"Find the red book on the table next to the sofa over there"... GIST develops AI robot technology to find clear objects in 3d space using sentences

- Professor Ue-Hwan Kim's team from the Department of AI Convergence develops 'Context-Nav,' an AI robot navigation technology that understands both the overall context of a sentence and the positional relationships between surrounding objects

- Achieves 2.3 times higher accuracy than existing technologies without additional training... Suggests potential for application in service robots

- Scheduled to be presented at the international AI conference 'CVPR 2026'

▲ (From left) Professor Ue-Hwan Kim of the Department of AI Convergence, and Won Shik Jang, integrated master's and doctoral student

The Gwangju Institute of Science and Technology (GIST, President Kichul Lim) announced that a research team led by Professor Ue-Hwan Kim of the Department of AI Convergence has developed "Context-Nav," an AI robot navigation technology capable of understanding and accurately locating objects described by humans in sentences within a 3D space.

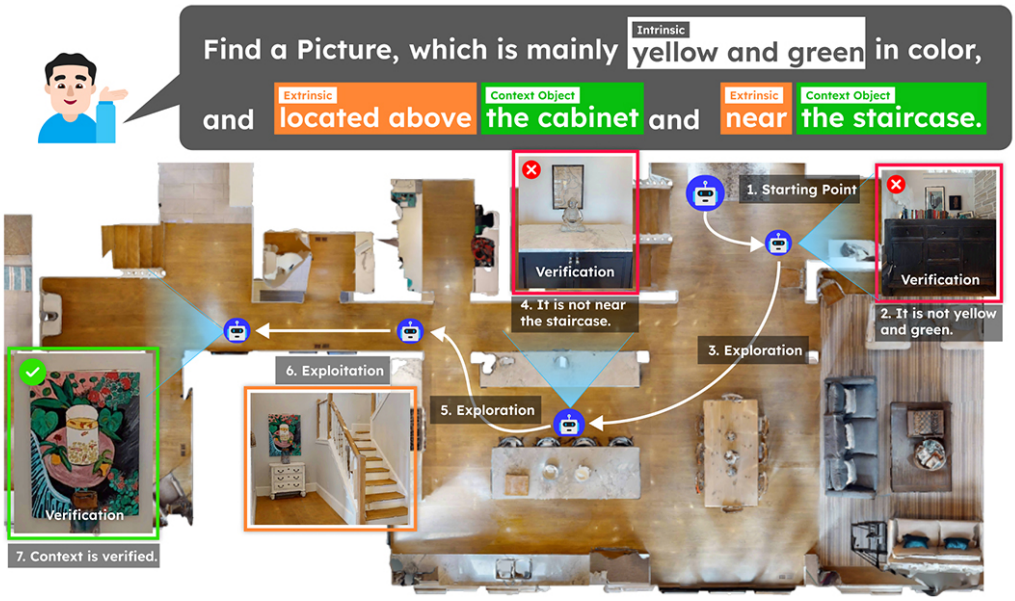
This technology demonstrates potential for expansion into various service robot fields by analyzing not only physical characteristics such as color and shape but also the relative positions of objects.

Autonomous robots (service robots) that perform various tasks such as cleaning, delivery, and guidance in indoor environments must comprehensively grasp the positional relationships with surrounding objects to understand human verbal instructions and execute them accurately.

Previous research primarily used "reinforcement learning," a method in which robots find optimal behaviors through repeated trials and failures to learn action strategies on their own; however, this method required vast amounts of data and high training costs.

Furthermore, there is a limitation in that existing methods rely solely on short, attribute-based information about objects, such as "chair" or "cup," and fail to adequately reflect the context within the sentence—including the position ▲ relative to surrounding objects, ▲ relative direction and arrangement, and ▲ situational clues—provided by humans through lengthy descriptions.

In particular, because spatial relationships (left, right, front, and back) vary depending on the observer's viewpoint or position, there is a high probability that the robot will mistake an incorrect candidate for the actual target object under existing methods.



