HKIE THE HONG KONG INSTITUTION OF ENGINEERS 香港工程師學會

Environmental awareness in machines: a case study of automated debris removal using Generative Artificial Intelligence and Vision Language Models

Jolly P C Chan[1], Heiton M H Ho[1], T K Wong[1], Lawrence Y L Ho[2], Jackie Cheung[2] and Samson Tai[3]
[1] Drainage Services Department, the HKSAR Government, Hong Kong, People's Republic of China
[2] Revotech Limited, Hong Kong, People's Republic of China
[3] Hong Kong Baptist University, Hong Kong, People's Republic of China

ABSTRACT

Water channels play a crucial role in stormwater management, but the build-up of debris in their grilles can lead to flooding, endangering humans and animals, properties, and critical infrastructure nearby. While automated mechanical grab systems are necessary for efficient debris removal, their deployment in outdoor environments has been non-existent due to safety concerns. Here we report the successful use of Generative Artificial Intelligence (GenAI) and a Vision Language Model (VLM) to endow an automated mechanical grab with "awareness", which allows it to differentiate between non-living and living objects, deciding whether to initiate or abort grabbing actions. The existing approaches such as YOLOv7 only achieve a sensitivity of 86.94% (95% CI: 83.44% to 89.93%) in detecting humans and specified animals. They systematically miss crouching workers and animals facing away from the cameras. Grounding DINO (VLM) can achieve a sensitivity of 100% (95% CI: 99.17% to 100.00%) and a specificity of 85.37% (95% CI: 77.86% to 91.09%). Together with BLIP-2 (GenAI), it acquires "awareness", allowing it to detect animals beyond those specified. This opens up possibilities for the application of GenAI/VLM in automation sectors where human-machine mingling occurs, such as manufacturing, logistics, and construction. This innovation can potentially improve the safety and efficiency in these domains.

KEYWORDS Artificial intelligence; machine learning model; object detection; computer vision; flood management; debris clearance
CONTACT Jolly P C Chan pcchan02@dsd.gov.hk
Received 4 March 2024

1. Introduction

Climate change has emerged as a global concern, significantly influencing weather patterns and causing shifts in hydrological regimes. The increased occurrence of extreme weather events, including heavy rainfall, has led to a rise in flood incidents worldwide (Hennessy et al., 2022), and Hong Kong is no exception. The annual rainfall figures from the Hong Kong Observatory showed that there was an increasing rate of 2.3 mm per year during 1884 to 2023 (Hong Kong Observatory, no date).

Water channels are essential components of stormwater and flood management systems. During heavy rainfall, debris including branches, leaves and other waste materials follows the waterflow and become accumulated at the grille systems in water channels. The accumulation of debris, if not timely cleared, can impede the water flow, leading to localised flooding and posing significant threats to infrastructure and properties nearby, as well as public safety (Fathy et al., 2020).

The case under study is the Kong Yiu Channel, a man-made concrete-lined channel located near Lin Ma Hang along the Hong Kong-Shenzhen border. The channel is equipped with a grille system designed to intercept debris, including tree branches and shrubbery from the local ecology as well as traffic cones from nearby roads. Historically, manual labour was employed for debris removal. However, in August 2021, the water channel experienced two instances of flooding due to debris accumulation. This highlights the inadequacy of manual clearing methods, as debris tends to wash down from upstream and accumulate in the grille during heavy rainfall. Moreover, working in a flowing water channel poses safety risks, particularly in unfavourable weather conditions. Another challenge arises from the presence of feral animals, such as cats, dogs, boars and monkeys, in the vicinity. These factors create difficulties in promptly addressing debris accumulation without compromising the safety and well-being of both people and animals.

To address these challenges, the Drainage Services Department installed an automated mechanical grab system with Generative Artificial Intelligence (GenAI) and a Vision Language Model (VLM) at the Kong Yiu Channel for debris removal. This is the first outdoor automated mechanical grab system in water channels in Hong Kong and possibly one of the first use cases of using GenAIs and VLMs for such a use case in the world.

The contribution of this paper is summarised as follows:

(a) to demonstrate a novel solution combining GenAI and a VLM to bestow "awareness" on an automated mechanical grab system so that it can differentiate between living and non-living objects for debris clearing in open water channels to ensure operation efficiency and safety; and

(b) to provide a reference model for integrating GenAI and a VLM to endow "awareness" to machinery in automation sectors where human-machine mingling occurs, such as manufacturing, logistics, and construction, thereby improving the safety and efficiency in these domains.

2. Literature review

In this section, we review the current practice of debris clearance in water channels, challenges in object detection and the selection of object detection models for evaluation.

2.1. Current practice of debris clearance

Debris accumulation in grilles or screens in water channels is a prevalent issue worldwide, including in Hong Kong. Traditionally, manual labour has been used to clear debris in open water channels. In the United Kingdom, the prevailing practice involves manual clearance of screens using CCTV, telemetry, or other sensors to detect blockages (Environmental Agency, 2022). However, these methods have limitations in terms of timeliness and worker safety.

