# A Study on Integrating YOLO and Large Multimodal Models for Improved Object Recognition

Ki Hong Im [a*], Pil Geun Jang [a] and Young Kyun Kim [a]
[a]*Korea Atomic Energy Research Insititute, 111, Daedeok-Daero 989Beon-Gil, Yuseong-Gu, Daejeon, KOREA*
[*]*Corresponding author: khim@kaeri.re.kr*

***Keywords :*** Object Recognition, YOLO, Large Multimodal Model

## 1. Introduction

In recent years, the YOLO model has emerged as a dominant tool in 2D object recognition due to its high accuracy and ability to process images in real-time. YOLO models tailored with custom datasets are particularly effective, enhancing recognition accuracy for specific items and finding widespread use across various industries. However, these custom-trained YOLO models excel at identifying known object classes but often struggle with objects outside these categories, leading to errors. This issue is especially pronounced with objects that closely resemble those in the training set, causing frequent misclassifications. Such limitations pose significant challenges when deploying YOLO for object recognition in real-world industrial settings.

To address these challenges, we propose an extended approach combining a Large Multimodal Model (LMM) with the existing YOLO model. By employing a corpus within the LMM framework, stable and accurate recognition for targeted object groups can be achieved.

## 2. Model Preparation and Evaluations

### *2.1 Model preparation for recognition task*

This study aimed to perform object recognition for seven object classes as shown in Fig. 1. To achieve this, a total of over 7,000 images, including both original and augmented data of the target object classes, were used for training [1,2]. Using this custom data, YOLOv8 was trained enough to recognize each class of objects.

In this study, images were acquired for object recognition using an RGBD camera. The software employed in the experiment consists of a custom YOLOv8 model and a commercially available LMM.

Initially, real-time images were fed into the custom YOLO model to perform object recognition. Simultaneously, images were also fed into the LMM in parallel. The final object recognition decisions were determined based on the combined outputs of these two processes. The final decisions of object recognition were categorized as follows: objects that belong to a target class shows its class label; objects from untrained general classes, which were not of interest, were categorized as ‘unclassified’; and cases where no objects were detected at all were categorized as ‘no detection’. To efficiently record the object recognition process and experimental results, a dedicated custom GUI program was developed and utilized, facilitating visual representation.

Fig. 1. Example objects from the seven target classes used in this study

### *2.2 Enhanced recognition utilizing LMMs*

The following examples show cases where objects were correctly recognized, misclassified by YOLO, or where untrained objects were present. Fig. 2 shows an example of successful object recognition, where both YOLO and LMM classified the object as an 'apple'. In such cases, the final decision is categorized as 'classified'.

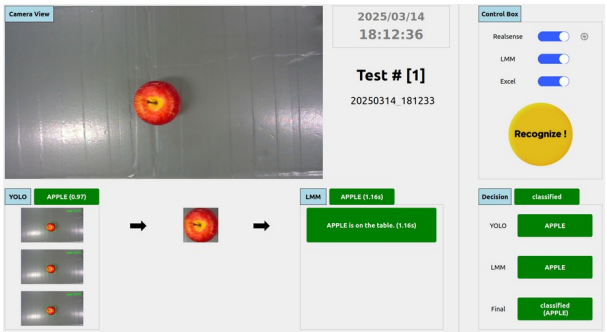


Fig. 2. Final decision: classified (APPLE recognized by both YOLO and LMM)

Fig. 3 shows an example of misclassification, where YOLO misclassified a human hand as an 'elephant'. In this case, LMM correctly identified the object as a ‘hand’. Consequently, the integrated decision from

YOLO and LMM categorized the final decision as 'unclassified'.


Fig. 3. Final decision: unclassified (YOLO misclassified as 'ELEPHANT', corrected by LMM)

Fig. 4 also shows an example of misclassification, where YOLO failed to detect the object, resulting in 'no detection' while attempting to recognize a drill. In this case, LMM correctly identified the object as a 'Drill'. Consequently, the integrated decision from YOLO and LMM categorized the final decision as 'unclassified'.

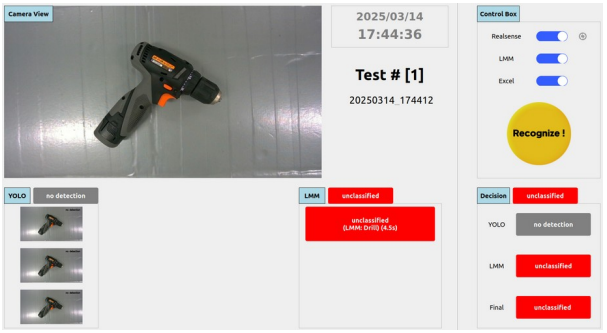


Fig. 4. Final decision: unclassified (YOLO failed to detect the object, identified by LMM as 'Drill')

Table I compares object recognition performance between using YOLO alone and the integrated YOLO and LMM approach for objects belonging to untrained classes. The experiment was conducted using the objects shown in Fig. 5, with each object tested 50 times. The results demonstrate a significant reduction in misclassifications by employing the integrated YOLO and LMM approach.

Table I: Misclassification Comparison Between YOLO and the Integrated YOLO-LMM Approach

| Actual Object | YOLO Recognition | YOLO Misclassification Rate | YOLO + LMM Misclassification Rate |
|---|---|---|---|
| Bracket | Iphone | 54% | 0% |
| Coffee Cup | Glass, Bulb | 56% | 0% |
| Glove | Grape | 24% | 4% |
| Hand | Elephant | 62% | 0% |
| Stapler | Iphone | 36% | 0% |

Fig. 5. Examples of objects from the five untrained classes used in the experiment

### *2.3 Test Results*

The integration of YOLO and LMM for object recognition clearly addressed the misclassification issues that occurred when using the custom YOLO model alone. While the custom YOLO model alone failed to recognize objects belonging to untrained classes, the integrated YOLO and LMM approach successfully identified these objects. Furthermore, the LMM effectively distinguished and filtered out misclassifications arising from objects that closely resemble those in the trained classes.

## 3. Conclusions

In this study, we proposed an integrated approach combining YOLO and LMM to achieve stable object recognition for specific target object groups. Using the integrated YOLO and LMM approach, common issues such as misclassification and reduced accuracy, often observed in custom YOLO-based object recognition, were effectively mitigated. Future research will focus on refining the YOLO-LMM model to improve object recognition performance tailored to meet specific objectives required in diverse environments.

## ACKNOWLEGEMENT

This work was supported by Robot Industry Core Technology Development Programs of the Ministry of Trade, Industry & Energy of KOREA(20018270)

## REFERENCES

[1] Im, Ki Hong, Park, Jongwon, Lee, Jinyi and Kwak, Siwoo, "Automation and Post-Processing of 3D Object Point Cloud Model Generation Using Robotic Manipulators," Proc. of the Korean Society of Mechanical Engineers Autumn Meeting, 2024, Jeju, Korea.
[2] Im, Ki Hong, Park, Jongwon, Lee, Jinyi and Nam, Yun Jun, "An Automated Generation of 3D Point Cloud Training Data for Object Recognition using Depth Cameras Mounted on Robotic Arms," Proc. of the Korean Nuclear Society Spring Meeting, 2024, Jeju, Korea.