▲ *A 3D context-based robot search process utilizing long sentence descriptions. The robot searches by utilizing not only the object's unique characteristics, such as color and shape, but also relative position information with other objects within the long natural language description.*

To address these issues, the research team proposed a method that utilizes the entire long sentence description provided by humans in the robot's search process, enabling it to understand both the characteristics of the target object and the 3D spatial relationships between surrounding objects.

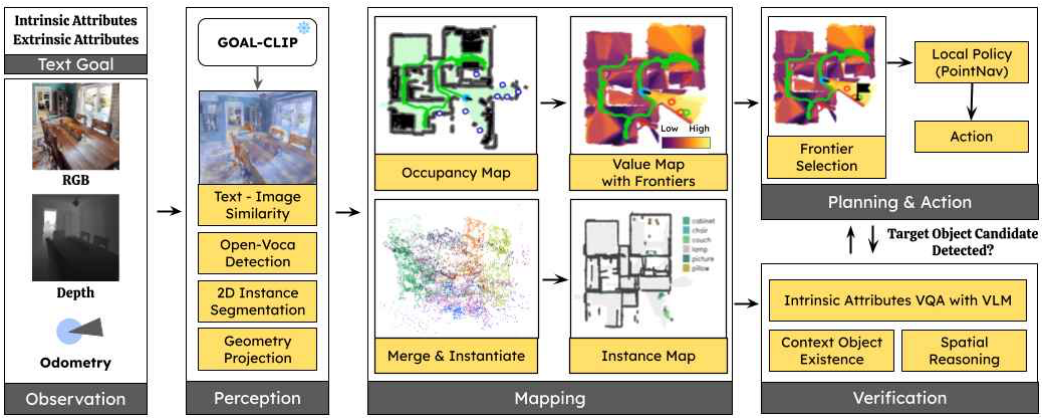
For example, if a person explains, “Find the red book on the table next to the living room sofa,” the robot interprets this sentence not merely as simple object information, but as location information within a three-dimensional space.

First, it recognizes the surrounding environment using an RGB camera* and a depth sensor, and then checks in real-time areas that are highly likely to match the description. It then calculates the suitability of candidate spaces to determine how well they match the goal and records this as a score on a ‘value map.’

Subsequently, it determines an efficient search path centered on the location with the highest score. When a candidate object is discovered, it utilizes a vision language model*, which understands both image and text information, to verify its attributes, and precisely verifies its positional relationships (up, down, left, right, front, and back) with surrounding objects through three-dimensional spatial inference.

> ** RGB camera: A camera that recognizes color like the human eye by recording the intensity of red, green, and blue light separately. It is used to distinguish the color and shape of objects in captured video to understand what the robot is looking at. In existing robot learning methods, data was collected solely from this video information.*

> ** vision language model: Refers to an artificial intelligence model trained to process images (visual information) and text (linguistic information) simultaneously and understand the relationships between them.*



▲ ***Structure diagram of AI robot navigation technology (Context-Nav). It illustrates the process of a robot searching for a target object by creating a 3D map of the surrounding environment using an RGB camera, depth sensor, location information, and target description. When an object is detected, the AI checks its color, shape, and location to determine if it is the target; if it is, it stops, otherwise, it continues the search.***

