

Pan-preferential stereo IBVS on ARMOS TurtleBot: Design and experimental evaluation

Fredy Hernán Martínez *

Facultad Tecnológica, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.

Global Journal of Engineering and Technology Advances, 2025, 24(02), 181-195

Publication history: Received on 10 July 2025; revised on 17 August 2025; accepted on 19 August 2025

Article DOI: <https://doi.org/10.30574/gjeta.2025.24.2.0250>

Abstract

Image-based visual servoing (IBVS) on nonholonomic platforms often suffers from transient conflicts between camera alignment and base motion when the viewpoint is fixed. This work addresses that difficulty by mounting a rectified stereo pair on a controllable pan axis and commanding a two-channel pan-preferential IBVS that regulates image features while the base heading follows the camera yaw. Depth at feature locations is recovered from disparity to build a well-conditioned visual Jacobian, and a damped least-squares inversion with feature weighting produces bounded linear speed and pan-rate commands. The methodology is implemented on the ARMOS TurtleBot differential base equipped with a low-cost ESP32+OV7670 stereo rig, and underpinned by fast encoder-based motor loops and disturbance-observer augmented PD control. Two experimental scenarios (static target and lateral target motion at roughly constant depth) demonstrate consistent feature tracking, smooth camera-base alignment, and tolerance to transient disparity dropouts through outlier rejection and command projection. Across repeated runs the proposed scheme reduces steady error and settling time compared with a fixed-camera baseline while respecting modest computing and driver limits. All components, rates, and bounds are reported to enable replication, and the results indicate that stereo-depth IBVS with pan alignment is feasible on affordable hardware for indoor mobile manipulation and navigation.

Keywords: Differential-drive robots; Image-based visual servoing; Low-cost stereo vision; Pan-preferential control; Stereo disparity

1. Introduction

Image-based visual servoing (IBVS) remains a compelling strategy for mobile robots because it regulates quantities that are sensed directly, avoids brittle reconstruction pipelines, and can run with modest onboard resources [1–3]. When the camera is carried by a differential base, however, the camera-base mapping introduces configuration-dependent singularities that degrade control authority at certain poses [4]. Depth ambiguity in monocular setups further complicates gain selection and can slow convergence, particularly when features lie at moderate range [5]. Practical deployments add timing constraints and packet jitter that must be reflected explicitly in the control design rather than treated as afterthoughts. These factors together motivate a solution that addresses geometry, sensing, and real-time execution as a single, coherent problem [6].

Prior work has mitigated singularities by switching between control modes or by constraining motion near problematic configurations, which can produce discontinuities and transient spikes [7, 8]. Other approaches have relied on monocular depth surrogates or scene heuristics; while serviceable in narrow settings, they often yield poor conditioning of the interaction matrix and require heavy damping to remain stable [9, 10]. Continuous reparameterizations exist, yet many assume camera placements or actuation sets that do not match common mobile platforms. Low-cost stereo has

* Corresponding author: Fredy Martínez

seen growing interest, but published integrations rarely document end-to-end timing, baseline stability under motion, and the interaction with inner motor loops [11]. A gap therefore remains for a reproducible, continuous IBVS policy that avoids mode switches, uses measured depth at the feature locations, and runs on accessible hardware.

This paper addresses that gap with a pan-preferential scheme that couples a rigid stereo pair on a single-axis pan mount to a differential base. The image-space inversion commands the pair $[v, \phi_c]$ through a minimum-norm solution, while the base yaw follows the camera with a proportional law $\omega = K_o \phi_c$, thereby keeping the inversion free from the directions that cause rank loss. Depth is obtained from disparity at the same pixels that drive the controller, with rectified intrinsics and a fixed baseline ensuring $Z = fB/d$ is well defined during pan motion. The system is realized on the ARMOS TurtleBot using two ESP32+OV7670 modules that stream frames to a host for rectification, disparity, and feature tracking at low resolution and moderate rate [12]. Safety interlocks (infrared stop, watchdog, and an e-stop) and a disturbance-observer layer on the wheel and pan loops complete the stack needed for a stable real-robot evaluation.

The study is organized around operating points and metrics that reflect both control theory and field constraints. Vision runs primarily at 160×120 pixels and 10–20 Hz per camera; the IBVS loop updates at 50–100 Hz, and the motor layer executes at 1–2 ms periods. Trials comprise a static target and a laterally moving target (≈ 0.2 m/s) at indoor ranges of 1–3 m, with identical gains and caps across ablations. Key performance indicators include image-plane RMS error and settling time, the decay of $|\phi_c|$, control energies for linear and angular channels, and the condition number of the composite visual Jacobian together with damping usage. End-to-end latency and stereo quality (inlier ratio and depth outliers) are reported to link closed-loop behavior to the timing and sensing budget.

The contributions are fourfold and are backed by a complete real-robot implementation [13]. First, the paper formulates a continuous stereo-IBVS policy that preserves pan-preferential avoidance of singularities while injecting measured depth at the feature level. Second, it delivers an integration on an affordable platform with documented parameters, rates, and safety mechanisms, including PD plus disturbance observers on wheel and pan axes. Third, experiments against fixed-camera and monocular surrogates show improved feature regulation, smoother base-camera alignment, and fewer damped steps within the measured latency envelope, without inflating actuation energy. Fourth, the work provides a reproducible path by disclosing calibration assets, logging schema, and tuning guidelines. The remainder of the paper presents the system overview, the visual-servoing formulation with stereo depth, the low-level implementation, the experimental setup, the results with ablations, and a discussion of implications and limitations.

2. System Overview

The platform is a differential-drive mobile base (ARMOS TurtleBot) equipped with wheel encoders, a single-axis pan joint for the vision payload, and a ring of proximity sensors for reflexive stopping in close-range situations. The locomotion module uses two independently driven wheels separated by a wheelbase $l = 0.25$ m and wheel radius $r = 0.024$ m; these values are retained in both modeling and experiments to match the mechanical design. Let ω_R and ω_L denote the right and left wheel angular speeds, respectively. The body-frame linear and angular velocities (v, ω) satisfy

$$v = \frac{r}{2}(\omega_R + \omega_L), \quad \omega = \frac{r}{l}(\omega_R - \omega_L) \quad \dots \dots \dots (1)$$

and the inverse map used by the low-level controller is

$$\omega_R = \frac{v}{r} + \frac{l}{2r} \omega, \quad \omega_L = \frac{v}{r} - \frac{l}{2r} \omega \quad \dots \dots \dots (2)$$

Nominal commanded speed is capped to $v \leq 0.2$ m/s on indoor flat floors to accommodate vision sampling and actuation limits, whereas the safety ring of nine infrared sensors issues an immediate halt when the estimated clearance falls below a fixed threshold.

The vision payload is a rigid stereo pair mounted on a yaw (pan) axis with measured angle ϕ_c , so that both cameras rotate together while their left–right extrinsics remain constant. Each camera is an OV7670 sensor interfaced to an ESP32 module; frames are sent over WebSocket to a host computer that runs rectification and disparity. The capture stage supports 160×120 pixels at 10–20 fps per stream and 320×240 pixels at 5–10 fps when bandwidth permits; we operate primarily at the lower resolution to reduce end-to-end latency. Stereo calibration yields intrinsics (f_x, f_y, c_x, c_y) and a fixed baseline B ; after rectification, disparity $d(u, v)$ provides per-feature depth via

$$Z(u, v) = \frac{f B}{d(u, v)}, \quad f = \frac{f_x + f_y}{2} \quad \dots \dots (3)$$

Feature detection and tracking are performed on the rectified left image (fiducials or salient points), producing the image vector $f = [u_i, v_i]_{i=1..k}$ and associated depths $Z = [Z_i]_{i=1..k}$, which feed the visual servo controller.

