

# ArthroCut: Autonomous Policy Learning for Robotic Bone Resection in Knee Arthroplasty

Xu Lu<sup>†</sup>, Yiling Zhang<sup>†</sup>, Wenquan Cheng, Longfei Ma, Fang Chen and Hongen Liao\*

**Abstract**—Despite rapid commercialization of surgical robots, their autonomy and real-time decision-making remain limited in practice. To address this gap, we propose *ArthroCut*, an autonomous policy learning framework that upgrades knee arthroplasty robots from assistive execution to context-aware action generation. *ArthroCut* fine-tunes a Qwen-VL backbone on a self-built, time-synchronized multimodal dataset from 21 complete cases (23,205 RGB-D pairs), integrating preoperative CT/MR, intraoperative NDI tracking of bones and end effector, RGB-D surgical video, robot state, and textual intent. The method operates on two complementary token families—Preoperative Imaging Tokens (PIT) to encode patient-specific anatomy and planned resection planes, and Time-Aligned Surgical Tokens (TAST) to fuse real-time visual, geometric, and kinematic evidence—and emits an interpretable action grammar under grammar/safety-constrained decoding. In bench-top experiments on a knee prosthesis across seven trials, *ArthroCut* achieves an average success rate of 86% over the six standard resections, significantly outperforming strong baselines trained under the same protocol. Ablations show that TAST is the principal driver of reliability while PIT provides essential anatomical grounding, and their combination yields the most stable multi-plane execution. These results indicate that aligning preoperative geometry with time-aligned intraoperative perception and translating that alignment into tokenized, constrained actions is an effective path toward robust, interpretable autonomy in orthopedic robotic surgery.

## I. INTRODUCTION

Surgical robots have transformed many clinical domains by improving precision, reducing variability, and enabling minimally invasive procedures. In orthopedics, robotic systems have become particularly prominent in joint replacement, where millimeter-scale accuracy is critical for implant positioning and long-term outcomes [2], [3]. Robotic platforms are now widely adopted in total and partial knee arthroplasty, demonstrating that computer-assisted execution can reduce outliers and improve implant alignment compared to manual techniques [4].

Despite these advances, current knee surgery robots are predominantly assistive rather than autonomous [5]. They

This work is supported in part by the National Natural Science Foundation of China (U22A2051, 82027807), National Key Research and Development Program of China (2022YFC2405200), the Natural Science Foundation of Beijing Municipality (L252126, L252201), and the Science and Technology Commission of Shanghai Municipality (24511104100).

Xu Lu, Yiling Zhang, Wenquan Cheng, Longfei Ma and Hongen Liao are with the School of Biomedical Engineering, Tsinghua University, Beijing 100084, China.

Fang Chen and Hongen Liao are with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Yiling Zhang is also with the Longwood Valley MedTech, Beijing 100176, China.

<sup>†</sup> Denotes equal contribution

Corresponding author: Hongen Liao, liao@tsinghua.edu.cn.

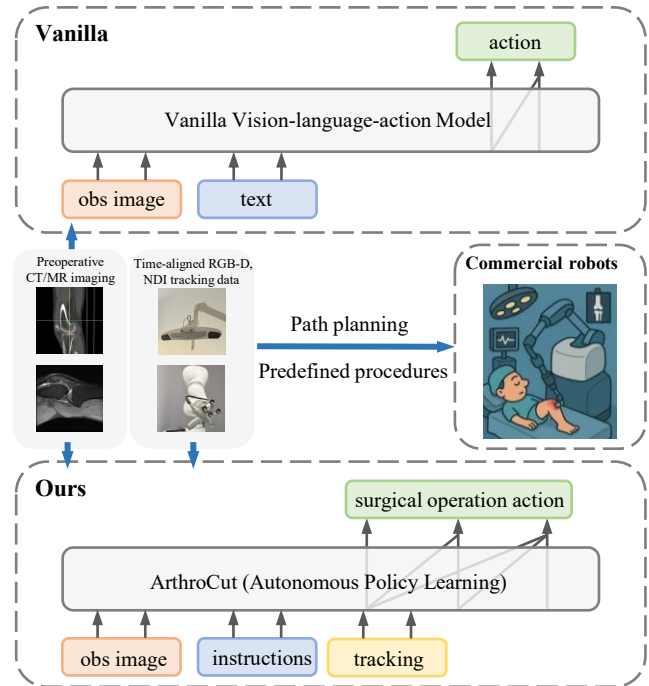


Fig. 1. Comparison between vanilla VLA, commercial orthopedic robots, and the proposed ArthroCut framework. Conventional VLA models (top) directly map task inputs to robotic actions but fail to incorporate detailed surgical scene understanding, limiting their robustness in complex environments. Commercial orthopedic robots (middle) use intraoperative data to follow pre-templated paths but lack real-time adaptation and personalization [1]. In contrast, ArthroCut (bottom) integrates preoperative imaging with temporally aligned multimodal intraoperative data, enabling accurate, real-time decision-making that improves both surgical precision and personalization.

constrain saw paths and guide resections but primarily execute preplanned trajectories under continuous surgeon supervision, with limited ability to perceive intraoperative changes or adapt plans online [6]. At the same time, the operating room provides rich, complementary signals—including preoperative imaging, optical tracking, and intraoperative surgical video—that are rarely modeled jointly. Harnessing these heterogeneous streams to interpret the scene and generate cutting actions could reduce manual burden, improve consistency, and enable safer, more efficient procedures.

