WALL-E: A Controlled Evaluation of Motion Control, Heading Estimation, and Object Ranging for a Low-Cost Inspection Rover

Rohan Chandra
Whiting School of Engineering

Johns Hopkins University
Baltimore, United States
rchand22@jh.edu

Aiperi Aya Yusupova
Whiting School of Engineering

Johns Hopkins University
Baltimore, United States
ayusupo1@jh.edu

***Abstract*—A mobile inspection robot depends on the reliability of the sensing and actuation beneath its decision-making interface. This paper presents WALL-E, a low-cost differential-drive rover, and a controlled protocol for evaluating three practical design choices: calibrated timed motion versus encoder-targeted stopping, wheel-only versus encoder–inertial heading estimation, and ultrasonic versus calibrated monocular object ranging. Fixed scripts issue all experimental commands, separating component performance from language-model variability. The planned study includes 15 paired motion blocks, 16 paired heading observations, and 64 paired ranging placements, with external physical ground truth, frozen tuning parameters, and explicit accounting for failures. Evaluation distinguishes accuracy on valid measurements from operational reliability and reports application-specific tolerances alongside uncertainty intervals. This manuscript describes the system and revised experimental protocol; main-study results are pending. The results shown are based on a one-session pilot run, not the main study. RQ3 was not piloted because the ultrasonic sensor and camera were not yet working. Its intended contribution is a reproducible, platform-specific characterization that supports subsequent closed-loop inspection and physical re-observation experiments.**

Keywords: mobile robotics, experimental evaluation, differential drive, sensor fusion, object ranging.

I. Introduction

An inspection instruction such as "approach the bin and take a photograph" depends on more than semantic understanding. A mobile robot must execute an appropriate displacement, estimate its orientation, and determine the distance to the intended object. Errors in these capabilities can invalidate an inspection even when the high-level instruction is correct. Consequently, a successful agent demonstration does not by itself establish that the underlying robot behaves reliably.

The goal of WALL-E is to carry out such requests through a ROSA ReAct agent, which turns a natural-language instruction into tool calls that move the robot, look at a target, and measure its range. This paper tests the components those tools depend on.

Low-cost differential-drive platforms offer an accessible setting for studying these dependencies. Their inexpensive motors and sensors support useful demonstrations, but wheel slip, motor asymmetry, geometric calibration, inertial drift, and object-dependent range returns complicate the relationship between a software command and physical behavior. These effects motivate evaluation of the assembled platform against external measurements rather than evaluation against its own odometry.

This paper describes WALL-E and a controlled comparison of three design choices. First, does encoder-targeted stopping reduce physical endpoint error relative to a calibrated timed command? Second, does wheel–inertial fusion improve heading estimates on the same physical turns? Third, how do ultrasonic and calibrated monocular ranging compare in valid-reading accuracy, measurement availability, and latency across controlled objects, distances, and angles?

The study separates these questions from language-model planning. Fixed scripts issue commands, both heading estimators observe the same turns, and both ranging methods measure the same placements. Physical ground truth is collected independently of the estimates. Operational failures remain visible rather than being replaced by an arbitrary error equal to a success threshold.

The intended contributions are: (1) a documented low-cost sensing and actuation architecture; (2) a paired evaluation protocol connecting component errors to inspection tolerances; and (3) a reporting framework that distinguishes valid-output accuracy, failure frequency, and acquisition cost. These are systems-evaluation contributions, not new control, filtering, or depth-estimation algorithms. This paper reports the protocol together with results from a one-session pilot run. The main study has not been run, so we make no comparative claims.

The rest of this paper is organized as follows. Section II reviews related work, and Section III describes the system. Section IV gives the experimental protocol and the pilot. Section V reports the pilot results, Section VI discusses limitations, and Section VII concludes.

II. Related Work

A. Odometry and Motion Execution

Borenstein and Feng identify unequal wheel diameters and uncertainty in effective wheelbase as important systematic error sources for differential-drive odometry, and develop practical calibration procedures [1]. Their work demonstrates why nominal chassis geometry alone is insufficient for accurate dead reckoning. WALL-E's motion comparison addresses a complementary operational question: whether an odometry-based stopping condition improves externally measured endpoint accuracy relative to a calibrated timed command. Encoder displacement supplies feedback but remains vulnerable to slip and scale error; it is therefore not used as ground truth.

B. Wheel–Inertial State Estimation

Moore and Stouch describe a generalized extended Kalman filter implementation for robot state estimation [5]. The associated `robot_localization` framework provides the implementation used by WALL-E [6]. This study compares wheel-only and fused heading estimates from identical turns. Such pairing isolates differences in estimation from differences in executed movement. Because neither estimate controls the turn's stopping condition, the experiment measures heading estimation rather than the effectiveness of a heading-feedback controller. An improvement in this experiment would motivate, but would not establish, improved route-following performance.

