INNO PRESS

May 2025

Journal of Emerging Applied Artificial Intelligence

Volume 1 / ISSUE 1

Issue 1 – Foundations of Emerging Applied Artificial Intelligence

The Journal of Emerging Applied AI (JEAAI) is pleased to present its inaugural issue, establishing a dedicated forum for high-quality, peer-reviewed scholarship at the intersection of artificial intelligence theory and real-world application. This first issue reflects the journal's foundational mission: to advance and disseminate research that demonstrates the transformative potential of AI technologies across sectors and disciplines.

This opening volume features contributions that exemplify the journal's emphasis on rigorously developed, practically deployed AI systems. The selected articles cover a spectrum of domains—including healthcare, robotics, transportation, education, and sustainability—demonstrating the breadth of AI's impact when translated from conceptual innovation to applied implementation.

With a commitment to methodological soundness, interdisciplinary relevance, and societal benefit, JEAAI aims to become a leading platform for scholars, practitioners, and innovators who are engaged in solving real-world problems through intelligent systems. The journal's scope encompasses original research, technical reports, case studies, and critical perspectives, all grounded in applicability and reproducibility.

We invite the academic and professional community to engage with JEAAI as contributors, reviewers, and readers, and to join us in shaping a future where applied artificial intelligence drives meaningful and responsible progress.

Editor-in-Chief

Section Editors

Dr. Meng Liu
Research Scientist, NVIDIA

Dr. Liu's research interests include graph neural networks, clustering, and multimodal learning. He has published over 20 papers in leading venues such as *Advanced Science*, *IEEE TPAMI*, *IEEE TKDE*, *CVPR*, *ICML*, and *ICLR*. His work includes an ESI Hot Paper and a Highly Cited Paper, with over 1,000 citations. He has received several awards, including Best Paper at the 2024 China Computational Power Conference and a DAAD AInet Fellowship.

Dr. Zhongbin Luo
Professor-level Senior Engineer, China Merchants Chongqing Communications Research & Design Institute. Master's Supervisor, Chongqing Jiaotong University & Shijiazhuang Tiedao University

Dr. Luo's research focuses on intelligent transportation, traffic safety, and vehicle–road collaboration. He has led over ten national and provincial research projects, holds 11 invention patents, and serves as an expert reviewer for journals such as *IEEE Access* and *PLOS ONE*.

Dr. Ruichen Xu
Postdoctoral Fellow, Department of Civil & Environmental Engineering,University of Missouri, Columbia, USA

Dr. Xu's research interests include hydrological ecology, AI-based flood forecasting, and sediment–pollutant dynamics. He has led or contributed to more than ten projects in China and the U.S. and has published over 20 peer-reviewed papers. He holds patents in environmental monitoring and serves as a reviewer for journals like *Journal of Hydrology* and *Ecological Indicators*.

A/Prof. Jinghao Yang

Assistant Professor, Electrical and Computer Engineering, The University of Texas Rio Grande Valley, USA

Dr. Yang has taught in the U.S. and specializes in applying machine learning to intelligent manufacturing systems. His research bridges intelligent sensing, control, and adaptive design with industrial applications, contributing to smart production technologies and data-driven innovation.

A/Prof. Xinyu Cai
Associate Professor, Jiaxing University

A/Prof. Xinyu Cai, Associate Professor at Jiaxing University, holds a Ph.D. in Economics and serves as a master's supervisor. He is a Certified Information Systems Auditor (CISA) and an expert with the Ministry of Education's Graduate Evaluation Program. His research focuses on human capital, employment and wage systems, and large-scale AI applications in sustainable development. A/Prof. Cai has led major national and provincial research projects and published over 20 academic papers, including in Nature sub-journals and top Chinese core journals. He has received multiple research awards and serves on national academic committees related to AI and human resources.

Yihan Zhao
PhD Candidate, University of Auckland, New Zealand

Yihan Zhao holds a Master's degree from Peking University and is currently a PhD candidate at the University of Auckland. Her research explores the intersection of communication, culture, and technology, with a focus on how algorithms reshape cultural expression and the subjectivity of marginalized communities. She previously served as an Assistant Research Fellow at the Development Research Centre of the State Council in China, contributing to national research projects. She has curated and coordinated panels for the China Development Forum, facilitating high-level dialogue on AI, sustainability, and governance.

Shen (Jason) Zhan
Graduate Researcher, University of Melbourne, Australia

Jason Zhan holds an Honours degree in Civil and Environmental Engineering from the University of Auckland and is currently a PhD researcher in the Teaching & Learning Lab at the University of Melbourne. He combines industry and academic experience, with a background in structural engineering and teaching. His research focuses on employability assessment and curriculum design in engineering education, with growing interest in the role of AI in authentic assessment and personalized learning.

Contents

Extraction and Recognition of Robotic Apple Picking Image Features Based on YOLOv5 Detection Models

Wenqian Hong[1], Hao Wu[2], and Yirong Jiang[2*]

[1]School of Mathematics and Statistics, Guilin University of Technology, Guilin, Guangxi 541004, P.R. China

[2]School of Mathematics and Sciences, Guangxi Minzu University, Nanning, Guangxi, 530006, P.R.China

*Corresponding author: 20240021@gxmzu.edu.cn

Abstract

As China has emerged as one of the leading exporters of apples globally, the shortage of agricultural labor has posed a significant challenge for the apple industry's growth. To address the issue of image recognition for robotic apple picking in complex orchard settings, this paper combines computerized image processing and deep learning concepts to propose a detection model based on YOLOv5. By implementing image preprocessing techniques and optimizing the loss function, the study successfully achieves accurate extraction and recognition of apple image features. The experimental findings demonstrate the high performance and accuracy of the proposed method in apple picking tasks, offering valuable support for the advancement of robotic automated picking systems. Future research endeavors will focus on further refining the algorithm to enhance efficiency in real-world production settings. The improved OLOv5 Detection Models proposed in this article can be applied in fields such as industrial detection, intelligent traffic signal control, and sports events.