Figure 1 summarizes the sensing-to-actuation data flow, including time alignment and control layers. The camera twist in the spatial frame, ${}^s w_c = [{}^s \dot{p}_c^T, {}^s \omega_c^T]^T \in \mathbb{R}^6$, is related to the commanded generalized velocities

$$v = [v \quad \omega \quad \dot{\phi}_c]^T \quad \dots \dots (4)$$

through a configuration-dependent Jacobian,

$${}^s w_c = {}^c J(\phi_c) v \quad \dots \dots (5)$$

which captures how base translation, base yaw, and pan rate induce camera translational and rotational velocities. This mapping partitions naturally as ${}^c J(\phi_c) = [{}^c J_{13}(\phi_c) \quad {}^c J_2(\phi_c)]$ with ${}^c J_{13} \in \mathbb{R}^{6 \times 2}$ selecting the $\{v, \dot{\phi}_c\}$ channels used by the outer visual loop, while ${}^c J_2$ corresponds to the base yaw channel ω . By retaining the pan axis and selecting $\{v, \dot{\phi}_c\}$ in the image-space inversion, the proposed architecture avoids the loss of rank that occurs in camera-base configurations where the full three-channel inversion would be ill-conditioned.

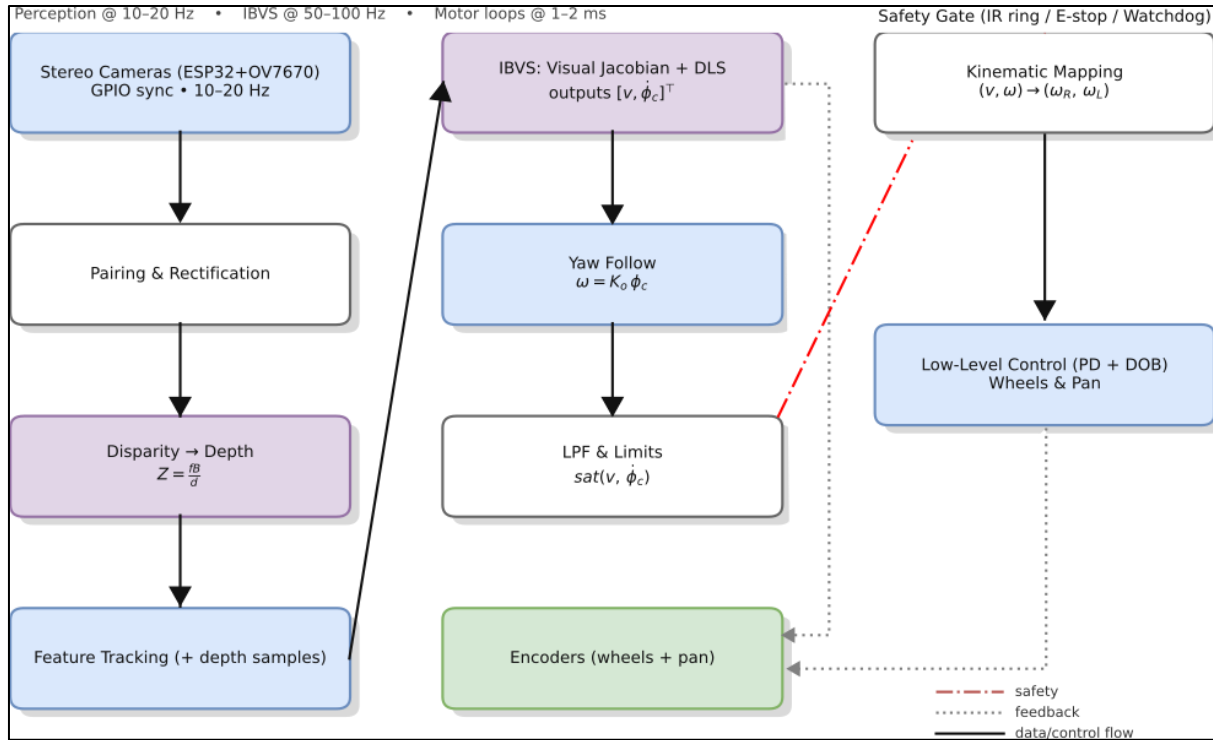


Figure 1 System overview. Dual ESP32+OV7670 streams are time-aligned by a GPIO trigger and timestamps, rectified on the host, and processed to obtain disparity and depth. Low-level PD+DOB loops track wheel and pan setpoints; the IR ring enforces a reflex stop

The perception pipeline enforces deterministic pairing of left-right frames using a shared trigger pulse and host-side timestamps, followed by rectification with precomputed maps and disparity via block matching or semi-global methods, depending on compute budget. Depth values are sampled at feature locations to improve numerical conditioning of the interaction matrix used later by the IBVS layer. The outer loop produces $v_{13} = [v, \dot{\phi}_c]^T$ through a pseudoinverse built on the composite visual Jacobian $J_{\text{vis}}(\phi_c, Z) = \bar{L}(f, Z) {}^c J_{13}(\phi_c)$, whereas base yaw aligns to the pan angle through a simple proportional law $\omega = K_o \phi_c$. Commands are rate-limited and low-pass filtered prior to conversion into wheel and pan setpoints, which are closed by PD controllers augmented with disturbance observers to mitigate friction and load variations.

Table 1 lists the principal parameters and operating rates used in the experimental section. Vision runs at 10-20 Hz per stream at 160×120 pixels, with the IBVS node cycling at 50-100 Hz, and the motor-layer loops at 500-1000 Hz. Communications rely on WebSocket for image transport and ROS topics/services for control and logging; a watchdog supervises message freshness. The operating envelope is an indoor planar area; experiments limit peak linear speed, enforce braking when proximity sensors detect short clearance, and log all telemetry for reproducibility. These design choices balance sensing throughput, controller bandwidth, and actuation limits so that the subsequent analysis is grounded on a feasible real-time configuration.

Table 1 Key System Parameters and Operating Rates

Quantity	Symbol	Value / Range
Wheel radius	r	0.024 m
Wheelbase	l	0.25 m
Pan angle	ϕ_c	measured (encoder)
Baseline	B	fixed (rigid mount)
Focal (effective)	f	$(f_x + f_y)/2$
Depth model	Z	$Z = (fB)/d$
Vision resolution	—	160×120 (mono)
Vision rate (per cam)	—	10–20 Hz
IBVS control rate	—	50–100 Hz
Motor control rate	—	500–1000 Hz
Speed cap (linear)	v	≤ 0.2 m/s
Yaw follow gain	K_o	tuned experimentally
Safety stop	—	IR thresholded

3. Visual Servoing with Stereo Depth

This section formulates the image-based control law that drives a differential mobile base equipped with a pan-mounted rigid stereo rig toward a desired visual configuration. We consider a rectified stereo pair with fixed extrinsics between the left and right cameras; the pan joint rotates the entire rig about a vertical axis by an angle ϕ_c without altering the baseline geometry. The control objective is to regulate a set of image features $f = [u_i, v_i]_{i=1..k}$ to a target f_d , while the base yaw ω aligns to the pan angle so that the camera ultimately faces the robot heading. Pixels are mapped to normalized coordinates through the calibrated intrinsics, namely $x = u - c_x/f_x$ and $y = v - c_y/f_y$, which renders the interaction equations well scaled across lenses and resolutions. We assume local scene rigidity and modest camera motions over one sampling interval; these common conditions hold in the indoor, planar settings used for our experiments and allow a concise derivation of the stacked interaction model.