C. Monocular and Ultrasonic Ranging

Depth Anything V2 provides learned monocular depth estimation [7]; its official implementation also supplies separately fine-tuned metric-depth models [2]. WALL-E uses an indoor metric model with an additional platform-specific calibration. A metric checkpoint must be distinguished from a relative-depth model: numerical depth output is not automatically an independently validated physical range.

An HC-SR04 estimates reflector distance using acoustic time of flight [3]. Its return is not inherently associated with an image-selected object. Conversely, a selected image region may include background pixels or fail under lighting changes. The proposed comparison therefore evaluates two complete ranging pipelines and records invalid outputs explicitly. Off-axis placements are an intentional stress condition, not evidence that the ultrasonic sensor necessarily observes the intended target at every angle.

D. Positioning of This Study

These methods are established individually. The present study contributes an application-specific comparison on an assembled inspection rover with shared conditions and independent reference measurements. It does not claim a new fusion algorithm, open-world object recognition, or field generalization. Its value depends on reproducible configurations, transparent failure accounting, and evidence about which design choices matter for this platform.

III. System Overview

A. Hardware and Execution Architecture

Fig. 1 shows the full system. WALL-E uses a two-wheel differential-drive chassis controlled by a Raspberry Pi Pico 2 W. A TB6612FNG driver actuates two TT 1:90 gear motors with quadrature encoders. A BNO055 supplies inertial measurements in IMU fusion mode without a magnetometer. An ESP32-CAM with an OV3660 faces forward, and an HC-SR04 is mounted below it. The source build places the camera approximately 8.26 cm above the floor.

The source configuration reports nominal 65 mm wheels and 130 mm separation, while its odometry configuration uses 63.5 mm and 127 mm. These values are not treated as interchangeable: actual wheel geometry, encoder interpretation, and deployed odometry parameters must be verified and recorded in the configuration manifest before main data collection. The reported encoder convention is 12 pulses per motor revolution, quadrature decoded at $\times 4$ through a 90:1 gearbox, yielding 4320 counts per output revolution; this convention must be checked by rotating each wheel through a measured revolution.

The Pico publishes encoder and IMU measurements through micro-ROS. Host-side ROS 2 nodes perform odometry, filtering, image processing, command execution, and logging. The reported host is an Apple M3 Pro laptop with 36 GB memory, running an Ubuntu 22.04/ROS 2 Humble container. Depth inference is CPU-only. A pre-study manifest will record exact container, package, model-revision, and firmware identifiers rather than relying on mutable model aliases. The pilot code and data are in the `pilot-run` branch of the project repository: https://github.com/rohanchandra-cpe/walle-litter-bot/tree/pilot-run.

The HC-SR04 is mounted directly below the camera with its transducer face in the same vertical plane as the camera lens. Its TRIG pin is on GP6, and its ECHO pin is on GP7 through a 1 kΩ/2 kΩ divider, because the ECHO output is 5 V and the Pico inputs are 3.3 V. The Pico and HC-SR04 run from a 5 V BreadVolt supply; the motors run from a separate supply consisting of 6 AA batteries, providing 9V.

Power wiring and driver supply limits must be checked against the installed components before energizing the system. The reported six-AA motor supply and logic supply are logged separately; battery chemistry determines charging or replacement procedures.

B. Motion and Heading Pipelines

Velocity commands reach the motor-control firmware through a fixed script interface. RQ1 compares a timed straight command with a wheel-odometry-targeted command. The original configurations also differ in velocity profile: the timed command uses a fixed requested speed, whereas the targeted command ramps its speed. Accordingly, the primary comparison concerns *controller configurations*; it cannot attribute any observed difference solely to encoder feedback.

Wheel-only heading is computed from differential wheel displacement. An EKF combines wheel forward speed and yaw rate with IMU heading and yaw rate in two dimensions. The IMU heading is relative to the start pose. The