Index Terms—YOLOv5 detection model, digital image processing, apple feature extraction and recognition, labelImg

1 Introduction

China, as the world's largest exporter of apples, has most local farmers growing apples in their own orchards. During the apple picking season, a significant number of workers are required to harvest ripe apples. The rapid urbanization in China, along with an aging agricultural workforce and a large portion of young individuals seeking work opportunities elsewhere, has resulted in labor shortages during this crucial season[1]. To address this issue, China has been developing apple-picking robots since 2011, making notable advancements. However, existing robots often struggle to accurately identify various obstacles in the orchard environment, such as 'leaf shading', 'branch shading', 'fruit shading', and 'mixed shading'. Without precise judgment based on the actual conditions, harvesting directly can lead to significant damage to the fruits, as well as potential harm to the robotic arm and the workers. This can negatively impact both harvesting efficiency and fruit quality, resulting in substantial losses. Furthermore, the accurate identification and classification of harvested fruits is crucial for subsequent sorting, processing, packaging, and transportation. Moreover, the similarity in color, shape, and size between apples and other fruits poses a challenge for apple-picking robots in accurately distinguishing between them.

At present, many scholars have studied this problem. Relevant studies mainly include:Wang Dandan etal[2] further analyzed the problems in the vision system of apple picking robot to provide reference for the in-depth study of the vision system of apple picking robot.Ka-pach etal[3] investigated the apple color detection method, but the algorithm detection effect is not ideal for immature apples or the situation of having branches and leaves blocking, and apples are similar to the background, and so on. Cao Chunqing etal [4] realized accurate recognition and 3D localization of apples in multiple natural scenes by fusing YOLOv3 and binocular vision algorithms. Zhao De'an etal [5] proposed a YOLO deep convolutional neural network-based localization method for robotic apple picking in complex backgrounds, using optimized YOLOv3 deep convolutional neural network to locate apples, and achieved apple recognition and localization in complex environments. Cao Zhipeng etal [6] used YOLOv4 neural network can recognize apples better, but the recognition speed of YOLOv4 is low, which can't meet the demand of real-time picking. The above methods perform well in recognizing apple targets, but they require high computational resources.

From the above research, we found that these methods are difficult to achieve fast and accurate identification and localization. It will be interesting to study what happens to a method if it not only enables quick recognition but also ensures precise localization. The proposed approach involves developing an apple image recognition model using the depth-separable convolutional YOLOv5 model, with optimized loss functions to enhance speed and accuracy in recognizing apples in complex environments. This advancement further facilitates the practical implementation of domestic apple picking robots. By analyzing labeled apple images and extracting relevant features, a high recognition rate, speed, and accuracy are achieved. Moreover, data analysis of the images enables automatic calculation of the number, location, maturity, and estimated quality of apples, thereby improving fruit recognition rates. The results obtained exhibit high precision and recall

rates of 99.08% and 99.3%, respectively. This research holds significant implications for the advancement of robotic apple picking technology in China.

2 Image Preprocessing

2.1 Description of the experimental dataset

To ensure the complexity of the apple image data, this paper uses the dataset of the 13th APMCM Asia-Pacific Regional Undergraduate Mathematical Modeling Competition problemA, 2023.There are three subsets of this dataset and they are: subset 1, subset 2 and subset 3. The basics are as follows:

Subset 1 is a ripe apple image dataset containing 200 images of ripe apples, each with a size of 270 * 180 pixels. Some screenshots of Subset 1 are shown in Fig.1:

Figure 1: Mature apple image datasets

Subset 2 is a fruit picking image dataset containing 20705 images of different picking fruits, each with known labels and classifications, with a size of 270 * 180 pixels. Some of the screenshots in Subset 2 are as requested in Fig.2, which includes apples, cactus fruits, pears, plums and tomatoes.

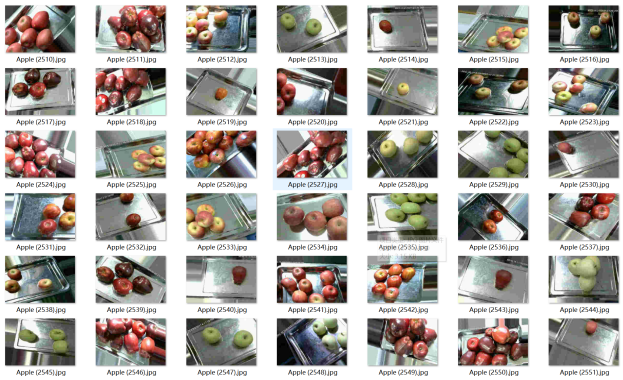

Figure 2: Fruit picking image dataset

Subset 3 shows the labeled dataset containing 20705 images of different picked fruits, each with a size of 270 * 180 pixels, but with unknown labels and classifications. Some screenshots of Subset 3 are shown in Fig.3:

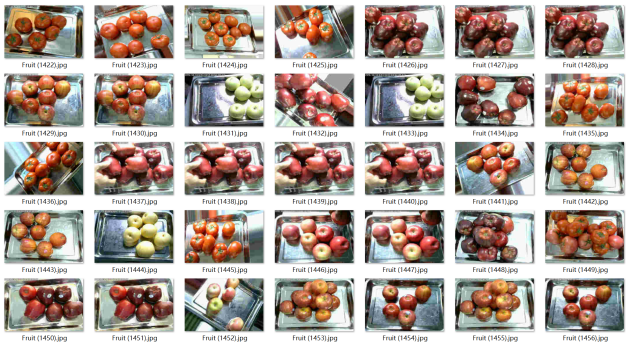

Figure 3: Tagged data sets

2.2 Original Image Information

In the apple images given in Subset 1, there are a total of 200 images, all of which have pixels of 270 x 185. The image file was taken at basically the same time and with sufficient light. However, the images are divided into four parts; a portion of the ripe red apples have problems such as part of the image being blurred, overexposed or underexposed, and shadows being blocked by leafy branches; a portion of the apples have people blocking them; a portion of the images are images of immature green apples or blossom bones; and a portion of the images are of other fruits. Some of the images are shown in Fig.4:

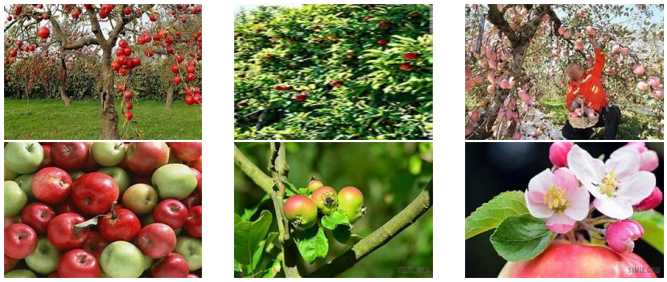

Figure 4: Example of an image from part of the subset

The above types of images will affect the feature extraction of apples, so it is necessary to pre-process the apple image in Subset 1 with image denoising, image enhancement, color space conversion, etc. to enhance the contrast of the image.