To effectively address the challenge of debris clearance and minimise the risk of flooding, an automated mechanical grab system is necessary. Initially, a time-based clearing approach was considered, but it lacks adaptability to changing conditions as debris accumulation rates vary significantly over time. Initiating a grabbing action without the presence of debris consumes energy, causes unnecessary wear and tear, and poses safety risks to nearby individuals and organisms.

An alternative solution is a demand-based system that utilises traditional computer vision. When debris is detected, the grab system is directed to the location of the debris and clears it. Although this approach is more energy-efficient than time-based control systems, it still lacks "awareness" and may mistakenly grab objects that are not debris, such as people or animals, leading to serious safety risks. A tragic incident occurred in 2023 in South Korea, where a worker was crushed to death by a robot that failed to differentiate him from the boxes of food it was handling (Atkinson, 2023).

2.2. Objection detection models

To ensure the safe operation of the automated mechanical grab system, the detection head is required to accurately differentiate living objects from non-living objects, which are often debris or trash. The most common approach is to use a traditional object detection model (Zou et al., 2023).

However, traditional object detection models face several challenges. A major one is detecting objects that are partially covered or outside the frame. These objects may only be partially visible within the camera's field of view or obstructed by other objects, making accurate detection difficult. Another one is the detection of small objects. Small objects present a significant challenge due to their limited visual information and low resolution. The most troublesome one is the variability in object appearance, as objects can vary significantly in terms of appearance, shape, size, lighting contrast, and orientation. For example, an algorithm not trained to differentiate between variations may mistake a person wearing a safety vest for a traffic cone. Furthermore, these models struggle with understanding abstract concepts, such as "living things".

In this respect, VLMs are also proposed for evaluation. VLMs (Li et al., 2022) are multimodal models that combine computer vision and natural language processing. They leverage pre-training on image datasets like ImageNet (Deng et al., 2022) and COCO (Lin et al., 2014) to learn general visual representations. Textual prompts are used to provide contextual comprehension of images, offering either generalised descriptions or specific object localisations (Zang et al., 2023). This language-based and context-aware interface enhances object detection applications, reducing the development costs for adaptation.

Based on the leaderboard in Object Detection on "Papers with code" (MetaAI, no date), a website that offers side-by-side comparisons of machine learning models' performance on various benchmarks, such as Grounding DINO (Pre-training data: OI, Cap4M, RefC, O365, GoldG, COCO; Swin-L Transformer. C = 192, layer numbers = {2, 2, 18, 2}) and YOLOv6, achieve Average Precision (AP) 63.0 and AP 57.2, respectively. Therefore, we chose Grounding-DINO and YOLOv7, the most up-to-date version of YOLO, for comparison.

You Only Look Once (YOLO) (Jiang, 2022) originally developed by Redmon et al. (Redmon et al., 2015), is a family of state-of-the-art real-time object detection architecture utilising Convolutional Neural Networks (CNNs), which are commonly deployed in commercial scenarios to detect humans and animals. YOLO reframes object detection from a classification problem to a regression problem for predicting specially separated bounding boxes and the associated class probabilities. The one-stage detection feature makes YOLO fast and efficient; however, in terms of generalisation, the models are trained on a specific dataset and may not generalise well to objects that have not been trained on it, which means that it may not be able to detect if a man is partially visible or blocked by other objects in the image.

Grounding-DINO is an open-set object detector which combines the Transformer-based detector DINO with grounded pre-training GLIP (Liu et al., 2023). Open-set object detection refers to the ability to predict all the objects of interest within an image, then classify all the objects as semantic classes, which are in turn identified using textual prompt inputs (Dhamija et al., 2020). Since both YOLOv7 and Grounding-DINO lack the capability to understand abstract concepts and cannot develop situational awareness beyond their training data, they cannot confer

any "awareness" on the detection head. Therefore, BLIP-2, a GenAI model, was also chosen for evaluation.

In a nutshell, Grounding-DINO, YOLOv7 and BLIP-2 were chosen for evaluation and comparison. Figure 1 provides how they differ from each other.

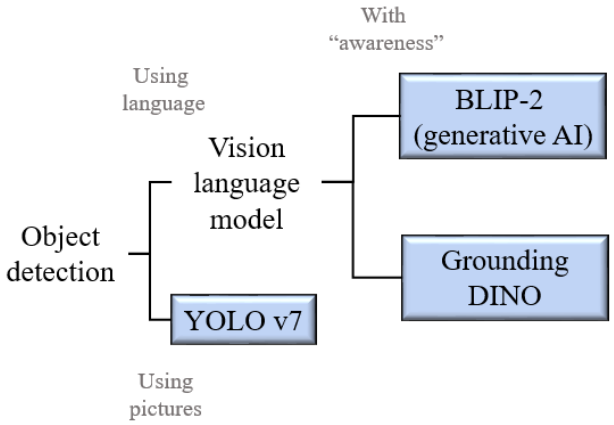


Figure 1. YOLOv7, BLIP-2 and Grounding-DINO.

