Hardware-Software Co-Design of a Dual-Vision, Infinite-Roll Hand for Edge-Native Grasping

Mingyu Li[1,*], Zhixian Gao[1], Zhangchi Guo[1]

***Abstract*— Conventional robotic grasping in semi-structured service and workstation environments is often constrained by two physical bottlenecks: visual blind spots in rearward coverage and the bounded rotational range of traditional wrists. This paper presents an edge-native hand-wrist platform that addresses these coupled limitations through electromechanical co-design. We integrate a dorsal-ventral dual-camera relay for complementary front-rear perception with a direct-drive slip-ring mechanism enabling infinite continuous wrist roll. To support real-time operation on resource-constrained edge processors, the system features a reparameterized zero-shot detector that supports language-guided grasping while removing online text-encoding overhead via offline feature reparameterization. Kinematic and spatial validations in a predefined workstation-style setup indicate that the integrated platform can execute task-prior-driven grasp acquisitions for both forward and rearward targets entirely on the edge, without additional task-specific fine-tuning or cloud offloading.**

I. Introduction

Dexterous robotic manipulation in semi-structured service and workstation environments often requires front-rear target acquisition around the end-effector [1], [2]. While recent advancements in data-driven grasp synthesis [3], [4] and continuous control [5] have remarkably enhanced forward-facing manipulation, prevailing arm-hand systems remain fundamentally constrained when attempting to acquire targets located posterior to the wrist joint. This multi-directional manipulation scenario exposes intrinsic hardware topology limitations that cannot be adequately resolved through algorithmic compensation alone.

More precisely, two fundamental physical constraints inherently circumscribe the grasping envelope. First, conventional eye-in-hand perception exhibits an intrinsic topological deficiency: a palm-mounted (ventral) camera provides exclusively forward-facing visual coverage [6]. This induces a complete perceptual void in the rearward hemisphere—a structural absence in the sensor's field-of-view (FOV) that often necessitates complex and inefficient visibility maximization maneuvers [7]. Second, standard wrist actuators enforce strict rotational bounds. Acquiring a rearward object necessitates a wrist rotation of approximately $180°$. When the requisite joint displacement exceeds the mechanical limit, an irrecoverable kinematic deadlock or singularity ensues [8], [9], rendering the target physically inaccessible irrespective of the supervisory control algorithms employed [10].

Prevailing paradigms predominantly attempt to address these physical constraints through software-level remediation. However, external multi-camera configurations compromise system self-containment and portability—attributes essential for field-deployable manipulation. View synthesis approaches [11] similarly prove inadequate for this specific task, as they presuppose an initial visual anchor from the very hemisphere that is structurally absent. Analogously, while sophisticated inverse kinematic solvers can optimize trajectories to circumvent singularities [8], [9], they are mathematically incapable of expanding the inherent reachable workspace delineated by fixed mechanical joint bounds.

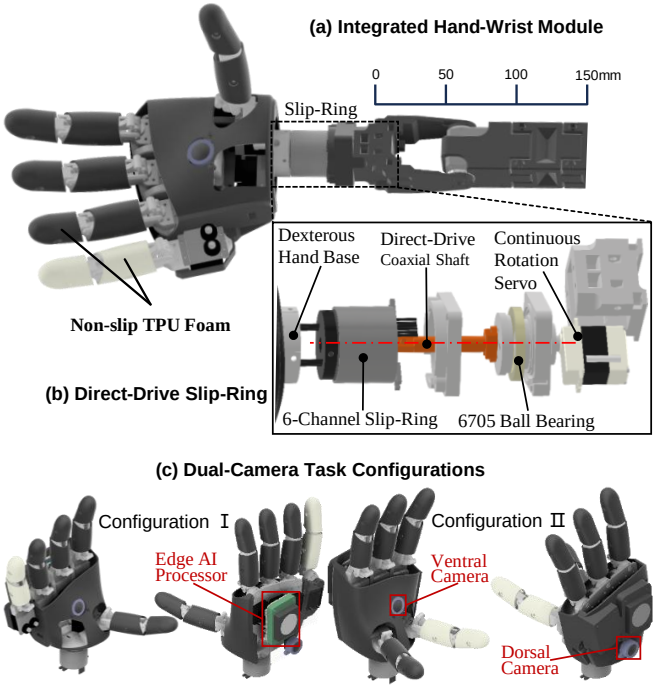

Fig. 1. Hardware architecture of the edge-native hand-wrist platform. **(a) Integrated Hand-Wrist Module:** The dexterous hand mounted on the distal forearm stub, highlighting the decoupled J1 joint. **(b) Direct-Drive Slip-Ring:** A magnified, exploded cross-sectional view of the wrist, detailing the 6-channel slip-ring and coaxial shaft that physically enable continuous rotation. **(c) Dual-Camera Task Configurations:** Configuration I integrates the AI Processor natively on the palm, while Configuration II highlights the dorsal (rear-view) and ventral (front-view) visual sensors for complementary front-rear coverage.

Furthermore, deploying a language-guided grasping pipeline on self-contained edge devices introduces a critical computational constraint. Conventional closed-set detectors [12], [13], while amenable to real-time execution via quantization [14], are strictly confined to predefined training categories, thereby precluding generalization to novel objects. Conversely, state-of-the-art Open-Vocabulary Detectors (OVD) afford zero-shot generalization [15]–[17] by

[1]M. Li, Z. Gao, and Z. Guo are with Tongji University, Shanghai, China.
*Corresponding author: 2352659@tongji.edu.cn.

