FlexiCup: Wireless Multimodal Suction Cup with Dual-Zone Vision-Tactile Sensing

***Abstract*—Conventional suction cups lack sensing capabilities for contact-aware manipulation in unstructured environments. This paper presents FlexiCup, a multimodal suction cup with wireless electronics that integrate dual-zone vision-tactile sensing. The central zone dynamically switches between vision and tactile modalities via illumination control, while the peripheral zone provides continuous spatial awareness. The modular mechanical design supports both vacuum (sustained-contact adhesion) and Bernoulli (contactless lifting) actuation while maintaining the identical dual-zone sensing architecture, demonstrating sensing-actuation decoupling where sensing and actuation principles are orthogonally separable. We validate hardware versatility through dual control paradigms. Modular perception-driven grasping achieves comparable success rates across vacuum (90.0%) and Bernoulli (86.7%) modes using identical sensing and control pipelines, validating the sensing architecture's effectiveness across fundamentally different pneumatic principles. Diffusion-based end-to-end learning achieves 73.3% and 66.7% success on contact-aware manipulation tasks, with ablation studies confirming 13% improvements from multi-head attention coordinating dual-zone observations. Hardware designs, firmware, and experimental videos are available at the companion website: https://anonymous.4open.science/api/repo/FlexiCup-DA7D/file/index.html?v=8f531b44.**

***Index Terms*—Multimodal sensing, Suction manipulation, Vision-tactile sensing, Policy learning**

I. INTRODUCTION

SUCTION cups have been widely adopted in robotic manipulation for handling diverse object geometries [1]–[5]. Effective manipulation in unstructured environments requires target identification, obstacle detection, contact verification, and surface adaptation. Traditional suction systems rely on preprogrammed trajectories without sensing feedback [6], [7].

Recent efforts have integrated sensing into suction cups to enable contact-aware manipulation. Camera-based tactile systems [8], [9] achieve high-resolution contact imaging but dedicate their entire optical field to the tactile membrane, sacrificing peripheral spatial awareness. Systems using external cameras [2], [10] can observe the workspace but suffer from occlusion during contact. These architectural constraints force a trade-off: existing sensory suction systems must choose between high-resolution contact sensing and spatial context awareness.

The visuotactile sensing community has demonstrated that coordinating vision and tactile observations enables robust manipulation under varied conditions [11]–[14]. However, these systems typically employ gripper-based designs [15], [16] that require opposing contact surfaces or graspable features, limiting their applicability to scenarios such as handling flat featureless objects or accessing confined spaces with restricted side clearance.

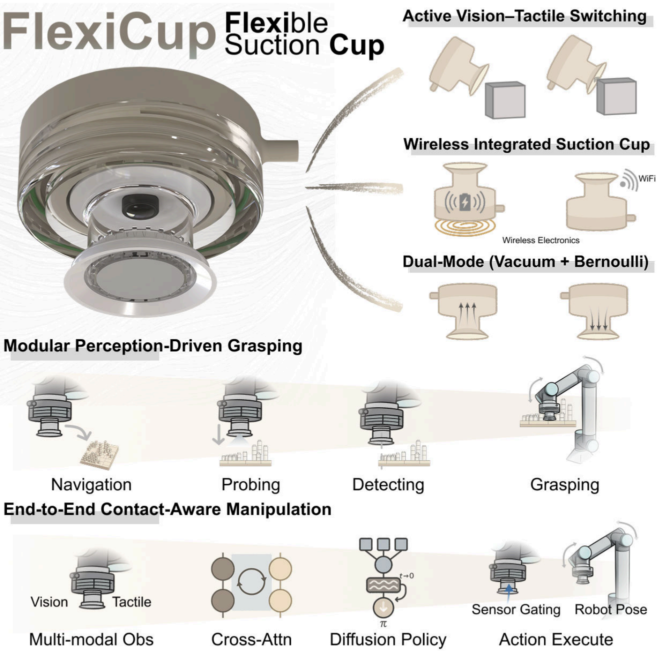

Fig. 1. FlexiCup overview showing hardware features (top), modular perception-driven grasping approach (middle), and end-to-end learning framework (bottom). The system integrates active vision-tactile switching, wireless operation, and vacuum-based contact manipulation through a modular hardware architecture.

Motivated by these advances in multimodal sensing, we present FlexiCup, a wireless multimodal suction cup that addresses the architectural limitations in existing sensory suction systems through dual-zone vision-tactile sensing while operating in manipulation scenarios where gripper-based approaches face geometric constraints (Fig. 1). A key insight is decoupling sensing from actuation: the dual-zone vision-tactile architecture operates across both vacuum and Bernoulli pneumatic principles. The architecture implements dual-zone sensing: a central zone switches between vision and tactile modalities via illumination control, while a peripheral zone maintains continuous spatial awareness.

This architecture enables adaptive manipulation requiring coordinated visual search, obstacle avoidance, and contact verification—capabilities that existing sensory suction systems cannot simultaneously provide. By overcoming the morphological constraints of gripper-based counterparts on flat or confined targets, FlexiCup validates its efficacy through experiments on varying obstacle densities, confined space extraction, and multi-phase manipulation sequences requiring continuous feedback loop.

The key contributions of this work include:

TABLE I
COMPARISON OF REPRESENTATIVE SUCTION CUP PERCEPTION SYSTEMS

Work	Mechanism	Modality	Perception Type	Force	Method	Deployment
Doi et al. [6]	Vacuum	Tactile	Capacitive	—	Stroke-based State Decision	Wired
Aoyagi et al. [7]	Vacuum	Tactile	Piezoresistive	—	Multi-cup Force Reasoning	Wired
Lee et al. [2]	Vacuum	Tactile	Pressure-based	19 N	Model-based Haptic Search	Wired
Shahabi et al. [17]	Vacuum	Tactile	Piezoresistive	—	ML-based Property Regression	Wired
van Veggel et al. [8]	Vacuum	Tactile	**Vision-based**	9.35 N	CNN-based Pose Correction	Wired
Yue et al. [18]	Vacuum	Tactile	Pressure-based	—	Hierarchical Embodied Control	Wired
Jang et al. [19]	Vacuum	Tactile	Pressure-based	2.8 N	Closed-loop Force Regulation	Wired
Ours	**Vacuum / Bernoulli**	**Vision + Tactile**	**Vision-based**	**41.5 N**	**Modular Perception Driven / Diffusion-based Imitation Learning**	**Wireless**

- Sensing-actuation decoupling through modular design, enabling identical sensing and control pipelines across vacuum and Bernoulli modes.
- Dual-zone vision-tactile architecture with wireless electronics, implementing illumination-controlled modality switching between contact imaging and spatial context.
- Dual-paradigm validation combining modular grasping (90.0% vacuum, 86.7% Bernoulli) and diffusion-based learning with multi-head attention, confirmed by ablation studies and baseline comparisons.