2.3 Image preprocessing methods

The steps of image preprocessing are shown in Fig.5:

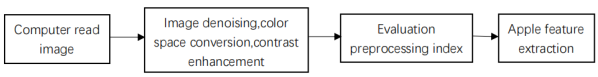

Figure 5: Steps in image preprocessing

Among them, we use median filtering method, inverse sharpening mask method of image denoising, enhancement, to

avoid the image blurring and other problems to interfere with the later for the color, shape and other features of the apple extraction.

2.4 Rating system

In order to judge the advantages and disadvantages of the preprocessed images, standard deviation standrad and structural similarity SSIM are selected as the evaluation indexes of the preprocessing results. SSIM index is mainly from the brightness, contrast and structure of three aspects to measure the degree of similarity between the two images, the value range is [0,1]. The SSIM index mainly measures the degree of similarity between two images from three aspects: brightness, contrast and structure, and the value range is [0,1], the larger the SSIM value is, the more similar the structure of the two images is. The calculation formula is shown in equation (2.1):

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (1)$$

where x and y denote the enhanced and real images, respectively, μ denotes the pixel mean of the image, and σ is the variance of the image. c_1 and c_2 are 0.0001 and 0.0009, respectively.

The standard deviation measures the degree of variation in the pixel values of an image and assesses the contrast and sharpness of the image. Here pixel mean can be used to measure the overall brightness of an image, the higher the pixel mean, the sharper the image. Suppose that there is a data set of $X_i, \quad i \in \{1, 2, \ldots, n\}$ The standard deviation of this data set is:

$$std = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2}. \quad (2)$$

2.5 Pre-processing results

(i) Image denoising

Upon comparing a portion of the original image with the denoised image as illustrated in Fig.6:, it is evident that the denoised apple image exhibits greater clarity than the original image. Additionally, the shape contour is more distinctly separated from the background in the denoised image.

(ii) Image Enhancement Processing

It can be seen in Fig.7that the difference between the apple and the background after image enhancement is obvious, and at the same time, it can better retain the detailed information in the original image to generate a higher quality image, and does not appear more serious distortion.

(iii) Image conversion to RGB format

As can be seen from Fig.8, there is little difference between the RGB-converted image and the original image in terms of sharpness and contrast.

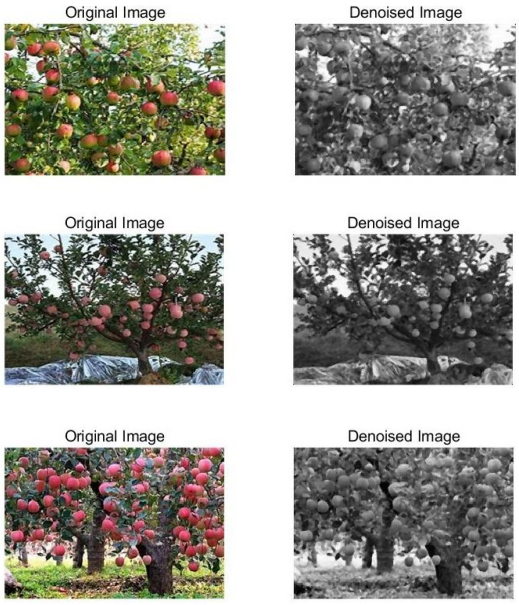

Figure 6: Image denoising

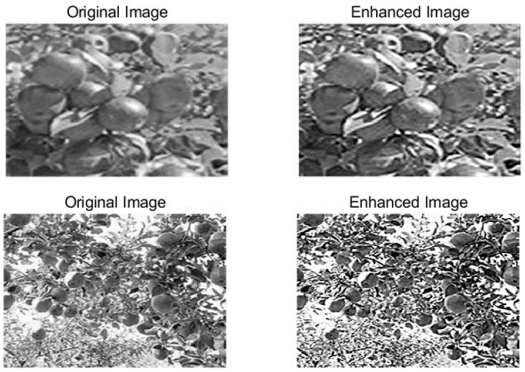

Figure 7: Comparison graph after image enhancement process

Figure 8: Comparison of original image and RGB image

2.6 Image Processing Evaluation

In order to quantitatively analyze the three preprocesses, the SSIM values of each of the three preprocesses were calculated during the experiments to evaluate the enhancement of the im-

ages. The experimental results are shown in Table 1, where the data are selected from the dataset in Subset 1.

Table 1: SSIM metrics for three preprocessing methods

Preprocessing methods	SSIM
Denoising	0.5143
Image enhancement	0.83345
RGB	0.13506

Table 1 lists the results of structural similarity calculations for different treatments, from which it can be seen that the quality of the generated images is improved by using the three preprocessing to enhance the apple images. In terms of SSIM metrics, image enhancement improves 0.319 and 0.698 compared to the two models of direct grayscale processing and image denoising, indicating that the images generated by the image enhancement process are less affected by noise, and the visual effect and brightness of the images are significantly improved.

Table 2: Standard deviation comparison

Preprocessing methods	Average ixels values of the original image	Average pixel value of processed image
Denoising		112.39
Image enhancement	106.34	112.84
RGB		110.34

Table 2 lists the standard deviation results of the different treatments, compared with the original image pixel mean value of 106.34, all have obvious improvement, in which the image denoising and image enhancement in the pixel mean value of the difference of only 0.45, it shows that the image resolution of these two processing methods is higher. After comparison and contrast above, we finally chose the image after image enhancement processing in improving the image brightness, contrast and clarity is better, on the basis of which the apple feature extraction operation is carried out.