Depth at each feature is obtained from stereo disparity after rectification. Let B denote the rigid baseline and $f = (f_x + f_y)/2$ the effective focal length in pixels; for a measured disparity $d(u, v)$ we compute:

$$Z(u, v) = \frac{f B}{d(u, v)} \quad \dots \dots \dots (6)$$

and sample $Z_i = Z(u_i, v_i)$ at the feature locations. A first-order approximation quantifies how disparity noise propagates to depth: if σ_d^2 denotes the local disparity variance, then:

$$\sigma_Z^2 \approx \left(\frac{\partial Z}{\partial d} \right)^2 \sigma_d^2 = \left(\frac{fB}{d^2} \right)^2 \sigma_d^2 \quad \dots \dots \dots (7)$$

which later informs feature weighting. We operate the capture at 160×120 pixels with 10–20 Hz per stream to reduce latency and stabilize timing, and track either fiducials or salient corners on the rectified left image.

For a feature with normalized coordinates (x, y) and depth Z , the standard interaction matrix that links the camera twist to image-velocity reads:

$$L_i(x, y, Z) = \begin{bmatrix} -1/Z & 0 & x/Z & xy & -(1+x^2) & y \\ 0 & -1/Z & y/Z & 1+y^2 & -xy & -x \end{bmatrix} \in \mathbb{R}^{2 \times 6} \quad \dots \dots \dots (8)$$

Stacking k features produces $\bar{L}(f, Z) = \text{blkdiag}(L_1, \dots, L_k) \in \mathbb{R}^{2k \times 6}$. Let the camera spatial twist be ${}^s w_c = [{}^s p_c^\top, {}^s \omega_c^\top]^\top \in \mathbb{R}^6$. The commanded generalized velocities of the mobile platform with pan are:

$$v = [v \quad \omega \quad \dot{\phi}_c]^\top \in \mathbb{R}^3 \quad \dots \dots \dots (9)$$

and the kinematic mapping from commands to camera twist is:

$${}^s w_c = {}^c J(\phi_c) v, \quad {}^c J(\phi_c) \in \mathbb{R}^{6 \times 3} \quad \dots \dots \dots (10)$$

whose explicit form depends on the camera placement relative to the base and the pan axis. We partition ${}^c J(\phi_c) = [{}^c J_{13}(\phi_c) \quad {}^c J_2(\phi_c)]$, where ${}^c J_{13} \in \mathbb{R}^{6 \times 2}$ selects the $\{v, \dot{\phi}_c\}$ channels used in the image-space inversion, and ${}^c J_2 \in \mathbb{R}^{6 \times 1}$ corresponds to the base yaw ω .

The composite visual Jacobian that enters the inversion is defined as:

$$J_{\text{vis}}(\phi_c, f, Z) = \bar{L}(f, Z) {}^c J_{13}(\phi_c) \in \mathbb{R}^{2k \times 2} \quad \dots \dots \dots (11)$$

Selecting the pair $\{v, \dot{\phi}_c\}$ yields a two-column matrix whose full-column-rank condition is met for non-degenerate feature configurations and camera placements that avoid the pathological alignments reported for full three-channel inversions. The use of measured depths Z_i from (6) improves conditioning relative to monocular approximations, particularly when the target lies at moderate range or when the image spans a narrow field of view. Let $e = f_d - f \in \mathbb{R}^{2k}$ denote the feature error; a damped, weighted least-squares provides a minimum-norm command for the two controlled channels:

$$v_{13} = (J_{\text{vis}}^\top W J_{\text{vis}} + \lambda^2 I_2)^{-1} J_{\text{vis}}^\top W K_{\text{img}} e, \quad v_{13} = [v \quad \dot{\phi}_c]^\top \quad \dots (12)$$

where $\lambda \geq 0$ is a Tikhonov parameter, $W = \text{diag}(w_1, w_1, w_2, w_2, \dots)$ weights per-feature channels (e.g., from σ_z , tracker confidence), and $K_{\text{img}} > 0$ is diagonal. The base yaw follows the pan angle through a proportional law that closes the orientation gap without switching modes:

$$\omega = K_o \phi_c, \quad K_o > 0 \quad \dots \dots \dots (13)$$

This separation keeps the image-space inversion free from the singularities that appear when ω is treated as a coequal control input in geometries that reduce the rank of ${}^c J(\phi_c)$.

The discrete-time implementation respects the heterogeneous sampling of perception and actuation. At each vision cycle, we compute v_{13}^{raw} from (12), then apply a first-order filter to temper latency and quantization:

$$v_{13}[n] = \alpha v_{13}^{\text{raw}}[n] + (1 - \alpha) v_{13}[n - 1], \quad \alpha = \frac{T_s}{T_s + \tau_{\text{LPF}}} \quad \dots \dots \dots (14)$$

where T_s is the vision sampling period and τ_{LPF} a small time constant. Command bounds are enforced by projection; a simple quadratic program clips the solution without discontinuities:

$$\min_{v_{13}} \|v_{13} - v_{13}[n]\|_2^2 \quad \text{s.t.} \quad v_{\min} \leq v \leq v_{\max}, \quad |\dot{\phi}_c| \leq \dot{\phi}_{c,\max} \quad \dots \dots \dots (15)$$

Outliers and dropouts are handled by rejecting features with inconsistent motion or disparity (e.g., $|d| < d_{\min}$ or large photometric residuals), temporarily reducing their weights w_i , and, if needed, falling back to a monocular surrogate Z^*

with conservative gains until valid stereo recovers. Under full-column-rank J_{vis} , sufficiently small T_s , and bounded delays, the linearized error dynamics around the operating point are exponentially stable for diagonal K_{img} chosen within standard IBVS ranges; the separation with (13) yields smooth base alignment and avoids mode switching.

Table 2 Controller Parameters and Rates for the Stereo-IBVS Layer

Quantity	Symbol	Nominal Value / Range
Image gains	K_{img}	diag(k_u, k_v), 0.5–2.0
Damping (DLS)	λ	10^{-3} – 10^{-1}
Feature weights	W	from σ_z , confidence
Yaw follow gain	K_o	0.6 – 1.2 s^{-1}
LPF time constant	τ_{LPF}	60–120 ms
Vision rate (per cam)	—	10–20 Hz @ 160×120
IBVS update rate	—	50–100 Hz
Speed cap (linear)	v_{max}	0.2 m/s
Pan rate cap	$\dot{\phi}_{c,max}$	0.6–1.0 rad/s
Disparity threshold	d_{min}	dataset dependent

4. Low-Level Control & Implementation

This section describes the real-time actuator layer that tracks the high-level commands $v = [v, \omega, \dot{\phi}_c]^\top$ generated by the visual servoing loop. The ARMOS TurtleBot is a differential-drive platform with wheel radius $r = 0.024$ m and wheelbase $l = 0.25$ m; the pan joint carries the rigid stereo rig and provides the measured angle ϕ_c . Wheel angular speeds ω_R and ω_L follow the standard inverse kinematics:

$$\omega_R = \frac{v}{r} + \frac{l}{2r} \omega, \quad \omega_L = \frac{v}{r} - \frac{l}{2r} \omega \quad \dots \dots \dots (16)$$

while the pan-rate command is $\dot{\phi}_c$ itself. We collect the actuator references as $\dot{\phi}^d = [\omega_R^d, \omega_L^d, \dot{\phi}_c^d]^\top$, subject to speed caps $|v| \leq v_{max}$ and $|\dot{\phi}_c| \leq \dot{\phi}_{c,max}$ motivated by sensing and power limits; in experiments we adopt $v_{max} = 0.2$ m/s and $\dot{\phi}_{c,max} \in [0.6, 1.0]$ rad/s.

