Robotica (2025), 1–19
doi:10.1017/S0263574725102646

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

A multi-modal embodied robot framework for English as a second language learning in preschoolers: design and evaluation

Anastasiya Rybakova[1] and JongSuk Choi[2]

[1]Center for Humanoid Research, Korea Institute of Science and Technology, Seoul, South Korea
[2]AI&ROBOTICS, University of Science and Technology, Daejeon, South Korea
Corresponding author: JongSuk Choi; Email: cjs@kist.re.kr

Received: 2 June 2025; **Revised:** 29 August 2025; **Accepted:** 5 September 2025

Keywords: service robots; novel applications of robotics; control of robotic systems; serial manipulator design and kinematics; man-machine systems

Abstract

A multi-modal embodied robot framework was developed and evaluated to support English as a Second Language (ESL) learning in preschoolers through physical interaction and adaptive engagement. The system integrates a 4-DOF OpenManipulator-X robot with a tablet-based educational application, forming a unified instructional platform that delivers synchronized auditory, visual, and kinesthetic stimuli. Designed to improve lexical retention and motivation in early learners, the framework enables task-based interaction through pick-and-place vocabulary reinforcement, collaborative drawing, and tablet-mediated language tasks, coupled with a real-time emotion recognition module to adjust instructional cues.

An experimental design within the subject was used with 21 Korean preschool children (ages 4–8), comparing robot-assisted language learning (RALL) with traditional teacher-led language learning (TLLL) in matched tasks involving vocabulary learning, math reasoning, color categorization, and spelling recall. Each session was conducted under controlled classroom conditions and analyzed using both quantitative and qualitative metrics, including engagement frequency, task precision, and structured post-session surveys.

The results demonstrate significantly higher participation and task completion rates in the RALL condition, with vocabulary acquisition outcomes comparable to TLLL ($p > 0.05$). Children exhibited increased motivation and sustained interaction when guided by the robot and the application, suggesting that embodied adaptive systems can effectively support early second language learning. The study contributes validated design principles for integrating physical embodiment, affective responsiveness, and multi-modal instructional delivery in educational robotics. Implications are discussed for the scalable deployment of robot-assisted systems in preschool contexts, emphasizing child-centered interaction and developmental appropriateness within RALL environments.

1. Introduction

Second language acquisition during early childhood is widely recognized as a critical contributor to long-term cognitive development, academic success, and cross-cultural competence [1, 2]. The preschool period is particularly significant, as heightened neuroplasticity during this developmental window allows children to acquire linguistic structures with greater ease and long-term retention. In non-English-speaking countries, such as South Korea, there is increasing emphasis on early English instruction to equip children with foundational communicative skills. However, traditional teacher-led English as a Second Language (ESL) instruction often lacks the flexibility, sensory engagement, and adaptivity required to meet the cognitive and emotional needs of preschool learners [3, 4].

Recent advancements in human–robot interaction (HRI) and educational robotics have created new opportunities to enrich early language learning through embodied, multi-modal, and socially interactive

systems. Robot-assisted language learning (RALL) systems, in particular, have shown promise in fostering vocabulary acquisition, pronunciation, and engagement through verbal and physical interaction [5, 6]. Social robots enable real-time instruction through speech, gaze, gesture, and physical embodiment, simulating aspects of human tutoring. Despite growing empirical support for their use in language learning, most existing RALL systems are limited in adaptivity, rely on pre-scripted content, and lack integration across modalities. Moreover, few have been validated through longitudinal or ecologically valid classroom studies with preschool-aged children who have no prior English exposure [7, 8].

Multi-modal learning approaches grounded in dual coding theory and embodied cognition emphasize the role of synchronized visual, verbal, and kinesthetic input in enhancing memory, comprehension, and motivation [9, 10]. Pre-school learners, in particular, benefit from sensory-rich environments where abstract linguistic concepts are reinforced through concrete interaction [12]. While some recent studies have introduced robots capable of gesture-based instruction or emotion expression, the combination of embodied interaction, real-time affective responsiveness, and digital reinforcement remains underexplored in real-world classroom environments.

To address this gap, the present study introduces and evaluates a multi-modal embodied robot framework for ESL instruction in preschool-aged children. The system integrates a 4-DOF robotic manipulator (OpenManipulator-X) with a tablet-based educational application, forming a synchronized instructional platform that provides verbal guidance, physical manipulation, collaborative drawing, and adaptive emotional feedback. Learning tasks were designed to reinforce vocabulary retention and cognitive development through embodied interaction, including pick-and-place activities, color matching, shape drawing, and spelling reconstruction games.

A within-subject experimental design was conducted in a Korean preschool setting with 21 children aged 4–8. Participants engaged in both RALL and teacher-led language learning (TLLL) instructions across two experimental rounds, allowing for comparative analysis of task performance, engagement levels, and learner feedback. Quantitative and qualitative data were collected through task metrics, structured surveys, and behavioral observations.

This work contributes to the field of educational robotics and child-robot interaction (CRI) by presenting a validated, scalable, and developmentally appropriate framework for early language instruction. By integrating affect-sensitive behavior, physical embodiment, and multi-modal instructional delivery, the proposed system advances the pedagogical and technical state of RALL and offers insights into the design of adaptive robotic tutors for preschool learners.

2. Related works

2.1. Early childhood second language acquisition

Second language acquisition (SLA) in early childhood is grounded in the critical period hypothesis, which posits that language learning capacity is maximized during the first decade of life due to increased neural plasticity [1]. Empirical studies confirm that preschool students acquire syntactic structures, phonological distinctions, and vocabulary with greater ease and long-term retention when exposed early to L2 environments [2, 3]. In particular, ESL instruction in non-English-speaking countries is often introduced in early education settings to leverage this developmental advantage. However, TLLL in preschools frequently fails to sustain attention or offer differentiated input tailored to individual learners' cognitive states [4, 5].

Sociocultural learning theory emphasizes that linguistic development occurs most effectively in socially mediated contexts where children co-construct meaning through guided interaction [6]. In early childhood education, this suggests the need for responsive and interactive instruction that supports scaffolding and immediate feedback. Despite this, ESL instruction often lacks multi-modal input, embodied experiences, and adaptivity – factors that are essential for young learners who rely heavily on sensorimotor processing.

2.2. Robot-assisted language learning

RALL has emerged as a promising approach for addressing the cognitive and attentional limitations of conventional ESL teaching. Social robots leverage physical embodiment, gesture, and speech to engage children in verbal and non-verbal interaction. Studies have shown that young learners exhibit increased engagement, verbal output, and motivation when interacting with socially assistive robots compared to screen-based applications [7, 8]. RALL systems also offer consistency and repeatability in content delivery, an advantage over human instruction in environments where teacher resources are limited.

Multiple studies have demonstrated the potential of robots to reinforce vocabulary acquisition, assist pronunciation, and support turn-taking behavior [9, 10]. However, most existing RALL systems are characterized by limited adaptability. Interaction flows are typically pre-scripted, with little response to real-time emotional states or levels of comprehension of learners [12]. Moreover, physical embodiment is often underutilized: many platforms rely solely on verbal instruction or animated facial expressions without integrating gesture, movement, or object manipulation [13]. In some systems, robots act merely as animated storytellers or question prompts, rather than as embodied, task-oriented tutors.

Few studies have implemented RALL in authentic preschool environments with participants who have no prior exposure to English. Most are carried out under laboratory conditions, with short intervention periods and limited ecological validity [14]. Furthermore, system evaluation is often limited to short-term vocabulary gains, rather than comprehensive measures of engagement, interaction quality, or long-term retention.

2.3. Multi-modal interaction in language learning

Multi-modal interaction – defined as the integration of verbal, visual, and kinesthetic stimuli – has long been identified as a critical factor in enhancing early language learning outcomes. Based on dual coding theory [15] and supported by the principles of embodied cognition [16], multi-modal instruction reinforces semantic encoding by associating spoken words with concrete sensory input. In the context of preschool ESL learners, combining auditory instruction with visual imagery, touchable objects, and physical actions improves attention, comprehension, and recall [16, 17].