II. Related Work

Suction manipulation has advanced through diverse sensing modalities (Table I). Early systems employed capacitive [6] and piezoresistive [7], [17] sensing for basic contact detection, while pressure-based approaches [2], [18], [19] enabled model-based haptic search and force regulation. Recent camera-based tactile sensors [8], [20], building on foundational vision-based tactile technologies [21]–[23], achieve high-resolution deformation imaging but dedicate their entire optical field to contact sensing, precluding simultaneous spatial awareness. Integration challenges remain: electronic tactile arrays can compromise pneumatic sealing, while maintaining perception throughout contact-to-non-contact transitions requires architectural solutions beyond single-modality designs.

Control paradigms have similarly evolved from classical planning [2], [24] to learning-based methods [8], [17], [25]–[27], yet deployment constraints persist—existing systems require wired connections for high-bandwidth sensing. This work addresses these limitations through wireless electronics, dual-zone vision-tactile architecture enabling modality switching, and modular dual-mode design supporting both classical and diffusion-based control while achieving 41.5 N force capacity.

III. Hardware Design

The FlexiCup hardware addresses the trade-off between contact imaging and spatial awareness by realizing dual-zone vision–tactile sensing in a wireless, modular suction morphology. It implements a layered pneumatic–optical–electronic stack and interchangeable bottom housings that decouple sensing from actuation while remaining compatible with both vacuum and Bernoulli principles; the following subsections detail the guiding design principles and system architecture.

A. Design Principles

FlexiCup integrates vision and tactile modalities within a unified optical framework, employs wireless electronics to eliminate electrical tethering, and implements modular mechanical design enabling dual-mode operation with complementary actuation mechanisms—vacuum for contact-based adhesion and Bernoulli for contactless lifting—while maintaining identical core electronics and dual-zone sensing architecture.

B. System Architecture and Integration

The system adopts a modular layered architecture integrating pneumatic actuation, electronic control, and optical sensing (Fig. 2(a)). The modular design separates task-specific pneumatic actuation in the bottom housing from shared electronic control and optical sensing in the top layer, enabling rapid reconfiguration between suction modes.

The bottom housing provides the pneumatic interface and contact surface, with a central groove accommodating the PDMS membrane for sealing and sensing. Pneumatic airways route around the membrane perimeter, physically separating actuation from sensing. The airway geometry differs between Vacuum and Bernoulli configurations, while the membrane mounting interface remains consistent. The bottom housing engages with the top through threaded connection with fluoroelastomer O-ring seal.

The top housing mounts a PCB assembly centered on an ESP32S3 microcontroller (3.7 V, 300 mAh battery, 12.5 μH wireless charging coil), streaming 640×480 images at 30 Fps over Wi-Fi. The battery provides approximately 30 minutes of continuous operation, with thermal management facilitated by pneumatic airflow and 2-hour wireless charging at 200 mA (Fig. 2(b,c)). This electronic module remains identical across both suction modes, enabling wireless operation with electrical decoupling from the robot arm.

The top housing integrates an optical sensing system below the electronic module, comprising an OV5640 camera with 180° fisheye lens and LED arrays enabling dual-zone perception through the tactile membrane interface.

The pneumatic control system employs distinct configurations: vacuum mode utilizes a vacuum pump (750 W, 140 L/min) at -90 kPa maximum pressure, achieving 41.5 N normal force at -80 kPa, while Bernoulli mode relies on an air compressor (800 W, 65 L/min) with supply pressure up to

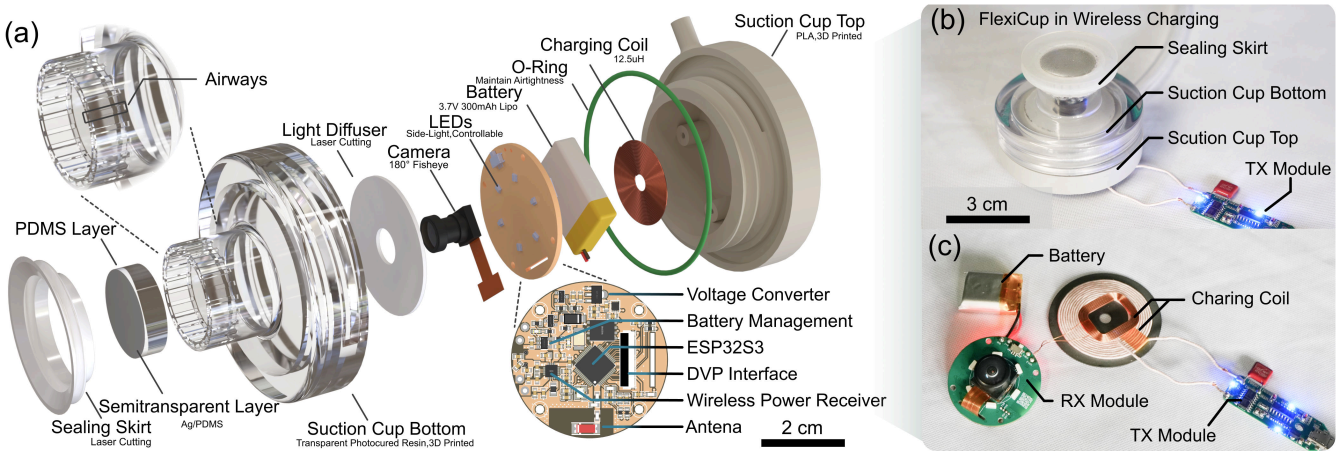

Fig. 2. System architecture showing (a) modular component integration, (b) wireless charging configuration, and (c) standalone charging module.