Recent progress in vision–language–action (VLA) learning suggests a path toward such autonomy. Robotics-oriented foundation models now represent actions as tokens, enabling end-to-end policies that map multimodal observations directly to control commands [7]–[9]. In parallel, surgical autonomy has progressed from task-level automation to sys-

tems capable of language-conditioned, hierarchical policies and recovery from perturbations, with landmark demonstrations of supervised and laparoscopic autonomous soft-tissue anastomosis [10], [11]. Nevertheless, no prior work has integrated preoperative volumetric imaging, intraoperative surgical scenes, optical tracking, and robot state into a single action-generative model for orthopedic surgery. As shown in Fig. 1, vanilla VLA models (top) adopt a direct mapping strategy from task inputs to robotic actions. While efficient in simplified settings, they lack the ability to parse complex surgical scenes, which severely constrains their robustness and generalizability in high-variability operative environments. Commercial orthopedic robots (middle) represent a step forward by incorporating intraoperative scene information; however, they largely rely on pre-templated surgical paths and standardized execution workflows. As a result, these systems can deliver high geometric accuracy but remain limited in their capacity for real-time adaptation, contextual awareness, and patient-specific customization. In contrast, the proposed ArthroCut framework (bottom) integrates preoperative imaging with temporally aligned multimodal intraoperative data. This fusion enables the model to continuously contextualize surgical actions within the evolving operative field

To address this gap, we introduce **ArthroCut**, a multi-modal autonomous policy learning framework for robotic bone resection in knee arthroplasty. ArthroCut is trained on a custom dataset that synchronizes preoperative CT/MR volumes, intraoperative NDI pose streams of both patient anatomy and robotic end-effector, RGB-D surgical video, natural language instructions, robot state, and ground-truth motion commands. We fine-tune a GPT-style vision-language model to directly generate robot action primitives.

Our key contributions conclude:

- We construct the first aligned, multimodal dataset for autonomous knee arthroplasty, pairing **21** patients’ preoperative CT/MR volumes with **23,205** RGB-D image pairs and **21** full-length intraoperative NDI tracking logs, alongside robot state, language intent, and action commands.
- We develop a VLA-based policy that fuses Preoperative Imaging Tokens (PIT) and Time-Aligned Surgical Tokens (TAST) and autoregressively generates textual motion instructions via an interpretable action grammar with quantized parameters under grammar/safety-constrained decoding.
- In benchtop tests on a knee prosthesis (six resections), we attain a SR of **0.86** and SPL of **0.75** across the six resections, and reduce robot execution time by **11.95%** vs. a vanilla baseline under identical constraints, improving accuracy and throughput for sequential, multi-plane osteotomies.

## II. RELATED WORK

### A. Orthopedic surgical robots

Robotic assistance in knee arthroplasty has been widely investigated as a means to enhance surgical precision, repro-

ducibility, and patient-specific outcomes. Numerous clinical studies have demonstrated that robotic systems improve component positioning and reduce alignment outliers relative to conventional jig-based techniques [12]. Both image-based and imageless platforms have shown improved accuracy in achieving mechanical alignment and restoring limb axes in total knee arthroplasty (TKA) [5]. Similarly, robotic-assisted unicompartmental knee arthroplasty (UKA) has been associated with improved functional outcomes, faster recovery, and radiographic precision comparable or superior to manual UKA [13], [14]. Furthermore, comparative analyses suggest that robotic systems outperform navigation-assisted workflows in coronal-plane accuracy, highlighting their potential to reduce intraoperative variability [15].

Beyond alignment, robotic systems may also contribute to reduced intraoperative bone resection errors, more consistent soft-tissue balancing, and a lower incidence of revision surgery in selected patient cohorts [16]. Nevertheless, these benefits come with limitations: existing platforms typically rely on constraint-based execution of preplanned cuts, offer limited intraoperative perception, and rarely integrate heterogeneous data streams such as volumetric imaging, real-time tracking, and video. This lack of multimodal integration constrains adaptability when intraoperative conditions deviate from preoperative plans. By contrast, our approach explicitly fuses preoperative CT/MR with intraoperative RGB-D, optical tracking, and robot state to generate verifiable action primitives that can adapt trajectories online, moving toward autonomy that is both data-driven and clinically grounded.

### B. Vision–Language–Action models for robotics

Vision–language–action (VLA) learning frames robot control as sequence modeling over multimodal tokens, enabling end-to-end policies that directly map complex sensory inputs to action outputs [17]–[27]. Robotics transformers co-fine-tuned on web-scale vision–language datasets and robot trajectories, such as RT-2, demonstrate semantic generalization and the ability to generate action tokens grounded in high-level instructions [7]. Embodied multimodal large language models (e.g., PaLM-E) further show how continuous sensory inputs (RGB, depth, proprioception) can be interleaved with language to support perception, reasoning, and control in a unified framework [8]. Multimodal prompting frameworks (VIMA) extend this paradigm by conditioning robot actions on diverse input modalities, enabling compositional task execution [9]. Program-synthesis methods (e.g., Code-as-Policies) leverage LLMs to generate executable action code, broadening the expressivity of robot policies [28]. In addition, affordance-aware LLM planning approaches couple LLMs with symbolic planners and affordance detectors, thereby improving long-horizon reasoning and task success [29].

Despite these advances, most VLA systems remain focused on tabletop manipulation and household robotics benchmarks, with limited consideration for high-stakes medical environments. Critically, they seldom incorporate preoperative volumetric medical imaging or enforce geometry-