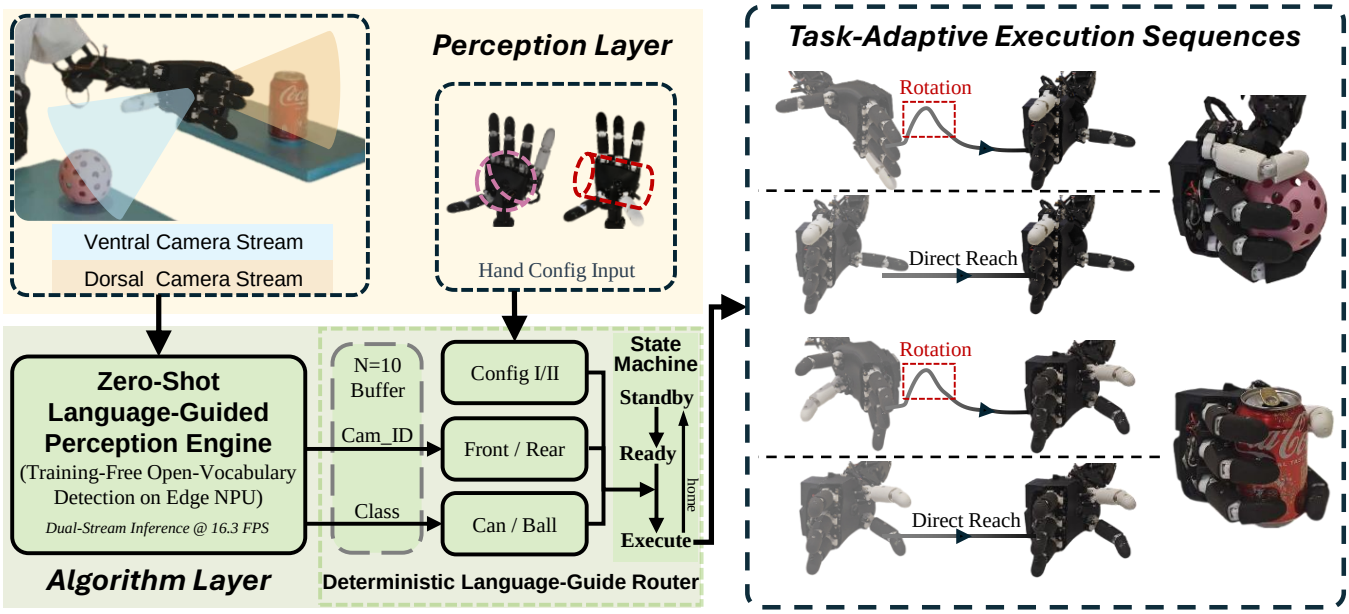

Fig. 2. System architecture of the proposed hardware-software co-designed grasping platform. The **Perception Layer** captures dual-camera visual streams and accepts a predefined task prior (Hand Config). Within the **Algorithm Layer**, an edge-native zero-shot language-guided engine processes the streams at 16.3 FPS to extract spatial (`Cam_ID`: ventral/front or dorsal/rear) and semantic (`Class`) cues. A deterministic language-guided router filters these inputs via an $N = 10$ consensus buffer and fuses them with the task prior to drive a finite state machine (Standby → Ready → Execute). This decoupling routes the manipulator toward either direct ventral reaches or dorsal-triggered continuous rotations using pre-calibrated grasping primitives.

leveraging large-scale vision-language models [18]. However, they necessitate concurrent image and text encoding at inference time. The computational burden of executing high-capacity text encoders online remains prohibitive for resource-constrained edge neural processing units (NPUs) [19].

To address these intertwined challenges, this paper proposes a hardware-software co-design framework that mitigates these bottlenecks through coordinated mechanical and computational design. By introducing a dual-camera relay architecture (ventral/front and dorsal/rear) and integrating a conductive slip-ring for infinite continuous wrist rotation—advancing beyond recent infinite-twist mechanisms [20], [21]—the rearward workspace becomes directly observable and mechanically accessible. Concurrently, a reparameterized [22] language-guided perception engine [23]—requiring no task-specific training—is deployed as a purely visual pipeline on the edge NPU through offline injection of text priors into the convolutional backbone. Algorithmically, this work does not claim a new VL-PAN module, a new structural reparameterization scheme, a new YOLO-family detector, or a new INT8 quantization method. Rather, to the best of our knowledge, the contribution lies in the first system-level integration and edge deployment of these established components within a dual-view, infinite-roll dexterous hand, such that semantic cues and camera provenance jointly support front-rear language-guided grasping on a resource-constrained NPU. The resulting system constitutes a self-contained dexterous hand prototype for front-rear perception and language-guided grasping under real-time edge constraints.

The principal contributions of this work are threefold:

- A dual-camera relay topology that furnishes complementary forward and rearward visual coverage, thereby substantially reducing the perceptual blind spot inherent to single-camera eye-in-hand configurations.
- An infinite-roll wrist mechanism enabled by a conductive J1 slip-ring, which effectively mitigates kinematic deadlocks and permits direct flip-over grasping without recourse to complex, latency-inducing motion replanning.
- An edge-native language-guided grasping pipeline—requiring no task-specific fine-tuning—that uses zero-shot semantic cues and deterministic spatial routing to support task-prior-driven execution on a resource-constrained NPU.