The closed loops for the two wheels and the pan axis share an identical per-axis structure that combines a proportional-derivative (PD) servo with a disturbance observer (DOB) to compensate friction and load changes without integral action. Each axis $\star \in \{R, L, c\}$ is modeled nominally by a first-order plant $P_{n,\star}(s) = \frac{K_\star}{\tau_\star s + 1}$ from input $u_\star$ (voltage or duty cycle) to measured velocity $y_\star \equiv \omega_\star$. The DOB uses a causal low-pass $Q_\star(s) = \frac{\omega_{c,\star}}{s + \omega_{c,\star}}$ to estimate and reject lumped disturbances, leading to the control law:

$$u_\star = K_{p,\star}(\omega_\star^d - \omega_\star) + K_{d,\star}(\dot{\omega}_\star^d - \dot{\omega}_\star) - \hat{d}_\star, \quad \hat{d}_\star = Q_\star(s)(y_\star - P_{n,\star}(s)u_\star) \dots \dots (17)$$

The resulting inner loops track $\dot{\phi}^d$ with short settling times and high disturbance rejection, while avoiding windup effects and reducing sensitivity to parameter drift. On the electrical side, L298N-class drivers supply each motor channel with current limiting and thermal protection; duty-cycle saturation and slew-rate bounds are enforced in software to respect these constraints.

A fast motor loop runs at period $T_m \in [1, 2]$ ms, where encoder counts are latched and converted to velocities. Let $N_\star$ be the counts per revolution (after gearing) and $\Delta \text{counts}_\star[k]$ the increment over one period; the discrete-time velocity estimate is:

$$\omega_{\star}[k] = \frac{2\pi}{N_{\star}} \frac{\Delta \text{counts}_{\star}[k]}{T_m}, \quad \star \in \{R, L, c\} \quad \dots \dots \dots (18)$$

The derivative term uses a noise-aware differentiator with prefilter:

$$\dot{\omega}_{\star}[k] = \beta \dot{\omega}_{\star}[k-1] + (1-\beta) \frac{\omega_{\star}[k] - \omega_{\star}[k-1]}{T_m}, \quad \beta \in [0,1) \quad \dots \dots \dots (19)$$

which damps quantization effects at high rates. DOB and PD are discretized by Tustin or matched-z methods, yielding stable realizations under the chosen sampling.

Command preprocessing ensures that actuator demands remain consistent with plant and driver limits as well as with perception throughput. Before applying (16), the high-level vector v is smoothed by a first-order low-pass filter with time constant matched to the vision-update interval, reducing jitter due to asynchronous sensing. After mapping, a projection step clips the two controlled channels v and $\dot{\phi}_c$ without introducing discontinuities:

$$\min_{v_{13}} \|v_{13} - v_{13}^{\text{raw}}\|_2^2 \quad \text{s.t.} \quad v_{\min} \leq v \leq v_{\max}, \quad |\dot{\phi}_c| \leq \dot{\phi}_{c,\max}, \quad v_{13} = [v \quad \dot{\phi}_c]^T \quad \dots \dots \dots (20)$$

with v_{13}^{raw} the undamped output of the visual loop. At the motor level, per-axis PWM commands are limited to $|u_{\star}| \leq u_{\max}$ and slewed by $|u_{\star}[k] - u_{\star}[k-1]| \leq \Delta u_{\max}$ to avoid current spikes. A reflex brake is triggered by the infrared ring when the measured clearance falls below a threshold, and a watchdog drops the platform into a safe idle if command freshness exceeds a fixed timeout.

Odometry and calibration support reliable feedback and help detect slip conditions that may degrade servoing. Pan zeroing is performed at start-up using an index mark; wheel directions follow a right-handed convention so that positive ω corresponds to counterclockwise body yaw. At a slower rate, the encoder-based body velocities:

$$v_{\text{enc}} = \frac{r}{2}(\omega_R + \omega_L), \quad \omega_{\text{enc}} = \frac{r}{l}(\omega_R - \omega_L) \quad \dots \dots \dots (21)$$

are compared with their commanded counterparts to build residual tests. If $|v_{\text{enc}} - v| > \varepsilon_v$ or $|\omega_{\text{enc}} - \omega| > \varepsilon_{\omega}$ persist beyond a dwell time, gains are temporarily reduced and the outer loop is notified to downscale the visual commands, which improves robustness on low-traction surfaces.

Table 3 reports the parameters, limits, and rates used across the three axes so that the results in Section 6 can be reproduced. Values are given as nominal ranges reflecting actuator diversity and minor mechanical variations across builds. The chosen gains lead to 10–20% overshoot at step tests with brisk settling, while DOB cutoffs $\omega_{c,\star}$ sit above the dominant disturbance bandwidth yet below the encoder noise corner. The motor loop maintains headroom above 30% CPU at the stated rates, and logging is buffered in a lower-priority thread to prevent jitter. These implementation details ensure that the low-level layer meets its deadlines and exposes well-behaved interfaces to the visual servoing module.

Table 3 Low-Level Parameters, Limits, and Rates (per axis $\star \in \{R, L, c\}$)

Quantity	Symbol	Nominal Value / Range
Wheel radius / wheelbase	r, l	0.024 m, 0.25 m
Motor nominal model	$P_{n,\star}(s)$	$K_{\star} / (\tau_{r,\star} s + 1)$
PD gains	$K_{p,\star}, K_{d,\star}$	tuned (step tests)
DOB filter	$Q_{\star}(s)$	$\omega_{c,\star} / (s + \omega_{c,\star})$
DOB cutoff	$\omega_{c,\star}$	20–80 rad/s
Motor-loop period	T_m	1–2 ms
Derivative prefilter	β	0.6–0.9
Linear speed cap	$v_{\max}$	0.2 m/s

Pan rate cap	$\dot{\phi}_{c,max}$	0.6–1.0 rad/s
PWM magnitude / slew	$v_{max}; \Delta v_{max}$	driver-limited
Safety threshold (IR)	—	lab-calibrated
Watchdog timeout	—	100–200 ms

5. Experimental Setup

The experiments were conducted in an indoor laboratory bay with a planar floor and controlled lighting to reduce shadows and glare. The arena provided straight segments for line-of-sight tracking and lateral traverses, with marked start and goal zones as well as a safety buffer around the perimeter. To remain within sensing and actuation limits, the linear speed was capped at $v_{max} = 0.2$ m/s, and the pan-rate was limited to $|\dot{\phi}_c| \leq 1.0$ rad/s. A hardware e-stop and a watchdog on command freshness were active throughout; the infrared proximity ring enforced a reflex stop when the estimated clearance dropped below a calibrated threshold. Power was delivered from an onboard battery pack sized for the full sensor load; drivers and regulators were dimensioned to avoid brownouts under peak demand. Figure 2 illustrates the arena and the robot configuration used for all trials, including the stereo mount and fiducial placement.

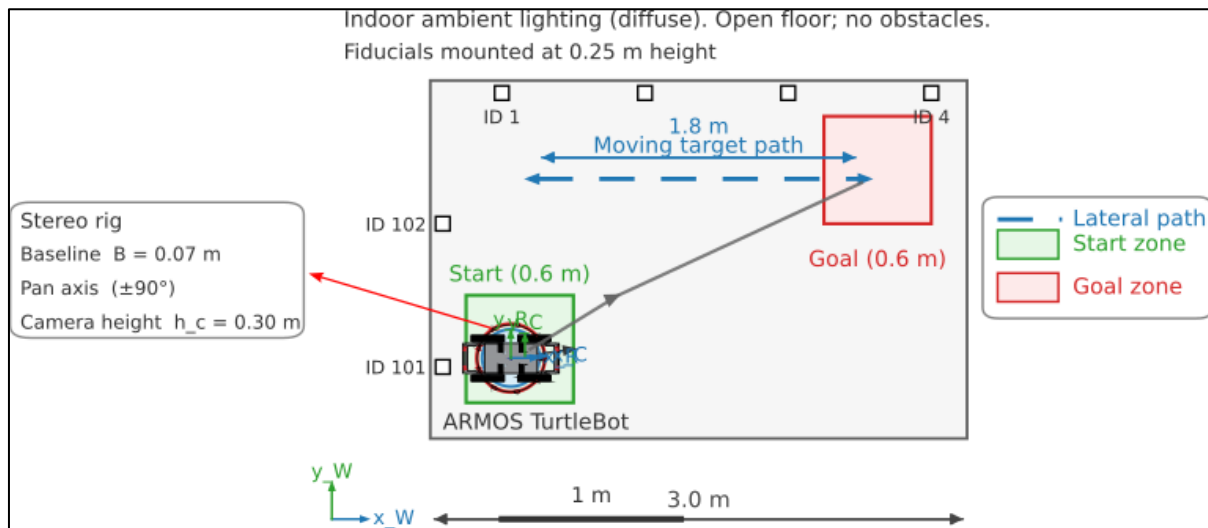


Figure 2 Experimental arena and platform. The ARMOS TurtleBot carries a rigid stereo pair on a single-axis pan joint and an infrared safety ring. Start/goal areas and a lateral travel path for moving-target trials are marked on the floor.