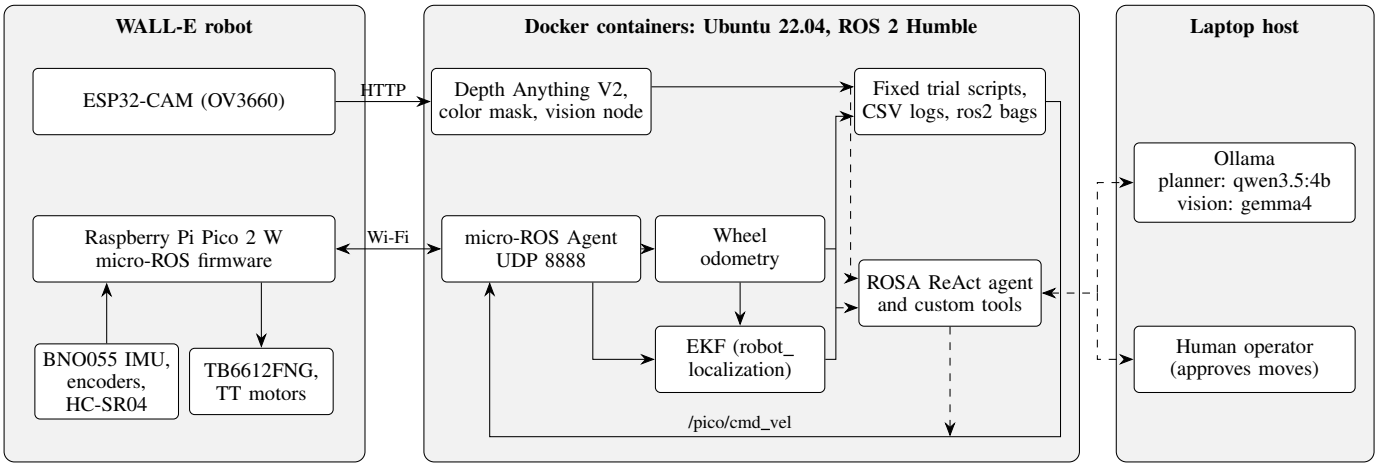


Fig. 1. WALL-E architecture. Solid links are the experiment path: fixed scripts issue every command and record all data. Dashed links are used only by the ROSA agent, which is excluded from all three comparisons. Ultrasonic pings reach the host on the same `/pico/wheel_ticks` message as the encoder counts.

covariance and measurement-selection configuration is frozen before testing. Because IMU heading and angular rate share sensor information, the experiment must not describe them as independent ground-truth sources or assume that fusion necessarily improves accuracy.

C. Ranging Pipelines and Geometry

The ultrasonic pipeline takes the median of five valid pings spaced 60 ms apart; a 30 ms echo timeout is logged for each ping. A placement returns a valid aggregate only if at least three pings are valid. The temperature-adjusted sound speed is

$$c(T) = 331.3 + 0.606T \quad \text{m/s}, \qquad \hat{r}_u = ct/2. \tag{1}$$

The image pipeline uses a fixed color mask to select the controlled target, followed by a median depth in a predefined window within that mask. It does not use an LLM detector during the experiments. The checkpoint is `depth-anything/Depth-Anything-V2-Metric-Indoor-Small-hf`. Preprocessing, including the source draft's requested 364-pixel short side, actual tensor dimensions, window size, mask thresholds, and model revision, is fixed in the manifest.

The target range is defined as *horizontal distance* from a shared sensor reference plane to the nearest target surface. For camera axial depth z, a calibrated horizontal bearing β gives

$$\hat{r}_c = z\sqrt{1 + \tan^2\beta} = z/\cos\beta. \tag{2}$$

This conversion assumes a validated axial-depth convention and camera intrinsics. It is not a general three-dimensional slant-range formula. The camera is aligned approximately horizontally, and the image window is chosen near the ultrasonic measurement height where target geometry permits. Differences in sensor origin, pitch, target surface, and returned reflector are documented. Calibration absorbs a fixed scale error but cannot guarantee object association.

D. Scope of the Agent Interface

A ROSA ReAct agent can invoke the custom tools in Table I in a demonstration, using a planning model at temperature 0 with a 16,384-token context and a separate vision model, both served by Ollama. The agent code is in `walle-pilot/rosa_agent/` in the repository. The agent is excluded from all three comparisons. The experiments issue no natural-language judgments, measure no tool-selection accuracy, and do not evaluate end-to-end autonomous inspection. This boundary makes component failures interpretable and establishes prerequisites for later physical re-observation studies.

IV. Experimental Protocol

A. Research Questions and Estimands

RQ1 (primary): Does encoder-targeted stopping reduce absolute forward endpoint error relative to calibrated timed motion? Both request 50 cm. Let x_i be externally measured forward displacement and y_i the lateral displacement. Endpoint error is $e_{d,i} = x_i - 50$ cm. The paired estimand is

$$\Delta_d = \mathbb{E}[|e_{d,A}| - |e_{d,B}|]. \tag{3}$$

Positive values favor B. The smallest effect of interest is 2 cm. A successful run must complete without a gate violation and satisfy $|e_d| \le 5$ cm and $|y| \le 5$ cm.

RQ2 (secondary): Does the EKF reduce absolute heading error relative to wheel odometry on the same turn? With externally measured terminal yaw ψ_{true},

$$e_\psi = \text{wrap}_{[-180,180)}(\psi_{\text{est}} - \psi_{\text{true}}). \tag{4}$$

The estimand is $\Delta_\psi = \mathbb{E}[|e_{\psi,w}| - |e_{\psi,f}|]$. The smallest effect of interest is 4°, and the application tolerance is 10°. These are engineering criteria, not guarantees about accumulated route error.

