require the termination condition to hold for K_{align} consecutive frames to avoid flicker.
- **Grasp (pose hold and full closure of the forceps).** *Inputs:* o_t, s_t, forceps tip pixel $\mathbf{p}_t$, cotton piece centroid $\mathbf{c}_t$, forceps visual proxy g_t.
  Objective: maintain the aligned pose and fully close the forceps.
  Termination (both):
  1) *Pose hold*: $\text{PD}_t^{\text{align}} \le \varepsilon_{\text{hold}}$.
  2) *Full closure*: $g_t \le \varepsilon_{\text{gap}}$.

  Guards: limit Δz_t (avoid pushing the target off-plane during closure).
  Hysteresis: enforce closure for K_{grasp} frames to confirm a secure grasp.
- **Transfer (retrieve and place over the red plate).** *Inputs:* o_t, s_t, forceps tip pixel $\mathbf{p}_t$, plate region R_{plate}.
  Objective: carry the grasped cotton piece to the plate area at a near-initial height.
  Progress signals:
  $$\text{PD}_t^{\text{plate}} = \text{dist}(\mathbf{p}_t, R_{\text{plate}}), \qquad h_t := z_t \quad \text{(proxy for height)}.$$
  Termination (both):
  1) *Planar proximity*: $\text{PD}_t^{\text{plate}} \le \varepsilon_{\text{plate}}$.
  2) *Height window*: $\|h_t - h_{\text{ref}}\|_2 \le \delta_h$.
- **Release (pose hold and full opening over the plate).** *Inputs:* o_t, s_t, forceps tip pixel $\mathbf{p}_t$, cotton piece centroid $\mathbf{c}_t$, plate region R_{plate}, forceps visual proxy g_t.
  Objective: release the cotton piece above the plate.
  Termination (any of):
  1) *Full opening*: $g_t \ge \varepsilon_{\text{open}}$ while $\mathbf{p}_t$ is over R_{plate}.
  2) *Successful drop*: the cotton piece lands with $\mathbf{c}_t$ within R_{plate}.

  Guards: maintain tip pose ($\text{PD}_t^{\text{plate}} \le \varepsilon_{\text{hold}}$) during opening; bound Δw_t to prevent overshoot.
  Hysteresis: validate landing over K_{rel} frames to avoid misclassifications.

Phase transitions and supervision. During training, the final frame satisfying each phase's termination condition is labeled with the `Finished` token for that phase. At inference time, the decoder emits `Finished` to advance to the next phase. Thresholds $(\varepsilon_{\text{align}}, \varepsilon_{\text{move}}, \varepsilon_{\text{hold}}, \varepsilon_{\text{gap}}, \varepsilon_{\text{open}}, \varepsilon_{\text{plate}}, \delta_h)$ are tuned on validation data and kept fixed during testing. This formulation ensures each skill has textual objectives, explicit success criteria, and runtime termination logic.

D. Training Objective

We train the model with standard autoregressive supervised fine-tuning over the discretized action-text sequence. Let $\mathbf{y}_t = \{y_{t,1}, \ldots, y_{t,N_t}\}$ denote the ground-truth token sequence representing the action at time step t. Conditioned on the multimodal context $(O_{t-2:t}, S_{t-2:t-1}, A_{t-2:t-1}, I)$, the objective is

$$\mathscr{L} = -\sum_{i=1}^{N_t} \log p_\theta(y_{t,i} \mid y_{t,<i}, O_{t-2:t}, S_{t-2:t-1}, A_{t-2:t-1}, I).$$

When the demonstrated action corresponds to phase completion, the target sequence includes the literal token `Finished`, so completion behavior is optimized under the same autoregressive objective.

V. Experiment

A. Experiment Setup

We randomly placed each cotton piece in one of three regions within the brain phantom to evaluate NeuroVLA. Aligned with the skill objectives, the debulking task is divided into four phases: (1) Align: the instrument tip approaches the cotton piece from the initial position until clear contact is made; (2) Grasp: after the forceps tip contacts the cotton piece, the forceps is fully closed to grasp the cotton piece; (3) Transfer: after grasping, the robot retrieves and moves the cotton piece above the red platform; (4) Release: once the cotton piece is above the red platform, the forceps is opened to release it.