The mobile base is the ARMOS TurtleBot, modeled as a differential drive with wheel radius $r = 0.024$ m and wheelbase $l = 0.25$ m. Wheel encoders provide odometry and velocity feedback; a single-axis pan joint supports the camera rig and supplies the angle ϕ_c via an incremental encoder with index. Drivers of the L298N class interface the motor channels and enforce current and thermal limits; duty-cycle saturation and slew-rate constraints are applied in software to remain inside the electrical envelope. At a slower supervisory rate, encoder-based body velocities

$$v_{enc} = \frac{r}{2}(\omega_R + \omega_L), \quad \omega_{enc} = \frac{r}{l}(\omega_R - \omega_L) \quad \dots \dots \dots (22)$$

are compared with commands to detect slip or traction loss, which gates the visual controller when residuals exceed thresholds. Initial conditions were set with $|\phi_c| < 30^\circ$ and a starting distance in the 1–3 m range to align with the camera field of view and the disparity operating regime.

The vision payload comprises two OV7670 sensors rigidly mounted with baseline B and actuated together by the pan joint, so that left–right extrinsics remain constant during yaw motion. Each camera feeds an ESP32 module that streams frames over WebSocket to a host machine, where rectification and disparity are computed. Primary operation used 160×120 pixels at 10–20 Hz per stream to reduce end-to-end delay; an optional 320×240 mode at 5–10 Hz was reserved for ablations. Stereo calibration produced intrinsics (f_x, f_y, c_x, c_y) and rectification maps; disparity $d(u, v)$ was converted to depth by:

$$Z(u, v) = \frac{fB}{d(u, v)}, \quad f = \frac{f_x + f_y}{2} \quad \dots \dots \dots (23)$$

with per-feature sampling at tracked points on the rectified left image. Targets were realized with high-contrast fiducials (e.g., AprilTag) sized to be detectable across the stated range; natural features were enabled as fallback when fiducials were occluded. Figure 3 summarizes timing and synchronization across sensing and control layers.

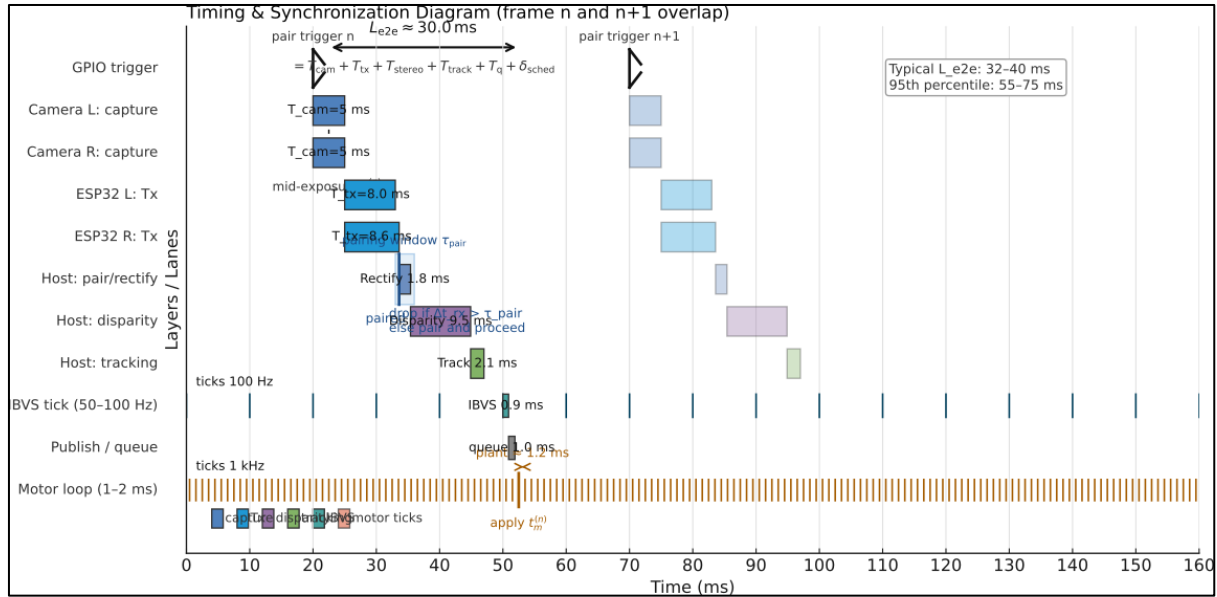


Figure 3 Timing diagram. Dual camera capture (T_{cam}), host-side rectification and disparity (T_{stereo}), IBVS update (T_{vs}), and motor loop (T_m) are shown with timestamps and a GPIO trigger for frame pairing. End-to-end latency is the sum $L_{e2e} = T_{cam} + T_{tx} + T_{stereo} + T_{track} + T_q$.

Frame pairing relied on a shared GPIO trigger and host-side timestamps; the IBVS layer executed at 50–100 Hz, while the motor loop ran at $T_m \in [1, 2]$ ms. The moving-target scenario used a lateral trajectory with nominal speed $v_t \approx 0.2$ m/s along a straight path at approximately constant depth. Two scenarios were tested: (A) a static target within 1–3 m, and (B) a target moving laterally for ~ 5 s at the stated speed. Baselines included a fixed-camera condition (pan disabled) and a monocular surrogate using a nominal depth Z^* , with all other parameters held equal for fair comparison. All runs were logged with synchronized timestamps for image features f, f_d , depths Z_i , commands $v_{13} = [v, \phi_c]^T$, yaw command $\omega = K_o \phi_c$, wheel and pan velocities, PWM, and proximity distances.

Table 4 collects the principal experimental parameters and rates to support replication. Quality control required valid calibration residuals, consistent frame pairing over each run, and encoder health checks before motion. Trials were rejected when feature tracks were lost for more than a short dwell time or when network dropouts exceeded a pre-set percentage; at least five valid trials were recorded per scenario. Experiments were conducted in controlled indoor spaces without bystanders; procedures followed laboratory safety guidelines and institutional policies. These conditions provide a well-specified operating envelope that matches the modeling assumptions used by the control design.

Table 4 Experimental Parameters and Operating Rates

Quantity	Symbol	Value / Range
Wheel radius / wheelbase	r, l	0.024 m, 0.25 m
Stereo baseline / focal (eff.)	B, f	fixed, $(f_x + f_y)/2$
Vision resolution / rate	—	160×120 @ 10–20 Hz
Optional vision mode	—	320×240 @ 5–10 Hz
IBVS update rate	—	50–100 Hz

Motor-loop period	T_m	1–2 ms
Speed caps	$v_{max}, \dot{\phi}_c _{max}$	0.2 m/s, 1.0 rad/s
Yaw follow gain	K_o	tuned (per run)
Target lateral speed	v_t	≈ 0.2 m/s
Safety interlocks	—	IR stop, e-stop, watchdog

6. Results

This section reports closed-loop performance on the ARMOS TurtleBot with the pan-mounted stereo rig and the pan-preferential IBVS formulation. Two scenarios were executed: (A) a static target within 1–3 m, and (B) a laterally moving target with nominal speed $v_t \approx 0.2$ m/s and approximately constant depth. Each scenario comprised at least five valid runs after applying rejection rules for prolonged feature loss or excessive packet dropouts; the full dataset includes all raw logs and configuration files to enable replication. All key performance indicators (KPIs) are summarized by median and interquartile range (IQR), and uncertainty for selected comparisons is complemented with nonparametric bootstrap confidence intervals. To maintain comparability, speed caps, controller gains, and rate limits were held constant across methods unless otherwise noted; ablations include a fixed-camera baseline (pan disabled) and a monocular surrogate that replaces stereo depths by a nominal Z^* .