II. Related Work

A. Visual Sensing for In-Hand Manipulation

Eye-in-hand camera configurations, rooted in the classical visual servoing paradigm [6], have become ubiquitous in dexterous manipulation due to their self-contained nature and freedom from external infrastructure [24], [25]. However, these systems inherently operate under the assumption that the end-effector's workspace aligns with the camera's limited FOV. Consequently, tasks requiring rearward reach suffer from severe visual occlusion and a complete absence of perceptual feedback. While external multi-view setups [26] can mitigate this, they compromise the portability essential for mobile deployment; similarly, view synthesis approaches [11] require an initial visual anchor from the very hemisphere that is structurally absent. Recent visibility-aware controllers [7] mitigate self-occlusions caused by the robot's own body, yet they cannot recover information from a sensor topology that never existed. Tactile sensors [27]

provide localized contact data but are insufficient for pre-grasp approach planning. To the best of our knowledge, no prior work has addressed this fundamental topological deficiency by integrating complementary, multi-directional cameras directly onto the dexterous hand chassis.

B. Joint Limits and Workspace Boundaries

Conventional approaches to kinematic joint limits rely predominantly on software-level singularity avoidance, employing techniques such as damped least-squares inverse kinematics (IK) [8], gradient projection [10], or redundancy resolution [9]. While efficacious within the existing mechanical envelope, these algorithmic strategies incur non-trivial computational overhead and remain fundamentally incapable of physically expanding the reachable workspace. When a target configuration resides strictly outside the joint boundaries, kinematic deadlocks become mathematically unavoidable. The mechanical design of dexterous hands has constituted a longstanding research challenge [2]. Seminal platforms such as the DLR Hand Arm System [28], the Utah/M.I.T. Dextrous Hand [29], and more recent compliant architectures including the Pisa/IIT SoftHand [30] have substantially advanced multi-DOF manipulation capabilities, yet their wrist joints remain constrained by conventional mechanical bounds. Although continuous-rotation mechanisms have been investigated in underactuated grippers [20] and cable-driven parallel robots [21], the integration of a conductive slip-ring into the wrist joint of a multi-fingered dexterous hand for front-rear dexterous grasping remains an open problem. The present work reconceptualizes joint limits not as immutable algorithmic constraints but rather as modifiable hardware design variables amenable to electromechanical intervention.

C. Open-Vocabulary Detection on Edge Devices

Extending the real-time detection paradigm established by YOLO [12], OVDs such as YOLO-World [23], Grounding DINO [15], and OWL-ViT [16] have broadened object recognition beyond predefined training categories by leveraging vision-language foundation models such as CLIP [18]. Nevertheless, their canonical two-stream architecture necessitates concurrent visual and textual encoding at inference time. This online text-processing overhead renders deployment on resource-constrained edge NPUs computationally intractable. While recent advances in model compression and quantization [17], [19], lightweight architectures tailored for mobile inference [13], efficient reparameterized network topologies [31], and integer-only quantization schemes [14] have reduced inference latency, these approaches typically retain the online text processing pathway. An edge-native deployment methodology that entirely eliminates the text encoder through offline feature reparameterization remains insufficiently explored within real-time robotic grasping systems.

III. Approach

This section presents the proposed co-design methodology addressing three fundamental bottlenecks: perception topology (Sec. III-A), kinematic joint limits (Sec. III-B), and edge-native language-guided detection (Sec. III-C). For each constraint, we provide a formal characterization and derive a minimal hardware-software co-design intervention to resolve it.

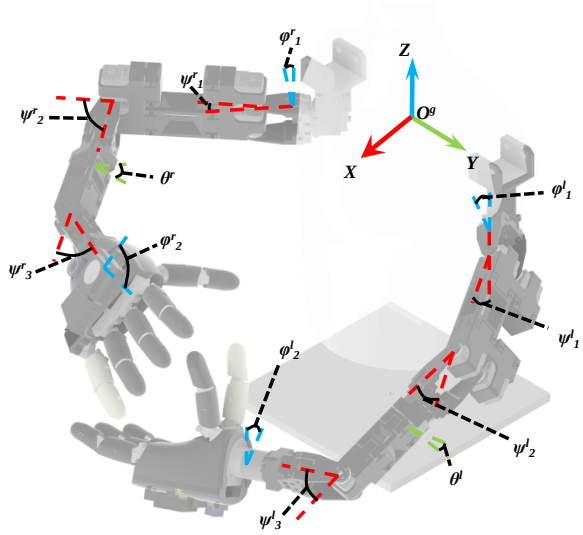

Fig. 3. Kinematic articulation framework of the edge-native manipulation platform. The global reference frame $\{O^g\}$ is anchored to the torso base. Spatial joint variables are denoted by φ, ψ, and θ, representing the yaw, pitch, and roll axes, respectively, with superscripts l and r indicating the left and right manipulators. Crucially, the distal wrist joints (denoted by θ^l and θ^r) encapsulate the direct-drive slip-ring mechanism. Unlike the strictly bounded proximal joints, these distal roll axes are mechanically decoupled to execute continuous rotation, supporting expanded front-rear target acquisition.

A. Perception Topology Completion

Let H denote the local coordinate frame of the dexterous hand, and $V \subset S^2$ represent the set of directional vectors observable by the hand-mounted visual sensors. In a conventional eye-in-hand configuration equipped with a single ventral camera, the observable space $V = V_{ven}$ strictly covers the frontal hemisphere. Consequently, the rearward workspace W_{rear} remains entirely unobservable, yielding

$$W_{rear} \cap V_{ven} = \emptyset. \tag{1}$$

This limitation is purely topological: a physical void in which no photon from the rearward workspace reaches the sensor array. Unlike dynamic environmental occlusions, such a structural blind spot cannot be compensated by any downstream algorithmic inference.