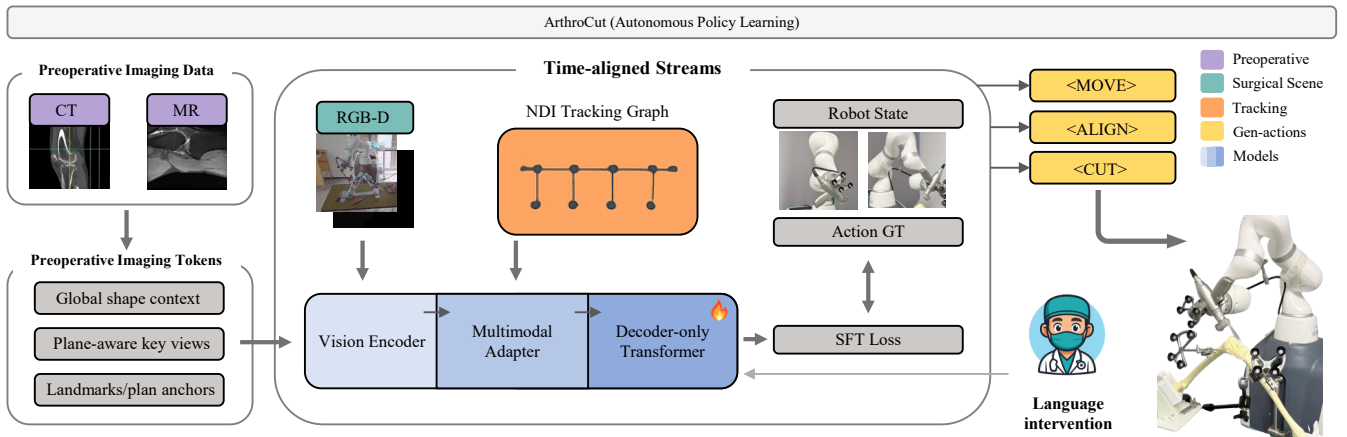


Fig. 2. **The overview of our proposed method for robotic bone resection in knee arthroplasty.** Preoperative imaging and intraoperative multimodal streams are encoded into tokens and processed by a transformer-based model to generate surgical actions (<MOVE>, <ALIGN>, <CUT>) for autonomous femoral cutting.

grounded safety constraints, which are essential in surgical applications. Our framework addresses this gap by combining cross-modal 3D anatomical grounding with tokenized surgical motion primitives, explicitly tailored to multi-plane bone resections in orthopedic surgery.

### C. Autonomy in surgical robotics

The spectrum of autonomy in surgical robotics ranges from pure teleoperation to shared control, conditional autonomy, and fully autonomous execution [30]. Landmark demonstrations include the Smart Tissue Autonomous Robot (STAR), which achieved supervised autonomous soft-tissue suturing in vivo with superior consistency to human surgeons [10], and the demonstration of autonomous laparoscopic small-bowel anastomosis capable of perturbation recovery in porcine models [11]. These milestones underscore the feasibility of autonomy in surgical contexts, particularly in soft-tissue procedures. Authoritative surveys and perspectives from leading journals emphasize the importance of safety assurance, rigorous verification/validation protocols, and maintaining meaningful human oversight as autonomy levels increase [31]. More recent frameworks such as SRT-H propose hierarchical policies conditioned on language, enabling adaptive stepwise planning and partial autonomy in minimally invasive procedures [32].

Nevertheless, prior work has largely concentrated on soft-tissue surgery and task-specific pipelines, with limited generalization to orthopedic procedures requiring rigid bone resection. Moreover, most approaches have not attempted to unify preoperative volumetric imaging with intraoperative tracking and multimodal sensing for real-time decision-making. In contrast, our method integrates multimodal information streams into an action-generative policy that produces verifiable motion primitives, enabling online adaptation and enhancing safety in orthopedic surgical autonomy.

## III. METHOD

In this section, we introduce **ArthroCut**, an autonomous policy-learning framework for robotic bone resection in knee

arthroplasty. As shown in Fig. 2, ArthroCut converts two input families into text-compatible tokens: (i) *Preoperative Imaging Tokens (PIT)* that summarize CT/MR anatomy and surgical intent (global shape, plane-aware key views, landmark/plan anchors), and (ii) *Time-Aligned Surgical Tokens (TAST)* that fuse synchronized intraoperative streams—RGB-D features, an SE(3) graph of NDI-tracked end-effector/bone poses, and robot kinematic/state vectors. A vision encoder and multimodal adapter interleave PIT and TAST, while a decoder-only transformer autoregressively emits a structured action string via an Action Grammar Tokenizer (e.g., <MOVE>, <ALIGN>, <CUT> with quantized parameters). A constraint decoder applies grammar and safety masks to ensure feasible, verifiable commands for multi-plane resections. Training uses next-token supervision with lightweight auxiliaries (plane grounding, SE(3) consistency, motion smoothness), aligning preoperative geometry with intraoperative evidence and yielding interpretable, stepwise policies.

### A. The Base Vision–Language–Action Model

We adopt Qwen2.5–VL–32B–Instruct [33] as the foundational multimodal policy backbone. Conceptually, the model follows a modern vision–language (VL) architecture: a high-capacity visual encoder  $\phi_{\text{vis}}(\cdot)$  processes images or video frames into patch-level features; a lightweight multimodal adaptor  $\psi(\cdot)$  projects these features into the shared language-embedding space; and a decoder-only transformer  $\mathcal{L}_{\theta}$  (a 32B-parameter autoregressive language backbone) performs unified reasoning over interleaved visual tokens and text tokens. The model has been instruction-tuned to follow natural-language prompts while grounding responses in visual evidence, supporting both multi-image inputs and temporally ordered frame batches. Inference is carried out via standard next-token prediction under a causal objective.

We select Qwen2.5–VL–32B for three reasons. First, it combines strong instruction following and robust multimodal grounding in a single, self-hostable framework. This makes it well suited to surgical settings that require long, interleaved

inputs—preoperative CT/MR, intraoperative RGB-D, tracker states, and live robot telemetry. Second, its vision–language pretraining provides transferable capabilities in OCR, spatial reasoning, and layout interpretation that are advantageous in operating-room environments (e.g., labels, rulers, guides, implants). Instruction tuning further aligns naturally with our textual action grammar, enabling stepwise, interpretable motion commands rather than opaque latent controls. Third, the 32B-parameter scale provides sufficient capacity for multi-plane geometric reasoning and temporal consistency without brittle task-specific heuristics. In practice, its mature tokenizer/IO stack, multi-image support, and stable generation controls (temperature, stop tokens) facilitate the seamless incorporation of surgical constraints (e.g., alignment checks, margin tolerances, safety pauses) directly into prompt templates.

### B. Token Representation for Time-Aligned Multimodal Surgical Data

We represent all inputs to the VLA backbone as *text-compatible tokens* that are compact, interpretable, and capable of preserving geometry, appearance, and control semantics. By synchronizing all modalities on a common timeline, the model can jointly reason over both preoperative and intraoperative evidence.