Some RALL studies have attempted to introduce multi-modal elements through gesture recognition, visual displays, or tablet-based interfaces [12, 13]. However, few systems implement true sensory integration, where verbal, visual, and kinesthetic feedback are synchronized in real time and tailored to the learner's affective state. Coordination across modalities is essential in preschool contexts, where embodied interaction and concrete associations support early semantic development. The present study addresses this gap by integrating physical pick-and-place tasks, robot-led drawing, and dynamic, emotion-aware feedback into a unified multi-modal framework for ESL instruction.

2.4. Emotion-adaptive systems in educational robotics

Affective responsiveness, the ability of an educational system to interpret and respond to the emotional states of learners, has become an increasingly important focus in CRI. Preschool learners are emotionally reactive and require frequent validation and encouragement to maintain learning focus. Real-time affective cues such as facial expressions, hesitation, or gaze aversion offer valuable indicators of readiness or disengagement of the learner [18]. Educational robots that detect and adapt to these cues have been shown to improve learning outcomes, user satisfaction, and session duration [19, 20].

Despite its relevance, few RALL systems incorporate real-time emotion recognition or affective adaptation. Interaction scripts are often fixed and feedback is triggered by static event thresholds rather than context-sensitive cues [18, 21]. Integration of emotion recognition in early ESL learning remains underexplored in RALL research, especially in developmental settings like preschools.

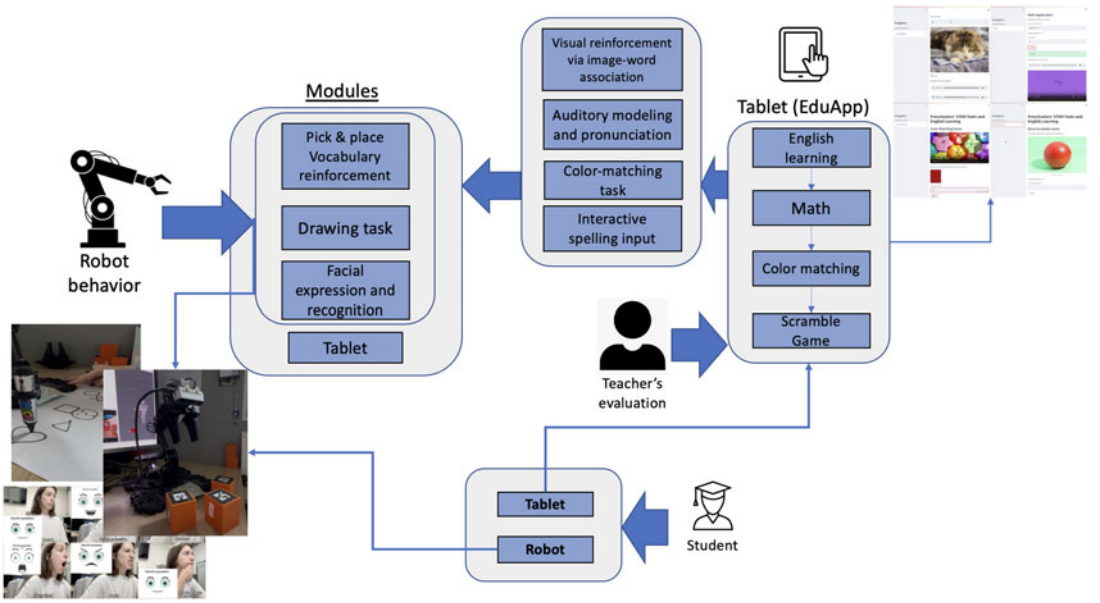

Figure 1. *System architecture and interaction flow of the proposed RALL framework integrating robot and tablet modalities.*

2.5. Research gap and system positioning

Although prior research affirms the value of RALL in improving early language learning, several persistent limitations remain. Most existing systems lack coordinated multi-modal integration and operate without affective adaptivity. Their implementations are often short-term, decontextualized, and do not reflect the complexity of real-world preschool settings. Few studies incorporate physical manipulation of objects or robot-led motor tasks that link vocabulary to embodied experience. Furthermore, systems are rarely evaluated using both quantitative and behavioral metrics under ecologically valid classroom conditions.

The present study addresses these limitations by developing a multi-modal, embodied, emotion-adaptive framework for ESL instruction in preschoolers. The proposed system integrates physical manipulation, collaborative drawing, and adaptive verbal cues in a unified robot-tablet platform. It is evaluated through an experimental design within the subject in an authentic Korean preschool setting, offering new empirical insights into the design and deployment of scalable, developmentally appropriate RALL systems.

3. System architecture and framework design

The developed RALL framework was designed to support early ESL acquisition in preschoolers through a multi-modal, emotionally responsive system. Built on Robot Operating System (ROS) middleware, the system integrates robot control, tablet-based instruction, emotion mirroring, and audiovisual feedback in a child-centered, classroom-ready configuration. The architecture emphasizes safety, developmental appropriateness, and real-time coordination between physical and digital modalities.

Figure 1 provides a visual overview of the system's architecture and interaction flow. On the left, the robot behavior modules include vocabulary-based pick-and-place tasks, shape drawing tasks to reinforce concept learning, and facial expression recognition that mirrors a child's emotion through an animated face display. These modules are tightly integrated with the tablet-based EduApp, which delivers four instructional tasks: English vocabulary learning, mathematics, color matching, and word scramble games. The central module outlines the cognitive components – such as visual and auditory

reinforcement, orthographic processing, and color-concept mapping – that bridge physical interaction and digital engagement. The bidirectional arrows indicate the system's synchronized communication flow between robot and tablet, ensuring seamless transitions during task execution. Additionally, teacher evaluations and real-time observations inform system calibration and session pacing. The bottom section of the figure displays real-world deployment, including robot-assisted drawing, augmented reality-based (AR-based) object recognition, and emotion detection via webcam input. Together, these elements illustrate the operational synergy between robotic embodiment and digital learning tools, forming a cohesive and developmentally appropriate learning environment for young ESL learners [22–24].

3.1. Hardware configuration

The physical components of the system were assembled to balance expressive interaction with usability and safety for children aged 4–8. The robotic arm used is the OpenManipulator-X (ROBOTIS) [25], a 4-DOF platform mounted on a standard classroom table. Motor ID14, which is part of the OpenManipulator-X arm, was equipped with a custom pen holder to enable the robot to draw predefined shapes (e.g., heart, house, sun, circle, square, and others) on paper and whiteboard surfaces using calculated Cartesian paths translated into joint-space trajectories. A U2D2 communication controller enabled real-time motor commands.

The robotic arm was controlled using a desktop computer running Ubuntu 20.04 and ROS Noetic. Communication between the robot and the computer was facilitated via the U2D2 interface, which allowed for real-time execution of motion commands. The robot's tasks – including both drawing and pick-and-place actions – were implemented by running custom-developed Python and C++ scripts based on the official OpenManipulator ROS packages.

For facial and emotional detection and expression, the system utilized a display with an integrated camera on the top that visualized animated facial expressions in response to the child's affective state. A Logitech C920 USB camera, positioned at the child's eye level in front of the table, captured facial data in real time for emotion classification. The tablet interface, implemented on an iPad (3rd generation, 2021, 10.2-inch), served as the primary platform for educational games and digital feedback, enabling a two-channel interaction environment: physical (robot) and digital (tablet). Two external speakers were connected to the main computing unit to deliver clear auditory output. Together, these components constituted a multi-modal configuration designed to facilitate socially engaging and developmentally appropriate CRI.

3.2. Software architecture and module integration

The system architecture was designed with modularity and real-time responsiveness as core principles. ROS was used to control robot actions and synchronize service calls between nodes responsible for drawing, pick-and-place, and timed motion sequences. The SetDrawingTrajectory and goal_task_client() services facilitated parameterized execution of drawing tasks, with inverse kinematics ensuring consistent pen-surface contact (Figure 2) [25]. Moreover, the pick-and-place task was implemented through the detection of AR markers using an Intel RealSense camera, enabling the robot to accurately identify initial cube positions and relocate them to predefined target locations (Figure 2).