3. Methodology

3.1. Introduction

The use of abstraction layers is proposed in order to facilitate a meaningful comparison of the performance of YOLOv7, Grounding-DINO, and BLIP-2. The concept of abstraction layers is commonly employed in software and complex system analysis and design. By utilising this concept, each abstraction layer can be developed, tested, and evaluated independently of other layers. It also provides modularity that enhances the scalability and transferability across the system.

A curated evaluation dataset comprising annotated images was utilised in order to assess the performance of the systems and enable a fair comparison. This dataset was designed to test the models' capabilities in detecting various objects and environmental conditions, simulating real-life scenarios. The dataset consisted of 567 still frame images extracted from video recordings captured by IP cameras installed in strategic locations around the channel. These images represent daily scenarios, including maintenance works, pedestrian activities, debris accumulation, animal intrusions, and empty channel conditions (Figure 2).

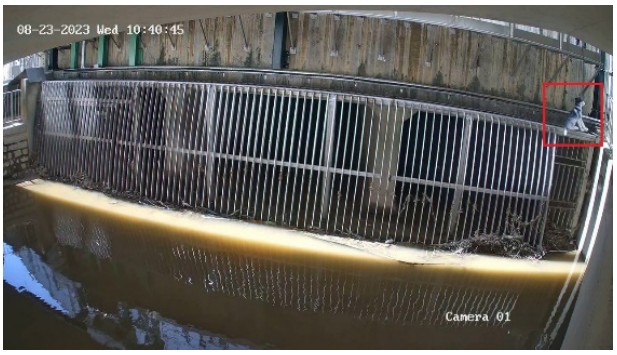
Evaluation metrics are crucial for quantitatively assessing the performance of the systems. Several key metrics have been proposed to provide comprehensive insights into different aspects of the models' performance. The chosen metrics include sensitivity, specificity, Matthews Correlation Coefficient (MCC), and inference time.

3.2. Abstraction layers

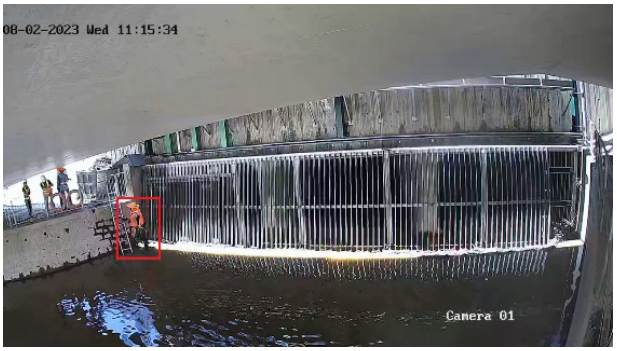
Figure 3 is an illustration of the abstraction layers of the automated mechanical grab system. The system is divided into five layers, namely Physical Layer, Control Layer, Localisation Layer, Detection Layer, and Application Layer. In this study, YOLOv7, Grounding DINO and BLIP-2 all reside in the Detection Layer.

Cameras are connected to the system via the Application Layer, which is responsible for converting the video feed into images and transmitting them to the Detection Layer. The Detection Layer continuously detects the accumulation of debris as well as the presence of people or designated animals. Once the Detection Layer identifies a build-up of debris surpassing a predefined threshold without the presence of people or animals, it signals the Localisation Layer to determine the precise location of the debris accumulation. The Control Layer receives instructions from the Localisation Layer regarding the location of the debris build-up and directs the actions of the Physical Layer, which is the mechanical grab system. In the event that individuals or animals are detected by the Detection Layer, the Control Layer triggers the cessation of the grabbing action by the Physical Layer. This algorithm is equipped with software safeguards to ensure safety while minimising the occurrence of false alarms and protracted periods of inactivity. The details regarding these safeguards will be presented in Section 4.4. Additionally, the Detection Layer also broadcasts warning messages through speakers and sends an SMS message to the operator. This coordinated response ensures the safety of the system and prevents any potential harm to people or animals present in the vicinity.

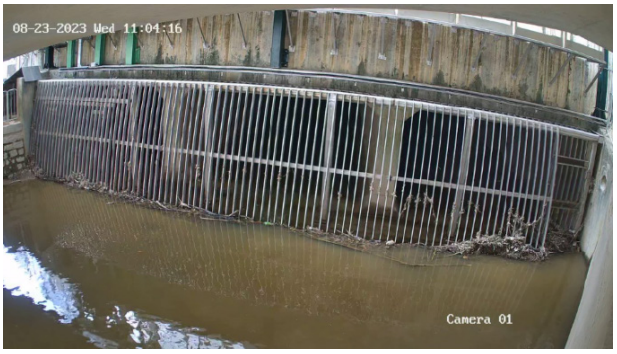
(a)



(b)



(c)



(d)

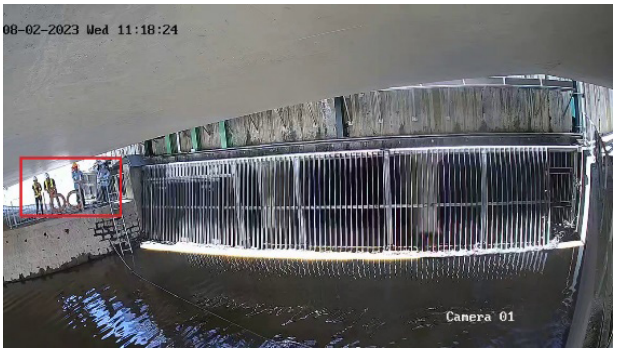


Figure 2. Examples of the image dataset extracted from video recordings: (a) animal intrusion; (b) maintenance work; (c) empty channel conditions; and (d) pedestrian activities.