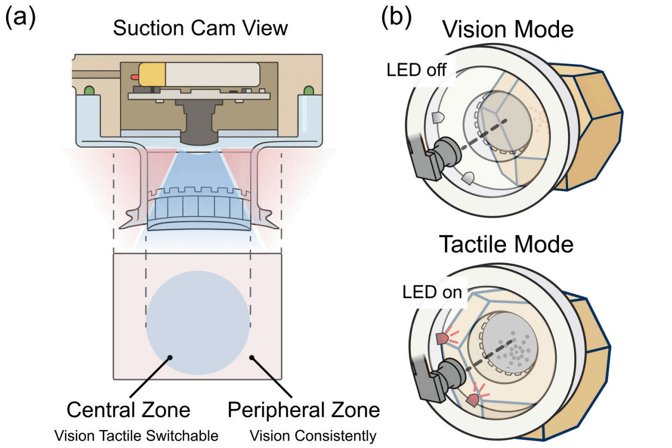

Fig. 3. Vision-tactile sensing showing (a) dual-zone camera view and (b) modality switching via illumination control.

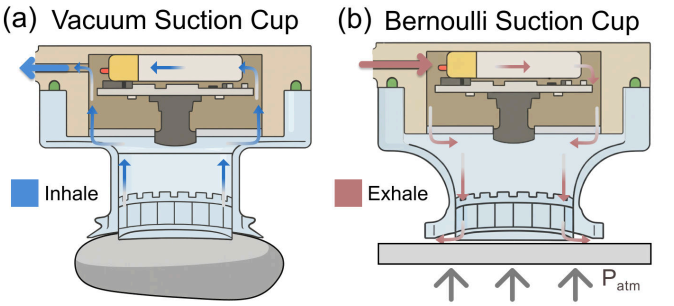

Fig. 4. Dual-mode suction mechanisms demonstrating complementary actuation principles. (a) Vacuum mode generates adhesion through inward airflow requiring sustained contact. (b) Bernoulli mode creates lifting force through outward airflow enabling contactless lifting. Both modes share the identical dual-zone vision-tactile sensing architecture.

0.8 MPa. Pneumatic actuation is controlled wirelessly through solenoid valves triggered by the onboard microcontroller. The standardized interface enables rapid reconfiguration by exchanging the bottom housing and connecting to the corresponding pneumatic source.

C. Vision-Tactile Sensing System

The 180° fisheye camera captures two functional zones: the central zone enables switchable vision-tactile sensing, while the peripheral zone provides continuous visual awareness (Fig. 3(a)).

The system switches modalities through illumination control in real-time by the ESP32S3 microcontroller with dynamic camera exposure and gain adjustments (Fig. 3(b)). In vision mode, LEDs remain inactive, allowing ambient light through the membrane for object detection. In tactile mode, LEDs illuminate the membrane internally, imaging deformations induced by contact forces. The semitransparent PDMS membrane exhibits ambient light sensitivity, addressed through high-intensity internal LEDs with optimized fisheye lens and camera exposure settings, though tactile features may become less pronounced under exceptionally strong environmental lighting.

D. Dual-Mode Suction Mechanisms with Complementary Sensing

Vacuum mode generates adhesion through negative pressure, requiring sustained contact that induces continuous membrane deformation, enabling dense tactile feedback throughout manipulation for deformable object handling and surface compliance detection (Fig. 4(a)).

Bernoulli mode creates contactless lifting through outward airflow (Fig. 4(b)), with the dual-zone architecture supporting visual perception and tactile verification during positioning. Both modes share identical sensing pipelines, confirming sensing-actuation decoupling where mode selection depends on task requirements rather than sensing constraints.

E. PDMS Membrane Design

The PDMS membrane provides tactile sensing and sealing through contact-induced deformation (Fig. 5(a)). The dual-layer system (Fig. 5(b)) employs a PDMS base layer (30:1 mass ratio, 70°C for 4 hours) providing compliance that captures surface details, and a semitransparent layer (Ag:PDMS

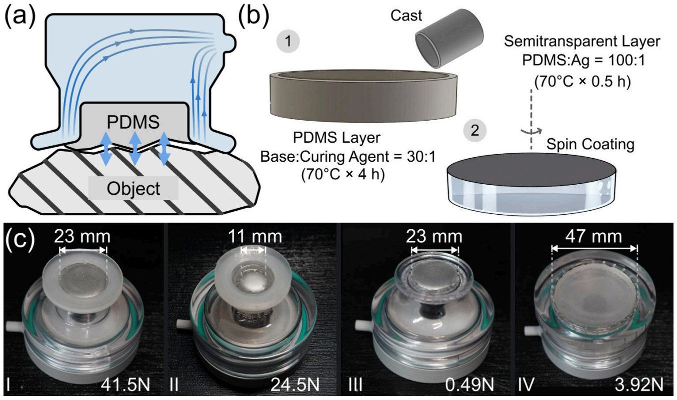

Fig. 5. PDMS membrane design and fabrication. (a) PDMS membrane deformation during contact. (b) Dual-layer membrane fabrication process. (c) Four modular configurations (I-IV).