1) *Preoperative Imaging Tokens (PIT)*: Preoperative CT/MR volumes  $V^{\text{CT}}, V^{\text{MR}}$  are summarized into three complementary token sets.

(1) **Global shape context**: A 3D encoder  $\phi_{3\text{D}}$  produces a pooled descriptor capturing overall anatomy:

$$\mathbf{c}_{\text{glob}} = \text{pool}(\phi_{3\text{D}}(V^{\text{CT}}, V^{\text{MR}})). \quad (1)$$

(2) **Plane-aware views**: For each planned resection plane  $P_m$ , a small set of aligned slices  $\{I_k^{(m)}\}$  are encoded and pooled:

$$\mathbf{c}_{\text{view}}^{(m)} = \text{pool}(\{\phi_{\text{vis}}(I_k^{(m)})\}). \quad (2)$$

(3) **Landmark/plan anchors**: Key anatomical landmarks  $\ell_i$  and plane parameters  $(\mathbf{n}_m, b_m)$  are discretized into symbolic bins:

$$\tilde{\ell}_i = \text{bin}(\ell_i), \quad \tilde{\mathbf{n}}_m = \text{bin}(\mathbf{n}_m), \quad \tilde{b}_m = \text{bin}(b_m). \quad (3)$$

All tokens are serialized into a compact block  $\mathbf{C}$  that encodes both global anatomy and surgical intent.

2) *Time-Aligned Surgical Tokens (TAST)*: Intraoperative data streams—RGB-D frames, optical tracking states, and robot kinematics—are aligned on a shared timeline  $\mathcal{T}$ . Each step  $t$  collects a local observation window  $W_t$  to capture short-term dynamics.

(1) **RGB-D scene encoder**: Stacked multi-frame inputs yield visual tokens:

$$\mathbf{v}_t = \phi_{\text{vis}}(\{I_\tau \oplus D_\tau\}_{\tau \in W_t}). \quad (4)$$

(2) **SE(3) graph encoder**: Relative transforms between tracked objects (end-effector, femur, tibia) are embedded as graph features:

$$\mathbf{n}_t = \text{S3GE}(\{T_t^{o \rightarrow o'}\}). \quad (5)$$

(3) **Robot state encoder**: Joint angles  $\mathbf{q}_t$ , velocities  $\dot{\mathbf{q}}_t$ , and torques  $\boldsymbol{\tau}_t$  are mapped into state tokens:

$$\mathbf{r}_t = \text{ASE}(\mathbf{q}_t, \dot{\mathbf{q}}_t, \boldsymbol{\tau}_t). \quad (6)$$

(4) **Action grammar tokenizer**: Ground-truth commands are serialized into motion primitives:

$$\mathbf{y}_t = \langle \text{MOVE} \rangle [\tilde{\mathbf{p}}] \langle \text{ALIGN} \rangle [\tilde{\boldsymbol{\pi}}] \langle \text{CUT} \rangle [\tilde{s}]. \quad (7)$$

Finally, the model input at time  $t$  concatenates preoperative context with intraoperative observations and the previous action:

$$\mathbf{Z}_t = \text{concat}(\mathbf{C}, \mathbf{v}_t, \mathbf{n}_t, \mathbf{r}_t, \mathbf{y}_{t-1}). \quad (8)$$

This formulation fuses patient-specific anatomy with real-time surgical evidence, enabling autoregressive generation of structured, interpretable action sequences.

### C. Training Procedures

**Setup and objective.** We fine-tune Qwen2.5-VL-32B-Instruct to autoregressively emit action tokens  $y_{t:t+H}$  given a serialized prefix that concatenates the preoperative context  $\mathbf{C}$ , the current observation block  $\mathbf{X}_t$ , and preceding actions (cf. III-B). Actions are represented by an Action Grammar Tokenizer (AGT) that encodes motion primitives with quantized parameters. Training uses teacher forcing, with supervision applied only on action tokens; observation tokens act as fixed prefixes. The primary loss is next-token SFT:

$$\mathcal{L}_{\text{SFT}} = - \sum_{\tau=t}^{t+H} \log p_\theta(y_\tau | \text{prefix}, y_{t:\tau-1}), \quad (9)$$

optionally combined with auxiliary terms (e.g., plane grounding, SE(3) consistency, action smoothness):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SFT}} + \sum_i \lambda_i \mathcal{L}_i.$$

Decoding is unchanged by auxiliaries.

**Sequence packing and prompts.** Each instance follows a compact instruction template that declares modality tags and task intent, then interleaves  $\mathbf{C}$ ,  $\mathbf{X}_t$ , and the previous action block. To maximize throughput, multiple temporal windows are concatenated up to the context budget with  $\langle \text{SEP} \rangle$  delimiters to avoid cross-window leakage. All numeric fields use the same quantization bins as at inference to prevent train–test mismatch. When latency compensation is needed, targets are shifted by  $\Delta t$  to align with system delay.

**Safety-aware token masking.** We apply grammar-constrained masking during training and decoding to enforce semantic and physical validity. The mask suppresses infeasible primitives (e.g., disallow  $\langle \text{CUT} \rangle$  before a valid  $\langle \text{ALIGN} \rangle$ ) and invalid parameter bins conditioned on instantaneous state and plan, while preserving gradients over admissible choices. Using the same constraints at inference ensures train–deployment consistency and restricts generation to the feasible action space.