A Streamlit-based EduApp (Figure 3), launched on the iPad, offered four structured learning tasks: vocabulary learning, math problems, color matching, and a word scramble game. The vocabulary content used in the application was derived from the Cambridge English: Pre A1 Starters Word List Picture Book to ensure age-appropriate and pedagogically aligned material [26]. The application was developed independently from the ROS node, but manually synchronized during sessions to maintain a coherent instructional flow.

OpenAI-generated Text-to-Speech (TTS) files [27] were generated for all vocabulary items, instructions, and motivational phrases. These were manually triggered by the operator in coordination with

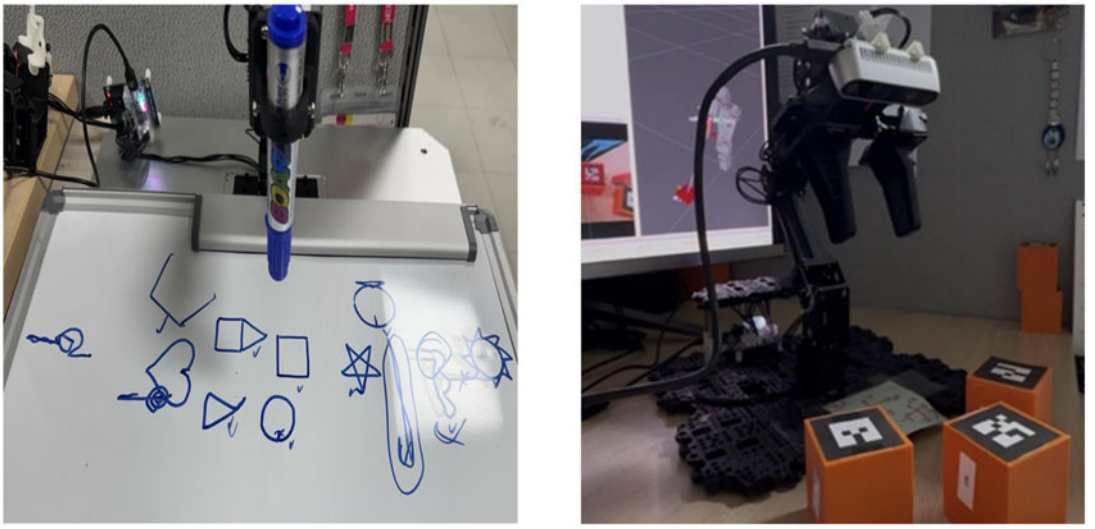

Figure 2. *Robot-executed drawing and vocabulary-based pick-and-place tasks using OpenManipulator-X to reinforce embodied ESL learning.*

task progression during the interaction between child and robot for drawing and drawing tasks, allowing adaptive pacing and tone modulation in real time.

3.3. Multi-modal interaction

Each learning session was divided into two sequential phases: robot-led physical interaction followed by tablet-based digital reinforcement. During the first phase, the robot executed pick-and-place tasks using AR-labeled vocabulary cubes or drew shapes symbolizing target words (e.g., "sun", "cloud") (Figure 2). Children were prompted to repeat the words or describe the object being drawn, anchoring vocabulary in sensorimotor experience.

In the second phase, children interacted with the tablet application, which offered tasks aligned with the robot-led activities. These included visual-matching tasks, math exercises, color-word associations, and scrambled word challenges (Figure 3) [29–32]. Each task featured animated feedback, embedded videos, and narration through TTS audio. The sequence was manually timed by the instructor to ensure smooth handoff between the physical and digital domains.

3.4. Emotional and facial expression and recognition

The system included a lightweight affective feedback loop based on facial expression recognition [34]. Using a webcam mounted at eye level, the system captured the real-time emotional state of children and classified expressions into five categories: neutral, happy, sad, angry, or surprised (Figure 4). The classification was implemented in Python using the DeepFace library [28], which internally supports models for real-time facial expression analysis. These detected emotions were displayed on a separate screen as an animated face that mimicked the child's emotional state, providing intuitive and immediate visual feedback.

Although the robot did not autonomously change behavior based on emotion input, the mirroring effect encouraged self-awareness and allowed the instructor to adapt the pace or offer verbal reinforcement accordingly. This mechanism was particularly effective in maintaining engagement during longer sessions and provided a foundation for future work on fully autonomous emotional adaptation.

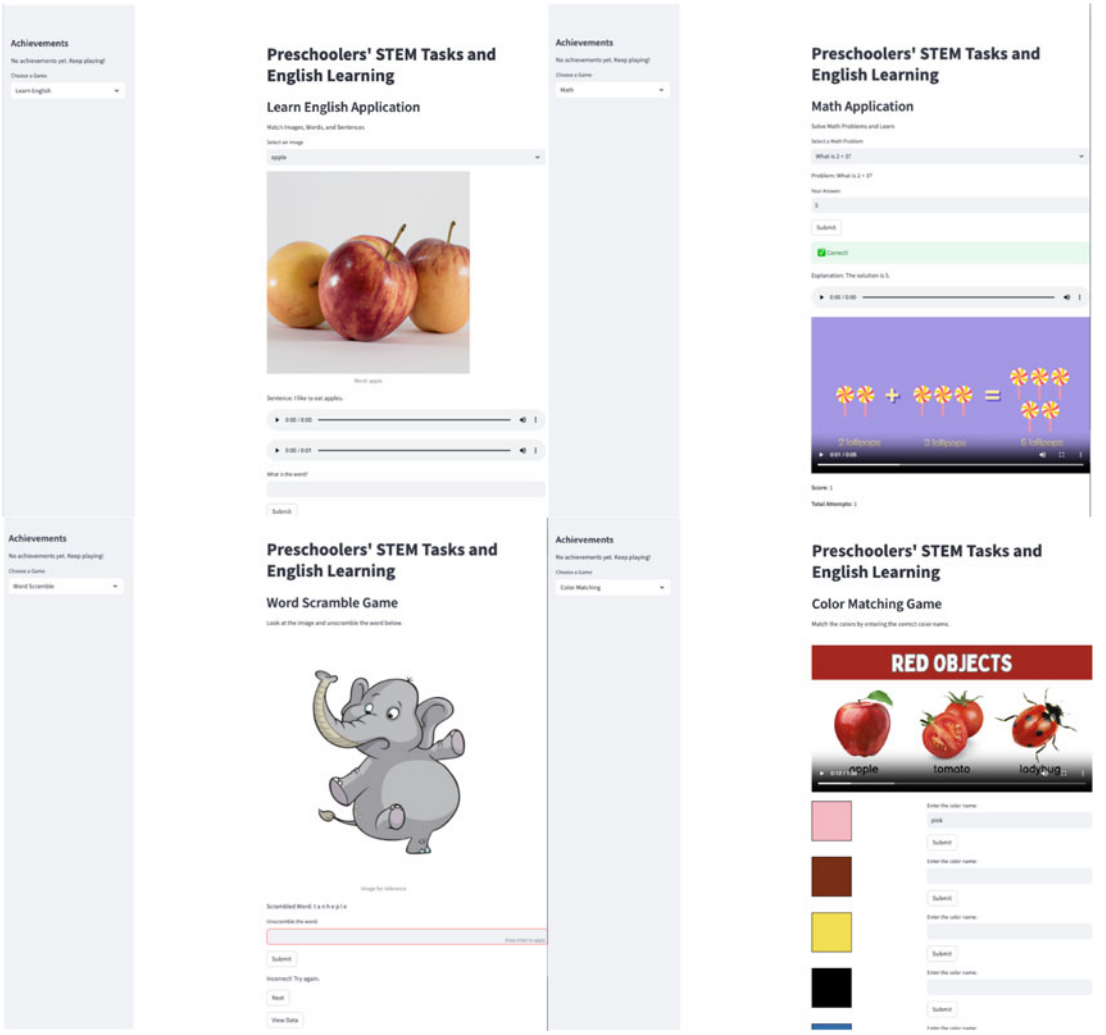

Figure 3. *Tablet-based EduApp interface with vocabulary, math, color-matching, and scramble tasks supporting robot-led instruction.*

The system emphasized emotional transparency and safety over complexity, aligning with the developmental needs of early childhood.