2.7 Apple feature extraction based on image processing

(i) Apple circumference

Perimeter is corrected by counting the number of pixels on an object's contour line, in the oblique direction, which produces errors specific to digitized images, by twice their number. When scanning the image from left to right and from bottom to top, a pixel value of 1 is found and its neighboring pixel values (8 neighborhoods) have a different value from it, the apple counter is added to 1, and the entire image is scanned to get the perimeter of the apple.Some of the results are shown in Fig.9.

(ii) Apple color

In this problem, color extraction is done on the basis of image enhancement, so the color features of apples are extracted from the grayscale histograms of the images, and since there

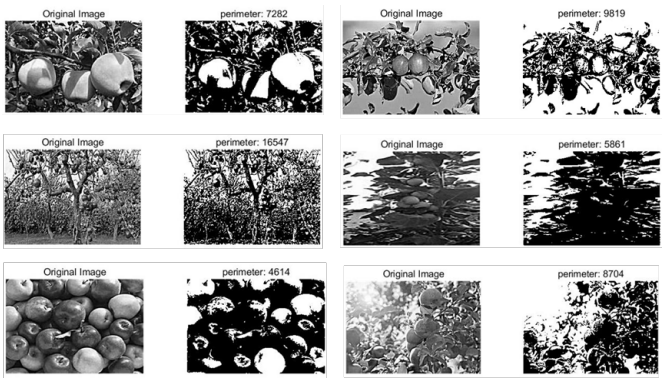

Figure 9: Perimeter features extracted after partial image enhancement

are 200 images in Subset 1, the grayscale values of each image are stored in a matrix. The feature matrix featureMatrix with size (numImages, 256) where numImages is the number of grayscale images in the image folder. This feature matrix holds the normalized grayscale histogram features of each grayscale image. It is shown in Fig.10:

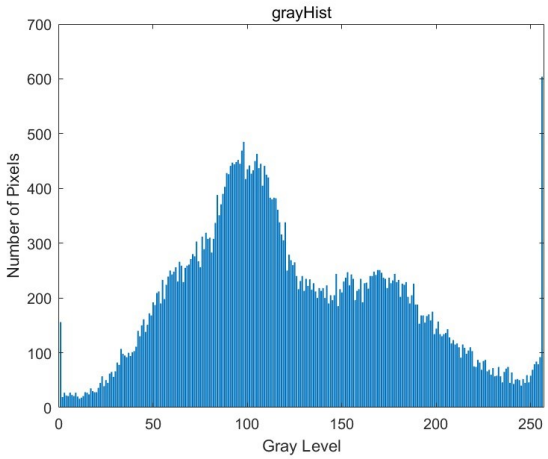

Figure 10: Grayscale histogram

As can be seen from the figure, the gray level of the apple is concentrated in the [50,150] interval and the number of pixels is in the [0,500] interval.

3 Fundamental model

3.1 A counting model for labeling apples based on the labelImg

Before using the YOLOv5 model to detect the position of the apple, we first need to label the position of the apple in the image, the labeling tool used here is labelImg, according to the title of the apple image information given by the use of the edge of the box is labeled, because the image is blurred or foliage obscured by the apple using a manual labeling method,

to improve the accuracy of the number of apples, the position, and to save the information of the labeling.

The data labeling steps are shown in Fig.11: The labeled

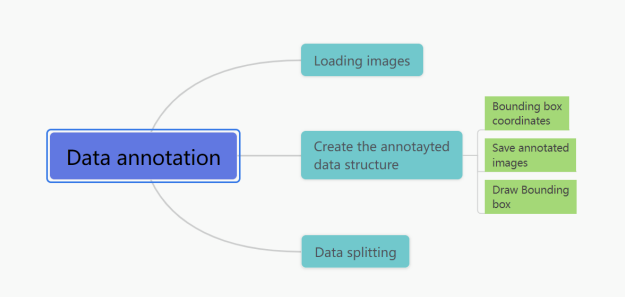

Figure 11: Data annotation content

results are shown in Table 3

Table 3: Number of apples in each image

Image number	Number of apples
1	4
2	7
3	15
4	15
...	...
196	21
197	11
198	52
199	2
200	21
Total number	2921

Note: The parts labeled in red indicate the same number of apples in the image.

3.2 Apple position detection model based on YOLOv5

(i) Model theory

Determination of apple location requires a target detection method, and in order to regressively predict the category and location information of the target object, we use the YOLOv5 target detection model.YOLOv5 uses an end-to-end mechanism to normalize the image and input it into a convolutional neural network. The network structure is mainly composed of four parts: the input, the feature extraction network, the Neck part and the prediction layer. The model structure is shown in Fig.12.

In this, the input side preprocesses the dataset; the feature extraction network performs the slicing operation on the image to achieve downsampling of the image without loss of information; Neck fuses the features of different latitudes; the prediction layer uses CloUzuo as the bounding box loss function, and the non-maximum value suppression algorithm filters the detected target frames. Through the above work, the best result of detecting the apple location is obtained.

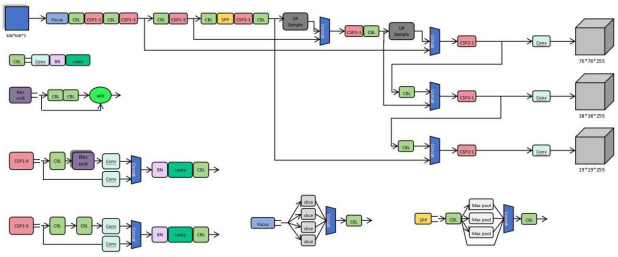

Figure 12: YOLOv5 Object Detection Model Structure

Considering the problems such as foliage occlusion, an ECA attention module is added after the CSP structure of each branch of the YOLOv5s neck network respectively, and the feature information extracted by the feature extraction network is used to perform adaptive learning of features in the spatial dimension to strengthen the fusion ability of features.The feature expression ability of the model in complex scenes is enhanced by adding the attention mechanism module to the neck network, which The interference of irrelevant information is suppressed, so that the model has better detection results in detecting the occluded apple target.