TABLE I
CUSTOM ROSA TOOLS. EACH RETURNS A SHORT TEXT RESULT. TOOLS THAT MOVE THE ROBOT REQUIRE HUMAN APPROVAL BEFORE STARTING, STOP ON A REFUSED OR FAILED CHECK, AND REPORT THE FUSED POSE. THESE TOOLS ARE NOT USED IN THE EXPERIMENTS.

Tool	Purpose	Inputs → output	Safety checks
`look`	Check a photo for one object	Target name → visible, position, size, path clear	Read only; a camera error tells the agent not to move
`describe_scene`	Describe the current view	None → description, path clear	Read only; same camera-error rule
`measure_distance`	Measure range with the depth model	Color and name → distance, bearing, range ahead	Read only; a depth error tells the agent not to move
`get_pose`	Report the fused pose	None → x, y, heading	Read only; reports a missing EKF topic
`drive_distance`	Drive straight by the fused estimate	Direction, meters → measured move	Up to 2 m; rejects invalid arguments
`drive_forward_if_clear`	Drive only if the camera sees no obstacle	Meters → result, pose	Up to 0.5 m; takes a photo first and refuses if blocked
`turn`	Turn in place by the fused estimate	Left or right, degrees → measured turn	Up to 180°; rejects invalid arguments
`go_to`	Turn toward a point and drive to it	x, y → measured move, miss distance	Refuses goals over 3 m away
`search_for`	Turn and look until an object is seen	Target name → found or not, pose	45° steps, one full circle at most
`approach`	Drive to a visible object	Target name → arrived, stopped, or lost	Looks before each move; 0.3 m steps, 20 moves at most; stops if blocked
`approach_to`	Stop a set distance from an object	Color and name, stop distance → result, pose	0.5–1.5 m stop distance; measures before each move; stops if depth disagrees with travel
`stop_robot`	Stop immediately	None → confirmation	Needs no approval

TABLE II
PLANNED COMPARISONS AND INDEPENDENT PAIRING UNITS.

RQ	Conditions	Allocation
RQ1	A: calibrated timed command; B: encoder-targeted stop	15 blocks, 30 runs
RQ2	Wheel-only and EKF heading on the same open-loop turn	16 turns, 8 each direction
RQ3	Ultrasonic and calibrated metric monocular ranging	4 objects × 8 ranges × 2 angles

RQ3 (exploratory): How do ultrasonic and monocular ranging trade off valid-reading error, availability, and latency across the specified placement grid? For a valid estimate,

$$\mathrm{AE} = |\hat{r} - r|, \qquad \mathrm{APE} = 100|\hat{r} - r|/r. \tag{5}$$

APE is not capped and can exceed 100%. The target tolerance is 15% APE; 5 percentage points is a descriptive effect-of-interest reference. A joint operational success requires a valid output and APE $\leq 15\%$. These criteria characterize the setup rather than establish collision safety.

B. Conditions and Trial Allocation

RQ1-A sends 0.15 m/s with zero angular command at 20 Hz for a shakedown-calibrated duration T_A. RQ1-B ramps between 0.05 and 0.15 m/s and stops at 0.5 m of wheel odometry, using the verified geometry rather than a hard-coded count target. A fixed 0.2 s coast setting is retained if confirmed in shakedown. RQ2 uses open-loop commands of ± 0.6 rad/s for a calibrated duration T_ψ to generate approximately 90° turns. Neither heading estimate determines when those turns stop.

RQ3 uses a red paper cup, upright blue 500 mL bottle, approximately 20 cm green bin, and yellow crumpled paper wrapper. Distances are 0.30, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, and 2.00 m; horizontal angles are 0° and 20°. Every placement is measured by both pipelines. A wrapper too short to intersect the acoustic beam is retained as a documented geometry challenge rather than silently repositioned after observing a failure.

RQ1 and RQ2 span three sessions: five RQ1 blocks per session and 5, 5, and 6 turns. RQ3 placements are balanced across sessions where practical and randomized within the placement schedule. A seed of 42 generates the schedule; the actual realized order and deviations are retained. A/B order within each motion block is randomized. Ranging acquisition order is counterbalanced to reduce exposure to time-varying conditions.

C. Environment and Calibration

The study uses one indoor area with a smooth wooden surface. A 1 m by 0.3 m lane and a printed protractor define the motion and heading areas. The scene in front of the range sensors is kept clear except for the intended target. Ceiling lighting and closed blinds are held consistent; temperature, approximate illuminance, supply voltage, and session identifiers are recorded. A phone light-meter reading is an environmental covariate, not a calibrated photometric reference.