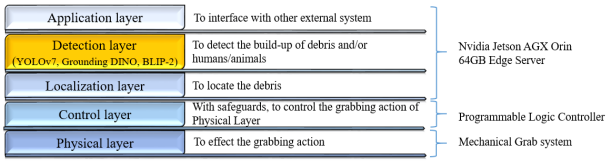


Figure 3. Abstraction layers of the automated mechanical grab.

The Physical Layer is the mechanical grab system at the Kong Yiu Channel (Figure 3). This system consists of a gripper unit, a travelling trolley with a driving speed of 15 m/min and supporting tracks. It is designed to travel to the location where it detects an accumulation of debris up to a predefined level, grab the debris, and transfer it to the garbage skip at the side. The maximum loading of the grab is 500 kg. The Control Layer controls the operations of the mechanical grab system through a Programmable Logic Controller (PLC).

YOLOv7 and Grounding-DINO operate across both the Detection Layer and Localisation Layer, capable of performing both detection and localisation tasks. On the other hand, BLIP-2 solely handles detection and relies on another model within the Localisation Layer for object localisation. In this study, we used Grounding-DINO as the localisation model for BLIP-2. These two layers, together with the Application Layer, are all operated from the Nvidia Jetson AGX Orin 64GB edge server.

All three models were fine-tuned using the same set of 600 images, which included images featuring workers, animals, and scenes containing only debris in the working area. Some training images were obtained from CCTV footage of workers engaged in desilting activities near the grille, while others were sourced from the web or generated by AI, using the grille area as the background.

To enhance the performance, Low-Rank Adaptation (LoRA) and parameter-efficient fine-tuning were adopted, utilising default hyperparameters without optimisation.

Figure 4. Mechanical grab system at the Kong Yiu Channel.

3.3. Evaluation dataset

An annotated image dataset was designed and curated to test the capabilities of AI models in detecting the presence of people and specified animals, namely cats, dogs, wild boars, and monkeys, which are reasonably foreseeable to appear at the Kong Yiu Channel. The dataset comprised a diverse selection of annotated images, covering various object classes and environmental conditions to simulate real-life scenarios. The images were extracted from video recordings captured by strategically placed 720p IP cameras installed on either side of the channel and one camera positioned on the top of the maintenance walkway. A total of 567 still frame images were extracted from four sets of recordings in June, July, and October, representing the daily scenarios in the vicinity of the channel. These scenarios encompassed maintenance works, pedestrian activities, object throwing, rainy conditions with flooding, and an empty channel without live objects.

The images in the dataset were annotated with the objects present. Each scenario was represented in similar proportions within the total dataset. Notably, the scenario of an empty channel, which is prevalent in reality, was given due consideration. To reflect real-world conditions, still frame images of the channel were captured under different lighting scenarios, ranging from morning to dusk across different months. Of the 567 images, 444 contained people and animals, while 123 depicted an empty channel without any live objects. To simulate animal intrusions into the grille system, deliberate efforts were made to introduce commonly found debris and trash items.

The annotated image dataset addresses several challenges commonly encountered in object detection tasks. These challenges include objects that are partially out of frame, small-sized objects and different variations within the same object category. By incorporating such variations and complexities, the dataset provides a comprehensive evaluation of the AI models' performance in realistic scenarios.

3.4. Evaluation metrics

When evaluating the performance of the automated mechanical grab system, it is essential to select appropriate evaluation metrics that provide a comprehensive assessment of its capabilities. In this context, sensitivity, specificity, MCC, and inference time were chosen as the key evaluation metrics. Each metric serves a specific purpose in evaluating different aspects of the system's performance, ranging from accuracy to computational efficiency. Accuracy and F1 score, though commonly used metrics in AI applications, were not selected due to specific limitations.

Sensitivity and specificity metrics were chosen as the most important metrics for evaluation. Sensitivity, also known as recall or true positive rate, measures the proportion of the presence of people or animals that are correctly identified by the system. On the other hand, specificity measures the proportion of the absence of people and animals that are correctly identified as negative by the system. Higher sensitivity ensures greater safety, while higher specificity reduces false alarms. These rates are instrumental in the design of effective "safeguards", which will be discussed in Section 4.4.