The geometric position of the apple is determined: the center of mass position of the bounding box is measured with the bounding box, and the center of mass position formula is as follows:

$$(x, y) = \left(\frac{\sum x_i}{N}, \frac{\sum y_i}{N} \right), \tag{3}$$

where, x_i, y_i are the coordinates of the pixels in the apple region and N is the total number of pixels in the region.

(ii) Experimental steps

Before using the YOLOv5 model to detect the apple position, we first have to label the position of the apple in the image using the labelImg tool. We use Subset 1 for training and classified Subset 2 for model testing and inspection. The steps are as follows:

Step1: Adding four folders in the data of YOLOv5 directory, Annotations folder is used to store xml files after labeling each image with labelImg; Images folder is used to store the original dataset images that need to be trained in jpg format; ImageSets folder is used to store files used for training, validation, and testing after the dataset has been divided into ImageSets folder is used to store the data set divided into files for training, validation and testing; Labels folder is used to store the labeled files in txt format after converting the labeled files in xml format;

Step2: Preparing the dataset, here we need the geometric position of the apple, so we use all the data in Subset 1 for training, validation and testing;

Step3: Organizing the results of the data obtained from the training and draw a 2D scatter plot of the geometric location;

Step4: Taking the obtained geometric position coordinates and calculate the area of the edge box, which is equivalent to the 2D area of the apple.

(iii) Evaluation system

In this paper, Precision (P) and Recall (R) are used as evaluation metrics to test the model performance. The evaluation metrics are calculated by the formulas respectively:

$$lP = \frac{TP}{TP + FP} \tag{4}$$

$$lR = \frac{TP}{TP + FN} \tag{5}$$

Where TP denotes the number of samples predicted by the model to be positive and are positive samples, FP denotes the number of samples predicted by the model to be positive but are negative samples, and FN denotes the number of samples predicted by the model to be positive but are negative samples.

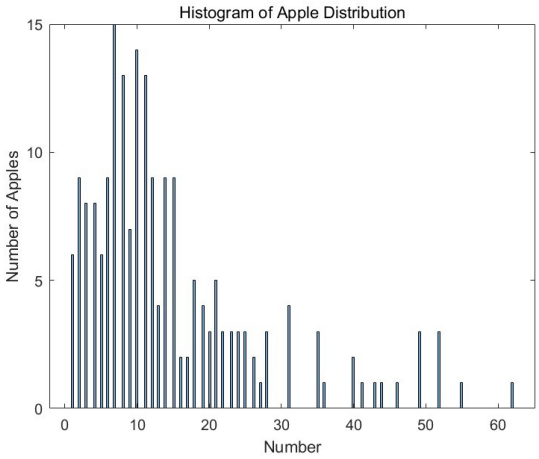

Figure 13: Histogram of Apple Distribution

4 Model prediction results

4.1 Calculate the number of apples

According to the apple image features, we use matlab digital image processing to extract the color and perimeter features of the apple; however, considering that some images have low pixels, which results in the apple features not being obvious in some pictures, the image is first preprocessed before feature extraction, such as adjusting the brightness and contrast to make the apple color and edges in the dark distinctly recognizable, applying filters to remove the noise and performing apply filters to remove noise, color space conversion, and image enhancement to better distinguish apples and background, etc.; after preprocessing, the image is then extracted with its features.

For the calculation of the number of apples, we used labbelmg software to label the data set in Subset 1 according to the given labels, and some apples that were not clear due to the image were labeled and counted manually to get the number of apples. And keeping the number of apples in each image in the labeled dataset, the distribution histogram of all apples was plotted using matlab.

The labeling information is organized into an Excle table to summarize, and then matlab plots the number of apples in each image in Subset 1. The results are shown in Fig.13:

4.2 Estimated Apple Location

The minimum bounding box of the apple is plotted by labbelmg to get the coordinates of the center position of the labeled apple and the length and width size of the labeled bounding box to determine the geometric position of the apple, and the obtained positional data information of the apple is normalized so that the geometric coordinates of the apple can be determined quickly in the next problem of establishing the coordinates to plot the positional information of the apple, and after obtaining the geometric coordinates of the apple, its two-dimensional scatterplot is plotted in matlab to plot its 2D scatter plot. Geometric coordinates of apples Some of the geometric coordinates of apples detected by the YOLOv5 model are shown in Table 4:

Table 4: Geometric coordinates of some apples

Image number	Barycentric coordinate	Bounding box length	Bounding box width
40	(0.890741,0.381081)	0.085185	0.059459
	(0.875926,0.462162)	0.100000	0.059459
	(0.337037,0.681081)	0.074074	0.064865
	(0.246296,0.616216)	0.085185	0.064865
	(0.509259,0.591892)	0.070370	0.059459
	(0.748148,0.275676)	0.066667	0.054054
	(0.753704,0.143243)	0.048148	0.059459
	(0.175926,0.391892)	0.070370	0.048649

The obtained data of bounding box dimensions and center of mass position are normalized and a scatter plot is drawn with the lower left corner as the coordinate origin. The 2D scatter plot of the apple geometric coordinates is shown in Fig.14:

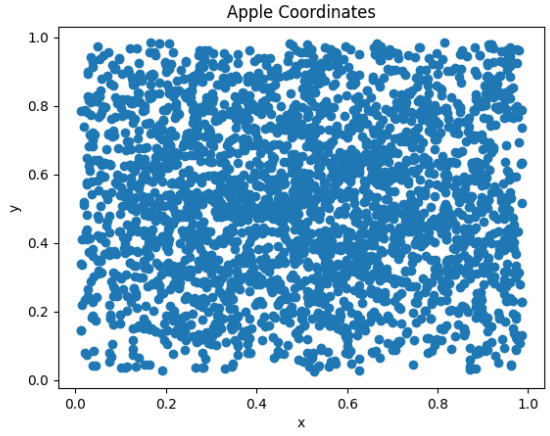

Figure 14: YOLOv5 Object Detection Model Structure

4.3 Estimated quality of apples

The relative area is used to estimate the mass of each apple, in which, for the two-dimensional area of apples, we use the area of the edge frame after the edge detection marking, and the information of the edge frame length and width data obtained to estimate the two-dimensional area of each apple, and the apple mass is estimated according to the average density of apples and the estimation formula. At the same time, taking into account the estimation error impact of the difference between the edge frame of the calculated area being rectangular and the oblate shape of the apple, we provide a certain confidence interval and error range for the estimation results, and produce a histogram of the mass distribution.