B. Training and Inference Implementation

The NeuroVLA model was finetuned based on the EndoVLA model [26]. Training was performed on a multi-GPU type consisting of one NVIDIA RTX A6000 (48 GB) and three NVIDIA RTX 3090 (24 GB) GPUs. The learning rate

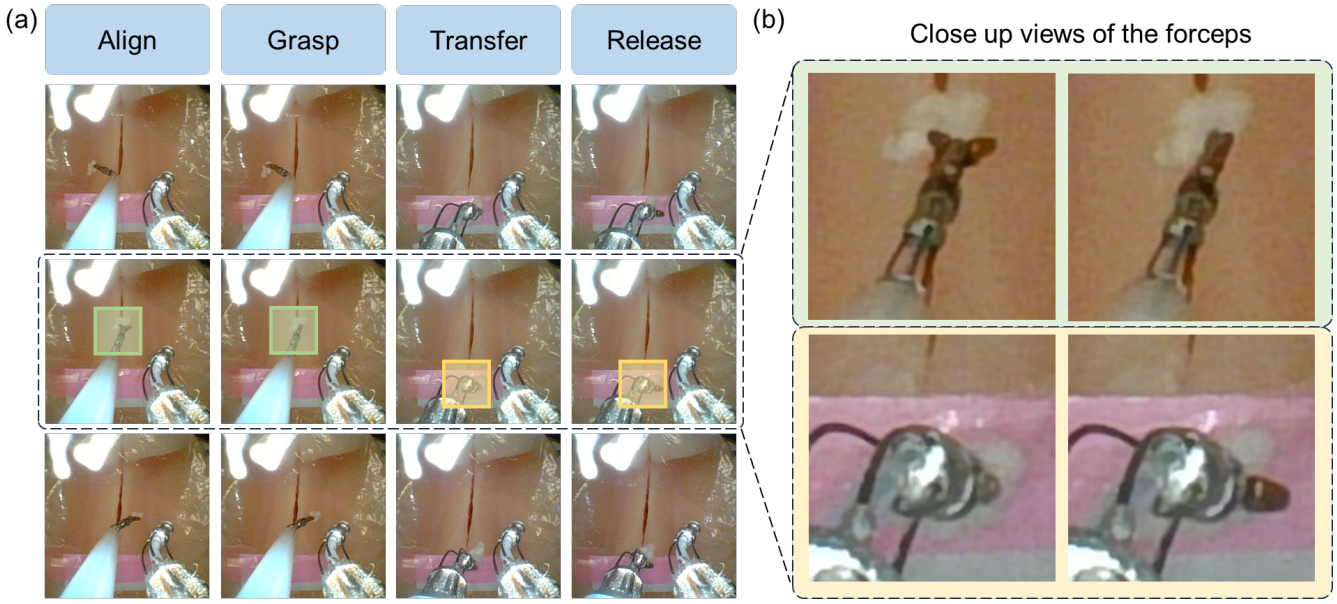

Fig. 4. Experiment results of successful skill finishes in three episodes. (a) The sequential endoscopic views of the robot performing four skills in the debulking tasks. (b) Close-up views of the forceps correspond to the areas marked with green and yellow rectangles in the images of the second row in (a). The first row of the close-up views shows the forceps details at the grasping position. The second row of the close-up views shows the forceps details at the release position.

was set to 0.0001 and optimized using a cosine learning rate scheduler with a warmup ratio of 0.1. Training lasted approximately 20 hours with a batch size of 4. Inference was performed on a single NVIDIA RTX 3090 (24 GB) GPU.

C. Experiment Results

A total of nine experiments were conducted across three phantom regions (left, middle, right), with three trials per region. In each experiment, the robot executed four skills: Align, Grasp, Transfer, and Release. Fig. 4 shows endoscopic images of experiments conducted in three different areas. The first row of the images shows experiments conducted in the left region of the phantom, the second row shows experiments in the central region, and the third row shows experiments in the right region. The Align column shows the moment when the forceps was aligned with the cotton piece after approaching from the initial state, and ready to close the forceps. The Grasp column shows the moment when the model indicated that the forceps was closed and was ready to transfer. The Transfer column shows the moment when the robot transferred the cotton piece above the red plate after retrieving from the grasping position and was ready to release. The Release column shows the moment when the model indicated that the forceps was fully opened.

We selected the advanced robotic VLA model Octo-1.5b [29] and the multimodal model QwenVL2.5-7b [30] and ran the same number of experiments under identical conditions. As shown in Table I, to quantitatively evaluate the performance of different models on the four skills, we propose two distinct evaluation metrics. For the align and transfer skills, we define the pixel distance (PD) as explained in Section. IV. In the align skill, the target location is the center of the cotton piece. To accurately determine the center of the cotton piece, we use the SAM2 model [31] on post-experiment images to obtain the ground truth mask of the cotton piece, and then calculate the centroid of the mask as the target position. In the transfer skill, the target location is the red platform area; thus, we define the PD for this skill as the minimum distance between the center of the forceps tip and the rectangular region representing the red platform.

TABLE I

QUANTITATIVE EXPERIMENT RESULTS COMPARISON WITH BASELINE MODELS.