Given the imbalanced nature of the dataset, the objective was to evaluate four outcomes of a binary classification task: maximising safety by increasing true positives and minimising false negatives, while also reducing nuisance alarms by maximising true negatives and minimising false positives. In light of these objectives, MCC was selected as the primary evaluation metric. MCC presents distinct advantages over other metrics, such as the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). While AUROC is most effective for balanced datasets focused on overall performance, and AUPRC is more suited for imbalanced datasets with a primary emphasis on the positive class, MCC comprehensively accounts for all four outcomes of the confusion matrix. This makes it particularly valuable for this project, as it effectively addresses the challenges posed by dataset imbalance and provides a balanced measure that accurately reflects the classification quality.

The inference time measures the duration that it takes for the system to process and make predictions on the video feed obtained from the cameras. This metric is crucial for assessing the efficiency and real-time applicability of the automated mechanical grab system. Evaluating the inference time provides insights into the system's computational efficiency, ensuring that it can operate effectively in real time or near-real-time scenarios.

Accuracy and F1 score, although commonly used metrics for binary classification tasks, were not chosen for this evaluation. While accuracy measures the overall correctness of the system's predictions, it may not provide a complete picture in the presence of imbalanced datasets or varying costs of correct or incorrect detection for people and animals. In this study, it may overemphasise the performance on the majority class. Similarly, the F1 score, which is the harmonic mean of precision and recall, does not consider true negatives and treats false positives and false negatives equally. This limitation makes the F1 score unsuitable in this context, as it does not adequately account for the consequences of incorrectly stating the absence of people or animals.

4. Results and discussion

4.1. Introduction

Based on the methodology introduced in Section 4, YOLOv7, Grounding DINO, and BLIP2 were evaluated, and the results are presented in this section. The results for sensitivity, specificity, MCC, and inference time are shown,

compared, and evaluated. Moreover, the integration of BLIP2 and Grounding-DINO for the purpose of providing "awareness' to the automated mechanical grab is discussed, with an explanation of the "safeguards" designed based on the evaluation results provided. Furthermore, an overview of the real-life performance of the automated mechanical grab at the Kong Yiu Channel is provided. Finally, the implications of combining GenAI with VLM are discussed.

4.2. Performance

Table 1 presents the values for sensitivity, specificity, MCC, and inference time. Since YOLOv7 and Grounding-DINO can perform the functions of both the Detection Layer and Localisation Layer simultaneously whilst BLIP-2 cannot, the inference time of BLIP-2 is presented with its combination with Grounding-DINO, where BLIP-2 handles the detection aspect and Grounding-DINO takes care of localisation. Figure 5 shows their inference time and Figure 6 shows the sensitivities and specificities of the three models.

Table 1. Performance of different models.

Model	Sensitivity	Specificity	MCC	Inference time (ms)
YOLOv7	86.94% (386/444)* (95% CI: 83.44% to 89.93%)	90.24% (111/123)# (95% CI: 83.58% to 94.86%)	0.695	35
Grounding DINO	100% (444/444)* (95% CI: 99.17% to 100.00%)	85.37% (105/123)# (95% CI: 77.86% to 91.09%)	0.906	258
BLIP2	99.10% (440/444)* (95% CI: 97.71% to 99.75%)	75.61% (93/123)# (95% CI: 67.05% - 82.90%)	0.818	6384^

* Sensitivity = True positive / (True positive + False negative)
# Specificity = True negative / (True negative + False positive)
^ The inference time of running both BLIP-2 and Grounding-DINO

Figure 5. Chart for inference time comparison.

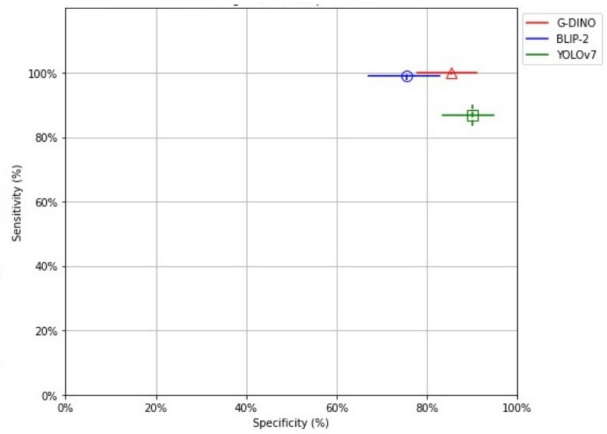


Figure 6. Sensitivities and specificities of models.

In terms of detection performance, Grounding-DINO stands out with the highest MCC, indicating superior overall results. Additionally, it is the only model in the evaluation to reach 100% sensitivity (95% CI: 99.17% to 100.00%). However, Grounding-DINO has a lower specificity (85.37%) compared to YOLOv7 (90.24%), which means that it may produce more false alarms. On the other hand, as expected, YOLOv7 has the lowest sensitivity (86.94%). This is particularly evident when the target, whether it be a person or animal, is partially out of view or when the lighting conditions cause significant reflections. YOLOv7 is also more susceptible to the variations in worker posture or viewing obstructions. As for BLIP-2, it exhibits a very high, although not 100%, sensitivity (99.10%) but suffers from the lowest specificity (75.61%).