According to the mass formula, theoretically, the larger the area, the greater the mass of the apple. Therefore, here we use the ratio of area to mass to solve for the area of the relative image, i.e. the relative area of the apple, to estimate the mass of the apple.

The following figure shows the histogram of apple mass distribution.

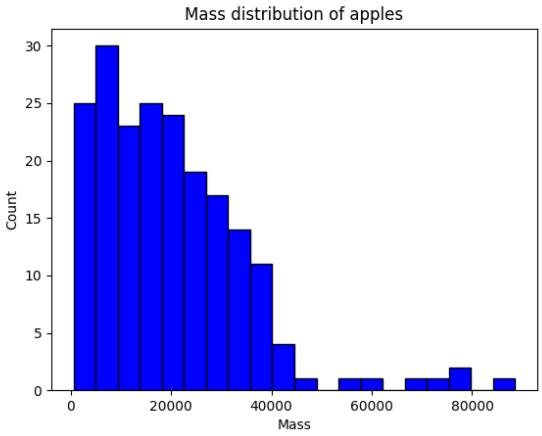

Figure 15: Histogram of apple ripeness distribution

As can be seen from the Fig.15, the apple quality is mainly concentrated in the [0, 2000] interval, and a small portion is distributed in the [60,000, 80,000] interval. The quality and size of the apple images in Subset 1 are basically the same.

4.4 Estimating apple ripeness

Before judging the estimated maturity of apples, we first determine the expression of apple maturity, according to the literature and common sense in daily life, we know that the color, size and texture can show the maturity state of apples, for example, lime green apples are immature and red apples are usually in a ripe state, by using these factors to determine the maturity of apples and coding the different maturity levels, and deepening the prominence of the image features to classify the ripeness of apples, again, we use YOLOv5 detection model for classification.

The maturity of apples is related to many factors and belongs to the multi-classification problem, which is usually categorized into different maturity levels, divided into four levels, and coded in the data annotation. The coding is defined as follows:

Table 5: Apple Data

Coding	Apple color	Tag number	Grade of maturity
1	All-red	15	mature
2	All green	17	Immature
3	Half red and half green	16	Semi-mature
4	(flower)bud	19	Extremely immature

Based on this criterion, the classification results are shown in Fig.16:

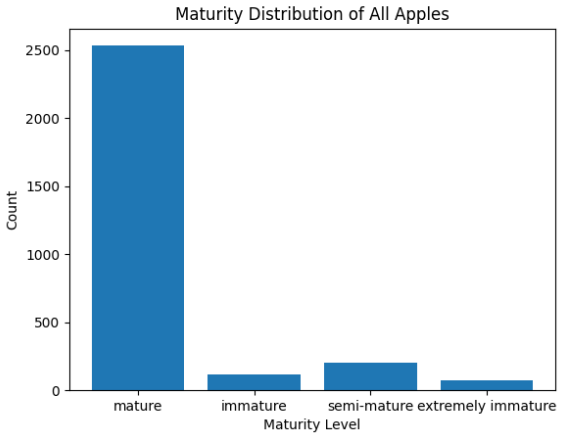

Figure 16: Histogram of apple ripeness distribution

As can be seen from the graph, the largest percentage of fully red apples, over 2,500, indicates a majority of fully ripe apples; followed by semi-ripe apples with the fewest flower bones.

4.5 Apple Recognition Based on YOLOv5 Models

We divide the classified labeled Subset 2 data as the training set and use Subset 3 as the test set to train the YOLOv5 detection model to achieve apple detection in Subset 3 data. The obtained results are shown below in Fig.17:

From the figure, it can be seen that [0,4140] has the highest number of ID number apples, close to 1400, and [8280-10350] has the lowest number of ID numbers, only about 800, with precision and recall rates of 99.08% and 99.3%.

5 Conclusions

(i) In feature extraction, the image quality is improved by preprocessing steps such as image denoising, enhancement and color space conversion, which makes the sub-

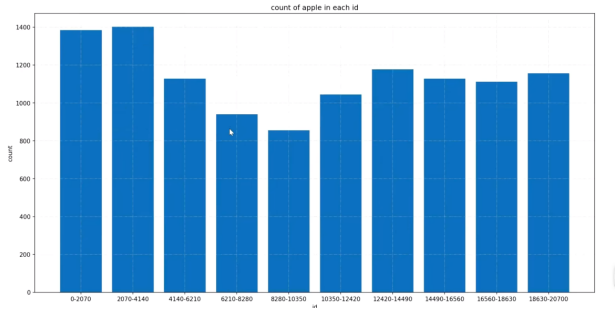

Figure 17: Histogram of apple ripeness distribution

sequent feature extraction and target detection more accurate and reliable.

(ii) Based on the detection model of YOLOv5, enhanced with an attention mechanism module, successfully extracts and recognizes features from robot apple picking images in complex scenes. This enhancement improves the model's feature expression ability, extraction capability of apple image features, and detection speed, enabling accurate detection of apple locations, numbers, maturity, and size.

(iii) The experimental results show that our proposed method has good performance and accuracy in apple picking tasks, which provides strong support for the development of robotic automated picking systems. Future research directions can further optimize the algorithm and improve the accuracy of detection and recognition to meet the needs of more efficient in real production.

Acknowledgment

The authors are grateful to the editor and the referees for their valuable comments and suggestions. This work was supported by College Student Innovation and Entrepreneurship Training Program Project (S202410608112).

