

“A single building in a satellite image has changed... Will AI notice?” GIST develops criteria to evaluate AI’s contextual awareness and image comprehension across diverse scenarios, such as disaster management and autonomous driving

- *A team led by Professor Ue-Hwan Kim of the Department of AI has developed an advanced benchmark comprising 4,996 image pairs across 51 scenarios in four categories... The benchmark moves beyond conventional word-matching methods to comprehensively evaluate accuracy, specificity, and contextual relevance*
- *Evaluation of GPT-5.2 revealed that while the generated sentences were natural, the model lagged significantly behind humans in its understanding of actual situations and the consistency of its explanations*
- *The benchmark is expected to contribute to performance improvements in various AI applications, such as disaster monitoring and autonomous driving, and has been accepted for presentation at the international conference ECCV 2026*



▲ (From left) Professor Ue-Hwan Kim (Department of AI) and Jae-Woo Kim (combined master’s/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 has developed "C3-Bench" (Context-Aware Change Captioning Benchmark). This new evaluation framework assesses whether an AI model can identify and accurately describe significant changes between two images based on a specific context.

Accurately comparing AI performance requires a standard to determine which model solves a given problem more precisely. C3-Bench functions as a test of an AI’s ability to comprehend changes; it presents two images along with a specific objective—

indicating what to look for—and evaluates whether the AI selects and describes changes relevant to that objective.

Simply identifying a large number of differences between two photos does not necessarily mean the AI truly understands the changes within them. This is because the significance of a change depends on the specific context or what is being examined.

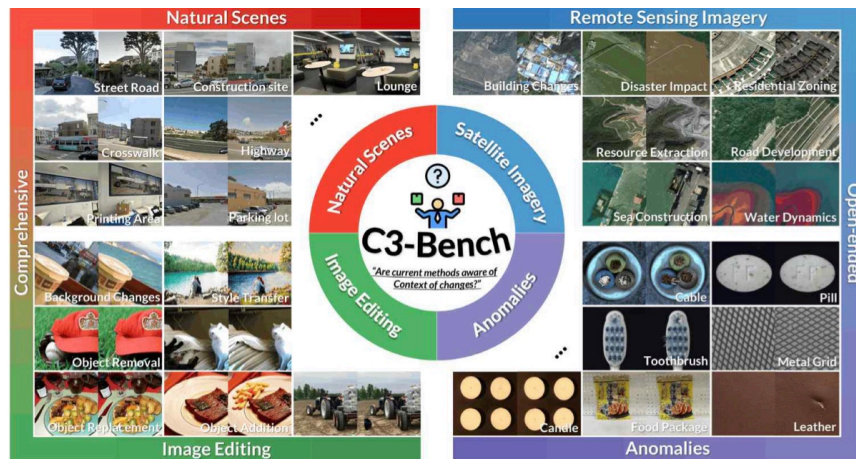
For instance, when comparing two photos of a railway, the accumulation of snow or a decrease in cloud cover might be considered significant changes if the goal is to observe weather variations. Conversely, if the objective is to assess railway safety, the appearance of a new train on the tracks would be the critical change, while weather-related differences might be deemed unimportant.



▲ *Examples of how the "correct answer" shifts depending on the context. In a pair of railway images, for instance, changes in snow and cloud cover are the key focus from a weather perspective, whereas the appearance of a train on the tracks is the critical change from a railway safety perspective.*

However, existing evaluation datasets often focused on specific scenes or types of change, or lacked clear guidelines on which changes were significant in a given situation. Consequently, it was difficult to determine whether an AI model that performed well on specific data could accurately identify relevant changes across diverse real-world scenarios.

To address these limitations, the research team selected 51 realistic change scenarios across four categories, including ▲ natural scenes, ▲ remote sensing imagery (satellite/aerial), ▲ image editing, and ▲ anomaly detection, and constructed "C3-Bench" using 4,996 image pairs accompanied by human-written descriptions of the changes.



▲ *Overview of C3-Bench. It encompasses 51 change contexts—such as roads, construction, disasters, resource development, and product defects—categorized into natural scenes, remote sensing imagery, image editing, and anomaly detection, and comprehensively evaluates the ability to describe changes appropriate to each context.*

For each scenario, specific criteria were established regarding the "changes to identify," "differences to ignore," and the "standards to follow when describing changes."