3.5. Design priorities and constrains

The framework was built around three guiding principles: developmental appropriateness, modular synchrony, and scalability. Tasks were short, goal-oriented, and visually reinforced. Robot movement was fine-tuned through pilot testing with adults and adjusted based on children's observed engagement levels. The use of manual synchronization (rather than full automation) was a deliberate choice to respect natural instructional pacing and maintain operational control in a dynamic classroom environment.

Although the robot and tablet were not fully integrated via shared control architecture, their coordinated deployment created a unified learning experience. The design demonstrated that meaningful robot-assisted ESL learning is achievable without high computational complexity – provided that interaction is structured, feedback is responsive, and engagement is multi-sensory.

Figure 4. *Emotion recognition process and real-time animated feedback display used to reflect the child's affective state.*

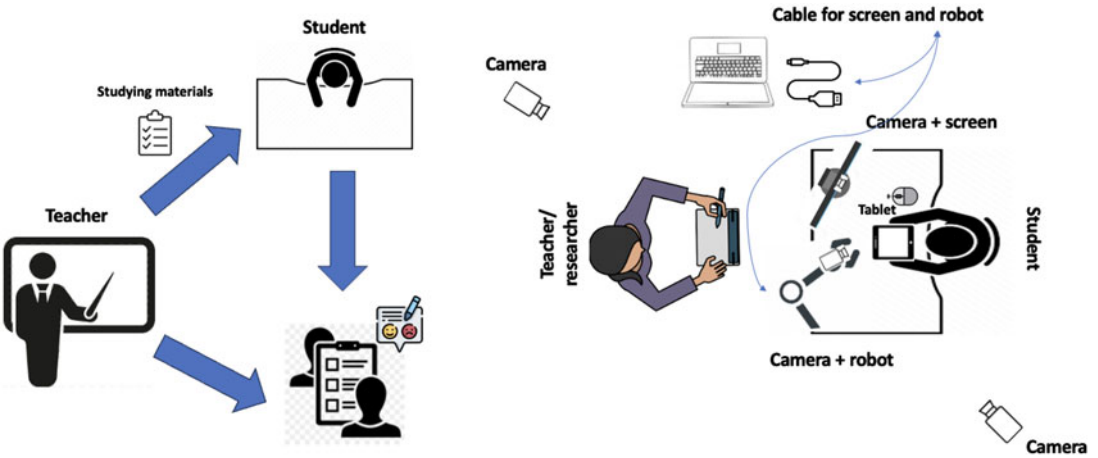

Figure 5. *Classroom-based instructional settings for TLLL (left) and RALL (right) conditions during live sessions.*

4. Experimental methodology

This section outlines the experimental setup employed to evaluate the effectiveness of the proposed multi-modal RALL framework for preschool ESL learning. The study was designed to compare the performance, engagement, and motivational responses of young learners under two instructional conditions: TLLL and RALL. A crossover methodology within the subject was applied to control for inter-individual variation while assessing learning outcomes in a real-world classroom context (Figure 5).

4.1. Participants

A total of 22 Korean preschool children, aged 4–8 years, were initially recruited for the study. Fourteen children participated in Round 1 (9 sessions), and an additional eight were enrolled in Round 2 (4 sessions). However, one child from Round 2 was excluded from the final analysis due to insufficient

participation (attendance below 80% of the sessions), resulting in a final sample of 21 children. None of the children had received prior formal instruction in English.

To provide a clearer developmental profile, the final sample comprised 3 children aged 4 years, 8 children aged 5 years, 4 children aged 6 years, 5 children aged 7 years, and 1 child aged 8 years, spanning the transition from preschool to early primary education. Of the 21 children, 16 were boys and 5 were girls. The participants were recruited from a daycare center affiliated with the Korea Institute of Science and Technology (KIST), and all experimental procedures were approved by the Institutional Review Board (IRB) of the institute (approval number KIST-202412-HR-001). Written informed consent was obtained from the children's guardians, and verbal assent was obtained from each child prior to participation. To ensure instructional focus and minimize group-induced distractions, children were divided into small sub-groups during sessions. Tasks were conducted in a dedicated CRI space, with alternating robot-led and tablet-based activities designed to maintain engagement and reduce fatigue.

In the South Korean educational system, children between the ages of 4 and 7 typically attend kindergarten, while formal elementary school begins at age 8. Therefore, the selected age range of 4–8 years represents the upper spectrum of early childhood education within this national context. This period is particularly formative for language acquisition, cognitive flexibility, and socio-emotional development. The RALL framework was intentionally designed to support learners across this transitional developmental span through differentiated, scaffolded interactions and age-appropriate multi-modal engagement strategies. By targeting this age group, the study aims to evaluate the framework's adaptability and instructional effectiveness during a critical window for early second language learning.

To ensure instructional focus and avoid group-induced distractions, children were divided into small subgroups during sessions. Tasks were conducted in a dedicated CRI space, with alternating robot-led and tablet-based activities designed to maintain engagement and reduce fatigue.

4.2. Experimental design

The study adopted a within-subject crossover design to enable direct comparison between the two instructional modalities – TLLL and RALL – under equivalent task conditions (Figure 5). Each child participated in both instructional modes over the course of multiple sessions, allowing matched data collection across conditions.

The experiment consisted of two rounds:

- Round 1 involved 14 participants and consisted of 9 sessions: 4 TLLL sessions, 4 RALL sessions, and 1 debrief and survey session (Figure 6).
- Round 2 involved 7 participants and included 4 sessions: 2 TLLL and 2 RALL. To control for order effects, the instructional sequence was reversed in Round 2, with RALL delivered before TLLL (Figure 7).

Each session lasted approximately 60–65 min, encompassing robot-assisted or teacher-led instruction, interactive tablet engagement, and observational or survey-based data collection. The instructional content and complexity of the task were kept constant under all conditions to ensure a fair comparison.

All instructional content was based on beginner-level ESL vocabulary selected from the Cambridge curriculum [26], adapted for preschool comprehension and engagement. Tasks were thematically aligned across both instructional modalities. The RALL condition, however, uniquely incorporated robotic embodiment through pick-and-place manipulation and drawing trajectories, in addition to synchronized verbal and digital feedback.

The implemented tasks included:

- ***Vocabulary Learning:*** In both conditions, English vocabulary items were introduced per session. In TLLL, the teacher explained each word verbally and used flashcards or printed visuals. In RALL, the robot vocalized the target words using pre-recorded TTS audio, while the tablet

Figure 6. *Round 1 experimental sessions comparing TLLL (left) and RALL (right) modalities in real classroom setting.*

Figure 7. *Round 2 instructional sequence: reversed order of RALL (left) and TLLL (right) conditions for crossover validation.*

displayed animated images of the objects. The children were encouraged to repeat each word aloud in both conditions.