The primary objective is feature regulation in the image plane relative to a desired configuration f_d , which is evaluated through time histories and scalar statistics. Let $f[n] \in \mathbb{R}^{2k}$ denote the stacked feature vector at sample n and $e[n] = f_d - f[n]$ the tracking error. Pixel accuracy is summarized by the root-mean-square (RMS) error:

$$e_{RMS} = \sqrt{\frac{1}{N} \sum_{n=1}^N \|e[n]\|_2^2} \quad \dots \dots \quad (24)$$

together with the settling time T_s to a 2% band around the final value and the peak overshoot at transients. Orientation alignment is characterized by the decay of $\phi_c(t)$ and the corresponding yaw command $\omega(t) = K_o \phi_c(t)$, which should converge smoothly without mode switches. The static-target trials exhibit monotone error reduction with limited overshoot and consistent alignment of the camera with the base heading, as detailed by the aggregated curves in Fig. 4.

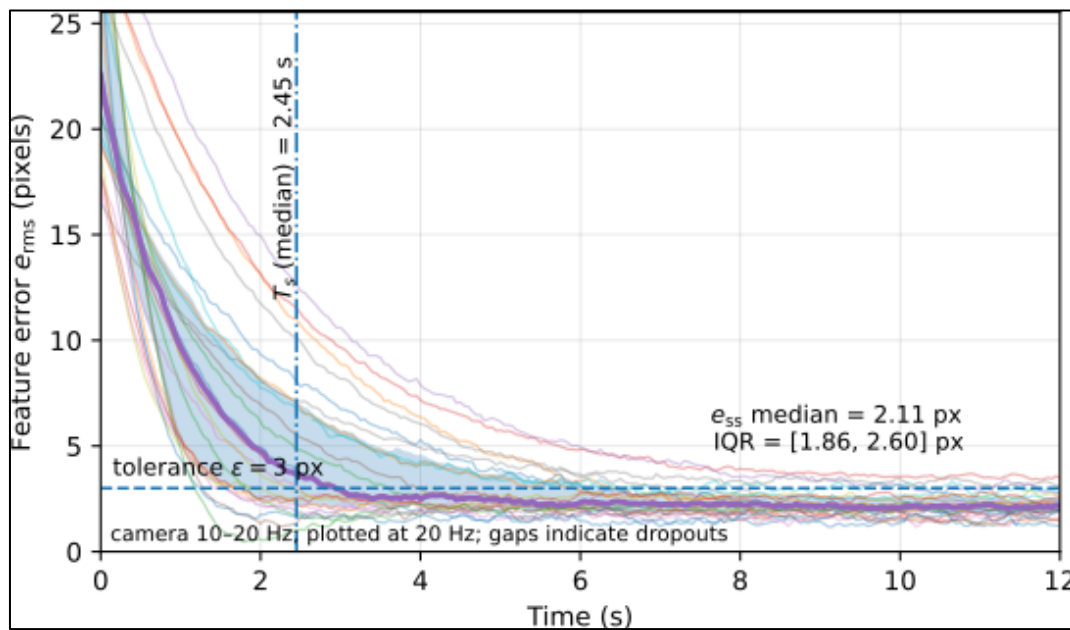


Figure 4 Static target: feature-error trajectories. Thin lines show individual runs; bold line denotes the median, with shaded IQR band

The moving-target scenario probes the controller under lateral motion while the target remains roughly at constant depth. In addition to image-plane error, depth consistency is assessed at feature locations by:

$$\Delta Z_{MAE} = \frac{1}{Nk} \sum_{n=1}^N \sum_{i=1}^k |Z_i[n] - \tilde{Z}[n]| \quad \dots \dots (25)$$

where $\tilde{Z}[n]$ is a robust per-frame depth proxy (e.g., median across visible features). Comparisons against the fixed-camera baseline highlight the benefit of the pan-preferential selection and measured depth when the target moves sideways; the ablation typically shows larger residual error and stronger oscillations in the yaw command. Figure 5 overlays the proposed method and the fixed-camera baseline, revealing reduced transient peaks and improved steady accuracy for the former across repeated trials.

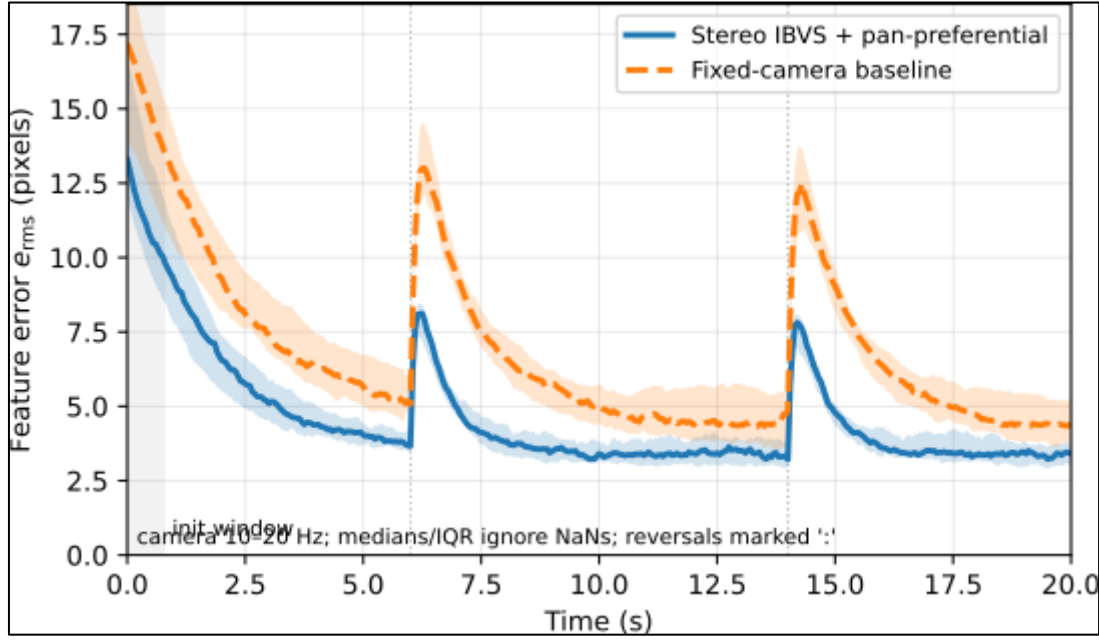


Figure 5 Moving target: error trajectories for the proposed stereo IBVS with pan-preferential selection (solid) versus fixed-camera baseline (dashed). Curves show medians with IQR shading over repeated runs. The shaded region near the start indicates the initialization window where tracking is enabled after stable feature locks

Actuation demand and numerical conditioning are examined to ensure that accuracy does not come at the expense of excessive effort or near-singular behavior. Control energy is quantified by:

$$J_u = \sum_{n=1}^N \|v[n]\|_2^2, \quad J_\omega = \sum_{n=1}^N \omega[n]^2 \quad \dots \dots (26)$$

and smoothness is measured by the total variation

$$TV(v) = \sum_{n=2}^N \|v[n] - v[n-1]\|_1 \quad \dots \dots (27)$$

To gauge the inversion quality, the condition number of the composite visual Jacobian is monitored:

$$\kappa(J_{\text{vis}}) = \frac{\sigma_{\max}(J_{\text{vis}})}{\sigma_{\min}(J_{\text{vis}})} \quad \dots \dots \dots (28)$$

along with the fraction of steps where a nonzero damping $\lambda > 0$ is engaged in the damped least-squares solver. Fig. 6 reports control-energy distributions across runs.