This approach ensured that, rather than simply listing visible differences between image pairs, the AI could be evaluated on its ability to accurately select and describe changes that were significant within the given context. The method for evaluating the AI's responses also differed from conventional approaches; instead of merely comparing word overlap with a reference sentence, the system comprehensively assessed the meaning of the response by examining ▲ accuracy, ▲ specificity, ▲ naturalness, and ▲ contextual relevance.

The team also verified how well the new evaluation method aligned with human judgment. By comparing results from the proposed method against direct human evaluations of the same AI responses, they found that their method more closely mirrored human judgment than existing automated evaluation techniques.

Subsequently, the research team tasked both humans and GPT-5.2 with describing changes across 400 image pairs, comparing their responses based on accuracy, specificity, naturalness, and contextual relevance. The results showed a 1.73-point gap in the composite score (out of 10), with humans scoring 7.45 and GPT-5.2 scoring 5.72. However, regarding sentence naturalness alone, GPT-5.2 scored 8.53, surpassing the human score of 8.32.

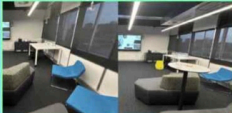
This demonstrates that an AI's ability to generate natural sentences, which is akin to human writing, is not necessarily the same as its ability to accurately identify significant changes within actual images.

The research team also tested whether the AI could consistently describe the same changes appropriately based on the sequence of the images (i.e., regardless of the order in which they were presented). For instance, if a description stated that an object "newly appeared" in the first photo, then reversing the order of the photos would require describing that same object as having "disappeared."

The research team evaluated this capability as "reversibility." The results showed that humans achieved a reversibility score of 93%, whereas GPT-5.2 scored only 61%. This demonstrates that while state-of-the-art AI excels at generating natural sentences, it still faces limitations in consistently grasping the direction of changes within images.

The team also analyzed the specific areas where AI tends to make errors when interpreting changes in photographs.

The results revealed that visual recognition errors, such as failing to properly identify objects in the images or misinterpreting their types and characteristics, accounted for the largest share (63.3%) of all errors. Other identified errors included interpreting insignificant differences as actual changes and misdescribing the significance of a change or the spatial relationships between objects.

Perceptual Error	Spatial Error	Inconsistency Error	Change Collapse Error
			
GT: The tree in the center have been scorched, and a villa in the upper right corner has been destroyed. Prediction: There is no identifiable change.	GT: The yellow stool added beside the rectangular white desk. Prediction: The white rectangular table has been replaced by a round table, and a yellow stool added.	Prediction: Remove the three dogs from the left side and replace the field on the left with a small pond. Prediction (reverse): Add three additional dogs in the grassy field.	Prediction: A new surface crack has developed near the right edge of the pill. Prediction (reverse): There is no identifiable change.

▲ *Examples of common errors made by the latest AI (GPT-5.2). From left: perceptual errors (missing objects or attributes), spatial errors (misunderstanding spatial relationships), inconsistency errors (descriptions failing to match when image sequences are reversed), and "change collapse" errors (detecting changes from only one direction). Red text indicates incorrect parts of the descriptions.*

Professor Ue-Hwan Kim stated, "This study systematically demonstrates that an AI's 'ability to provide natural descriptions' is not necessarily the same as its 'ability to accurately understand context-specific changes.'" He added, "We hope C3-Bench will serve as a common foundation for future research aimed at evaluating and improving

AI's ability to understand changes in diverse real-world environments, such as monitoring changes at disaster or accident sites, remote sensing using satellite and aerial imagery, autonomous driving, and anomaly detection."

This research was supervised by Professor Ue-Hwan Kim of the Department of AI and led by Jae-Woo Kim (a student in the combined master's and doctoral program) as the first author. It was supported by the Ministry of Science and ICT (MSIT) and the National Research Foundation of Korea (NRF) through the Outstanding Young Researcher Support Program; the Institute for Information & Communication Technology Planning & Evaluation (IITP) through the project on self-directed visual intelligence technology based on problem-hypothesis and self-supervision; and the MSIT's GIST-InnoCORE program.

The research findings — C3-Bench: A Context-Aware Change Captioning Benchmark — have been accepted for presentation at the European Conference on Computer Vision (ECCV) 2026, scheduled to be held in Malmö, Sweden, from September 8 to 12, 2026. ECCV is a major international conference in the fields of computer vision and machine learning, organized by the European Computer Vision Association (ECVA).

Meanwhile, GIST announced that it is accepting inquiries from companies and organizations interested in the technology commercialization and industrial application of this research achievement through the Technology Commercialization Center at the Institute for Startup Innovation (hgmoon@gist.ac.kr).