Before testing, at least five shakedown executions per motion command determine T_A, T_ψ, and stopping settings. These runs do not enter the main comparison. Camera intrinsics and the common sensor reference are checked before fitting depth calibration on a separate object at eight distances spanning approximately 0.35–1.85 m and distinct from test distances. The exact calibration grid is recorded. Selection between a scale-only and scale-plus-power fit uses leave-one-out calibration error, not test errors. Extrapolation to 0.30 and 2.00 m is explicitly reported.

All parameters are frozen before main data collection: geometry, encoder convention, velocity profiles, stop thresholds, EKF settings, image masks, inference preprocessing, and depth calibration. A source commit and configuration manifest

accompany every session. This revised protocol supersedes the earlier draft only if adopted before main testing; any change after data inspection must be labeled a post hoc amendment.

D. Independent Ground Truth

For RQ1, a fixed chassis pin marks the midpoint of the wheel axle. A steel tape with millimeter graduations measures terminal forward and lateral displacement relative to the starting marks. For RQ2, a chassis-aligned pointer is read against a printed protractor with 1° graduations; overhead video provides an audit trail. The uncertainty of hand measurement is measured empirically and is not assumed to equal the instrument graduation.

For RQ3, the reference is the horizontal distance from the shared sensor plane to the target's nearest surface at the specified angle. Surface shape and measurement height are recorded. The reference geometry is the same for both methods; a center-point image depth is not silently compared to a different surface without noting the mismatch.

Before the main study, both measurers independently repeat five representative poses five times. The resulting variability describes measurement repeatability. A randomly selected 20% of main observations is double-measured, and absolute inter-measurer disagreement is reported. The measurer does not see numerical robot estimates or the strategy label, although complete blinding may be impossible when motion profiles are visibly different. Reference values are entered before estimates are revealed for scoring. Paired reference measurements are retained rather than replaced selectively after inspecting error.

E. Execution and Failure Handling

At each session, check supply voltage, common ground, driver readiness, sensor publication, and emergency-stop operation. Hold the robot stationary for IMU settling and record calibration status; a 30 s wait alone does not establish calibration. Complete two designated warm-up motion runs, excluded in advance.

For each motion trial, align the axle pin and pointer, reset wheel and fused initial poses, start logging, and execute the scheduled command. Allow up to 10 s for completion. After the robot rests, record both estimates and external ground truth. A 500 ms command-loss watchdog stops the motors. For each range placement, measure the reference, acquire both outputs in the scheduled order, and save the image, mask, depth map, individual ping status, aggregate result, and timing.

Operational failures remain outcomes. A lane exit, collision, stall, emergency stop, or timeout sets run success to zero. If a terminal physical pose can be measured, retain its actual error even when it exceeds tolerance. An unavailable endpoint is marked missing rather than assigned an invented 5 cm error. Missingness and failure reasons are reported by condition. A range timeout or unusable target mask produces an invalid measurement, not a fabricated range or capped APE.

A trial invalidated by external interference or loss of the reference measurement is flagged, retained in the audit log, and repeated at the session end under the same assigned condition. Such repeats replace neither genuine hardware failures nor poor numerical outcomes. If device log loss prevents recovery, its cause is disclosed and the affected endpoint identified. Exclusion rules are frozen before testing.

F. Pilot Run and Resulting Changes

We ran a pilot on the real robot on November 3, 2026, using the session 1 schedule. Pilot data are reported separately and are not pooled with the main study. The trial log is `walle-pilot/data/logs/motion_log.csv` in the repository, and the analysis outputs are in `walle-pilot/data/results/`.

Shakedown. Tuning runs came first and are excluded from every comparison. With placeholder durations, strategy A commanded for 3.33 s traveled a mean of 92.0 cm (SD 2.6, $n = 5$), and strategy B reached a mean of 49.2 cm (SD 0.6, $n = 5$). Wheel odometry showed a steady speed near 0.27 m/s under the 0.15 m/s command, because the firmware maps velocity directly to PWM without feedback. Turns commanded for 2.62 s appeared to reach only 69°, but encoder counts and the gyroscope showed about 430°: the robot had spun one extra full rotation that neither the protractor nor the wrapped heading could show. Lengthening the turn, as the wrapped value suggested, would have been wrong.

Pilot configuration. We set $T_A = 1.773$ s and $T_\psi = 0.497$ s, and kept the 0.2 s coast setting. Odometry used a 64.8 mm wheel diameter, 127 mm wheel separation, and 4320 counts per revolution. The session manifest records configuration hash `62cbfd1a2c`. No source commit was recorded, because the working copy was not a git checkout.