Integrating a dorsal camera on the hand chassis constitutes a necessary and sufficient topological completion. It is necessary because, absent a rearward-facing sensor, zero-order observations are fundamentally unattainable. It is sufficient because a dorsal camera with a field of view $V_{dor} \supseteq W_{rear}$ yields a comprehensive perceptual sphere:

$$V = V_{ven} \cup V_{dor} \supseteq W_{front} \cup W_{rear}. \tag{2}$$

In the present implementation, each camera provides a 206° diagonal FOV (Table I), yielding substantial overlap between

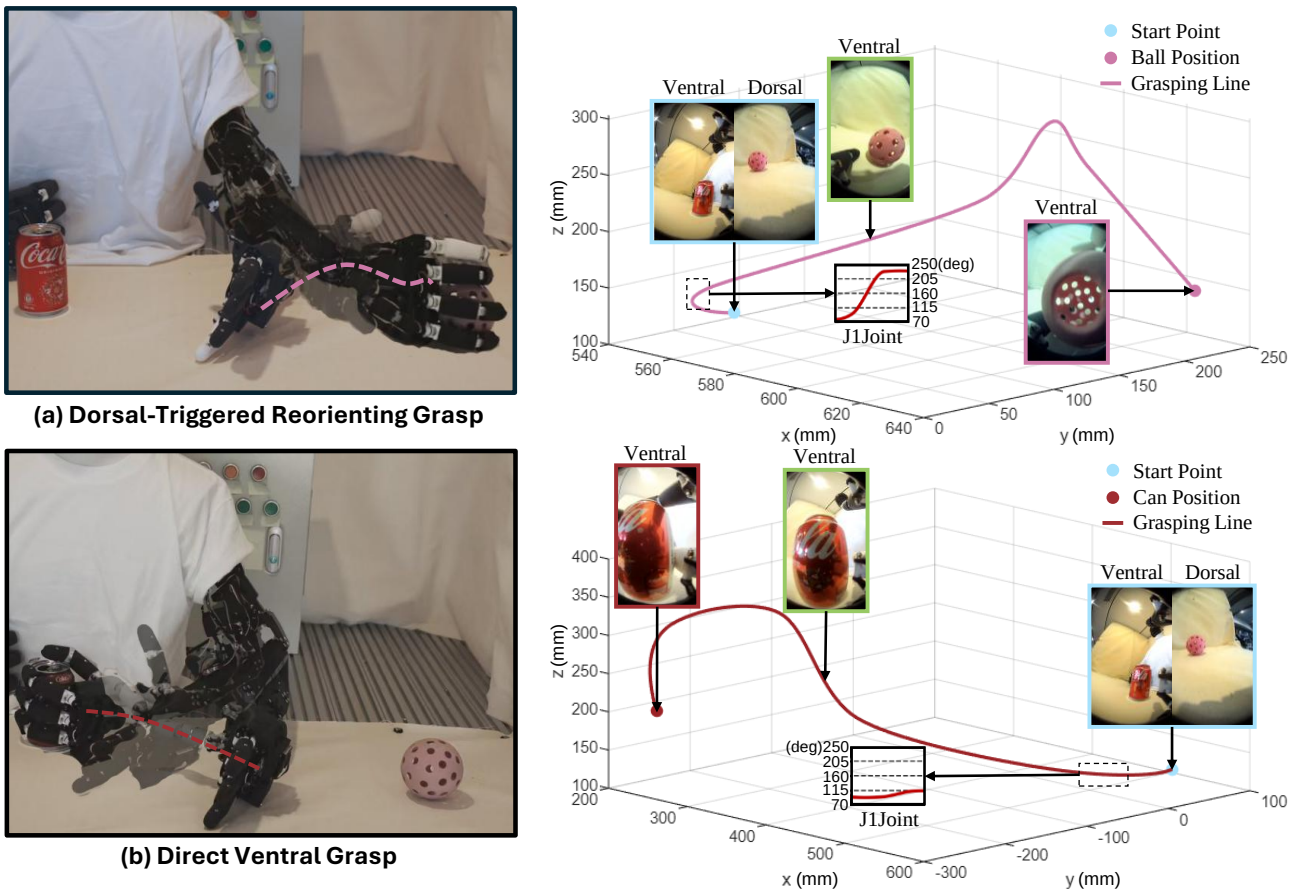

Fig. 4. Integrated spatial and kinematic validation of the edge-native grasping system executed under predefined task priors. **(a) Dorsal-Triggered Reorienting Grasp:** The target (pink sphere) is detected in the rearward hemisphere. The 3D Cartesian trajectory maps the end-effector's spatial path, while the magnified kinematic inset captures the continuous 180° J1 roll executed to physically unbind the workspace. **(b) Direct Ventral Grasp:** The target (Coke can) is authenticated in the frontal hemisphere, triggering a standard forward reaching execution with minimal J1 variation. Across both sequences, the aligned dual-camera streams confirm the robust visuo-spatial routing and deterministic state machine execution processed natively on the edge NPU.

V_{ven} and V_{dor} and thereby satisfying the sufficiency condition in practice.

Crucially, both cameras and the edge processor are rigidly co-located on the distal side of the wrist slip-ring, rotating synchronously with the hand. This co-rotation eliminates the need to transmit high-bandwidth video across the rotary joint; only low-bandwidth power and serial-bus signals traverse the six slip-ring channels (Table I), exemplifying the tight electromechanical coupling central to the proposed co-design.