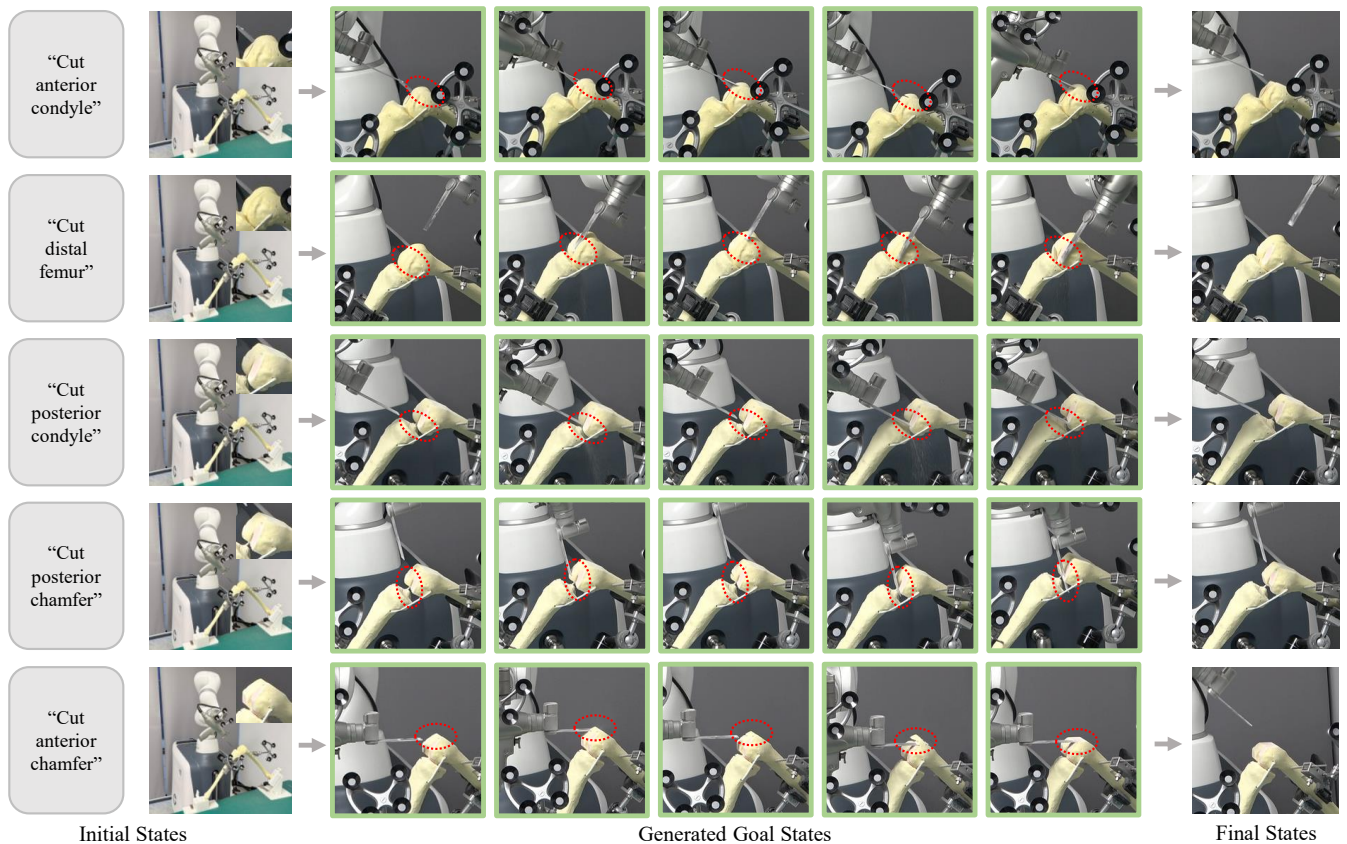


Fig. 3. **Example of task execution for femoral cutting across five planes using ArthroCut.** For each task: Left—text instruction with the initial state (top-right inset of the second column shows the pre-cut bone surface); Middle—generated intermediate goal states; Right—final state upon task completion. Full execution trajectories are provided in the supplementary video.

## IV. EXPERIMENTS

### A. Dataset Creation

We collected a patient-level, time-synchronized multi-modal dataset from **21** complete primary knee arthroplasty procedures performed with a commercial orthopedic robot under attending surgeons. Each case includes a full intra-operative timeline and paired preoperative CT–MR imaging for the operated knee. During surgery, six clinically standard resections are performed: tibial, distal femur, anterior condyle, posterior condyle, anterior chamfer, and posterior chamfer. These phases provide natural supervision for multi-plane planning and execution.

An RGB–D camera records the surgical field, yielding **23,205** synchronized RGB/Depth pairs across all cases. Intraoperative optical tracking (NDI) logs the SE(3) poses of the end-effector, femur, and tibia; the robot controller logs joint state, velocities/torques and tool–center-point pose; and the console exports time-stamped surgical events and motion commands. All streams are hardware time-stamped and *software-aligned* onto a common reference grid (25 ms step) before tokenization (cf. III-B).

This dataset is designed to (i) ground preoperative geometry and plan intent (PIT) and (ii) train a time-aligned, multimodal policy to generate tokenized robot actions for multi-plane bone resections. It provides the necessary cov-

erage of tibial and femoral cuts as well as anterior/posterior condyle and chamfer planes to evaluate *sequential* planning and execution.

### B. Evaluation Metrics

We adopt a single, task-level success criterion and report two metrics: *Success Rate (SR)* and *Success weighted by Path Length (SPL)*. An episode is deemed successful ( $S_i=1$ ) if all six resections (tibial, distal femur, anterior/posterior condyle, anterior/posterior chamfer) are executed and the mean surface deviation between the executed and planned cut surfaces, averaged over the six planes, is within a fixed tolerance  $\delta$  mm; otherwise  $S_i=0$ . The surface deviation per plane is computed as the bidirectional Chamfer distance (in mm) between uniformly sampled points on the executed and planned surface patches within the resection window.

### C. Evaluations Results

We evaluate on a *physical bench test* using a knee prosthesis and bone analogs, executing the complete sequence of six resections over **7** independent runs. Success is defined in IV-B: an episode is successful if, after all six planes are cut, the *mean surface deviation* from plan (bidirectional Chamfer distance over the resection patches) averaged across the six planes is  $\leq \delta=1.5$  mm. We report **SR** and **SPL** with grammar/safety-constrained decoding for all methods. For

TABLE I

**BENCH EVALUATION ON A KNEE PROSTHESIS (SR).** SUCCESS THRESHOLD  $\delta=1.5$  MM MEAN SURFACE DEVIATION ACROSS THE SIX PLANES. ALL METHODS USE THE SAME ACTION GRAMMAR AND CONSTRAINED DECODING. FT: FINE-TUNE.