Model	**Skill Performance**			
	Align PD ↓ (px)	**Grasp SR** ↑ (%)	**Transfer PD** ↓ (px)	**Release SR** ↑ (%)
Baseline Models				
Octo	79.72	33.30	65.46	66.67
Qwen2.5-VL	44.61	55.56	49.45	66.67
Ours	**29.10**	**88.89**	**21.55**	**100.00**

For the grasp and release skills, we use the success rate (SR) of closing/opening of the forceps as the evaluation metric. Note that, unlike normal grippers mounted at the end of general robotic arms, the kinematics of the forceps of a parallel continuum robot is affected by the bending of the instrument tip, and does not provide contact force feedback (or requires additional sensors to do so). Therefore, the state cannot be simply defined as binary (0 or 1). To achieve more precise control of the forceps, we use the opening and closing state of the forceps in the images to drive the forceps control.

This makes the closing and opening of the forceps a task of controlling a linear joint movement. For the grasp skill, we define a successful grasp as successfully grasping the cotton piece and not dropping during subsequent movements. For the Release skill, we define a successful release as the forceps being fully opened or the cotton piece dropping as a result of opening the forceps.

As shown in Table I, NeuroVLA demonstrates significant improvements in both the align and transfer skills, achieving pixel distances of 29.10 and 21.55 pixels for the Align and Transfer skills, respectively. For the grasp and release skills, NeuroVLA also attains higher success rates (SR) compared to other models. Although grasp and release are relatively simple one DoF skills, Qwen2.5-VL and Octo display poorer scene understanding and task completion. In particular, these baseline models often judge the grasp as complete before the forceps are fully closed. In contrast, NeuroVLA better preserves grasp robustness and can successfully initiate grasps from more distal forceps starting positions.

D. Ablation Study

TABLE II
ABLATION STUDY ON SCENARIO-UNDERSTANDING CONTEXT INSTRUCTION AND FINE-TUNING.

Model	**Skill Performance**			
	Align PD ↓ (px)	**Grasp SR ↑** (%)	**Transfer PD ↓** (px)	**Release SR ↑** (%)
Ablation Variants				
W/O scenario understanding context	66.22	55.56	59.81	55.56
W/O fine-tuning	84.62	11.11	79.45	22.22
Ours	**29.10**	**88.89**	**21.55**	**100.00**

The scenario understanding context interprets the surgical scenario as skill-based instructions and guides the action generation. To validate its effectiveness, we compared three variants: the full NeuroVLA, NeuroVLA without the scenario understanding context instruction, and NeuroVLA with the scenario understanding context instruction but without AutoDebulk dataset fine-tuning, on execution of the four skills. As shown in Table II, the NeuroVLA model maintained the best performance, indicating that exclusion of the scenario understanding context reduces the model's ability to recognize surgical objects and spatial relations required for accurate control. The unfinetuned variant also produced lower success rates for grasp and release compared with the fully finetuned model. Notably, even when equipped with the scenario understanding context, the model lacking AutoDebulk finetuning underperformed the finetuned variant without the scenario underderstanding across all skills, suggesting that domain-specific, robot-task data are critical for correct decision-making in surgical control.

VI. CONCLUSION

This work collected unique data from debulking operations performed on a brain phantom and used this data to train a VLA model capable of sequentially completing the four skills required for the debulking task. Experiments conducted in the phantom demonstrated that the proposed NeuroVLA model with the scenario understanding context exhibits superior understanding of surgical scenes and can sequentially guide robotic skills. On this basis, we plan to incorporate additional scene and robot labels to further enhance the robot's ability to accomplish tasks in more realistic and dynamic environments. Furthermore, by acquiring more bimanual operation data, we aim to leverage the capabilities of dual-arm robots to accomplish a wider range of surgical tasks.

REFERENCES

[1] W. A. Awuah, F. T. Adebusoye, J. Wellington, L. David, A. Salam, A. L. W. Yee, E. Lansiaux, R. Yarlagadda, T. Garg, T. Abdul-Rahman, *et al.*, "Recent outcomes and challenges of artificial intelligence, machine learning, and deep learning in neurosurgery," *World neurosurgery: X*, vol. 23, p. 100301, 2024.

[2] N. A. Shlobin, J. Huang, and C. Wu, "Learning curves in robotic neurosurgery: a systematic review," *Neurosurgical review*, vol. 46, no. 1, p. 14, 2022.

[3] K. Price, J. Peine, M. Mencattelli, Y. Chitalia, D. Pu, T. Looi, S. Stone, J. Drake, and P. E. Dupont, "Using robotics to move a neurosurgeon's hands to the tip of their endoscope," *Science Robotics*, vol. 8, p. eadg6042, Sept. 2023.

[4] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, V. J. Santos, and R. H. Taylor, "Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, p. eaam8638, Mar. 2017.