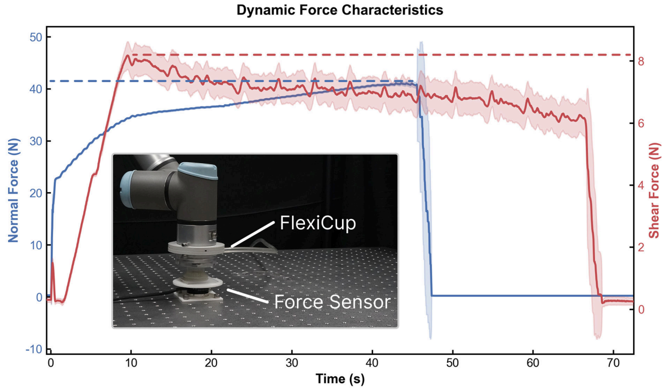

Fig. 6. Force characteristics from vacuum mode. Blue curve: normal force evolution during vertical pull-off from contact to detachment. Red curve: shear force evolution during horizontal drag from contact to detachment. Inset: experimental setup with FlexiCup mounted on robot arm and force sensor fixed on platform.

100:1, 70°C for 0.5 hours) providing the reflective surface for photometric tactile imaging. Four modular bottom configurations (I-IV, Fig. 5(c)) implement varying membrane diameters for vacuum operation. Configurations I-II are optimized for deformable objects and tactile sensitivity. Adhesion force ranges from sub-Newton to over 40 N across configurations. Configuration selection requires only bottom housing replacement while all sensing components remain unchanged.

F. Force Characteristics from Vacuum Mode

To characterize suction performance, we conducted automated force measurements using a 6-axis force/torque sensor on a smooth acrylic surface at -80 kPa vacuum pressure. The test protocol included normal pull-off tests measuring detachment force and horizontal drag tests measuring shear resistance, with each test repeated 20 times. The averaged force profiles (Fig. 6) reveal transient behaviors during attachment and detachment. Results demonstrate mean maximum normal force of 41.5 N and shear force of 8.34 N, exceeding theoretical predictions (F = P × A ≈ 33.2 N) due to structural compliance increasing the effective sealing area.

G. Multimodal Sensing Validation

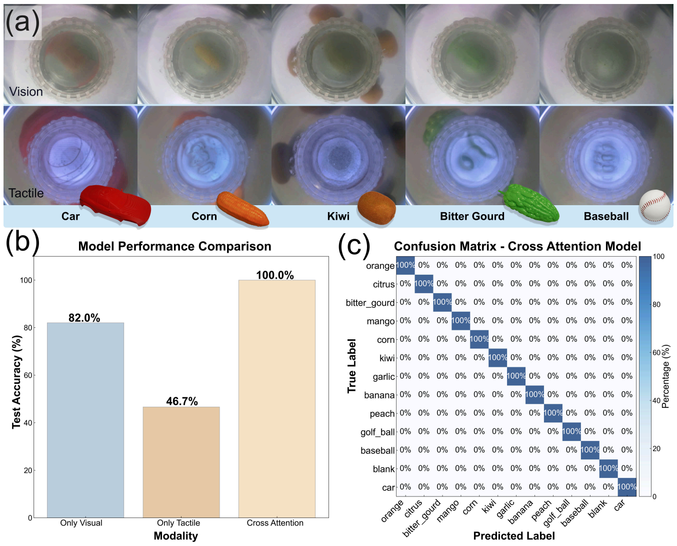

Fig. 7. Multimodal recognition validation showing (a) vision-tactile sensing comparison across objects, (b) accuracy comparison (vision-only: 82.5%, tactile-only: 46.7%, multi-head attention: 100%), and (c) confusion matrix.

We conducted object recognition experiments across thirteen objects (Fig. 7(a)). Using ImageNet-pretrained ResNet-18, vision-only achieves 82.5%, tactile-only 46.7%, and multi-head attention fusion 100% accuracy (Fig. 7(b)). The fusion model employs multi-head attention (8 heads, 512-d) trained on 1,300 paired samples across 13 categories with AdamW optimization. Results demonstrate that multimodal integration provides complementary information where vision captures global features and tactile reveals local contact details. The confusion matrix (Fig. 7(c)) confirms classification across all objects when both modalities are coordinated.

IV. Modular Perception-Driven Grasping

To validate that the dual-zone vision-tactile sensing architecture operates effectively across both vacuum and Bernoulli actuation principles, we conduct experiments using two control paradigms: modular perception-driven approach (this section) and learning-based policies (Section V). The modular approach demonstrates that both modes can leverage identical sensing pipelines and control logic with only pneumatic parameter adjustments, operating autonomously using FlexiCup's onboard dual-zone camera. All experiments use FlexiCup mounted on a UR3 robot arm (Fig. 8(a)).

A. Modular Control Framework

The modular perception pipeline leverages YOLOv8n for peripheral target detection and ResNet-34 for central tactile verification, demonstrating hardware versatility across classical planning methods. The YOLOv8n detector processes peripheral vision for target localization and workspace boundary detection. When LED illumination is activated, the ResNet-34 segmentation model (pretrained on ImageNet) analyzes the central tactile imprint to identify contact regions and verify surface flatness. With LED off, a YOLOv8n-seg model