- ***Pick-and-place:*** To reinforce vocabulary through embodied action, the robot manipulated labeled cubes or toy objects associated with the vocabulary items (e.g., "sun", "apple", "cloud"). The robot identified the correct item using predefined AR marker IDs and performed a pick-and-place movement to place it in a matching bin or zone. This task involved spatial reasoning, object-word association, and joint attention.
- ***Drawing Task:*** Using a pen attached to the gripper of the OpenManipulator-X, the robot performed shape-drawing tasks on A4 paper and later switched to the whiteboard fixed in the workspace. Each shape corresponded to a target vocabulary word (e.g., heart, house, smiley, sun). Drawings were executed using pre-programmed joint trajectories derived from 2D coordinates and implemented via the SetDrawingTrajectory ROS service. The children were prompted to identify the drawn object and verbally describe it.
- ***Math and Color-Matching Tasks:*** Children engaged in simple arithmetic and color-categorization games. These were implemented as drag-and-drop activities in the tablet interface during RALL, and as verbal or card-based games during TLLL. The robot narrated instructions in RALL sessions using synchronized audio and visual cues.
- ***Word Scramble Game:*** To assess vocabulary retention and spelling recognition, the children completed a scramble game in which they reconstructed target words from jumbled letters. The

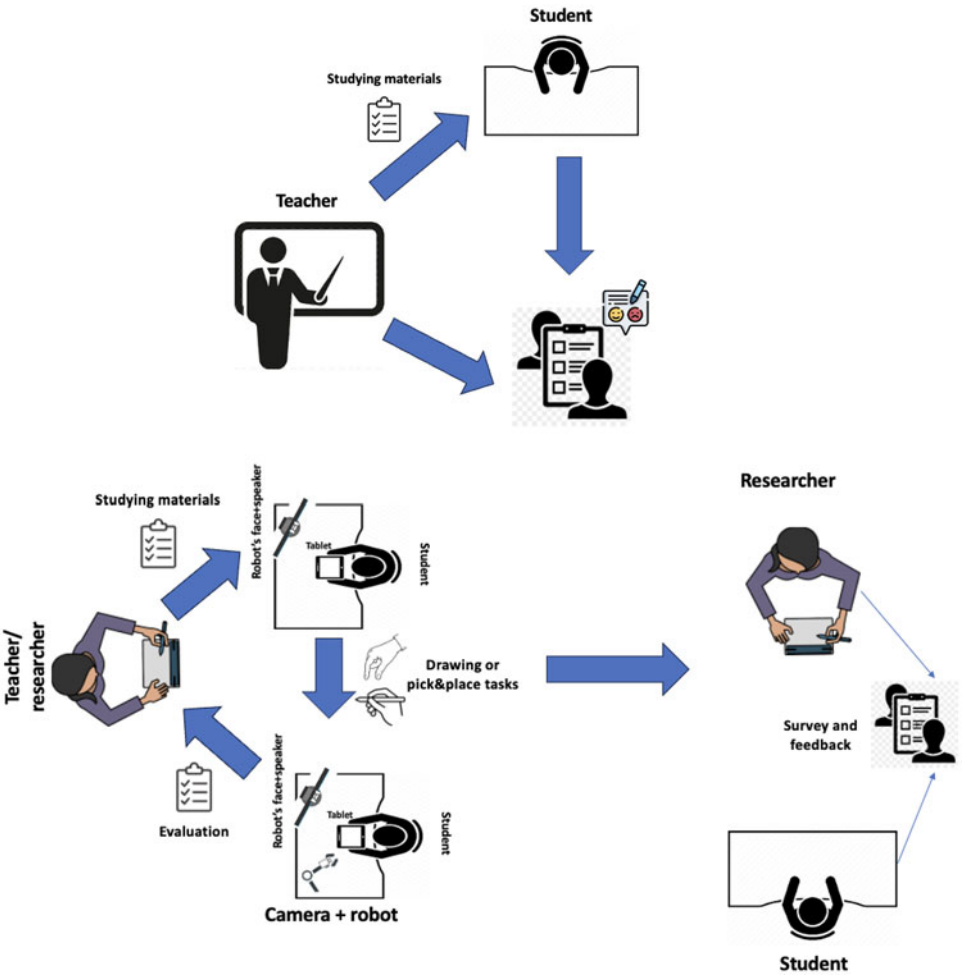

Figure 8. *Structured post-session user experience survey conducted with children following TLLL and RALL sessions.*

RALL version used the tablet interface with input and submit functionality, while the TLLL version involved teacher-facilitated oral guidance.

Each session in RALL began with robot-led physical interaction (drawing or pick-and-place), followed by digital tablet-based engagement. In contrast, TLLL sessions followed a linear format with verbal instruction, guided repetition, and teacher-directed activities. The tasks were intentionally short and interactive, with break intervals to maintain attention.

4.3. Data collection and evaluation measures

To evaluate the effectiveness of the proposed framework, a mixed-methods data collection strategy was used, which incorporated both quantitative and qualitative measures. The study focused on capturing instructional outcomes across the two conditions – RALL and TLLL – by assessing learner performance, engagement, emotional response, and subjective feedback.

Following the completion of each instructional condition, children participated in structured, one-on-one oral surveys administered in English by a trained Korean research assistant (Figure 8). These

surveys were designed to be developmentally appropriate for young learners and employed a simplified 5-point Likert-type scale, ranging from 1 (“not at all”) to 5 (“very much”). The assistant provided consistent verbal scaffolding, supportive hand gestures, and visual cues to help children understand each item and respond meaningfully. This support enabled reliable data collection across the full 4–8 age range. To ensure accessibility, responses were collected with the aid of consistent verbal prompts and hand gestures. The survey assessed self-reported understanding, ease of learning, confidence, and enjoyment across four task categories: vocabulary learning, mathematics, color matching, and scrambled word reconstruction. In the RALL condition, additional items probed children’s impressions of robot behavior, ease of tablet interaction, and the perceived helpfulness of the integrated system.

All instructional tasks – including vocabulary learning, mathematics, color matching, and word reconstruction – were conducted primarily in English, aligning with the ESL learning objectives of the study. For tasks with less inherent linguistic content (e.g., color matching and mathematics), instructions and prompts were delivered in simple English, accompanied by visual aids and occasional Korean clarification to support comprehension. This hybrid linguistic strategy was intended to balance meaningful English exposure with cognitive accessibility. Children were encouraged to respond in English when possible, although Korean responses were accepted without penalty. This flexible approach fostered a supportive, low-pressure environment conducive to language development and task engagement.

The mathematics task was structured as an interactive problem-solving activity in which children were presented with simple visual number problems and prompted to input their answers using the tablet interface. Instructions and questions were delivered in simple English (e.g., “What is 3 + 2?”), accompanied by intuitive visual cues to support comprehension. Upon submission, children were shown a short educational animation in English that explained the solution through gamified storytelling. This format helped reinforce early arithmetic reasoning while simultaneously providing consistent exposure to instructional English in an engaging, age-appropriate manner.

The color-matching task focused on supporting receptive and productive vocabulary skills through guided multi-modal interaction. Each session began with a short English-language animated video introducing basic color names and associated objects (e.g., “red apple”, “blue sky”). Following the video, children were shown an image and asked in English, “What color is it?” They were then required to type the correct English color name (e.g., “yellow”, “green”) and submit their answer. Immediate feedback indicated whether their response was correct or incorrect, reinforcing both spelling and semantic understanding. While the task was primarily conducted in English, minimal Korean support was offered when necessary to ensure comprehension. This sequence enabled children to first internalize the vocabulary and then apply it in a low-pressure, interactive setting, maintaining alignment with the study’s ESL learning objectives.

In parallel, behavioral observations were recorded during each session using structured annotation protocols. Observational data focused on attentional markers (e.g., gaze direction, verbal responsiveness), behavioral cues indicating cognitive effort or confusion, and spontaneous emotional expressions. The observer also noted levels of physical engagement, including participation in robot-led drawing and pick-and-place tasks.

In the RALL sessions, real-time affective data were captured using a webcam positioned at eye level with the participant. The system classified facial expressions into five categories: neutral, happy, sad, angry, and surprised – using a lightweight convolutional neural network (CNN)-based model trained on child-appropriate datasets. The implementation was supported by OpenCV and TensorFlow libraries, enabling efficient image capture and classification on the local system. Although the robot did not autonomously adapt to emotional input, mirroring of facial expressions on an adjacent display served to increase emotional awareness and indirectly supported session pacing. These emotional logs were used as supplemental data to interpret overall engagement and response.

Quantitative analysis was conducted using paired-sample t-tests to examine differences between instructional conditions for each task and survey item. Repeated measures ANOVA was used to assess instructional effects between rounds and to assess whether instructional order influenced the outcomes. This method is particularly suitable for within-subject designs, as it accounts for the correlation between

repeated observations from the same individuals. By partitioning the variance into within- and between-subject components, ANOVA provides a robust framework for examining changes over time or across multiple interventions, while controlling for individual variability. Where appropriate, the effect sizes were calculated using Cohen's d and f to provide estimates of practical significance.