[5] J. C. Norton, P. R. Slawinski, H. S. Lay, J. W. Martin, B. F. Cox, G. Cummins, M. P. Desmulliez, R. E. Clutton, K. L. Obstein, S. Cochran, and P. Valdastri, "Intelligent magnetic manipulation for gastrointestinal ultrasound," *Science Robotics*, vol. 4, p. eaav7725, June 2019.

[6] J. Zhu, L. Lyu, Y. Xu, H. Liang, X. Zhang, H. Ding, and Z. Wu, "Intelligent soft surgical robots for next-generation minimally invasive surgery," *Advanced Intelligent Systems*, vol. 3, no. 5, p. 2100011, 2021.

[7] 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.

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

[9] M. Yip, "The robot will see you now: Foundation models are the path forward for autonomous robotic surgery," *Science Robotics*, vol. 10, p. eadt0684, July 2025.

[10] T. Haidegger, S. Speidel, D. Stoyanov, and R. M. Satava, "Robot-assisted minimally invasive surgery—surgical robotics in the data age," *Proceedings of the IEEE*, vol. 110, no. 7, pp. 835–846, 2022.

[11] J. W. B. Kim, J.-T. Chen, P. Hansen, L. X. Shi, A. Goldenberg, S. Schmidgall, P. M. Scheikl, A. Deguet, B. M. White, D. R. Tsai, R. J. Cha, J. Jopling, C. Finn, and A. Krieger, "SRT-H: A hierarchical framework for autonomous surgery via language-conditioned imitation learning," *Science Robotics*, vol. 10, p. eadt5254, July 2025.

[12] F. Celotto, N. Ramacciotti, A. Mangano, G. Danieli, F. Pinto, P. Lopez, A. Ducas, J. Cassiani, L. Morelli, G. Spolverato, *et al.*, "Da vinci single-port robotic system current application and future perspective in general surgery: a scoping review," *Surgical Endoscopy*, vol. 38, no. 9, pp. 4814–4830, 2024.

[13] R. Martinez-Perez, L. C. Requena, R. L. Carrau, and D. M. Prevedello, "Modern endoscopic skull base neurosurgery," *Journal of neuro-oncology*, vol. 151, no. 3, pp. 461–475, 2021.

[14] Y. Shi, J. Li, D. Song, B. Zhang, Z. Zhang, T. Fukuda, and C. Shi, "Advances in transanal quasi-single-port surgery robotic systems: A comprehensive review," *IEEE/ASME Transactions on Mechatronics*, 2025.

[15] H. Gao, X. Yang, X. Xiao, X. Zhu, T. Zhang, C. Hou, H. Liu, M. Q.-H. Meng, L. Sun, X. Zuo, Y. Li, and H. Ren, "Transendoscopic flexible parallel continuum robotic mechanism for bimanual endoscopic submucosal dissection," *The International Journal of Robotics Research*, Nov. 2023.

[16] A. Lee, T. S. Baker, J. B. Bederson, and B. I. Rapoport, "Levels of autonomy in fda-cleared surgical robots: a systematic review," *NPJ Digital Medicine*, vol. 7, no. 1, p. 103, 2024.

[17] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 651–679, 2021.

[18] N. Simaan, R. M. Yasin, and L. Wang, "Medical technologies and challenges of robot-assisted minimally invasive intervention and diagnostics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, 2018.

[19] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Science translational medicine*, vol. 8, no. 337, pp. 337ra64–337ra64, 2016.

[20] F. Zhong and Y.-H. Liu, "Integrated planning and control of robotic surgical instruments for task autonomy," *The International Journal of Robotics Research*, vol. 42, pp. 504–536, June 2023.

[21] C. Shin, P. W. Ferguson, S. A. Pedram, J. Ma, E. P. Dutson, and J. Rosen, "Autonomous tissue manipulation via surgical robot using learning based model predictive control," in *2019 International conference on robotics and automation (ICRA)*, pp. 3875–3881, IEEE, 2019.

[22] 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*, pp. 2165–2183, PMLR, 2023.

[23] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "\pi_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[24] G. Team, "Galaxea g0: Open-world dataset and dual-system vla model," *arXiv preprint arXiv:XXXX.XXXXX*, 2025.

[25] 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.

[26] N. C. KIT, L. Bai, G. Wang, Y. Wang, H. Gao, K. yuan, C. Jin, T. Zeng, and H. Ren, "EndoVLA: Dual-phase vision-language-action for precise autonomous tracking in endoscopy," in *9th Annual Conference on Robot Learning*, 2025.

[27] S. Ramos, S. Girgin, L. Hussenot, D. Vincent, H. Yakubovich, D. Toyama, A. Gergely, P. Stanczyk, R. Marinier, J. Harmsen, *et al.*, "Rlds: an ecosystem to generate, share and use datasets in reinforcement learning," *arXiv preprint arXiv:2111.02767*, 2021.

[28] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 691–699, IEEE, 2018.

[29] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[30] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[31] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.