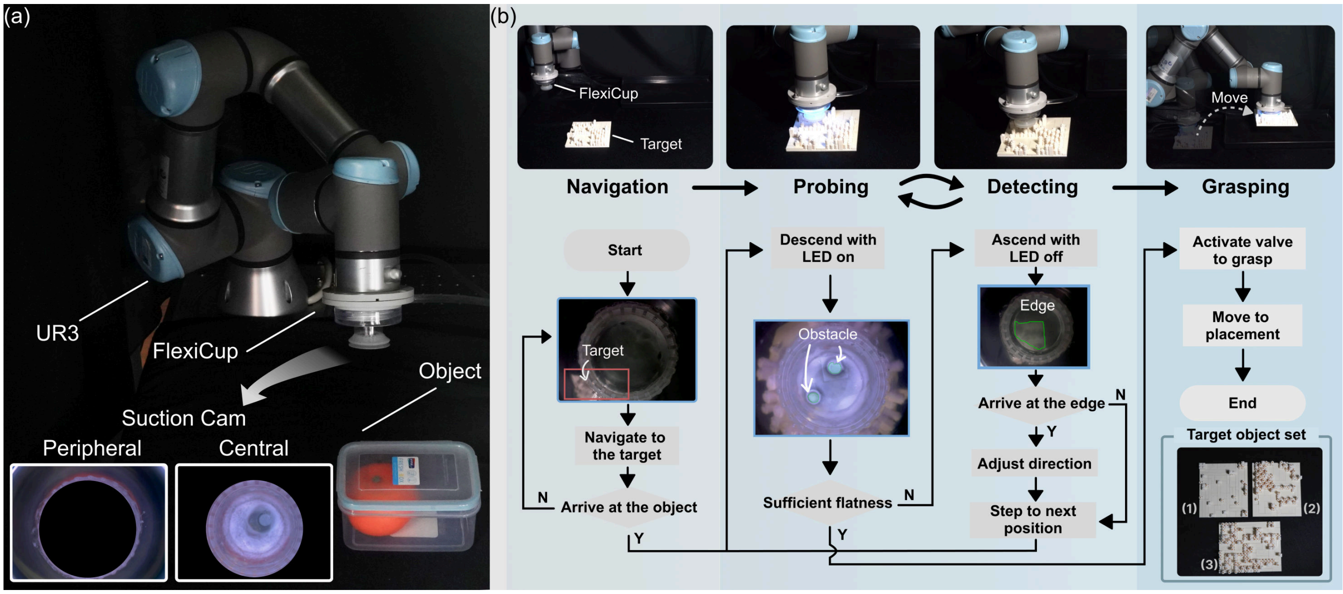

Fig. 8. Experimental setup and modular control framework for perception-driven grasping. (a) FlexiCup mounted on UR3 robot. (b) Control pipeline integrating YOLOv8 and ResNet-34 for dual-mode validation. Target object set (bottom right): (1)-(3) with 25%, 50%, 75% obstacle coverage.

performs edge detection on the central visual stream to guide obstacle avoidance during spatial search.

As illustrated in Fig. 8(b), the control pipeline integrates these modules with iterative feedback: peripheral vision detects the target and guides approach; upon arrival, tactile segmentation verifies contact and surface flatness; if unsuitable, the system transitions to vision mode and uses edge detection to guide spatial stepping; finally, valve activation initiates suction and transport.

B. Experimental Configuration

The validation employs identical sensing and control pipelines across both modes, with only bottom housing and pneumatic parameters varying. The three perception modules were trained offline on task-specific datasets: the peripheral vision detector achieved 87.7% mAP50 on 226 images, the tactile segmentation model achieved 98.3% IoU on 18,929 contact samples, and the edge detection module achieved 99.5% mAP50 on 1,954 images, all supporting real-time inference for closed-loop control.

Test materials include custom LEGO boards with 25%, 50%, 75% obstacle coverage. We conduct 30 trials per mode (vacuum and Bernoulli), totaling 180 trials, with success criteria including navigation to target, obstacle-free region identification, stable attachment verified by tactile feedback, and transport without object loss.

C. Results and Analysis

Table II shows vacuum (90.0%) and Bernoulli (86.7%) achieve comparable success rates, validating three key aspects: (1) Sensing universality—the dual-zone architecture operates effectively across contact-based and non-contact principles; (2) Control portability—the modular perception pipeline requires only pneumatic parameter adjustments without algorithmic modifications; (3) Mechanical modularity—bottom housing reconfiguration maintains full sensing and processing capabilities.

TABLE II
HARDWARE MODALITY VALIDATION: SUCCESS RATES ACROSS PNEUMATIC PRINCIPLES

Test Material	Vacuum	Bernoulli
(1) 25% obstacle coverage	90.0%	86.7%
(2) 50% obstacle coverage	93.3%	90.0%
(3) 75% obstacle coverage	86.7%	83.3%
Mean	**90.0%**	**86.7%**

The comparable success rates validate sensing-actuation decoupling: both modes achieve effective manipulation using the shared dual-zone architecture, with performance differences attributable to actuation mechanisms rather than sensing limitations.

Failure cases for both modes occur when the discrete search strategy (1 cm step size) exhaustively traverses the workspace without identifying suitable attachment regions. Success rates peak at moderate obstacle density (mean 91.7%) compared to low and high densities (88.4% and 85.0%), as scattered obstacles can fragment surfaces into regions smaller than the suction cup footprint, affecting both vacuum and Bernoulli modes comparably.

Bernoulli's slightly lower success rate (86.7% vs 90.0%) stems from reduced adhesion force during the contactless lifting phase: while the shared sensing pipeline successfully identifies suitable regions, the weaker aerodynamic force occasionally fails to maintain grasp on challenging configurations where vacuum's contact-based adhesion succeeds.

However, Bernoulli provides unique capabilities: a complementary semiconductor wafer handling experiment demonstrates that vacuum-picked wafers exhibit visible contact smudging (∼3.5 N peak contact force), while Bernoulli-picked wafers remain pristine with near-zero contact force, confirming contactless handling essential for delicate surfaces (see companion website).

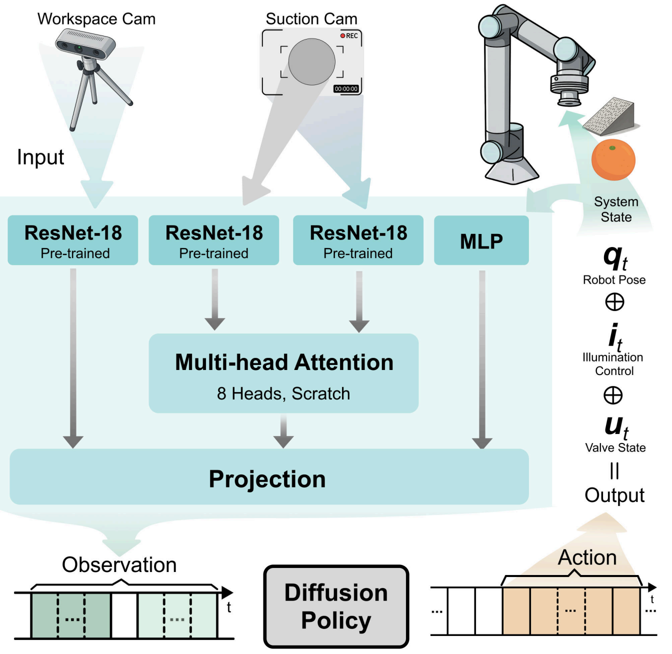

Fig. 9. Diffusion policy framework processing multimodal observations (workspace camera, peripheral view, central view, system state) through parallel encoders and multi-head attention to generate contact-aware control actions.