When considering the inference time, both YOLOv7 and Grounding-DINO demonstrate fast performance (35 ms and 258 ms, respectively), making them practically indistinguishable in terms of suitability for most applications. However, BLIP-2 has a considerably longer inference time (6384 ms). This slower speed is still acceptable for the automated mechanical grab at the Kong Yiu Channel: with its cameras capturing a wide view of the scene of its fenced-off area, it is highly unlikely for any living objects, if possible at all, to be able to reach the mechanical part without being detected within a few seconds. However, this slow speed may render BLIP-2 not suitable for applications that require high-speed or real-time detection unless a more efficient edge server than Nvidia Jetson AGX Orin 64GB edge server is used.

4.3. Awareness via Generative AI

Although based on the results in Section 4.2, Grounding-DINO seems to be the best option to adopt. However, recognising that Grounding-DINO (and also YOLOv7) lacks the capability to detect all living objects, it being limited to the objects that are specified in the prompts, i.e., people, cats, dogs, wild boars, and monkeys, and acknowledging that BLIP-2, while theoretically capable

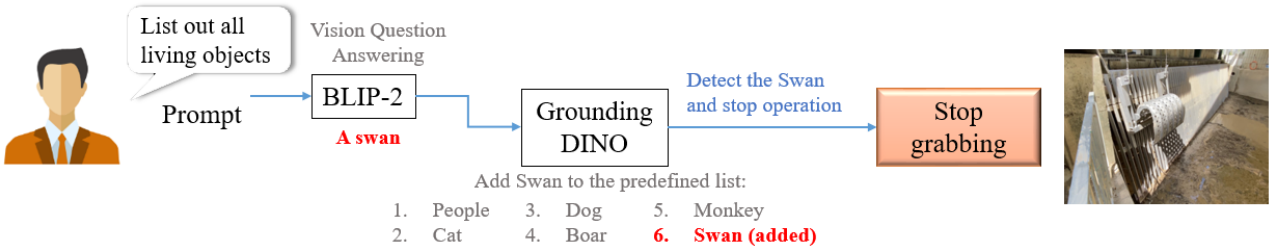


Figure 7. Integration of visual question answering of BLIP-2 with Grounding-DINO.

of detecting all living objects, suffers from poor specificity due to its tendency to "hallucinate", we propose a solution that combines the strengths of both models.

By integrating the creative powers of BLIP-2 with the object detection capability of Grounding-DINO, which outperforms YOLOv7, we aimed to endow the automated mechanical grab with "awareness". We prompted BLIP-2 to utilise its "Visual Question Answering" capability to generate a comprehensive list of living objects. The programme then refines this list by removing objects that BLIP-2 may incorrectly identify as living objects, such as bridges, debris, or grilles. The resulting curated list is then incorporated into Grounding-DINO's existing list of objects under detection, which includes people and the four anticipated animal categories (Figure 7).

Through this approach, the automated mechanical grab acquires its own "awareness" of all living objects, in addition to being programmed to detect specific objects. This concept can be applied to various scenarios beyond the current context.

Although the "awareness" process may exceed 6 seconds, it is important to note that this duration is only activated when the automated mechanical grab is specifically initiated for a grabbing operation. After providing the refined object list to Grounding-DINO, the detection time will return to sub-second levels until the completion of the grabbing operation.

4.4. Safeguards

Safeguards are put in place to ensure the automated mechanical grab operates with minimal false alarms and non-grabbing incidents while maintaining a high safety standard. By combining control algorithms based on statistical confidence measures and fallback procedures like timed alerts for prolonged non-grabbing action, the system aims to balance operational efficiency with safety. When the grabbing action is initiated, the system begins scanning with six scans in a detection block, which occurs within two seconds. If any of the scans in a detection block detect the presence of humans or animals, the grabbing action is immediately stopped. Otherwise, the grabbing action is initiated.

To reduce false alarms and minimise disruptions for operators, the control algorithm requires that all six scans in a detection block must be positive before an alarm is issued. However, if any of the scans detect the presence of humans or animals, the grabbing action will be halted. The false alarm rate is calculated as "1 - specificity." In the worst-case scenario, where the test maintains over 95% confidence (as shown in Table 1), the specificity is around 23%. This results in a very low rate of incorrect identifications-about 0.01% (= 0.236). It's important to note that alarms are triggered only after the grabbing action begins, meaning this 0.01% false alarm rate equates to roughly one false alarm for every 10,000 grabbing actions. To mitigate extended periods of non-grabbing caused by false detection, an additional safeguard is implemented. If no grabbing action is initiated within a specified adjustable time frame, an alert is issued to the operator. Considering the system's specificity rate of 78% (> 95% confidence), the worst-case scenario delay when the grabbing action is triggered by demand-based detection is 10 seconds.

To mitigate safety risks, a single detection in a detection block of six triggers an immediate stoppage. Considering that the rate of presence of people/animals incorrectly identified is non-existent in testing but statistically it is < 1% (> 95% confidence), the risk of grabbing in the presence of people/animals is < 10-10 %, which is virtually non-existent.