Remark 1*: (Necessity of Co-Rotation)* The world-frame FOV transforms as ${}^{W}V(q_1) = \mathbf{R}_z(q_1)\,{}^{H}V$, where $\mathbf{R}_z(q_1) \in SO(3)$ denotes the wrist roll rotation. Had the cameras been mounted proximally on the forearm, ${}^{H}V$ would decouple from the hand's manipulation frame during a 180° flip, recreating the original perceptual gap. Co-rotation ensures that V_{ven} and V_{dor} remain invariant w.r.t. the hand frame H under arbitrary wrist configurations. This invariance is what permits the language-guided router (Sec. III-C) to treat the camera source label as a reliable spatial proxy without requiring online extrinsic recalibration.

B. Joint-Space Unbounding

Standard robotic wrists impose strict mechanical roll limits, typically bounded within a feasible set $\mathcal{Q}_{bound} = \{q_1 \in \mathbb{R} \mid -\alpha \leq q_1 \leq \alpha\}$. Executing a rearward flip-over grasp necessitates a rotational displacement of approximately $\theta_{flip} \approx 180°$ from a given initial pose $q_{1,init}$.

Proposition 1. *For a revolute joint constrained by $\mathcal{Q}_{bound}$, a target flip configuration $q_1^* = q_{1,init} + \theta_{flip}$ is strictly unreachable if $\theta_{flip} > \alpha - q_{1,init}$.*

Proof: Kinematic feasibility dictates that the target configuration must satisfy $q_1^* \in \mathcal{Q}_{bound}$, requiring $q_1^* \leq \alpha$. Substituting the target expression yields $q_{1,init} + \theta_{flip} \leq \alpha$, which simplifies to:

$$\theta_{flip} \leq \alpha - q_{1,init}. \tag{3}$$

For any commercially typical wrist whose total range satisfies $2\alpha < 360°$, there necessarily exist feasible initial poses $q_{1,init} \in \mathcal{Q}_{bound}$ from which a required displacement $\theta_{flip} \approx 180°$ deterministically violates this bound. The resulting impossibility is inherently geometric, not algorithmic.

To resolve this kinematic deadlock, a conductive slip-ring is integrated into the wrist joint, eliminating the mechanical roll constraint and yielding an unbounded configuration space: $q_1 \in (-\infty, +\infty)$.

Crucially, this mechanism is rendered practically viable by the adopted serial bus-servo architecture. Since all distal

digit actuators communicate over a single serial bus, only six slip-ring channels—carrying two supply rails, their returns, and a shared serial bus (Table I)—traverse the rotary joint. Had independent PWM control been adopted, the requisite channel count would have exceeded any compact slip-ring, rendering infinite rotation unviable. Thus, early communication topology choices directly relieve downstream mechanical constraints. To physically realize this without backlash-prone gear transmissions, a direct-drive coaxial shaft couples the J1 bus servo to the inner rotor of the slip-ring, as illustrated in the exploded inset of Fig. 1(b).

C. Edge-Native Zero-Shot Language-Guided Grasping

With complementary front-rear visual coverage and continuous wrist-roll capability established, the remaining requirement is a perception pipeline that can support language-guided grasping for novel target categories without fine-tuning while satisfying real-time inference constraints on the edge NPU. Here, "language-guided grasping" refers to using semantic cues to authenticate a target and trigger a corresponding pre-calibrated grasping primitive under a predefined task prior, rather than synthesizing arbitrary grasps end-to-end. We adopt a "train-multimodal, infer-unimodal" paradigm leveraging a lightweight open-vocabulary detector. During offline preparation, a frozen CLIP text encoder transforms semantic prompts into dense embeddings, which are fused into a Vision-Language Path Aggregation Network (VL-PAN) [32] via structural weight reparameterization [22]. This effectively encodes semantic priors directly into the convolutional weights, yielding a purely visual, text-free network for online inference. The resulting pipeline comprises three stages: a lightweight YOLOv8s visual backbone whose VL-PAN neck incorporates the language priors via reparameterization, a set of hardware-aware operator adaptations for NPU compatibility (detailed below), and a decoupled detection head that emits zero-shot class predictions and bounding boxes entirely without online text processing.

a) Hardware-Aware Operator Adaptation: Deploying the open-vocabulary detector onto the resource-constrained RK3588 NPU necessitates replacing the network's pervasive SiLU activation function:

$$f(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}}, \tag{4}$$

with the hardware-native ReLU:

$$g(x) = \max(0, x). \tag{5}$$

While SiLU provides a smooth non-linearity, its exponential component inherently mandates floating-point arithmetic. On integer-only edge NPUs, evaluating Eq. (4) requires complex Look-Up Table (LUT) approximations that substantially degrade parallel throughput. Conversely, ReLU requires only a sign-bit comparison. Critically, this operator replacement is not applied in isolation; it is performed jointly with the offline weight reparameterization described above, during which the network structure is already being reorganized to absorb CLIP embeddings. The restructured backbone then undergoes post-training quantization (PTQ) to INT8 precision via symmetric linear quantization:

$$X_{\text{int8}} = \text{round}\left(\frac{X_{f32}}{S}\right), \tag{6}$$

where scale factors S are calibrated on a small, task-agnostic calibration set (e.g., generic COCO images) to account for the modified activation profile; no target-domain data is required. Across this combined pipeline—reparameterization, operator adaptation, and calibrated quantization—the measured end-to-end cost is a 1.5 percentage point mAP degradation on the COCO validation set, offset by a $2.4\times$ improvement in inference throughput on the RK3588 NPU. Through this joint elimination of online text encoding and NPU-incompatible operators, the engine is compressed into a deterministic $\sim$15 MB INT8 module achieving 16.3 FPS natively.

b) Orthogonal Decoupling of the Decision Space: The dual-camera architecture naturally decouples the workstation task logic into two orthogonal dimensions. **Path Selection (Where):** Detections from the ventral/front and dorsal/rear cameras trigger the Forward Path and Flip-over Path, respectively. **Grasp Configuration (What):** Rather than relying on unconstrained algorithmic morphological synthesis, the specific hand configuration (e.g., Config I or II) is provided to the system as a predefined task prior. The semantic output from the detector is utilized to authenticate the target entity, while the extracted spatial source triggers the corresponding kinematic routing.