Data availability statement

The data that support the findings of this study are available from Asia-Pacific Student Mathematical Modeling Contest Organizing Committee, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors(corresponding author:229744332@qq.com) upon reasonable request and with permission of Asia-Pacific Student Mathematical Modeling Contest Organizing Committee.

Ethical Approval

Not applicable.

Competing interests

There is no conflict of interests.

Authors' contributions

The authors contributed equally to this paper.

References

[1] SONG Zhe, WANG Hong, LILI Hui, et al,"Main problems, development trend and solutions of apple industry in China," Jiangsu Agricultural Science,vol.44,no.9,pp.4-8,2016.

[2] WANG Dandan, SONG Huaibo, HE Dongjian,"Research progress of apple picking robot vision system," Journal of Agricultural Engineering,vol.33,no.10,pp.59-69,2017.

[3] KAPACHK,BARNEAE,MAIRONR,etal,"Computervision for fruit harvesting robots-state of the art and challengesahead," International journal of computational vision androbotics,vol.3,no.2,pp.4-34,2018.

[4] Cao Chunqing, Zhang Wuping, Li Fuzhong, et al,"Research on fusion algorithm for multi-target apple recognition and localization in natural scenes," Hubei Agricultural Science,vol.61,no.7,pp.145-151,2022.

[5] ZHAO De'an, WU Rendi, LIU Xiaoyang, et al,"Robotic apple picking localization in complex context based on YOLO deep convolutional neural network," Journal of Agricultural Engineering,vol.35,no.3,pp.164-173,2019.

[6] Cao ZP,"Research on actuator and target detection of apple picking robot," Kunming University of Science and Technology,2023.

[7] Wang Yong, Tao Zhaosheng, Shi Xinyu, et al,"Target detection method of different maturity apples based on improved YOLOv5s,"Journal of Nanjing Agricultural University,2023.

[8] ZHANG Shifu,"Research on apple target recognition and localization algorithm based on deep learning," Zhejiang University of Technology,2020.

[9] Song Yang,"Research on image enhancement and apple detection method based on deep learning," South-Central University for Nationalities,2022.

[10] ZHANG Tao; LI Zhisheng,"Apple object detection based on BG and RTHTR image processing," Electronic Design Engineering,vol.31,no.10,pp.135-140,2023.

[11] HU Shilin, CHEN Wei, ZHANG Jingfeng, et al,"Target detection method of apple picking robot based on improved YOLO v5," Agricultural Mechanization Research,2024.

Saliency-Driven Multi-Scale Feature Discrepancy Fusion for Fine-Grained Video Anomaly Detection

Xukui Qin
Department of Computer Science, The George Washington University, Washington, United States
*Corresponding author: kuschqin@gmail.com

Abstract

Video Anomaly Detection (VAD), a critical task in intelligent surveillance systems, plays a vital role in public safety, traffic management, and emergency response. However, detecting small-scale and transient anomalies in complex scenes remains a significant challenge due to the scarcity of anomaly samples and the difficulty in capturing fine-grained features. To address these issues, this paper proposes a novel dynamic feature enhancement framework built upon the Masked Autoencoder (MAE) architecture. At the core of the proposed framework is the Multi-Scale Discrepancy Saliency Fusion (MDSF) module, which explicitly models and dynamically amplifies channel-wise feature discrepancies between teacher and student networks, thereby enhancing the saliency of anomalous regions. Furthermore, MDSF integrates multi-scale semantic features through a saliency-guided fusion strategy, enabling the model to effectively capture anomalies across varying spatial and temporal resolutions. The proposed method is trained in an end-to-end manner without requiring pre-trained weights and is evaluated on standard benchmark datasets, including UCSD Ped2, Avenue, and ShanghaiTech. Experimental results demonstrate that the proposed MDSF module significantly improves detection accuracy while maintaining low computational complexity, highlighting its practical value and strong generalization capabilities for real-world video anomaly detection tasks.

Index Terms— Video Anomaly Detection, Masked Autoencoder, Feature Enhancement, Multi-Scale Fusion, Distillation, Attention.

1 Introduction

With the rapid advancement of deep learning techniques [1, 2, 32, 14, 26, 10, 11], video anomaly detection (VAD) has emerged as a critical component in intelligent surveillance systems, playing a pivotal role in ensuring public safety, managing traffic flow, and enabling efficient emergency response. These systems are increasingly deployed in complex and dynamic environments, such as urban traffic networks, public venues, and critical infrastructure, where the timely identification of abnormal events is essential. Despite the remarkable progress achieved in VAD, existing methods often struggle to accurately capture the subtle, fine-grained features of anomalies, especially those occurring at small scales or within highly cluttered and dynamic backgrounds. This limitation is further exacerbated by the scarcity and diversity of anomalous samples in real-world data, which hampers model generalization and limits their robustness in practical scenarios [20, 9, 25].

In recent years, self-supervised learning frameworks based on the Masked Auto-Encoder (MAE) architecture have demonstrated considerable promise for VAD tasks [21, 18]. MAE models are typically trained by reconstructing masked regions of normal video samples, enabling the network to learn the spatiotemporal patterns of normal events without requiring explicit anomaly annotations. At the testing stage, anomalies—due to their deviation from the learned normal feature distribution—tend to induce higher reconstruction errors, thereby facilitating indirect anomaly detection. This paradigm, often referred to as "reconstruction error-based anomaly detection", has achieved widespread adoption; however, it still faces several fundamental limitations. First, real-world anomaly events often involve challenges such as illumination variations, motion blur, and occlusions, which can corrupt the normal feature learning process, leading to unstable reconstruction errors. Second, global reconstruction objectives are susceptible to background noise and dynamic scene variations, reducing the saliency of localized anomaly signals. Third, conventional MAE-based approaches fail to fully exploit the rich feature discrepancy information between teacher and student networks, resulting in limited sensitivity to subtle anomalies and suboptimal generalization in complex scenes.