4.5. Real-life performance at the Kong Yiu Channel

During the on-site validation process, a red traffic cone was utilised as debris to assess the performance of the automated mechanical grab system. The system successfully identified the cone's precise location and promptly dispatched the grab to retrieve it. Additionally, a realistic toy dog sized to resemble an actual dog, was employed to validate the safety function. Upon detecting the toy dog, the grabbing action was immediately halted. This safety mechanism was similarly activated when a worker approached the detection zone.

The automated mechanical grab underwent a "stress test" during Super Typhoon Saola and during the rainstorm on 7 and 8 September 2023, which was the most severe rainstorm in Hong Kong's records with most of the territories receiving more than 600 mm of rainfall in 24 hours. Despite these extreme conditions, nearby areas remained free from flooding incidents. The system

effectively collected twigs and other debris. There were no safety incidents reported since the system's commencement of operation. With its proven performance, similar grabs should be deployed in other water channels.

4.6. Implications

The challenge of debris accumulation and flooding in open channels extends far beyond the Kong Yiu Channel and even transcends the boundaries of Hong Kong. It is, in fact, a global concern that countries worldwide are actively addressing using modern technologies. For instance, the United Kingdom employs a combination of CCTV, telemetry, and other sensors along watercourses to guide their manual clearance efforts (Environmental Agency, 2022).

The introduction of the combination of GenAI and a VLM with demand-based debris clearing represents a significant paradigm shift. This innovative approach enables debris clearance based on real-time demand, all while prioritising the safety of both people and wildlife. It signifies a substantial advancement in ensuring efficient and safe debris removal processes not only in Hong Kong, but also worldwide. The invention will be showcased and promoted to mainland China and globally through regular communications in the field, international conference and institutes, setting the stage for potential collaborations and expansion.

We also demonstrate the capability of utilising GenAI and a VLM to provide an automated system with "awareness". This enables the system to detect objects beyond the predefined categories, an approach typically found in traditional AI object detection systems. Industries characterised by human-machine interaction, including manufacturing, logistics, and construction, stand to benefit significantly from this breakthrough. The integration of GenAI/VLM technology in these sectors holds the promise of improved performance, enhanced safety, and increased efficiency. As the power and affordability of GenAI and VLMs as well as GPUs continue to advance with each passing day, the implementation of these technologies will only become more accessible and seamless.

5. Conclusion

This paper introduces a ground-breaking combination of GenAI and a VLM in the form of embodied AI to enhance an automated mechanical grab system for debris clearance in open water channels. By endowing the system with "awareness", it gains the capability to distinguish between non-living and living objects, enabling it to make informed decisions about whether to grab or to stop operation, thereby improving the efficiency and safety in debris removal.

The evaluation results demonstrate that the integration of Grounding-DINO and BLIP-2 successfully achieves this "awareness" for the automated mechanical grab and proves to be sufficient for the purpose of debris clearance in open water channels. This represents a paradigm shift as worldwide methods relying on manual labour to clear grilles or screens in open water channels can be replaced by automated systems, overcoming the efficiency and safety challenges associated with human involvement.

The implications of this paper extend beyond storm water management, offering promising applications in automation sectors characterised by human–machine mingling such as manufacturing, logistics, and construction. By providing systems with "awareness", the safety and efficiency of these sectors can be significantly improved. The ongoing advancements in GenAI and VLM technology, coupled with their increasing affordability, make it increasingly feasible to implement these innovative solutions in real-world scenarios.

Moving forward, it is crucial to continue exploring the full potential of GenAI and VLM technologies and fostering further innovation. By doing so, we can unlock new possibilities for a safer and more efficient future across various industries. Embracing and harnessing the power of these technologies will undoubtedly lead to transformative advancements and contribute to the continued progress of automation and safety worldwide.

Acknowledgements

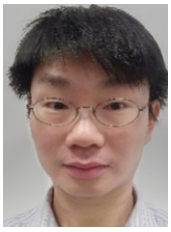
This paper is published with the permission of the Drainage Services Department of the Government of the Hong Kong Special Administrative Region of the People's Republic of China.

Notes on contributors

Ms Jolly P C Chan is an Electrical and Mechanical Engineer at the Drainage Services Department, the Government of the Hong Kong Special Administrative Region (HKSAR) of the People's Republic of China. She is currently pursuing the application of artificial intelligence in storm water and drainage services.

Ir Heiton M H Ho is a Senior Electrical and Mechanical Engineer at the Drainage Services Department, the Government of the Hong Kong Special Administrative Region (HKSAR) of the People's Republic of China. He has over 20 years of experience working with the Government of the Hong Kong Special Administrative Region (HKSAR) of the People's Republic of China in the Drainage Services Department. Ir Ho has had extensive experience managing the operation

and maintenance of the sewage treatment facilities at the Stonecutters Island Sewage Treatment Works, which is the largest of its kind in Hong Kong. Subsequently, he assumed a role in the E&M Project Division, where he is currently responsible for planning, designing, and managing various projects aimed at upgrading sewage treatment infrastructure and developing food waste co-digestion facilities.