This orthogonal decoupling ensures spatial planning and morphological configuration remain conceptually disentangled. An N-frame consensus buffer ($N = 10$) filters transient false positives: a detection is deemed *valid* only after N

Algorithm 1: Edge-Native Dual-Camera Language-Guided Routing

Input: Ventral/front-view I_v, dorsal/rear-view I_d, detector $\mathcal{M}$, task prior C_{task}

Initialize slip-ring wrist joint $q_1 \leftarrow 0$;
State $\leftarrow$ STANDBY;
while *State* = STANDBY **do**
  $D_v \leftarrow \mathcal{M}(I_v)$; // Front view
  $D_d \leftarrow \mathcal{M}(I_d)$; // Rear view
  if D_v *is valid* ***or*** D_d *is valid* **then**
    State $\leftarrow$ READY ; // Target locked
    if D_v *is valid* **then**
      Source $\leftarrow$ FRONT ; // Front path
    else
      Source $\leftarrow$ REAR ; // Rear path
    end
    State $\leftarrow$ EXECUTE;
    if Source $=$ REAR **then**
      Execute_Flip() ; // 180° roll
    end
    Execute_Grasp(C_{task}) ; // Apply prior
    State $\leftarrow$ HOME ; // Reset pose
    State $\leftarrow$ STANDBY ; // Loop closed
  end
end

consecutive confirmations, yielding a stabilization window of $\sim$613 ms at 16.3 FPS. This value was empirically selected to suppress single-frame misclassifications while maintaining sub-second reaction latency. The resulting logic feeds into the deterministic state machine formalized in Algorithm 1, where the predicate "is valid" encapsulates this temporal consensus criterion.

Notably, in dense environments where valid targets appear in both the ventral/front and dorsal/rear fields of view simultaneously (i.e., both D_v and D_d are valid), Algorithm 1 inherently prioritizes ventral/front acquisition. This deliberate priority serves two purposes: it avoids state-switching oscillations between competing paths, and it minimizes cycle time by omitting the additional 180° kinematic reorientation. The N-frame consensus criterion applied upstream substantially mitigates the risk of this priority being triggered by transient false positives in the ventral stream.

IV. Experimental Validation

To empirically validate the proposed hardware-software co-designed architecture presented in Sec. III, we design a focused experimental protocol that isolates and evaluates the system's topological, kinematic, and language-guided routing capabilities. The primary objective is to verify that the proposed hardware interventions (dual-camera relay and slip-ring) successfully unlock rearward manipulation with performance parity to forward grasping, rather than to benchmark detection accuracy across a broad object set. The evaluation is structured around three core hypotheses:

H1 (Topological Completeness). The dual-camera relay effectively provides visual coverage for both the frontal and previously occluded rearward workspaces, as visually corroborated by the concurrent perception streams in Fig. 4.

H2 (Kinematic Unbounding). The J1 slip-ring effectively mitigates joint-limit deadlocks, allowing rearward (flip-over) grasps to be executed with success rates comparable to forward grasps in the tested setup, as evidenced by the continuous 180° roll trajectory mapped in Fig. 4 (a).

H3 (Language-Guided Routing Autonomy). The zero-shot perception engine can provide the semantic cues required for language-guided grasping and trigger the corresponding pre-calibrated action primitive natively on the edge NPU, driving the distinct spatial execution paths contrasted in Fig. 4.

A. Experimental Setup and Protocol

The dexterous hand is initialized in a neutral, palm-forward pose. We establish two predefined manipulation zones: a frontal station (ventral/front camera FOV) and a rearward station (dorsal/rear camera FOV). To evaluate the language-to-action mapping, we utilize two distinct objects: a pink 3D-printed sphere (80 mm diameter) and a Coke can (330 mL). The complete hardware specifications of the platform are summarized in Table I.

To strictly isolate the evaluation of perception topology (H1) and kinematic unbounding (H2) from confounding variables introduced by continuous visual servoing, we deliberately decouple the system into a high-level language-guided router and low-level open-loop execution. This methodological choice ensures that observed failure modes can be unambiguously attributed to either the language-guided engine or the mechanical platform, rather than being obscured by closed-loop control artifacts. The grasping actions (Config I and Config II) are accordingly programmed as highly repeatable, pre-calibrated joint-angle primitives at the designated stations.

During each trial, objects are randomly assigned to either the front or rear stations. The system then runs without manual intervention: the visual engine detects the object's class and spatial source (ventral/front or dorsal/rear camera), and the deterministic state machine routes these discrete labels to the corresponding action primitive. A trial is marked successful if the object is stably held for ≥ 3 s post-lift. We conduct 40 trials in total (10 per Object $\times$ Path combination).

B. Results and Validation

Table II summarizes system performance across the orthogonal decision spaces (object class and spatial path).