V. End-to-End Contact-Aware Manipulation

While the modular perception-driven approach validates hardware reconfigurability, it relies on discrete state transitions and manual threshold tuning. We develop a learning framework based on diffusion policies [28], [29] mapping multimodal observations to action trajectories.

A. Multi-Modal Learning Framework

We adopt the diffusion policy framework for sequential suction operations with vision-tactile modality transitions.

The learning experiments utilize vacuum mode (Configuration I) exclusively, as contact-aware policy learning requires continuous tactile feedback from sustained contact and rich membrane deformation. While Section IV demonstrated that the dual-zone sensing architecture operates effectively across both actuation principles, vacuum's sustained contact provides dense tactile gradients essential for learning fine-grained manipulation skills involving contact detection, surface compliance interpretation, and valve timing coordination. In contrast, Bernoulli's contactless lifting during suction minimizes membrane deformation, providing limited tactile information for learning contact dynamics despite supporting tactile verification during pre-suction positioning.

We extend the framework with multi-head attention to fuse dual-zone observations (Fig. 9), coordinating colocated heterogeneous sensory streams during vision-tactile switching.

B. Multi-Modal Observation Encoding

The system processes observations through specialized encoders. The workspace view $I_t^{workspace}$ from the D435 camera provides global scene context, encoded via ImageNet pre-trained ResNet-18 to $f_t^{workspace} \in \mathbb{R}^{512}$. The dual-zone suction camera captures local observations: the central view $I_t^{central}$ switches between visual and tactile modes for contact sensing, and the peripheral view $I_t^{peripheral}$ maintains spatial awareness, both encoded through ResNet-18 to $f_t^{central}, f_t^{peripheral} \in \mathbb{R}^{512}$. The system state $s_t \in \mathbb{R}^8$ is processed through a 2-layer MLP to f_t^{state}.

C. Feature Integration and Policy Learning

The multi-head attention module (8 heads, 512 dimensions) coordinates the central and peripheral views, correlating contact details with spatial context during approach-to-manipulation transitions. Integration proceeds through two stages: workspace features combine with attended suction features, then concatenate with system state to form the complete observation representation.

The diffusion policy generates action trajectories $a_t = [q_t, i_t, u_t]$ controlling robot joints ($q_t \in \mathbb{R}^6$), illumination switching (i_t), and pneumatic valve (u_t). The action chunking mechanism (8-step history, 48-step horizon) generates coordinated sequences for multi-phase operations.

D. Experimental Evaluation

We evaluate two tasks: inclined transport (150 demos) and orange extraction (100 demos), collected via kinesthetic teaching at 30 Hz and downsampled to 10 Hz, with randomized positions, angles, and orientations, and random cropping augmentation (76×76 from 224×224). Inclined transport involves positioning above the surface (7 cm × 5.3 cm × 6 cm), searching for suitable contact regions, adjusting tilt angle guided by tactile feedback to match the inclined surface (5°, 10°, or 15°), verifying contact, and performing secure lifting. Orange extraction consists of transparent cover removal (container: 15 cm × 10.5 cm × 7 cm) in vision mode, realignment above the orange (approximately 6 cm diameter), then tactile-guided grasping with LED-enabled contact detection.

Five configurations (Ours, w/o Multi-Head Attention, w/o Peripheral View, w/o Central View, Workspace Camera Only) and a BC-RNN baseline were trained for 500 epochs with batch size 16 on RTX4090 GPU, with 30 trials per task totaling 300 experiments (Fig. 10). These ablations functionally approximate representative prior architectures: Workspace Camera Only corresponds to conventional vision-guided suction without tactile feedback [10]; w/o Peripheral View approximates single-zone camera-based tactile suction systems [8], [9]; and w/o Central View simulates external-camera-only setups [2] lacking intimate contact observation. Success is defined as completing the full manipulation sequence without object loss, with primary failure modes: (1) object slipping, (2) search budget exhaustion, and (3) phase transition failures.

Table III shows the full system achieves 73.3% (inclined transport) and 66.7% (orange extraction). Ablation analysis

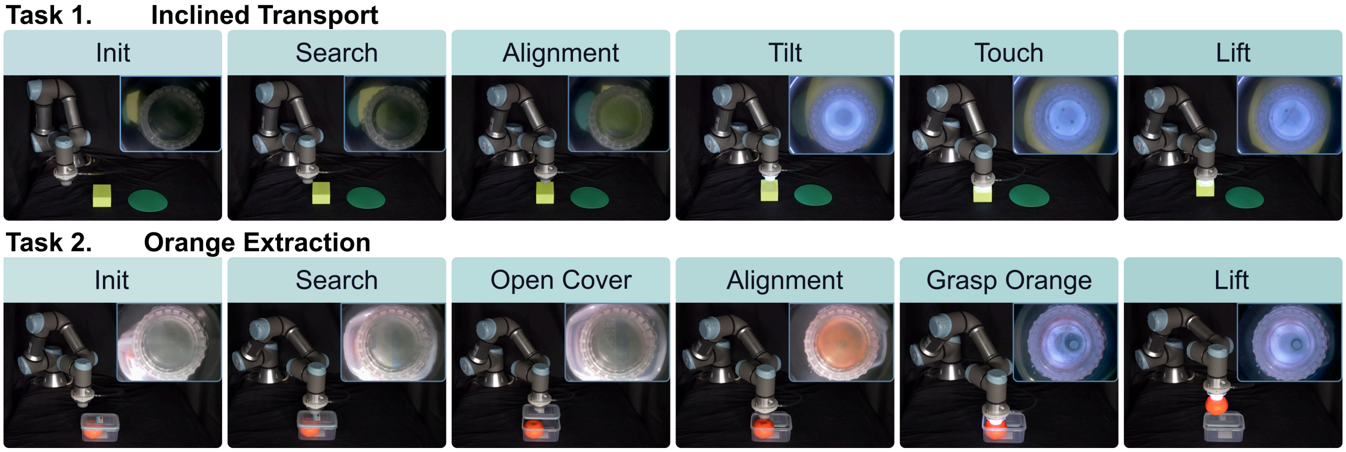

Fig. 10. Experimental task demonstrations showing inclined transport and orange extraction sequences with multimodal sensing integration.