Method	Metrics					
	Success Rate (SR) $\uparrow$					
	anterior chamfer	distal femur	anterior condyle	posterior condyle	tibial	posterior chamfer
OpenVLA (ft) [17]	0.29 (2/7)	0.29 (2/7)	0.29 (2/7)	0.00 (0/7)	0.29 (2/7)	0.00 (0/7)
Diffusion Policy [34]	0.47 (4/7)	0.29 (2/7)	0.47 (4/7)	0.29 (2/7)	0.00 (0/7)	0.00 (0/7)
RT-2 (ft) [7]	0.29 (2/7)	0.29 (2/7)	0.29 (2/7)	0.00 (0/7)	0.00 (0/7)	0.00 (0/7)
Octo (ft) [35]	0.29 (2/7)	0.00 (0/7)	0.00 (0/7)	0.29 (2/7)	0.00 (0/7)	0.00 (0/7)
Qwen2.5-VL-Instruct (vanilla) [33]	0.00 (0/7)	0.00 (0/7)	0.29 (2/7)	0.00 (0/7)	0.00 (0/7)	0.00 (0/7)
ArthroCut (Ours)	<b>1.00 (7/7)</b>	<b>1.00 (7/7)</b>	<b>1.00 (7/7)</b>	<b>1.00 (7/7)</b>	<b>0.57 (4/7)</b>	<b>0.57 (4/7)</b>

TABLE II

**BENCH EVALUATION ON A KNEE PROSTHESIS (SPL).** SUCCESS THRESHOLD  $\delta=1.5$  MM MEAN SURFACE DEVIATION ACROSS THE SIX PLANES. SPL IS COMPUTED WITH THE PROTOCOL AND REPORTED AS MEAN  $\pm$ SD OVER EPISODES. ALL METHODS USE THE SAME ACTION GRAMMAR AND CONSTRAINED DECODING. FT: FINE-TUNE.

Method	Metrics					
	Success weighted by Path Length (SPL) $\uparrow$					
	anterior chamfer	distal femur	anterior condyle	posterior condyle	tibial	posterior chamfer
OpenVLA (ft) [17]	0.22 $\pm$ 0.02	0.24 $\pm$ 0.02	0.24 $\pm$ 0.02	0.00 $\pm$ 0.00	0.20 $\pm$ 0.04	0.00 $\pm$ 0.00
Diffusion Policy [34]	0.42 $\pm$ 0.05	0.23 $\pm$ 0.03	0.43 $\pm$ 0.04	0.22 $\pm$ 0.03	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
RT-2 (ft) [7]	0.21 $\pm$ 0.02	0.18 $\pm$ 0.03	0.20 $\pm$ 0.03	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
Octo (ft) [35]	0.22 $\pm$ 0.03	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.22 $\pm$ 0.02	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
Qwen2.5-VL-Instruct (vanilla) [33]	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.24 $\pm$ 0.03	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
ArthroCut (Ours)	<b>0.92 <math>\pm</math> 0.03</b>	<b>0.93 <math>\pm</math> 0.02</b>	<b>0.95 <math>\pm</math> 0.02</b>	<b>0.91 <math>\pm</math> 0.04</b>	<b>0.39 <math>\pm</math> 0.08</b>	<b>0.42 <math>\pm</math> 0.07</b>

fairness, all baselines are retrained on our dataset with the same train/val split and receive the same action grammar; decoding uses identical constraints and temperature/top- $p$  settings.

**Baselines:** We compare against five strong policies: (1) **Diffusion Policy** (pixel-to-action diffusion with multi-frame RGB-D) [34], (2) **OpenVLA** (open-source VLA backbone fine-tuned on our tokens) [17], (3) **Octo** (large multi-task policy with vision-language prompts) [35], (4) **RT-2** (vision-language-action transformer with text-as-action fine-tuning) [7], and (5) **Qwen2.5-VL-Instruct (vanilla)** [33]—a backbone-matched baseline trained with our grammar but *without* PIT or the structure-aware S3GE encoder or safety masks. Our method **ArthroCut (Ours)** uses PIT+TAST tokens, S3GE, grammar masks, and the training protocol in III-C.

As shown in Table I, II, ArthroCut achieves the highest SR and SPL, attaining a 100% success rate across four primary osteotomy planes—anterior chamfer, distal femur, anterior condyle, and posterior condyle. Performance decreases on the anterior and posterior chamfer planes, where the minimal bone resection leads to prosthesis instability and consequently lower success rates. Among competing methods, the strongest baseline is the Diffusion Policy, while diffusion-style and generalist VLA approaches (e.g., RT-2, Octo) occasionally reach individual planes but frequently drift or require detours, impairing both reliability and efficiency. The vanilla Qwen2.5 baseline rarely completes tasks, highlighting

the critical role of our key innovations: PIT tokens (preoperative geometry), structure-aware tracking (S3GE), and grammar/safety masks for converting multimodal evidence into safe and executable cutting actions. Owing to its superior success rate across osteotomy tasks, ArthroCut also achieves the highest SPL among all evaluated methods.

#### D. Ablation Study

We introduce two complementary token families to bridge planning and execution in knee arthroplasty. Preoperative Imaging Tokens (PIT) provides a stable anatomical prior that reduces ambiguity in plane identification and aligns action generation with the preplanned resection objectives. Time-Aligned Surgical Tokens (TAST) fuse synchronized intraoperative evidence—RGB-D scene features, an SE(3) graph of NDI-tracked poses for end effector and bones, robot kinematic/state vectors, and the previous action—so the policy can reason over real-time, structure-aware relations and temporal continuity.