To complement the numerical results, open-ended responses from the surveys were thematically analyzed. Children frequently highlighted the robot voice, patient pacing, and visual gestures as helpful in remembering words and completing tasks. Comments such as "the robot waits until I'm ready" or "it helps me understand" reflected a high degree of perceived comfort with the system. These qualitative findings reinforced quantitative trends and were consistent in both rounds of instruction, suggesting that the positive reception of the robot was not simply a novelty effect but indicative of the alignment of the framework with the cognitive and emotional needs of preschool learners.

4.4. Ethical and developmental considerations

The experiment was carried out according to the ethical guidelines for research with minors. IRB (KIST-202412-HR-001) approval was secured, and informed consent was obtained from all legal guardians. Sessions were designed to be playful, low-stress, and developmentally appropriate. All interaction cues were delivered at a pace suitable for preschoolers, and robot movements were optimized for safety and clarity. Facial data used for emotion recognition was processed in real time and not stored beyond session classification logs, and all data were anonymized prior to analysis.

5. Results and analysis

This section presents the results of the comparative analysis between the traditional TLLL condition and the RALL condition. The findings are organized into quantitative and qualitative analyses, followed by comparative interpretations of learner outcomes. Statistical evaluations were conducted using paired-sample t-tests and repeated measures ANOVA to assess the significance of differences in learner performance, engagement, and reported motivation across both instructional modalities.

5.1. Task performance analysis

To evaluate the comparative impact of robot-assisted and teacher-led instruction on core learning outcomes, a series of paired-sample *t-tests* was conducted across eight targeted educational indicators, which include *Vocabulary Understanding, Vocabulary Confidence, Mathematics Understanding, Mathematics Confidence, Color Matching Understanding, Color Matching Confidence, Scramble Game Understanding, and Scramble Game Confidence*. Each of the 21 preschool-aged participants completed all instructional tasks under both conditions, RALL and TLLL. The experimental procedure spanned 13 instructional sessions, divided into two distinct rounds of implementation.

Given the within-subjects design, the use of paired t-tests was methodologically appropriate. Each child served as their own control, thus reducing inter-subject variability and enhancing the sensitivity of the analysis to detect condition-specific effects. For each learning metric, the null hypothesis posited no significant difference in mean scores between the two instructional conditions. Effect sizes were calculated using Cohen's d to interpret the practical significance of observed differences, supplementing statistical significance with pedagogically meaningful insight.

The statistical approach employed in this study allows for a robust comparison of instructional modalities, capturing both cognitive gains and the magnitude of learning improvements facilitated by embodied robot interaction.

The test statistic was computed using the standard formula for dependent means:

$$\bar{D} = \frac{\bar{X}rall - \bar{X}tlll}{n} \tag{1}$$

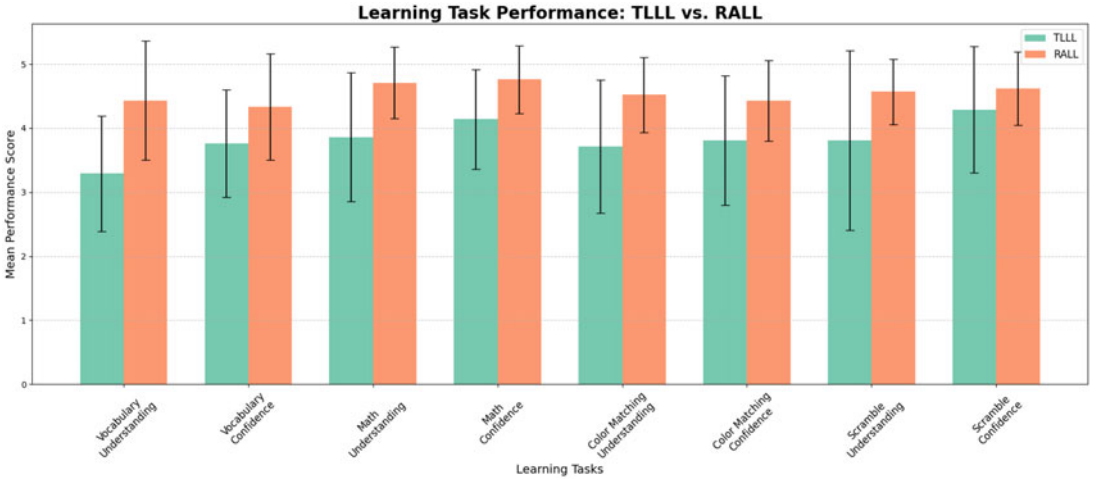

Figure 9. *Statistical comparison of mean task scores under RALL and TLLL conditions across four learning activities.*

where:

- $\bar{X}$ is the sample means for RALL and TLLL, respectively,
- n is the number of participants ($n = 21$).

For direct comparison between the two instructional modalities, paired-sample t-tests were used to evaluate mean score differences. The test statistic is defined as:

$$t = \frac{\bar{D}}{\frac{SD}{\sqrt{n}}} \tag{2}$$

where:

- t is the test statistic,
- $\bar{D}$ is the mean difference,
- SD is the standard deviation of the difference scores,
- n is the number of participants ($n = 21$).

In addition, Cohen's d was calculated:

$$d = \frac{\bar{D}}{SD} \tag{3}$$

To interpret the magnitude of learning differences between instructional conditions, Cohen's d was calculated for each paired-sample t-test. As a guideline, effect sizes were categorized following conventional thresholds: $d = 0.2$ (small), $d = 0.5$ (medium), and $d \geq 0.8$ (large), thereby providing insight into the educational relevance of each result.

In the Vocabulary Understanding task, children performed significantly better in the RALL condition ($M = 4.43$, $SD = 0.93$) compared to the TLLL condition ($M = 3.29$, $SD = 0.90$), with $t(20) = 3.68$, $p = 0.0015$, and a large effect size of $d = 0.80$. These findings suggest that the integrated verbal, visual, and interactive feedback provided by the robot and tablet interface supported superior vocabulary retention and semantic processing (Figure 9). While Vocabulary Confidence also favored the RALL condition ($M = 4.33$) over TLLL ($M = 3.76$), the difference approached but did not reach statistical significance ($p = 0.069$), though the effect size was still moderate ($d = 0.42$). This indicates a promising trend in learner self-efficacy that may benefit from further longitudinal studies.

Table I. *Task-based performance metrics (Paired t-test results).*

Task	M-TLLL	SD-TLLL	M-RALL	SD-RALL	*t*-value	*p*-value	Cohen's *d*
Vocabulary understanding	3.29	0.90	4.43	0.93	3.68	0.0015	0.80
Vocabulary confidence	3.76	0.84	4.33	0.83	1.92	0.069	0.42
Math understanding	3.86	1.01	4.71	0.56	3.70	0.0014	0.81
Math confidence	4.14	0.78	4.76	0.53	3.22	0.0037	0.72
Color-matching understanding	3.71	1.04	4.52	0.59	2.79	0.0112	0.61
Color-matching confidence	3.81	1.01	4.43	0.63	2.17	0.0433	0.47
Scramble understanding	3.81	1.40	4.57	0.51	3.01	0.0048	0.70
Scramble confidence	4.29	0.99	4.62	0.57	1.08	0.2935	0.24

A pronounced advantage for the RALL condition emerged in Math Understanding, where children scored significantly higher ($M = 4.71$, $SD = 0.56$) than in TLLL ($M = 3.86$, $SD = 1.01$), with $t(20) = 3.70$, $p = 0.0014$, and a large effect size of $d = 0.81$. Similarly, Math Confidence improved significantly in the RALL setting ($M = 4.76$, $SD = 0.53$) relative to TLLL ($M = 4.14$, $SD = 0.78$), with $t(20) = 3.22$, $p = 0.0037$, and $d = 0.72$. These findings confirm the pedagogical value of robot-mediated instruction that provides children with personalized pacing, multisensory input, and responsive scaffolding (Figure 9).

In the Color Matching Understanding task, children again performed better under RALL ($M = 4.52$, $SD = 0.59$) than under TLLL ($M = 3.71$, $SD = 1.04$), with $t(20) = 2.79$, $p = 0.0112$, and a moderate-to-large effect size of $d = 0.61$. These results suggest that the consistent delivery of the robot and the engaging visual interface helped standardize task execution among learners, reducing performance variability (Figure 9).