Validation of H1 & H2. The language-guided detection rates remained comparable between the Forward and Flip-over paths for each object class (90% vs. 90% for the sphere; 60% vs. 70% for the can), suggesting that rearward placement did not introduce an obvious additional perceptual penalty in

TABLE I
System Specifications of the Edge-Native Hand-Wrist Platform

Kinematics		**Perception & Computing**		**Mechatronics & Power**	
Parameter	*Specification*	*Parameter*	*Specification*	*Parameter*	*Specification*
Total DOF	11 (10 digits + wrist)	Camera array	2$\times$ SC233HGS (RGB, global shutter)	Slip-ring	6-channel (12 V, 7.4 V, GND, bus)
J1 actuator	Magnetic encoder servo	Image resolution	1088$\times$1280 @ 20 fps	Supply voltage	12 V DC / 7.4 V DC
J1 max torque	4.5 kg·cm	Lens FOV	206° diagonal	Total weight	0.76 kg
J1 range	Infinite (unbounded)	Synchronization	HW trigger, AE sync	Dimensions	150$\times$175$\times$278 mm
		Edge processor	Rockchip RK3588 SoC		
		NPU computing	6 TOPS (INT8)		

TABLE II

SYSTEM PERFORMANCE ACROSS ORTHOGONAL DECISION SPACES.

Object	Path	Trials	Language Guided Detection	Execution Success	Cycle Time
Pink Sphere	Forward	10	90%	90%	9.44 s
Pink Sphere	Flip-over	10	90%	90%	15.63 s
Coke Can	Forward	10	60%	60%	9.44 s
Coke Can	Flip-over	10	70%	60%	15.63 s
Aggregate	—	**40**	**78%**	**75%**	12.54 s

TABLE III

LATENCY BREAKDOWN OF THE EDGE-NATIVE LANGUAGE-GUIDED PERCEPTION ENGINE (TOTAL: 61.3 MS / 16.3 FPS).

Stage	Unit	Time (ms)	%
Image Acquisition	VPU/CPU	6.5	10.6
Pre-processing	RGA	5.2	8.5
Network Inference	NPU (INT8)	42.8	69.8
Post-processing	CPU	6.8	11.1
Total	**SoC**	**61.3**	**100**

the tested setup, supporting H1 at a proof-of-concept level. Furthermore, the execution success parity between Forward paths (75% avg.) and Flip-over paths (75% avg.) suggests that the slip-ring can support rearward flip-over trials with reliability comparable to forward trials within the tested workstation conditions. Within this prototype, the rearward grasp can therefore be treated as a practical execution mode rather than only a singularity-prone edge case (H2).

Validation of H3. The 75% aggregate end-to-end success rate over 40 trials—encompassing semantic perception, spatial routing, *and* open-loop mechanical execution—supports the feasibility of the proposed language-guided grasping framework in this proof-of-concept setting. Notably, this metric represents a composite measure, as any single-stage failure (detection, routing, or actuation) results in a complete trial failure. As profiled in Table III, the heterogeneous RK3588 SoC achieves 16.3 FPS (61.3 ms/frame) by distributing preprocessing to the dedicated RGA engine (5.2 ms), backbone inference to the NPU (42.8 ms), and NMS to the CPU (6.8 ms). Under this latency budget the decoupled visual backbone provided sufficiently stable target labels to trigger the appropriate pre-calibrated state. The $N = 10$ frame consensus buffer provided stable target selection in $\sim$613 ms, offering a practical stability-latency trade-off for the scripted execution pipeline (H3).

Per-Object & Failure Analysis. A notable performance asymmetry exists between the two objects: the pink sphere achieved a 90% language-guided detection rate, whereas the Coke can achieved 65% overall (60% on the Forward path and 70% on the Flip-over path). This gap appears to originate primarily in the target-authentication stage rather than in the spatial routing logic. The sphere's saturated appearance is comparatively easy for the reparameterized detector to validate, whereas the can's specular surface introduces view-dependent reflections that reduce semantic confidence under INT8 deployment. Because the current experiments use predefined stations and pre-calibrated open-loop primitives, the dominant limitation is not continuous localization accuracy but the stability of target validation together with the limited contact tolerance of the scripted grasp on a narrow cylindrical object. This underscores an expected trade-off in edge-native open-loop deployments: eliminating depth sensing and closed-loop correction preserves 16.3 FPS throughput but reduces robustness when object appearance and contact geometry are less forgiving. Future closed-loop visual servoing is expected to particularly benefit such objects.

C. Discussion and Limitations

The experimental results support the central premise of this proof-of-concept study: topological and kinematic bottlenecks in dexterous manipulation can be alleviated through targeted hardware interventions, thereby simplifying the edge-side software architecture required for front-rear execution in semi-structured service and workstation settings. The dual-camera relay and slip-ring decouple the decision space into orthogonal "Where" and "What" dimensions. This enables a lightweight deterministic state machine to govern the perception-and-execution pipeline without recourse to computationally intensive motion planning during runtime.

A salient architectural benefit is its *compositional scalability*. Because the language-guided router operates on discrete labels, extending the system to new grasp primitives requires merely appending entries to the language-to-action mapping table. This leaves the perception backbone and hardware unmodified, contrasting sharply with end-to-end learned policies that demand costly retraining.

***Remark 2**: (Generality of the Co-Design):* While instantiated here for front-rear grasping, the underlying principle—that structurally immutable hardware topology constraints (e.g., limited sensor coverage, bounded joint range) may be resolved through targeted electromechanical intervention—can potentially generalize to other manipulation scenarios where algorithmic compensation alone proves insufficient.