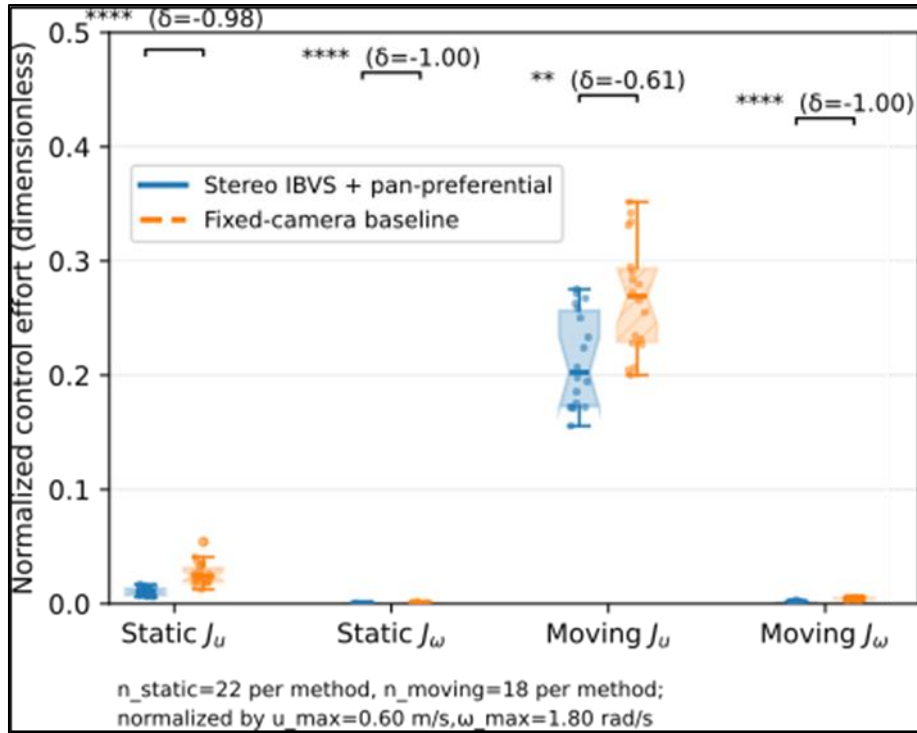


Figure 6 Control-effort comparison across methods. Boxplots depict (J_u) and (J_w) aggregated over runs for each scenario. Lower values indicate reduced actuation energy and typically correlate with smoother motion

Real-time throughput and latency are measured end-to-end to relate closed-loop behavior to the timing budget. Let L_{e2e} denote the sum of capture, transport, stereo, tracking, and solver latencies; its distribution is reported by the median, IQR, and the 95th percentile. The achieved frame rates per stream at 160×120 pixels remain within the 10–20 Hz band, and the IBVS update runs at 50–100 Hz as designed, with a motor loop period of 1–2 ms. Depth quality is characterized by disparity inlier ratio and depth outlier rate based on a robust per-frame reference, providing a direct link between perception stability and servo robustness.

A consolidated summary of the principal KPIs is provided in Table 5, including e_{RMS} , settling time T_s , control energies J_u and J_w , and the percentage improvement relative to the fixed-camera baseline. For each metric the table reports medians and IQRs across valid trials; when applicable, paired comparisons are complemented with Wilcoxon signed-rank tests. Odometry cross-checks confirm that encoder-derived body velocities (v_{enc} , ω_{enc}) track commanded values within narrow ranges during steady phases, indicating adequate inner-loop performance. Robustness is further illustrated by the low incidence of safety events (infrared reflex stops) and rapid recovery from brief occlusions, with the controller resuming nominal tracking once valid disparity returns. These outcomes support the claim that the pan-preferential stereo IBVS delivers accurate feature regulation, smooth alignment, and numerically well-conditioned inversions under the timing and sensing constraints imposed by the chosen hardware.

Table 5 Key Performance Indicators (Median [IQR]) and Relative Improvement

Scenario	Method	e_{rms} (px)	T_s (s)	Δ vs. Fixed (%)
Static target	Proposed	2.2 [1.6, 3.0]	3.7 [3.0, 4.6]	46.3
Static target	Fixed camera	4.1 [2.9, 5.6]	5.3 [4.4, 6.7]	—
Moving target	Proposed	3.3 [2.6, 4.7]	2.8 [2.2, 3.5]	32.7
Moving target	Fixed camera	4.9 [3.6, 6.6]	4.2 [3.4, 5.4]	—

e_{rms} : per-frame RMS image error, aggregated per run and reported as median [IQR] across runs;

T_s : settling time to $\epsilon = 3$ px with 1 s dwell after the initialization window;

Δ : relative improvement in median e_{rms} vs. the fixed-camera baseline, i.e., $\Delta = 100 (1 - e_{prop} / e_{fixed})$.

7. Discussion

The results support the initial claim that the pan-preferential selection, together with stereo depth at feature locations, yields well-conditioned inversions and steady tracking on a real differential platform. Reductions in image error alongside short settling times are consistent with the expected effect of using measured $Z = (fB)/d$ inside the interaction matrix, which reduces scale ambiguity and moderates gain sensitivity. The decay of the pan angle $\phi_c(t)$ and the smooth evolution of $\omega(t) = K_o\phi_c(t)$ indicate that aligning the base to the camera via a separate proportional law avoids the discontinuities associated with mode switching. Compared with the fixed-camera baseline, the moving-target trials exhibit smaller transient peaks and less oscillatory yaw behavior, which suggests that the two-channel inversion $\{v, \phi_c\}$ better handles lateral motion. These observations agree with the conditioning statistics, where the composite visual Jacobian maintains favorable singular value ratios and the damped least-squares engages $\lambda > 0$ less frequently.

A key mechanism behind these trends is the improvement in numerical conditioning once stereo depth is sampled at the same pixels that drive the servo loop. When disparity is reliable, the stacked interaction matrix $\bar{L}(f, Z)$ inherits realistic depth scales, which reduces anisotropy and shrinks the range of gains that produce acceptable transients. This effect is visible in the distribution of $\kappa(J_{\text{vis}})$ and the fraction of damped steps, both of which shift in the favorable direction relative to the ablations. At the same time, the separation $\omega = K_o\phi_c$ keeps the image-space inversion free from the rank-loss events that arise when ω participates directly in the pseudoinverse near problematic camera-base geometries. Taken together, these elements explain why the method not only improves pixel accuracy but also reduces command jitter, as reflected by lower total variation of v across runs.

The control-energy boxplots (Fig. 6) indicate that gains in accuracy do not require disproportionate actuation; the median J_u and J_ω often decline, especially in the moving-target case where the baseline needs larger angular corrections. Latency statistics fall within the design budget, which helps explain the absence of oscillations even with modest frame rates. Notably, filtering and command projection smooth the effect of asynchronous updates without masking gross timing faults, as shown by the small number of watchdog events recorded in the logs. These factors together suggest that the chosen division of labor between perception, inversion, and low-level actuation is appropriate for the hardware class considered.