Mr T K Wong is a Senior Electronic Engineer at the Drainage Services Department, the Government of the Hong Kong Special Administrative Region (HKSAR) of the People's Republic of China. He is currently responsible for research and development (R&D) in storm water and drainage services in the E&M Project Division. He has worked in different sectors of both the Electrical and Mechanical Services Department and the Drainage Services Department for more than 25 years. He has professional experience in biomedical engineering, operation, and maintenance (O&M) of the Stonecutters Island Sewage Treatment Works, etc.

Mr Lawrence Y L Ho obtained his M.Sc. degree in Genomics and Bioinformatics in 2020 from the Chinese University of Hong Kong. He is working on his Ph.D. degree at the Architecture and Civil Engineering Department in Hong Kong City University. He is a strong AI specialist with a proven track record in implementing data and AI solutions.

Mr Jackie Cheung earned his bachelor's degree in MSIM from Hang Seng University in 2021. Currently, he is employed as a mobile engineer at a leading health tech company, where he actively contributes to the development of cutting-edge technologies and innovation projects.

Ir Prof Samson Tai is the Professor of Practice at Hong Kong Baptist University. Besides his role as a faculty member, he is a member of the Board of Directors at ASTRI and ASTRI FinTech Limited in Hong Kong. Prior to his current academic position, Prof. Tai served as the Chief Technology Officer (CTO) at IBM Hong Kong. His exceptional contributions led to his appointment as an IBM Distinguished Engineer in 2013. Prof Tai has a bachelor's degree in computer science, and a master's and doctorate degree in Business Administration. He is a Chartered Engineer (CEng MIET), a Chartered I/T Professional (CITP MBCS), and a member of The Hong Kong Institution of Engineers (HKIE).

References

[1] Atkinson E (2023). Man crushed to death by robot in South Korea. BBC news. Available at: <https://www.bbc.com/news/world-asia-67354709>.

[2] Deng J, Dong W, Socher R, Li LJ, Li K and Li FF (2009). Imagenet: A large-scale hierarchical image database. *IEEE conference on computer vision and pattern recognition*, pp. 248-255. IEEE.

[3] Dhamija A, Gunther M, Ventura J and Boult T (2020). *The overlooked elephant of object detection: Open set*. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1021-1030.

[4] Environmental Agency (2022). Reducing flood risk - the maintenance work we do to keep rivers flowing. [online]. Available at: <https://environmentagency.blog.gov.uk/2022/03/28/reducing-flood-risk-the-maintenance-work-we-do-to-keep-rivers-flowing>.

[5] Fathy I, Abdel-Aal GM, Fahmy MR, Fathy A and Zeleňáková M (2020). The negative impact of blockage on storm water drainage network. *Water*, 12(7), 1974.

[6] Hennessy K, Lawrence J and Mackey B (2022). IPCC sixth assessment report (AR6): climate change 2022-impacts, adaptation and vulnerability: regional factsheet Australasia.

[7] Hong Kong Observatory (2022). Climate change in Hong Kong – rainfall. Hong Kong Observatory (HKO) Climate Change.

[8] Jiang P, Ergu D, Liu F, Cai Y and Ma B (2022). A Review of Yolo algorithm developments. *Procedia Computer Science*, 199, pp. 1066-1073.

[9] Li F, Zhang H, Zhang YF, Liu S, Guo J, Ni LM and Zhang L (2022). *Vision-language intelligence: Tasks, representation learning, and large models*. [online report]. Available at: <arXiv:2203.01922>.

[10] Li J, Li D, Savarese S and Hoi S (2023). *Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models*. [online report]. Available <at: arXiv:2301.12597>.

[11] Lin TY, Maire M, Belongie SJ, Bourdev LD, Girshick RB, Hays J, Perona P, Ramanan D Zitnick CL (2014). *Microsoft COCO: Common Objects in Context*. [online report]. Available at: <http://arxiv.org/abs/1405.0312>.

[12] Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J and Zhang L (2023). *Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection*. [online report]. Available at: <arXiv:2303.05499>.

[13] MetaAI (no date). Object Detection on COCO minival. Papers With Code. Available at: <https://paperswithcode.com/sota/object-detection-on-coco-minival>.

[14] Redmon J, Divvala S, Girshick R, Farhadi A (2015). *You Only Look Once: Unified, Real-Time Object Detection*. [online report]. Available at: <arXiv:1506.02640>.

[15] Zang Y, Li W, Han J, Zhou K, and Loy CC (2023). *Contextual Object Detection with Multimodal Large Language Models*. [online report]. Available at: <arXiv:2305.18279>.

[16] Zhang J, Khayatkhoei M, Chhikara P, and Ilievski F (2023). *Visual Cropping Improves Zero-Shot Question Answering of Multimodal Large Language Models*. [online report]. Available at: <arXiv:2310.16033>.

[17] Zou Z, Chen K, Shi Z, Guo Y, and Ye J (2023). *Object detection in 20 years: A survey*. In: Proceedings of the IEEE.