Session. At the session check the motor supply read 8.24 V, the logic supply 4.96 V, the ground offset 0 V, and the driver logic 2.8 V, with $21.7\,^\circ$C and 121 lux on the phone meter. After a 30 s still period, encoder, IMU, and odometry topics ran at 49 Hz and the EKF at 43 Hz. Two warm-up runs were excluded in advance. The remaining 15 trials were five RQ1 blocks and five RQ2 turns (three CCW, two CW). All 17 trials took about 12 minutes. One author both operated the robot and measured ground truth.

Deviations from the protocol. RQ3 was not piloted: the session check recorded no ultrasonic response on 10 of 10 pings, the camera was unreachable, and no depth calibration had been fitted. Ground-truth repeatability and double measurement were not performed, so label noise is unknown. The schedule's direction balance gave only two CW turns. We also deleted shakedown runs SD021–SD039 from the log because they used a faulty time duration, so the run numbers skip from SD020 to SD040 and the retained log does not show every run behind the final durations. These were tuning runs and would not have entered any comparison.

Changes to the plan. (1) Durations are now calibrated against measured motion, not the commanded speed. (2) Shakedown turns are checked against encoder totals, so a full extra rotation cannot pass as a short turn. (3) The ultrasonic and camera faults must be fixed and the depth model calibrated before any RQ3 data are collected. (4) Protractor repeatability, including CW readings, must be measured before the main study (Section V-C). (5) Bag recording needs a longer start-up wait: IMU data were complete for only two of the five turn

bags, although the logged estimates, which come from the live nodes, were unaffected.

G. Statistical Analysis

Accuracy and reliability are separate endpoints. For each condition, report attempted, valid, failed, and externally invalid trials; mean, median, standard deviation, and maximum measured error; and completion or detection fractions. Plot all paired observations, signed errors, run order, and engineering tolerances. Report time and movement costs descriptively.

For RQ1, the planned primary comparison is a one-sided paired t-test of $H_0 : \Delta_d \leq 0$ against $H_1 : \Delta_d > 0$. The analysis uses blocks with both endpoint measurements available, including operational failures with measurable endpoints. Report the mean paired difference with a two-sided 95% t interval. No normality-test result will trigger switching to a different primary test. Inspect the paired-difference plot and report a paired sign-flip sensitivity analysis, stating its symmetry/exchangeability assumptions and the limited sample size.

RQ2 uses the same paired mean-error framework on the 16 turns, with $H_0 : \Delta_\psi \leq 0$ against $H_1 : \Delta_\psi > 0$. If inferential claims are made for both RQ1 and RQ2, apply Holm correction across their two one-sided p-values [4]. Label the individual 95% intervals as unadjusted estimation intervals; do not describe them as simultaneous intervals.

Missing physical endpoints can make complete-pair comparisons optimistic. Therefore, report their number and causes, the complete-pair estimand, and failure fractions alongside every accuracy claim. If missingness is substantial or differs by condition, restrict conclusions to observable endpoints and avoid unconditional superiority claims. Success-rate differences and failure categories are descriptive at this sample size. A failure is never discarded merely to improve normality.

RQ3 is exploratory and descriptive. The 64 placements are a crossed grid involving only four physical objects; they are not 64 independent draws from an object population. Report paired valid-reading differences on the common-valid subset, method-specific valid-reading errors, joint success fractions, and latency. Break down results by object, distance, and angle and show failure locations. Common-valid results may select easier placements, so they are interpreted with availability rather than in isolation. No confirmatory test across 64 ostensibly independent placements or claim of generalization to unseen object classes is made.

The planned sample is a feasibility-scale study, not a demonstrated power guarantee. The original assumed paired SDs of 2.5 cm and 5° imply standardized effects of 0.8 for the specified RQ1 and RQ2 engineering differences, but actual variance, failures, and session dependence are unknown. No observed first-session result changes the planned sample or thresholds. Subsequent studies can use variance estimates to plan larger, independently replicated experiments.

A positive corrected test is evidence of a directional difference, not automatically engineering relevance. Interpret the estimate and interval relative to 2 cm or 4°. A nonsignificant result does not establish equivalence. All additional subgroup analyses and any amendments after data inspection are labeled exploratory.

H. Confounders and Reproducibility

Pairing reduces gradual battery and floor changes within motion blocks but cannot eliminate sudden mechanical faults. Log voltages, warm-up status, slips, sensor calibration, and network gaps. Fixed reset marks reduce initial-pose variability; repeated reference measurements quantify residual noise. Both heading estimators share wheel data, so their errors are correlated by design. Session-level dependence limits interpretation of trial-level intervals; report session-specific estimates and consistency in addition to pooled values.

The reproducibility package should contain command scripts, launch files, firmware, calibration data, the realized schedule, frozen configuration, raw logs, image/depth outputs, and the analysis script. The pilot's assets are public in the `pilot-run` branch of the project repository. Under `walle-pilot/data/`, `logs/` holds the trial, shakedown, and session logs; `bags/` the ROS bags; `schedule/` the seed-42 schedules; `manifests/` the session manifest; and `results/` the analysis outputs. No image or depth outputs exist yet, because RQ3 was not piloted.