Finally, the Scramble Game Understanding task revealed a significant benefit for the robot-assisted condition ($M = 4.57$, $SD = 0.51$) compared to TLLL ($M = 3.81$, $SD = 1.40$), with $t(20) = 3.01$, $p = 0.0048$, and a large effect size of $d = 0.70$. The nature of this task – requiring sequence assembly, phoneme-grapheme mapping, and letter recognition – likely benefited from the robot's real-time corrective cues and interactive control structure (Figure 9).

Together, these findings underscore the consistent advantage of RALL in multiple learning domains, highlighting its ability to improve both cognitive outcomes and learner confidence in early language and numeracy education (Table I).

5.2. Engagement and motivation metrics: t-test and ANOVA-based analysis

To rigorously evaluate the instructional impact of RALL in comparison to traditional TLLL, both paired-sample *t*-tests and one-way ANOVA were applied to assess learner engagement, task completion confidence, and learning motivation. Effect sizes were calculated using Cohen's *d* for *t*-tests and Cohen's *f* for ANOVA to determine both statistical and practical significance.

Given the within-subjects design, paired-sample *t*-tests provided direct comparison of performance within individuals, reducing variability between participants. In parallel, one-way ANOVA offered a variance-based assessment of instructional effects on group-level outcomes. The ANOVA test statistic was computed as:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{4}$$

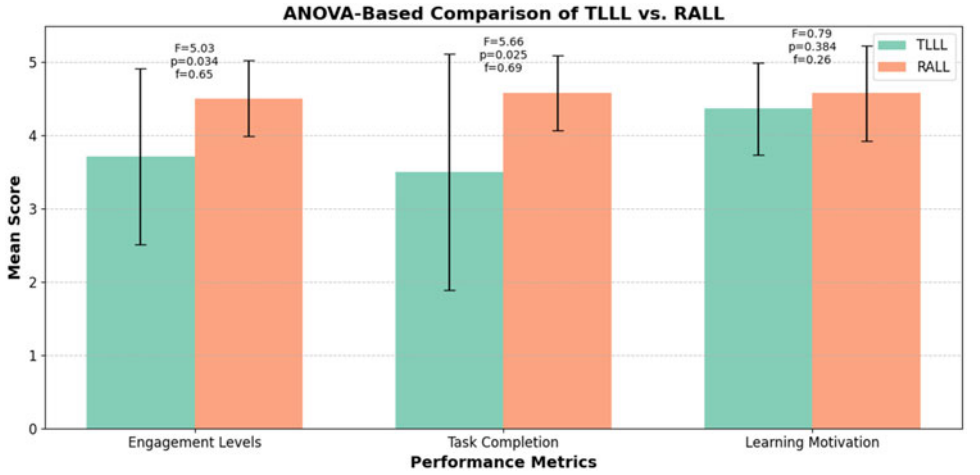

Figure 10. *Statistical comparison of learner engagement, task completion confidence, and learning motivation between TLLL and RALL conditions using ANOVA analysis.*

where:

- MS_{between} is the mean square between groups,
- MS_{within} is the mean square within groups,
- Degrees of freedom: $df_{\text{between}} = k - 1 = 1$, $df_{\text{within}} = N - k = 20$, with $k = 2$ groups and $N = 21$ participants.

Cohen's f was calculated as:

$$f = \sqrt{\frac{F}{N-k}} \tag{5}$$

The following standard conventions were interpreted for the effect sizes of small ($f \geq 0.10$), medium ($f \geq 0.25$), and large ($f \geq 0.40$).

Engagement Level: ANOVA results indicated a significant difference in favor of RALL ($M = 4.50$, $SD = 0.52$) over TLLL ($M = 3.71$, $SD = 1.20$), $F(1, 20) = 5.03$, $p = 0.034$, with a large effect size $f = 0.65$ (Figure 10).

Task Completion Confidence: RALL yielded significantly higher confidence scores ($M = 4.57$, $SD = 0.51$) than TLLL ($M = 3.50$, $SD = 1.61$), $F(1, 20) = 5.66$, $p = 0.025$, $f = 0.69$ (Figure 10).

Learning Motivation: Though RALL produced higher mean scores ($M = 4.57$, $SD = 0.65$) than TLLL ($M = 4.36$, $SD = 0.63$), the difference was not statistically significant: $F(1, 20) = 0.79$, $p = 0.384$, $f = 0.26$ (Figure 10).

These findings reinforce the value of RALL in enhancing engagement and learner confidence – two critical factors for sustained academic success. Although learning motivation did not differ significantly, the directional trend and the moderate size of the effect suggest that long-term exposure to robot-assisted instruction may reveal greater emotional and motivational impacts (Table II).

The use of both t-tests and ANOVA offers a robust dual-method framework that increases internal validity and accounts for individual- and group-level variance. The convergence of significant p-values and large f-values across key metrics affirms that robot-assisted learning offers pedagogical advantages in early childhood language education.

Table II. *Engagement and motivation metrics (ANOVA results).*

Metric	M-TLLL	SD-TLLL	M-RALL	SD-RALL	$F(1,20)$	p-value	Cohen's f
Engagement level	3.71	1.20	4.50	0.52	5.03	0.034	0.65
Task completion confidence	3.50	1.61	4.57	0.51	5.66	0.025	0.69
Learning motivation	4.36	0.63	4.57	0.65	0.79	0.384	0.26

6. Conclusion and discussion

This study presents a comprehensive evaluation of a multi-modal RALL framework designed for preschool ESL education. Drawing on both statistical and observational data, the findings substantiate the pedagogical potential of embodied, interactive robotic systems in early childhood language instruction. Children who participated in both learning modalities – RALL and TLLL – demonstrated significantly better performance during the RALL sessions. Their achievements in vocabulary understanding, mathematical reasoning, color matching, and scramble game tasks were consistently superior when guided by the robot-assisted system.

These enhancements were statistically validated, with moderate to large effect sizes observed across key learning dimensions, underscoring the educational relevance of the improvements. In addition to cognitive performance, the study also found that RALL encouraged greater behavioral participation and confidence in task completion. The children showed better attentional focus, more autonomous task execution, and increased affective involvement during robot-assisted sessions. Although learning motivation did not reach statistical significance, the observed upward trend suggests the potential for further improvement with extended use and deeper emotional adaptation.

Importantly, the current results align with findings from a prior pre-deployment study conducted with adult learners [33], which confirmed the usability and perceived value of the same RALL framework in supporting English language acquisition and engagement. The adult study provided crucial insights into user interface design, content delivery, and multi-modal coordination, which were instrumental in shaping the child-focused deployment. Together, these studies support the generalizability and scalability of the framework across age groups, reinforcing its pedagogical and technological robustness.

From a technological perspective, the study confirms the feasibility of integrating real-time perception (AR marker recognition via RealSense camera), physical interaction (pick-and-place using labeled cubes and drawing task), and expressive feedback (verbal cues, facial mirroring) into a unified and responsive educational platform. The modular system architecture also allows for easy extension to additional tasks, affirming its adaptability to diverse early learning scenarios.

The limitations of the study include a modest sample size and the short duration of experimental exposure, which limits generalizability and insight into long-term effects. In addition, the system required manual activation for speech delivery, restricting its autonomy. Future work should address these limitations through larger, longitudinal deployments, enhanced automation, and inclusion of culturally diverse cohorts to assess broader applicability.

In conclusion, this research provides strong empirical support for the integration of social robots into early childhood language education. The RALL framework not only improves task performance and learner engagement, but also offers a validated, scalable methodology to advance HRI in pedagogically meaningful ways. These contributions represent a pivotal step toward aligning educational robotics with the developmental and emotional needs of young second-language learners. Moreover, these findings are consistent with our previous study conducted with adult learners [33], which demonstrated similar improvements in engagement and learning outcomes, thus supporting the generalizability of RALL principles across age groups and instructional settings.