As shown in Table IV, the introduction of time-aligned surgical tokens (TAST) emerges as the dominant factor driving performance, elevating the vanilla policy from SR = 0.05/SPL = 0.04 to 0.68/0.61 (+0.63 and +0.57 absolute). Incorporating preoperative imaging tokens (PIT) alone provides a modest yet consistent improvement (0.12/0.10, +0.07/+0.06), reflecting the utility of anatomical priors even in isolation. When combined (ArthroCut), the two components achieve the highest performance with SR = 0.86 and

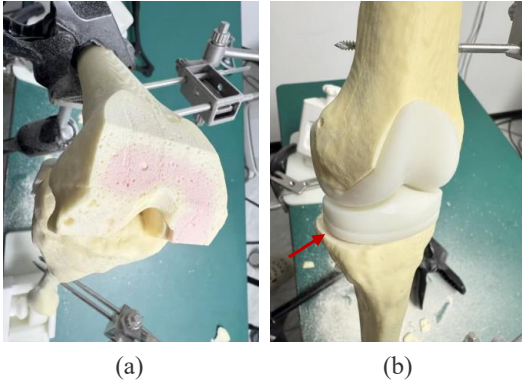


Fig. 4. **Schematic diagram of knee replacement verification after osteotomy.** (a) Femoral resections provide a visually flat surface for component seating; (b) the joint shown in full extension after component placement.

TABLE III  
COMPARISON OF INFERENCE LATENCY AND ROBOT EXECUTION TIME (MEAN) BETWEEN THE VANILLA MODEL AND THE PROPOSED ARTHROCut FRAMEWORK.

Method	Inference time (ms)	Robot exec. time (s)
Vanilla Model	267.14	121.89
ArthroCut (Ours)	302.52	107.33

$SPL = 0.75$ , representing gains of  $+0.18/+0.14$  over TAST alone and  $+0.81/+0.71$  over vanilla. These results highlight a synergistic effect: PIT anchors the model to planned anatomy and target planes, while TAST delivers real-time, structure-aware cues essential for robust and efficient multi-plane cutting.

#### E. Arthroplasty Verification and Efficiency Analysis

Beyond the quantitative success criteria, we qualitatively verified prosthesis fit on the post-cut knee surrogates. As illustrated in Fig. 4(a), the femoral resections seat the component with visually flat, coplanar surfaces and negligible step-offs across cut transitions, consistent with the high per-plane completion rates. Figure 4(b) shows the joint in full extension after component placement—a configuration critical for stability during weight bearing and gait. In this posture, the femoral component sits flush, whereas a small gap is visible at the tibial tray–bone interface. This observation accords with our per-plane statistics, where the tibial plane exhibits the lowest success rate among the six tasks.

Table III summarizes policy latency and robot execution time. With grammar- and safety-constrained decoding, ArthroCut achieves a mean per-action latency of 302.52 ms on a single NVIDIA A100; the dominant contribution to wall-clock time remains mechanical motion and material removal. Because cutting proceeds stepwise under safety gates (move  $\rightarrow$  align  $\rightarrow$  cut), policy computation constitutes only a small fraction of total episode time, indicating that current latency is unlikely to bottleneck clinical workflow. Nonetheless, further acceleration via distillation, low-bit quantization, and speculative decoding is under consideration.

TABLE IV  
ABLATION STUDY ON THE PROPOSED ARTHROCut FRAMEWORK. PIT DENOTES THE PROPOSED PREOPERATIVE IMAGING TOKENS AND TAST DENOTES THE TIME-ALIGNED SURGICAL TOKENS.

Method	Ablation		SR $\uparrow$	SPL $\uparrow$
	PIT	TAST		
Vanilla Model			0.05	0.04
Vanilla Model	$\checkmark$		0.12	0.10
Vanilla Model		$\checkmark$	0.68	0.61
ArthroCut (Ours)	$\checkmark$	$\checkmark$	0.86	0.75

#### V. CONCLUSION, LIMITATION AND FUTURE WORK

In this work, we introduced *ArthroCut*, an autonomous policy learning framework that translates time-aligned multimodal observations into structured and verifiable robot actions for multi-plane bone cutting in unicompartmental knee arthroplasty. The framework leverages two complementary token families: Preoperative Imaging Tokens (PIT), which compactly encode patient-specific CT/MR anatomy and surgical intent, and Time-Aligned Surgical Tokens (TAST), which fuse intraoperative RGB–D cues, SE(3) bone–tool relations from optical tracking, robot states, and the preceding action. Built on a Qwen2.5–VL–32B–Instruct backbone, the model performs unified next-token reasoning and generates an action grammar with quantized parameters under grammar and safety masks. Using a time-synchronized dataset of 21 cases (including 23,205 RGB–D pairs) and a physical bench evaluation with a knee prosthesis, *ArthroCut* consistently outperforms state-of-the-art methods on *Success Rate (SR)* and *SPL*. Ablation studies further confirm that TAST is the primary driver of execution reliability, while PIT provides essential anatomical grounding; their integration yields the largest improvements and the most stable multi-plane performance. Furthermore, ArthroCut achieves a mean per-action latency of 302.52 ms on a single A100 and reduces robot execution time by 11.95% relative to the vanilla baseline.

Despite these encouraging results, several practical constraints remain. First, the dataset—though fully synchronized and multimodal—remains of moderate scale and is derived from a single collection pipeline, potentially limiting generalization to diverse anatomies, implant types, or surgical workflows. Second, evaluation was performed primarily on prosthesis-based bench tests; while providing controlled realism, these settings do not fully capture the complexity of live operating rooms (e.g., soft-tissue interference, blood, and intraoperative variability). Third, the current policy assumes calibrated RGB–D sensing and stable optical tracking, rendering it sensitive to sensor drift or temporary marker loss. These limitations are engineering in nature rather than fundamental, and they highlight clear avenues for enhancing robustness and external validity. Looking forward, we believe recent advances in autonomous surgical robots and generative models [32], [36], [37] present promising opportunities to enhance surgical decision-making accuracy and surgical continuity.