Limitations & Future Work. The current architecture adopts open-loop execution at predefined stations—a deliberate design choice to isolate hardware validation variables—which precludes dynamic reaching for uncalibrated spatial poses. The present study should therefore be interpreted as a system-level proof-of-concept rather than a full benchmark of closed-loop grasping in unstructured scenes. Accordingly, lifecycle wear, contact resistance, and electrical noise under prolonged high-load slip-ring operation, as well as comparative ablations against single-camera perception, bounded-wrist hardware, and closed-loop servoing variants, fall outside the scope of the current stage and are reserved for future work. Additionally, while the monocular RGB global-shutter cameras ensure low-latency language-guided detection, they inherently lack explicit depth sensing for precise 3D geometric reasoning. Future work will integrate

a continuous 6D pose estimation head with closed-loop visual servoing [6] to fully exploit the unbounded workspace, alongside depth or tactile modalities to enrich the perceptual state.

V. Conclusions

This paper presented a hardware-software co-design framework for mitigating front-rear topological and kinematic bottlenecks in dexterous manipulation under semi-structured service and workstation conditions. By integrating a dual-camera relay and an infinite-roll slip-ring wrist, the platform expanded rearward perception coverage and alleviated wrist-roll limitations. Concurrently, an edge-native, reparameterized zero-shot detector enabled language-guided grasping on a resource-constrained NPU, with semantic cues used to trigger pre-calibrated action primitives. A deterministic routing framework maps these detections to spatial routing decisions and grasping actions via a consensus-filtered state machine running at 16.3 FPS. Real-world proof-of-concept experiments yielded a 75% end-to-end success rate across the tested front and rear workstation trials, without additional task-specific fine-tuning. Taken together, these results suggest that targeted electromechanical interventions at hardware bottlenecks can serve as a practical complement to algorithmic advances in edge-native manipulation systems.

References

[1] C. Piazza, G. Grioli, M. Catalano, and A. Bicchi, "A century of robotic hands," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 2, pp. 1–32, 2019.

[2] A. Bicchi, "Hands for dexterous manipulation and robust grasping: A difficult road," *IEEE Trans. Robot. Autom.*, vol. 16, no. 6, pp. 652–662, 2000.

[3] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Trans. Robot.*, vol. 30, no. 2, pp. 289–309, 2014.

[4] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. RSS*, 2017.

[5] M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, "Learning dexterous in-hand manipulation," *Int. J. Robot. Res.*, vol. 39, no. 1, pp. 3–20, 2020.

[6] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Trans. Robot. Autom.*, vol. 12, no. 5, pp. 651–670, 1996.

[7] K. He, R. Newbury, T. Tran, J. Haviland, B. Burgess-Limerick, D. Kulic, P. Corke, and A. Cosgun, "Visibility maximization controller for robotic manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 8479–8486, 2022.

[8] C. W. Wampler, "Manipulator inverse kinematic solutions based on vector formulations and damped least-squares methods," *IEEE Trans. Syst., Man, Cybern.*, vol. 16, no. 1, pp. 93–101, 1986.

[9] Y. Nakamura and H. Hanafusa, "Inverse kinematic solutions with singularity robustness for robot manipulator control," *J. Dyn. Syst. Meas. Control*, vol. 108, no. 3, pp. 163–171, 1986.

[10] A. Liégeois, "Automatic supervisory control of the configuration and behavior of multibody mechanisms," *IEEE Trans. Syst., Man, Cybern.*, vol. 7, no. 12, pp. 868–871, 1977.

[11] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. ECCV*, 2020, pp. 405–421.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF CVPR*, 2016, pp. 779–788.

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[14] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF CVPR*, 2018, pp. 2704–2713.

[15] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. ECCV*, 2024.

[16] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, "Simple open-vocabulary object detection with vision transformers," in *Proc. ECCV*, 2022.

[17] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[19] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[20] T. Nishimura, Y. Suzuki, T. Tsuji, and T. Watanabe, "1-degree-of-freedom robotic gripper with infinite self-twist function," *IEEE Robot. Autom. Lett.*, vol. 8, no. 2, pp. 1172–1179, 2023.

[21] M. Métillon, P. Cardou, K. Subrin, C. Charron, and S. Caro, "A cable-driven parallel robot with full-circle end-effector rotations," *ASME J. Mechanisms Robotics*, vol. 13, no. 3, p. 031014, 2021.

[22] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF CVPR*, 2021, pp. 13 733–13 742.

[23] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-time open-vocabulary object detection," in *Proc. CVPR*, 2024.

[24] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. J. Robot. Res.*, vol. 37, no. 4–5, pp. 421–436, 2018.

[25] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Proc. CoRL*, 2020.

[26] A. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Proc. CoRL*, 2018.

[27] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" in *Proc. CoRL*, 2017.

[28] M. Grebenstein, A. O. Albu-Schäffer, T. Bahls, M. Chalon, O. Eiberger, W. Friedl, R. Gruber, S. Haddadin, U. Hagn, R. Haslinger, H. Höppner, S. Jörg, M. Nickl, A. Nothhelfer, F. Petit, J. Reill, N. Seitz, T. Wimböck, S. Wolf, T. Wüsthoff, and G. Hirzinger, "The DLR hand arm system," in *Proc. IEEE ICRA*, 2011, pp. 3175–3182.

[29] S. C. Jacobsen, E. K. Iversen, D. F. Knutti, R. T. Johnson, and K. B. Biggers, "Design of the Utah/M.I.T. dextrous hand," in *Proc. IEEE ICRA*, 1986, pp. 1520–1532.

[30] M. G. Catalano, G. Grioli, E. Farnioli, A. Serio, C. Piazza, and A. Bicchi, "Adaptive synergies for the design and control of the Pisa/IIT SoftHand," *Int. J. Robot. Res.*, vol. 33, no. 5, pp. 768–782, 2014.

[31] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, M. Li, W. Cheng, W. Nie, B. Li, Y. Zhang, X. Liang, X. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[32] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF CVPR*, 2018, pp. 8759–8768.