Acknowledgements. This work was supported by the Korea Institute of Science and Technology (KIST) Institutional Program under Grant (2E33602) and by the Technology Innovation Program (RS-2024-00419883, Development of a Collaborative Robot System and Multi-modal Human–Robot Interaction Services for Supporting Young Children's Daily Activity Care; RS-2024-00507746) funded by the Ministry of Trade Industry and Energy (MOTIE, Korea).

Author contributions. Anastasiya Rybakova developed the framework, ran the experiment in both rounds, analyzed all collected data and wrote the article. JongSuk Choi as advisor reviewed the article.

Competing interests. The authors declare no conflicts of interest exist.

Ethical approval. The research was approved by IRB(KIST-202412-HR-001) team at KIST.

References

[1] E. H. Lenneberg, *Biological foundations of language* (John Wiley and Sons, New York, 1967). doi: 10.1080/21548331.1967.11707799.
[2] J. K. Hartshorne, J. B. Tenenbaum and S. Pinker, "A critical period for second language acquisition: Evidence from 2/3 million English speakers," *Cognition*, **177**, 263–277 (2018).
[3] S. Unsworth, "The impact of age, exposure and aptitude in child L2 development," *Bilingualism Lang. Cognit.* **24**(1), 1–13 (2021).
[4] S. D. Krashen, *Principles and Practice in Second Language Acquisition* (Pergamon Press, Oxford, 1982).
[5] C.-T. Hsin, M.-C. Li and C.-C. Tsai, "The influence of young children's use of technology on their learning: A review," *Edu. Technol. Soc.* **17**(4), 85–99 (2014).
[6] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes* (Harvard University Press, 1978).
[7] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati and F. Tanaka, "Social robots for education: A review," *Sci. Rob.* **3**(21) (2018).
[8] J. Kory and C. Breazeal, "Storytelling with robots: Learning companions for preschool children's language development," *Int. J. Soc. Rob.* **6**(4), 439–455 (2014).
[9] R. Van den Berghe, J. Verhagen, O. Oudgenoeg-Paz, S. Van Der Ven and P. Leseman, "Social robots for language learning: A review," *Rev. Edu. Res.* **89**(2), 259–295 (2019).
[10] W. Xueqing and X. Li, "Designing robot tutors for early childhood ESL learning: A comparative classroom study," *J. Educ. Rob.* **6**(1), 21–39 (2024).
[11] K. Y. Fung, K. C. Fung, T. L. R. Lui and K. F Sin, "Exploring the impact of robot interaction on learning engagement: A comparative study of two multi-modal robots," *Smart Learn. Environ.* **12**(12), 1–25 (2025). doi: 10.1186/s40561-024-00362-1.
[12] E. S. L. Ho and W. Y. Lim, "Limitations of scripted interaction in robot-assisted language tutoring," *Educ. Rob. Rev.* **5**(2), 51–66 (2020).
[13] F. Tanaka, T. Takahashi and K. Isshiki, "Pepper learns together with children: Development of an educational application," *International Conference on Humanoid Robots (Humanoids)*, 270–275 (2015). doi: 10.1109/HUMANOIDS.2015.7363546.
[14] T. Iio, K. Shinozaki and F. Tanaka, "Emotion-adaptive robots for early language instruction: Real-time facial cue modulation," *Int. J. Soc. Rob.* **16**(1), 91–108 (2024).
[15] A. Paivio, *Mental Representations: A Dual Coding Approach* (Oxford University Press, 1986). doi: 10.1093/acprof:oso/9780195066661.001.0001.
[16] A. Brown and R. Ellis, "Foundations of second language acquisition for young learners," *Int. J. Res. Dev.* **9**(6), 285–287 (2024). doi: 10.36713/epra2016.
[17] G. Blazejowska, L. Gruba, B. Indurkhya and A. Gunia, "A Study on the Role of Affective Feeedback in Robot-Assisted Learning," *Sensors* **23**(3) (2023). doi: 10.3390/s23031181.
[18] C. Breazeal, G. Gordon and L.-P. Morency, "Socially Aware Robot Tutors for Preschoolers: Emotional Responsiveness and Learning Gains," *Proc. HRI* (2020).
[19] J. Finders, E. Wilson and R. Duncan, "Early childhood education language environments: Considerations for research and practice," *Front. Psychol.* **14** (2023). doi: 10.3389/fpsyg.2023.1202819.
[20] D. J. Ackerman and A. H. Friedman-Krauss, "Preschoolers' executive function: Importance, contributors, research needs and assessment options," *ETS Res. Rep. Ser.* **2017**(1), 1–24 (2017). doi: 10.1002/ets2.12148.
[21] U. Park and M. S. Kim, "Robot Facial Expression Framework for Enhancing Empathy in Human–Robot Interaction," *IEEE International Conference on Robot and Human Interactive Communication* (2021).
[22] I. E. Ajaj, "The effectiveness of interactive teaching strategies in teaching English language," *J. AI-Farahidi's Arts*, 482–492 (2023). doi: 10.51990/jaa.15.52.2.25.
[23] M. Alimardani, J. Duret, A. L. Jouen and K. Hiraki, "Social robots as effective language tutors for children: Empirical evidence from neuroscience," *Front. Neurosci.-SWITZ* **17** (2023). doi: 10.3389/fnbot.2023.1260999.
[24] G. S. E. Broek, E. Wesseling, L. Huijssen, M. Lettink and T. Van Gog, "Vocabulary learning during reading: Benefits of contextual inferences versus retrieval opportunities," *Cognit. Sci.* **46**(4), (2022). doi: 10.1111/cogs.13135.

[25] ROBOTIS, OpenMANIPULATOR-X e-Manual. Available online: https://emanual.robotis.com/docs/en/platform/openmanipulator.
[26] Cambridge Assessment English, "Pre A1 Starters Word List Picture Book," (2018). For exams from 2018. Accessed July 14, 2025. *Available at:* https://www.cambridgeenglish.org/images/starters-word-list-picture-book.pdf.
[27] OpenAI TTS, "Text-to-Speech API Documentation". Availabale at: https://platform.openai.com/docs/guides/text-to-speech.
[28] S. I. Serengil, "DeepFace: A Lightweight Face Recognition and Facial Attribute Analysis Framework for Python". Available at: https://github.com/serengil/deepface.
[29] C. R. Cox and E. Haebig, "Child-oriented word associations improve models of early word learning," *Behav. Res. Methods* **55**, 16–37 (2023). doi: 10.3758/s13428-022-01790-y.
[30] S. Fernandes, L. Querido and A. Verhaeghe, "Learning to read in an intermediate depth orthography: The longitudinal role of grapheme sounding on different types of reading fluency," *Behav. Sci.* **14**(5), 2024. doi: 10.3390/bs14050396.
[31] E. Petrolo, S. Guerrera, M. G. Logrieco, L. Casula, S. Vicari and G. Valeri, "The role of executive functions in preschool children with autism spectrum disorder: A short narrative review," *Res. Dev. Disabil.* **157** (2025). doi: 10.1016/j.ridd.2024.104905.
[32] O. Engwall and J. Lopes, "Interaction and collaboration in robot-assisted language learning for adults," *Comput. Assisted Lang. Learn.*, 1273–1309 (2020). doi: 10.1080/09588221.2020.1799821.
[33] A. Rybakova and J. Choi, "Evaluating the Effectiveness of Social Robots in Enhancing English Language Acquisition and Educational Engagement: A Study with Adults and Its Implications for Korean Kindergarten Children," *HCI International 2025*, Lecture Notes in Computer Science (LNCS) (Springer, 2025) pp. 328–345. doi: 10.1007/978-3-031-93861-0_21.
[34] M. Shahab, A. Taheri, M. Mokhtari, A. AsemanRafat, M. Kermanshah, A. Shariati and A. F. Meghdari, "Manufacture and development of Taban: A cute back-projected head social robot for educational purposes," *Intell. Serv. Rob.* **17**, 871–889 (2024). doi: 10.1007/s11370-024-00545-2.

Cite this article: A. Rybakova and J. Choi, "A multi-modal embodied robot framework for English as a second language learning in preschoolers: design and evaluation", Robotica. https://doi.org/10.1017/S0263574725102646