## REFERENCES

- [1] Y. Wang, R. Li, and P. Zheng, "Progress in clinical application of artificial intelligence in orthopedics," *Digital Medicine*, vol. 8, p. 4, 2022.
- [2] J. Zhang, W. S. Ndou, N. Ng, P. Gaston, P. M. Simpson, G. J. Macpherson, J. T. Patton, and N. D. Clement, "Robotic-arm assisted total knee arthroplasty is associated with improved accuracy and patient reported outcomes: a systematic review and meta-analysis," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 30, no. 8, pp. 2677–2695, 2022.
- [3] F. Zhang, H. Li, Z. Ba, C. Bo, and K. Li, "Robotic arm-assisted vs conventional unicompartmental knee arthroplasty: a meta-analysis of the effects on clinical outcomes," *Medicine*, vol. 98, no. 35, p. e16968, 2019.
- [4] S. Petrillo, F. Migliorini, G. Moretti, and S. Romagnoli, "Accuracy of rosa knee system in bone cuts orientation during total knee arthroplasty: An observational study," *Journal of Clinical Medicine*, vol. 14, no. 15, p. 5205, 2025.
- [5] M. Bautista, J. Manrique, and W. J. Hozack, "Robotics in total knee arthroplasty," *The journal of knee surgery*, vol. 32, no. 07, pp. 600–606, 2019.
- [6] G. Zheng and L. P. Nolte, "Computer-assisted orthopedic surgery: current state and future perspective," *Frontiers in surgery*, vol. 2, p. 66, 2015.
- [7] 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*. PMLR, 2023, pp. 2165–2183.
- [8] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: an embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 8469–8488.
- [9] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," *arXiv preprint arXiv:2210.03094*, vol. 2, no. 3, p. 6, 2022.
- [10] 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.
- [11] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Léonard, M. H. Hsieh, J. U. Kang, and A. Krieger, "Autonomous robotic laparoscopic surgery for intestinal anastomosis," *Science robotics*, vol. 7, no. 62, p. eabj2908, 2022.
- [12] B. Kayani, S. Konan, J. R. Pietrzak, and F. S. Haddad, "Iatrogenic bone and soft tissue trauma in robotic-arm assisted total knee arthroplasty compared with conventional jig-based total knee arthroplasty: a prospective cohort study and validation of a new classification system," *The Journal of Arthroplasty*, vol. 33, no. 8, pp. 2496–2501, 2018.
- [13] B. Kayani and F. Haddad, "Robotic unicompartmental knee arthroplasty: current challenges and future perspectives," *Bone & Joint Research*, vol. 8, no. 6, pp. 228–231, 2019.
- [14] Y. Sun, W. Liu, J. Hou, X. Hu, and W. Zhang, "Does robotic-assisted unicompartmental knee arthroplasty have lower complication and revision rates than the conventional procedure? a systematic review and meta-analysis," *BMJ open*, vol. 11, no. 8, p. e044778, 2021.
- [15] F. Figueroa, E. Wakelin, J. Twiggs, and B. Fritsch, "Comparison between navigated reported position and postoperative computed tomography to evaluate accuracy in a robotic navigation system in total knee arthroplasty," *The Knee*, vol. 26, no. 4, pp. 869–875, 2019.
- [16] T. Bayoumi, L. J. Kleebblad, T. A. Borus, T. M. Coon, J. Douchis, J. T. Nguyen, and A. D. Pearle, "Ten-year survivorship and patient satisfaction following robotic-arm-assisted medial unicompartmental knee arthroplasty: a prospective multicenter study," *JBJS*, vol. 105, no. 12, pp. 933–942, 2023.
- [17] 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.
- [18] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, *et al.*, " $\pi_{0.5}$ : a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.
- [19] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1702–1713.
- [20] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: A 3d vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024.
- [21] Y. Guo, J. Zhang, X. Chen, X. Ji, Y.-J. Wang, Y. Hu, and J. Chen, "Improving vision-language-action model with online reinforcement learning," *arXiv preprint arXiv:2501.16664*, 2025.
- [22] P. Ding, H. Zhao, W. Zhang, W. Song, M. Zhang, S. Huang, N. Yang, and D. Wang, "Quar-vla: Vision-language-action model for quadruped robots," in *European Conference on Computer Vision*. Springer, 2024, pp. 352–367.
- [23] K. F. Gbagbe, M. A. Cabrera, A. Alabbas, O. Alyunes, A. Lykov, and D. Tsetserukou, "Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations," in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2024, pp. 2864–2869.
- [24] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, *et al.*, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robotics and Automation Letters*, 2025.
- [25] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.
- [26] R. Zhang, M. Dong, Y. Zhang, L. Heng, X. Chi, G. Dai, L. Du, Y. Du, and S. Zhang, "Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation," *arXiv preprint arXiv:2503.20384*, 2025.
- [27] J. Liu, H. Chen, P. An, Z. Liu, R. Zhang, C. Gu, X. Li, Z. Guo, S. Chen, M. Liu, *et al.*, "Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model," *arXiv preprint arXiv:2503.10631*, 2025.
- [28] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," *arXiv preprint arXiv:2209.07753*, 2022.
- [29] T. Birr, C. Pohl, A. Younes, and T. Asfour, "Autogpt+: Affordance-based task planning with large language models," *arXiv preprint arXiv:2402.10778*, 2024.
- [30] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastris, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 651–679, 2021.
- [31] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," p. eam8638, 2017.
- [32] J. W. Kim, J.-T. Chen, P. Hansen, L. X. Shi, A. Goldenberg, S. Schmidgall, P. M. Scheickl, A. Deguet, B. M. White, D. R. Tsai, *et al.*, "Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning," *Science robotics*, vol. 10, no. 104, p. eadt5254, 2025.
- [33] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.
- [34] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [35] 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.
- [36] W. Song, J. Chen, W. Li, X. He, H. Zhao, C. Cui, P. D. S. Su, F. Tang, X. Cheng, D. Wang, *et al.*, "Rationalvla: A rational vision-language-action model with dual system," *arXiv preprint arXiv:2506.10826*, 2025.
- [37] Y. Long, A. Lin, D. H. C. Kwok, L. Zhang, Z. Yang, K. Shi, L. Song, J. Fu, H. Lin, W. Wei, *et al.*, "Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery," *Science Robotics*, vol. 10, no. 104, p. eadt3093, 2025.