TABLE III
TASK SUCCESS RATE COMPARISON WITH STATISTICAL ANALYSIS

Configuration	Inclined Transport	Orange Extraction
Ours	**73.3%**	**66.7%**
w/o Multi-Head Attention	60.0%	53.3%
w/o Peripheral View	43.3%	36.7%
w/o Central View	46.7%	33.3%
Workspace Camera Only	23.3%	0.0%
Baseline	Fail	Fail

reveals the central view as critical for contact detection, with its removal reducing performance to 46.7% and 33.3% due to loss of tactile sensing. The peripheral view contributes spatial context essential for approach planning, with its absence degrading performance to 43.3% and 36.7%. Multi-head attention provides 13% improvements by coordinating dual-zone information during vision-tactile transitions. Workspace camera alone achieves only 23.3% and 0.0% success, confirming that intimate sensing is necessary for contact-critical manipulation. The BC-RNN baseline failed both tasks (0% success), frequently becoming stuck or failing to coordinate modality switching—consistent with [28], where recurrent baselines exhibited similar stuck behaviors with multimodal action distributions, confirming the advantage of diffusion policy's action chunking for coordinated suction manipulation.

Failure mode analysis reveals distinct patterns: for configurations lacking peripheral sensing, 60% of failures stem from phase transition difficulties, while for the full system, 70% result from physical constraints rather than perceptual limitations, demonstrating that the multimodal architecture effectively addresses perceptual challenges.

VI. DISCUSSION

Object Limitations. The diameter of the suction cup bounds the manipulable object range. Objects smaller than approximately 15–20 mm fail to establish adequate sealing due to insufficient contact area, while the 41.5 N vacuum force capacity limits practical payload to approximately 2–3 kg under dynamic manipulation. The system requires a reasonably flat contact region with principal curvature radii exceeding 20–25 mm; highly curved, deeply concave, or highly porous surfaces prevent membrane sealing. Surface roughness exceeding approximately 1–2 mm feature height also causes consistent seal failure.

Wireless Electronics. While pneumatic connections remain, wireless electronics address fragile MIPI CSI cables and complex slip rings by keeping the signal path (CMOS to MCU) onboard. A 30-minute streaming test confirms stable performance: 27.05 FPS, 43 ms latency, compatible with pneumatic seal dynamics (200–500 ms).

Material and Surface Constraints. The PDMS membrane is subject to gradual degradation through repeated contact, particularly with abrasive or sharp-edged objects. Over 500 pick-and-place cycles in our experiments showed no observable degradation, but industrial deployment would require periodic inspection. Additionally, material selection (membrane hardness, tear resistance) and surface preprocessing for porous regions remain important directions for extending the operational envelope.

Dynamic Performance. Dynamic stress tests at 3.0 rad/s joint speed and 3.0 rad/s² acceleration demonstrate stable suction and sensing performance while holding objects with varying mass distributions (orange, water-filled bottle with sloshing dynamics; see companion website).

Sensing and Modality Constraints. High-frequency switching is limited by CMOS stabilization during illumination changes, addressed through (1) finite-state machines switching at task boundaries, or (2) learned switching heuristics from demonstrations. The semi-transparent membrane is also sensitive to ambient lighting; high-intensity internal LEDs overpower external interference in tactile mode, but uncontrolled lighting can occasionally wash out tactile features.

Dual-Mode Complementarity. As demonstrated in Section IV, vacuum and Bernoulli modes share identical sensing pipelines but serve complementary domains: vacuum provides sustained contact for tactile-rich learning (Section V), while Bernoulli enables contactless handling for delicate objects (wafer experiment, companion website).

VII. Conclusions

This paper presented FlexiCup, a multimodal suction cup with wireless electronics integrating dual-zone vision-tactile sensing through illumination-controlled modality switching. The hardware supports dual suction modes with complementary actuation mechanisms—vacuum for sustained-contact adhesion and Bernoulli for contactless lifting—both sharing the identical dual-zone sensing architecture.

Hardware versatility was validated through dual control paradigms. Modular perception-driven grasping achieves 90.0% (vacuum) and 86.7% (Bernoulli) success rates, confirming that the dual-zone sensing architecture operates effectively across both actuation principles. Performance differences stem from actuation mechanisms rather than sensing limitations, with Bernoulli's reduced adhesion occasionally leading to object loss while offering unique contactless handling for delicate surfaces, as demonstrated through wafer experiments. End-to-end learning based on diffusion policies achieves 73.3% and 66.7% success on contact-aware manipulation tasks, with ablations confirming 13% improvements from multi-head attention. The integration enables continuous contact-aware manipulation adapting to surface variations without manual threshold engineering.

Current limitations and their implications are discussed in Section VI. The hardware design files, firmware source code, and experimental videos have been made publicly accessible at the companion website. Future work will focus on extending the framework to multi-contact scenarios, investigating computational optimization for embedded hardware, integrated flow-switching mechanisms for online mode reconfiguration, calibrated force sensing to complement the current deformation-based tactile observations, and integrating richer tactile modalities for adaptive suction control.

References

[1] T. M. Huh, K. Sanders, M. Danielczuk, M. Li, Y. Chen, K. Goldberg, and H. S. Stuart, "A multi-chamber smart suction cup for adaptive gripping and haptic exploration," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 1786–1793.

[2] J. Lee, S. D. Lee, T. M. Huh, and H. S. Stuart, "Haptic search with the smart suction cup on adversarial objects," *IEEE Transactions on Robotics*, vol. 40, pp. 226–239, 2024.