V. Results and Evaluation

Study status. Every number in this section is a pilot result from one session, unless it is labeled as a shakedown value. The main study has not been run. With five pairs per comparison, these results are descriptive: the planned tests are shown for completeness, and we claim no statistical significance. RQ3 has no pilot data.

A. Pilot Accounting

Table III summarizes the pilot, analyzed with the planned methods. All ten RQ1 runs and all five RQ2 turns completed with measurable endpoints; there were no failures, external invalidations, repeats, or missing values. All five blocks and five turns form complete pairs, out of 15 and 16 planned. The largest gap in wheel-encoder messages was 112 ms, well below the 500 ms watchdog.

B. RQ1: Endpoint Error

Both strategies landed well inside the 5 cm tolerance on every run (Table III, Fig. 2). The direction was opposite to our hypothesis: A was closer to 50 cm than B in four of five blocks, with $\Delta_d = -0.80$ cm. The interval's upper limit of $+0.24$ cm is below the 2 cm effect of interest, so the pilot gives no sign of a meaningful benefit from B. It does not show that A is better. A's duration was tuned earlier the same evening at nearly the same motor voltage, which is its most favorable case. A timed command should be most exposed to changes between sessions, and one session cannot test that.

B stopped short in four of five runs, with a mean signed error of -1.02 cm. Its odometry ended at a mean of 47.8 cm, so its stopping rule cut power earlier than the robot's coast needed; the tape read about 2.6% more than odometry, which recovered part of the shortfall. This bias is small, but it is larger than B's own spread and could be reduced by retuning



Fig. 2. Pilot results from session 1. Left: paired absolute endpoint error for RQ1, with the 5 cm tolerance dashed. Center: signed endpoint error by run order. Right: paired absolute heading error for RQ2 on the same turns, with the 10° tolerance dashed. Both CW turns (orange) have a protractor reading that disagrees with the robot's sensors, as discussed in Section V-C.

TABLE III
PILOT RESULTS FROM SESSION 1 (NOT MAIN-STUDY EVIDENCE). INTERVALS ARE UNADJUSTED TWO-SIDED 95% t INTERVALS; Δ IS POSITIVE WHEN THE SECOND OPTION IS BETTER.

RQ1: 50 cm straight	**A timed**	**B encoder**
Attempted / failed	5 / 0	5 / 0
Run success	5/5	5/5
Mean (SD) $\lvert e_d \rvert$, cm	0.26 (0.21)	1.06 (0.68)
Maximum $\lvert e_d \rvert$, cm	0.60	1.70
Mean signed e_d, cm	$+0.14$	-1.02
Mean (SD) $\lvert y \rvert$, cm	1.50 (1.33)	0.98 (1.08)
Mean completion, s	2.07	2.02
Δ_d (5 pairs), cm	-0.80, CI $[-1.84, +0.24]$	
Cohen's d_z	-0.95	
p / Holm p / sign-flip p	0.950 / 0.950 / 0.969	
RQ2: $\sim 90^\circ$ turn	**Wheel**	**EKF**
Mean (SD) $\lvert e_\psi \rvert$, deg	6.75 (6.98)	5.02 (4.02)
Maximum $\lvert e_\psi \rvert$, deg	14.70	9.61
Turns within 10°	3/5	5/5
Mean $\lvert e_\psi \rvert$, CCW ($n = 3$), deg	1.75	2.17
Mean $\lvert e_\psi \rvert$, CW ($n = 2$), deg	14.26	9.30
Δ_ψ (5 pairs), deg	$+1.73$, CI $[-2.82, +6.29]$	
Cohen's d_z	$+0.47$	
p / Holm p / sign-flip p	0.175 / 0.350 / 0.156	

before the main study. Both strategies also took about 2 s, so at the real speed B's ramp changed the motion very little. Both drifted slightly right, with a mean heading change near -3.5°, and neither exceeded 2.7 cm of lateral error.

C. RQ2: Heading Error

The EKF's mean error was 1.73° lower than wheel odometry's, but the interval spans zero and includes both a loss of 2.82° and a gain above the 4° effect of interest. The pilot does not settle RQ2.

This result depends on a reference measurement we doubt. The entire EKF advantage comes from the two CW turns. On those turns the protractor read exactly 75° both times, while wheel odometry gave 89° and 90° and the EKF gave 84° and 85°. The one CW turn with complete IMU data integrates to 88° of gyroscope rotation, against the 75° reading. Encoder counts were also nearly identical between CW and CCW turns, yet the three CCW readings of 86–89° agreed with wheel odometry within 4°. A true shortfall of about 13° on CW turns alone therefore seems unlikely, and a reading or alignment error is plausible. We have kept these readings as recorded. In a post hoc look at the three CCW turns only, the two estimators were indistinguishable (means of 1.75° and 2.17°).