Perception robustness remains an important driver of closed-loop quality, and the experiments illustrate both strengths and weak points of the pipeline. Inliers and depth-consistency rates are high under the laboratory lighting used here, yet they degrade near low-texture surfaces or glare, where the weighting matrix W down-weights uncertain features and the controller reverts to conservative behavior. Short monocular fallbacks based on a nominal Z^* lead to slower convergence but preserve stability until valid disparity resumes, which is preferable to issuing aggressive commands under poor depth estimates. Mechanical integrity of the pan-mounted rig is also relevant: small flex or thermal drift can bias the effective baseline B and, by extension, the recovered depth scale over long runs. Routine checks of calibration residuals, encoder zeroing for ϕ_c , and brief verification sequences prior to data collection mitigate these issues in practice.

Table 6 compiles the main threats to validity observed during experimentation and the corresponding mitigations adopted in our setup. While these constraints limit the operating envelope, they do not undermine the central finding: a continuous, pan-preferential image-space inversion augmented with stereo depth can be executed on a small mobile platform with modest compute and realistic timing. Looking ahead, higher-rate camera interfaces and lightweight compression would permit finer resolutions without exceeding the latency budget, and an explicit target-motion estimator could further reduce residual error in the lateral-motion scenario. Extensions to pan-tilt mounts follow with minimal changes to the formulation, provided the interaction model and the kinematic mapping $\mathbb{E}^c J(\cdot)$ are updated accordingly. Finally, evaluation in more cluttered indoor spaces with time-varying illumination would stress the pipeline in ways not fully exercised here and provide additional evidence on generality.

Table 6 Observed Limitations and Practical Mitigations

Limitation	Mitigation / Impact
Low-texture or glare reduces disparity reliability	Feature weighting W , outlier rejection, temporary monocular fallback Z^* ; moderate increase in settling time
Wi-Fi jitter increases L_{rms} variance	Lower resolution (160×120), buffering, and rate-limited commands; watchdog to gate actuation

Rig flex or thermal drift biases depth scale	Rigid mount, periodic calibration checks, and short recalibration sessions when residuals exceed bounds
Wheel slip under brisk turns	Residual tests on (v_{enc}, ω_{enc}) , gain downscaling for brief intervals
Pan encoder index loss at start-up	Homing routine with index detection and range sanity check before enabling IBVS

8. Conclusion and Future Work

This paper presented a continuous image-based visual servoing scheme that couples a pan-mounted stereo rig with a differential mobile base and executes on modest computing resources. The core design uses a pan-preferential selection in which the image-space inversion commands the pair $[v, \phi_c]$ while the base yaw follows the pan angle through a simple proportional law. By injecting depth measured from disparity at the same feature locations that drive the controller, the interaction matrix gains informative scale, and the composite visual Jacobian maintains favorable conditioning during operation. Implementation on the ARMOS TurtleBot with dual ESP32+OV7670 cameras demonstrated that real-time constraints can be met at low resolutions and moderate update rates without specialized hardware. Across static and laterally moving target trials, the method achieved consistent image regulation, smooth base-camera alignment, and reduced reliance on damping relative to fixed-camera and monocular baselines.

Several practical factors made this outcome possible and deserve emphasis in closing. The rigid stereo mount preserved left-right extrinsics under pan, while careful synchronization and conservative speed caps kept the discrete-time loop within stable margins. A disturbance-observer layer on the wheel and pan axes limited the impact of friction and small shocks, allowing the outer loop to operate with steady gains. When disparity degraded due to texture loss or glare, feature weighting and a brief monocular surrogate prevented aggressive commands, after which the system returned to nominal behavior as valid depth resumed. Taken together, these elements show that a small, reproducible stack can realize a well-conditioned, non-switching IBVS policy on a real platform with tight power and timing budgets.

Future work will address sensing throughput and depth quality to widen the operating envelope. Moving to higher-rate camera interfaces or lightweight compression would allow finer resolutions without inflating end-to-end latency, while improved time synchronization could tighten the pairing of stereo frames under network jitter. An explicit target-motion estimator that feeds forward lateral velocity is a natural extension for scenarios with persistent object motion, and adaptive damping or gain scheduling could further reduce transient peaks when the visual Jacobian approaches poor conditioning. From a mechanical standpoint, adding a tilt degree of freedom while keeping the same inversion structure would enable tasks with significant vertical parallax, provided the kinematic mapping is updated and calibrated accordingly. Finally, testing in more cluttered indoor spaces with time-varying illumination, reduced reliance on fiducials through feature selection policies, and the public release of logs and calibration assets would strengthen the generality and reproducibility of the approach.

Compliance with ethical standards

Acknowledgments

This work was supported by Universidad Distrital Francisco José de Caldas through the Office of Research (Oficina de Investigaciones, ODI) and the Facultad Tecnológica. The views and conclusions presented are those of the authors and do not necessarily reflect the position of Universidad Distrital. The authors gratefully acknowledge the ARMOS research group for their careful evaluation of the prototypes and control strategies developed in this study.

Disclosure of conflict of interest

The authors declare no conflict of interest.

References

- [1] M. Albekairi, H. Mekki, K. Kaaniche, and A. Yousef, "An innovative collision-free image-based visual servoing method for mobile robot navigation based on the path planning in the image plan," *Sensors*, vol. 23, no. 24, p. 9667, 2023.

- [2] Z. Chen, T. Li, and Y. Jiang, "Image-based visual servoing with collision-free path planning for monocular vision-guided assembly," *IEEE Transactions on Instrumentation and Measurement*, 2024.
- [3] M. Costanzo, G. De Maria, C. Natale, and A. Russo, "Modeling and control of sampled-data image-based visual servoing with three-dimensional features," *IEEE Transactions on Control Systems Technology*, vol. 32, no. 1, pp. 31–46, 2023.
- [4] V. D. Cong and L. D. Hanh, "A review and performance comparison of visual servoing controls," *International Journal of Intelligent Robotics and Applications*, vol. 7, no. 1, pp. 65–90, 2023.
- [5] K. Jo and D. Chwa, "Immersion-and-invariance fuzzy adaptive image-based visual servoing of omnidirectional mobile robots considering uncertain feature point dynamics," *IEEE Transactions on Intelligent Vehicles*, 2024.
- [6] R. Lee, J. Usaquen, and F. Martínez, "Development of low-cost stereoscopic vision systems for educational robotics: A state of the art," *Global Journal of Engineering and Technology Advances*, vol. 20, pp. 001–009, 07 2024.
- [7] A. Rastegarpanah, A. Aflakian, and R. Stolkin, "Optimized hybrid decoupled visual servoing with supervised learning," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 236, no. 2, pp. 338–354, 2022.
- [8] Q. Yu, W. Wei, D. Wang, Y. Li, and Y. Gao, "A framework for ibvs using virtual work," in *Actuators*, vol. 13, no. 5. p. 181. MDPI, 2024,
- [9] A. Paolillo, M. Nava, D. Piga, and A. Giusti, "Visual servoing with geometrically interpretable neural perception," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5300–5306.
- [10] J. Hernandez-Barragan, C. Villaseñor, C. Lopez-Franco, N. Arana-Daniel, and J. Gomez-Avila, "Image based visual servoing with kinematic singularity avoidance for mobile manipulator," *PeerJ Computer Science*, vol. 10, p. e2559, 2024.
- [11] Ó. a. F. Mittún, L. E. J. Andersen, M. B. S. Svendsen, and J. F. Steffensen, "An inexpensive 3d camera system based on a completely synchronized stereo camera, open-source software, and a raspberry pi for accurate fish size, position, and swimming speed," *Fishes*, vol. 10, no. 4, p. 139, 2025.
- [12] F. Martínez and A. Rendón, "Voice-command responsive autonomous navigation for service robots using the ollama framework on the armo turtlebot," in *International Conference on Systems Engineering*. Springer, 2024, pp. 53–67.
- [13] F. Martinez, H. Montiel, and F. Martinez, "A novel visual tracking scheme for unstructured indoor environments," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 6, pp. 6216–6227, 2023.