[3] P. Davis, B. Gray, and K. Caldwell, "An end effector based on the Bernoulli principle for handling sliced fruit and vegetables," *Robotics and Computer-Integrated Manufacturing*, vol. 24, no. 2, pp. 249–257, 2008.

[4] S. Liu, X. Wang, Y. Zhang, and H. Chen, "Design and tests of a noncontact Bernoulli gripper for rough-surfaced and fragile objects gripping," *Engineering Science and Technology, an International Journal*, vol. 23, no. 4, pp. 729–738, 2020.

[5] Y. Yoo, J. Eom, M. J. Park, and K.-J. Cho, "Compliant suction gripper with seamless deployment and retraction for robust picking against depth and tilt errors," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1492–1499, 2023.

[6] S. Doi, H. Koga, T. Seki, and Y. Okuno, "Novel proximity sensor for realizing tactile sense in suction cups," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 638–643.

[7] S. Aoyagi, M. Suzuki, T. Morita, T. Takahashi, and H. Takise, "Bellows suction cup equipped with force sensing ability by direct coating thin-film resistor for vacuum type robotic hand," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 5, pp. 2501–2512, Oct. 2020.

[8] S. van Veggel, M. Wiertlewski, E. L. Doubrovski, A. Kooijman, B. Mazzolai, and R. B. N. Scharff, "Optoelectronically innervated suction cup inspired by the octopus," *Advanced Intelligent Systems*, vol. 7, no. 4, p. 2400544, 2025.

[9] R. Yuan, J. Ren, Z. Peng, F. Chen, and G. Gu, "SuckTac: Camera-based tactile sucker for unstructured surface perception and interaction," arXiv preprint arXiv:2511.02294, 2025.

[10] C. Wang, J. van Baar, C. Mitash, S. Li, D. Randle, W. Wang, S. Sontakke, K. E. Bekris, and K. Katyal, "Demonstrating multi-suction item picking at scale via multi-modal learning of pick success," arXiv preprint arXiv:2506.10359, 2025.

[11] Z. Zhao, Y. Liu, W. Chen, and K. Zhang, "PolyTouch: A robust multi-modal tactile sensor for contact-rich manipulation using tactile-diffusion," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[12] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Beyond Sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding," arXiv preprint arXiv:2501.04693, 2025.

[13] Z. Zhao, W. Li, Y. Li, T. Liu, B. Li, M. Wang, K. Du, H. Liu, Y. Zhu, Q. Wang, K. Althoefer, and S.-C. Zhu, "Embedding high-resolution touch across robotic hands enables adaptive human-like grasping," *Nature Machine Intelligence*, pp. 1–12, 2025.

[14] P. Lin, Y. Huang, W. Li, J. Ma, C. Xiao, and Z. Jiao, "PP-Tac: Paper picking using tactile feedback in dexterous robotic hands," arXiv preprint arXiv:2504.16649, 2025.

[15] M. Wang, Y. Zhou, J. Luo, and S. Wang, "Large-scale deployment of vision-based tactile sensors on multi-fingered grippers," arXiv preprint arXiv:2405.08959, 2024.

[16] E. Del Bianco, D. Torielli, F. Rollo, D. Gasperini, A. Laurenzi, L. Baccelliere, L. Muratore, M. Roveri, and N. G. Tsagarakis, "A high-force gripper with embedded multimodal sensing for powerful and perception driven grasping," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, 2024, pp. 149–156.

[17] E. Shahabi, F. Visentin, A. Mondini, and B. Mazzolai, "Octopus-inspired suction cups with embedded strain sensors for object recognition," *Advanced Intelligent Systems*, vol. 5, no. 2, p. 2200201, 2023.

[18] T. Yue, C. Lu, K. Tang, Q. Qi, Z. Lu, L. Y. Lee, H. Bloomfield-Gadêlha, and J. Rossiter, "Embodying soft robots with octopus-inspired hierarchical suction intelligence," *Science Robotics*, vol. 10, no. 102, p. eadr4264, 2025.

[19] N. Jang, M. W. Lee, and D. Hwang, "A suction-based peripheral nerve gripper capable of controlling the suction force," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 4, pp. 1867–1876, Aug. 2021.

[20] T. Shiratori, Y. Kanazawa, J. Sakamoto, T. Takahashi, M. Suzuki, and S. Aoyagi, "Development of a vision-based tactile sensor with micro suction cups," *Sensors and Actuators A: Physical*, vol. 371, p. 115276, 2024.

[21] W. Yuan, S. Dong, and E. H. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[22] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, "DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[23] D. Ma, E. Donlon, S. Dong, and A. Rodriguez, "Dense tactile force estimation using GelSlim and inverse FEM," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5418–5424.

[24] T. Motoda, T. Kitamura, R. Hanai, and Y. Domae, "SuctionPrompt: Visual-assisted robotic picking with a suction cup using vision-language models and facile hardware design," arXiv preprint arXiv:2410.23640, 2024.

[25] A. Patel, D. Kumar, S. Lee, and J. Brown, "Deep reinforcement learning for tactile robotics: learning to type on a braille keyboard," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5678–5685.

[26] X. Song, Y. Li, Y. Zhang, Y. Liu, and L. Jiang, "An overview of learning-based dexterous grasping: recent advances and future directions," *Artificial Intelligence Review*, vol. 58, no. 10, pp. 1–44, 2025.

[27] S. An, Z. Meng, C. Tang, Y. Zhou, T. Liu, F. Ding, S. Zhang, Y. Mu, R. Song, W. Zhang, Z. Hou, and H. Zhang, "Dexterous manipulation through imitation learning: A survey," arXiv preprint arXiv:2504.03515, 2025.

[28] 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*, 2023.

[29] Z. Xue, S. Deng, Z. Chen, Y. Wang, Z. Yuan, and H. Xu, "DemoGen: Synthetic demonstration generation for data-efficient visuomotor policy learning," arXiv preprint arXiv:2506.08680, 2025.