D. RQ3: Ranging

No ranging results exist. As described in Section IV-F, both sensors failed the session check, so the pilot's only RQ3 finding is that the sensing chain was not yet ready.

E. What the Full Study Needs

The pilot's paired SDs were 0.84 cm for RQ1 and 3.67° for RQ2. At those values, 15 blocks and 16 turns would give over 99% power for the 2 cm and 4° effects of interest (one-sided $\alpha = 0.05$). Five pairs estimate an SD poorly, however. At the one-sided 95% upper bounds of those SDs (1.99 cm and 8.70°), power is 98% for RQ1 but only 54% for RQ2. RQ2's SD also depends on the disputed CW readings, so we are keeping the planned sample sizes, as the protocol requires. The main study needs all three sessions, balanced turn directions, measured ground-truth noise, a second measurer, and working RQ3 sensors.

F. Decision Criteria

A component recommendation will weigh error, uncertainty, reliability, and time cost together. Conflicting outcomes may support complementary use rather than replacing a sensor. Passing these indoor tolerances implies no safety certification or general field readiness. The pilot supports no component recommendation yet.

VI. DISCUSSION AND LIMITATIONS

The study's principal strength is interpretability: independent physical measurements, paired comparisons, and exclusion of agent variability make component errors easier to attribute. Its strongest possible result would be evidence that a specific configuration improves a meaningful endpoint without an unacceptable reliability or time penalty. A null or mixed result can also inform design if uncertainty and failures are fully reported.

The scope is limited to one rover, short motion primitives, one indoor smooth wooden surface, and four controlled objects. Repeated placements do not substitute for new objects, days, robots, or environments. Hand measurements and imperfect blinding introduce uncertainty; session dependence can make nominal trial-level intervals overconfident. The camera and ultrasonic sensor can sample different physical surfaces, especially on angled or irregular targets. Color masking standardizes target selection but excludes the recognition difficulty of natural litter.

The protocol establishes neither an active-perception policy nor learning under distribution shift. It instead provides a prerequisite for such work: evidence about whether a requested movement changes the robot's physical viewpoint predictably and whether its distance measurements are usable. A subsequent "learning when to look again" study should freeze these component settings, introduce independent scene splits, and test whether the benefit and accepted-error risk of acquiring another viewpoint can be predicted before movement.

VII. Conclusion

This paper presented WALL-E and a controlled protocol for evaluating motion execution, heading estimation, and object ranging on a low-cost rover. A one-session pilot showed that the motion protocol runs end to end: 17 trials took about 12 minutes with no failures. All findings so far are provisional. In this pilot, timed and encoder-targeted motion both met the 5 cm tolerance, with no sign that encoder targeting helps by 2 cm, and the heading comparison was inconclusive because of doubtful CW reference readings. The most useful lessons came from checks against independent measurements: a 0.15 m/s command actually produced about 0.27 m/s, and an apparent 69° turn was really about 430°.

The evidence is limited to five pairs per comparison, one session, one operator who also measured, and an unknown ground-truth error; RQ3 has no data at all. In the remaining weeks we will repair the ultrasonic and camera links, fit the depth calibration, measure protractor and tape repeatability, and fix the bag start-up timing. We will then freeze the configuration, commit it, and run all three main-study sessions before making any claim about which configuration is better.

Author Contributions

Rohan Chandra wrote the Experimental Protocol, ran the pilot, and wrote the Results and Evaluation and Conclusion sections. Aiperi Aya Yusupova completed all other sections.

References

[1] Johann Borenstein and Liqiang Feng. Measurement and correction of systematic odometry errors in mobile robots. *IEEE Transactions on Robotics and Automation*, 12(6):869–880, 1996.

[2] Depth Anything V2 Authors. Depth anything V2 for metric depth estimation. https://github.com/DepthAnything/Depth-Anything-V2/tree/main/metric_depth. Official implementation and metric-model documentation.

[3] ElecFreaks. Ultrasonic ranging module HC-SR04. Technical datasheet. Module specifications; verify against the installed module variant.

[4] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[5] Thomas Moore and Daniel Stouch. A generalized extended kalman filter implementation for the robot operating system. In *Intelligent Autonomous Systems 13*, volume 302 of *Advances in Intelligent Systems and Computing*, pages 335–348. Springer, 2016.

[6] robot_localization contributors. robot_localization: State estimation nodes. https://docs.ros.org/en/noetic/api/robot_localization/html/state_estimation_nodes.html. Official software documentation.

[7] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, 2024.