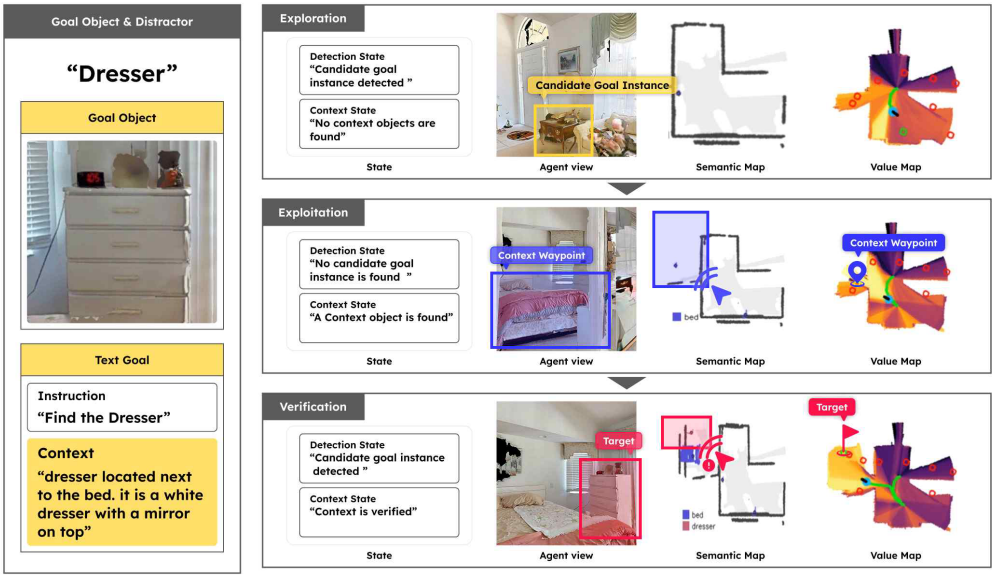
The research team's technology also demonstrated high performance in the 'Goal Finding Test (CoIN-Bench)*,' which evaluates a robot's understanding of long sentences containing detailed attributes such as relationships with objects, color, and shape.

Unlike conventional reinforcement learning, where robots learn the optimal behavior for finding a goal autonomously through repeated attempts and failures, the research team's technology achieved a success rate of 20.3% without additional training, demonstrating a performance approximately 2.3 times higher than existing reinforcement learning-based methods (8.9%).

In particular, the study proved that a strategy involving interpreting human descriptions in 3D space, reflecting the context of the entire sentence to move and search from the most probable locations, and then verifying positional relationships with surrounding objects is effective in significantly improving the accuracy of robot behavior.

Furthermore, it was confirmed that the more fully the entire long sentence description is incorporated into the search process, the less unnecessary movement occurs. Additionally, the process of verifying positional relationships in 3D from multiple viewpoints reduces instances of misidentifying the goal.

> ** Goal Finding Test (CoIN-Bench): A public test evaluating a robot's ability to locate a target object based on a long natural language description. It measures success rates and movement efficiency in environments that include detailed attributes such as relationships with objects, color, and shape, and is a reliable standard that allows researchers worldwide to compare performance under the same conditions based on evaluation sets released at academic conferences such as ICCV 2025.*



▲ ***An example of a step-by-step context-based search process. It illustrates the process of a robot searching for a "white chest of drawers with a mirror next to the bed," temporarily suspending initial candidates, moving to a room that matches the context within the sentence, and finally verifying the target object.***

Professor Ue-Hwan Kim stated, "This research presents a precise technology that goes beyond the level of robots merely observing the characteristics of an object itself, enabling them to understand surrounding context and 3D spatial relationships as well." He added, "Since it can be immediately applied to descriptions of new spaces or unfamiliar objects without separate training or adjustments tailored to specific tasks, it will become a core foundational technology that enhances the practical applicability of indoor service robots and intelligent robot systems in the future."

This research, supervised by Professor Ue-Hwan Kim of the Department of AI Convergence and conducted by Won Shik Jang, a student in the integrated master's and doctoral program, was supported by the Ministry of Science and ICT and the National Research Foundation of Korea's Excellent Young Researcher Support Program, the Institute for Information and Communication Technology Planning and Evaluation (IITP)'s "Development of Self-Directed Visual Intelligence Technology Based on Problem Hypothesis and Self-Supervision," and the National Science and Technology Council (NST)'s Global TOP Strategic Research Group Project.

The research results — Context-Nav: Context-Driven Exploration and Viewpoint-Aware 3D Spatial Reasoning for Instance Navigation — were pre-released on the international academic server 'arXiv' on March 18, 2026, and are scheduled to be

presented at the prestigious international conference in the field of AI, the Computer Vision and Pattern Recognition Conference (CVPR 2026).

CVPR 2026 will be held in Denver, Colorado, USA, from June 3 to 7.

Meanwhile, GIST stated that this research achievement takes into account both its academic significance and potential for industrial application, and that discussions regarding technology transfer can be conducted through the Technology Commercialization Office (hgmoon@gist.ac.kr).