To overcome these challenges, this paper proposes a novel module named Multi-Scale Discrepancy Saliency Fusion (MDSF), built upon the MAE architecture. The core innovation of MDSF lies in explicitly modeling and dynamically amplifying the channel-wise feature discrepancy between the teacher and student networks, allowing the model to highlight abnormal regions where reconstruction errors manifest. Furthermore, MDSF integrates multi-scale semantic features through a saliency-guided fusion strategy, enabling the model to capture fine-grained anomalies across different spatial and temporal resolutions. This design not only enhances the model's sensitivity to small-scale and transient anomalies but also mitigates the interference caused by background clutter. The proposed method is evaluated on benchmark datasets such as UCSD Ped2, Avenue, and ShanghaiTech, where it

demonstrates significant improvements in detection accuracy while maintaining low computational complexity, highlighting its potential for practical deployment in real-world intelligent surveillance systems.

The main contributions of this paper are as follows:

- We propose a novel Multi-Scale Discrepancy Saliency Fusion (MDSF) module based on the Masked Auto-Encoder (MAE) framework, which explicitly models and dynamically amplifies the channel-wise feature discrepancy between the teacher and student networks. This design significantly enhances the saliency of anomalous regions and improves the model's sensitivity to fine-grained anomalies.
- A multi-scale saliency-guided fusion strategy is introduced within MDSF, enabling the integration of hierarchical semantic features from shallow to deep layers. This approach facilitates the detection of small-scale, spatially localized anomalies and improves the model's robustness against background noise and dynamic scene variations.
- Extensive experiments on benchmark datasets (UCSD Ped2 [12], Avenue [15], and ShanghaiTech [16]) demonstrate that the proposed MDSF module achieves superior detection accuracy compared to existing methods, while maintaining low computational complexity. This confirms the effectiveness and practical potential of our approach for real-world video anomaly detection tasks.

2 Related Works

2.1 Video Anomaly Detection

Deep learning has significantly advanced video anomaly detection (VAD), enabling end-to-end spatiotemporal modeling from raw video data. Existing methods can be categorized into supervised, weakly-supervised, and unsupervised paradigms.

Supervised methods formulate VAD as a classification task using precisely annotated datasets [7, 4]. While achieving high accuracy, they are heavily dependent on costly frame-level annotations and lack generalization to unseen anomalies [22, 6].

Weakly-supervised methods use video-level labels and multi-instance learning (MIL) frameworks to reduce annotation costs [27, 29]. However, they struggle to capture fine-grained spatiotemporal features, limiting their sensitivity in complex scenes.

Unsupervised methods, which train solely on normal data without requiring anomaly annotations, have gained increasing attention due to their scalability and adaptability. Reconstruction-based models [5, 24]learn normal patterns and detect anomalies by identifying reconstruction errors, while prediction-based method [28] rely on temporal consistency. Hybrid models [17] combine both strategies for improved robustness. Recent works have explored discrepancy modeling between teacher-student networks [23], highlighting its potential for anomaly detection.

Despite these advances, unsupervised methods face challenges, including background noise interference and limited sensitivity to small-scale anomalies. Nonetheless, compared to supervised or weakly-supervised approaches, unsupervised learning is better suited for real-world VAD scenarios, where anomalies are rare, diverse, and costly to annotate.

Building on this, we propose a novel Multi-Scale Discrepancy Saliency Fusion (MDSF) module within the MAE framework, which explicitly models feature discrepancies and integrates multi-scale semantic information, thereby enhancing fine-grained anomaly detection in complex video scenes.

2.2 Attention Mechanisms in Computer Vision

The attention mechanism has become an essential component in modern computer vision systems, enabling models to dynamically focus on salient regions within input data. By adaptively reweighting spatial and channel-wise features, attention modules enhance the representational capacity of neural networks, improving performance across various tasks such as image classification, object detection, and semantic segmentation. One of the seminal works in this area is the Squeeze-and-Excitation (SE) block proposed by Hu et al. [8], which introduced channel attention by modeling inter-channel dependencies and recalibrating feature responses, leading to significant improvements in classification tasks. Building upon this, the Non-Local Neural Network by Wang et al. [30] pioneered the modeling of long-range dependencies through self-attention mechanisms, enabling networks to capture global contextual information across distant spatial locations. Furthermore, the Convolutional Block Attention Module (CBAM) proposed by Woo et al. [31] extended attention modeling to both channel and spatial dimensions, demonstrating superior performance in a wide range of vision tasks.

These advances have been widely adopted in diverse application scenarios [13, 3]. These works underscore the versatility and efficacy of attention mechanisms in computer vision, inspiring further exploration in designing robust, lightweight, and scalable attention modules for complex visual tasks. Building upon these insights, our work leverages the attention paradigm within the Multi-Scale Discrepancy Saliency Fusion (MDSF) module to enhance fine-grained anomaly detection in video surveillance. Specifically, we model the channel-wise feature discrepancies between the teacher and student networks as attention signals and dynamically amplify these differences across multiple spatial scales. This design allows the model to selectively highlight subtle, spatially localized anomalies while suppressing background noise, addressing key limitations in existing unsupervised anomaly detection frameworks.

3 Methodology

3.1 Overall Architecture

In this section, we introduce the proposed Multi-Scale Discrepancy Saliency Fusion (MDSF) module integrated within

the Masked Autoencoder (MAE) framework, designed specifically to address critical limitations of existing unsupervised video anomaly detection methods. The proposed framework consists of three main components: (1) a Teacher-Student network for feature extraction and reconstruction, (2) the MDSF module for dynamic discrepancy amplification and multi-scale fusion, and (3) an anomaly scoring mechanism.

The central motivation behind MDSF is to explicitly measure and dynamically enhance the channel-wise discrepancy between the teacher and student network features, thereby highlighting regions exhibiting high reconstruction errors indicative of anomalies. Additionally, MDSF incorporates multi-scale semantic feature fusion guided by saliency maps, enabling the detection of subtle anomalies while effectively suppressing background noise.

3.2 Teacher-Student Network Feature Encoding and Reconstruction

The Teacher-Student structure in our model leverages the reconstruction capabilities of a robust teacher network to guide a relatively lightweight student network. Specifically, given an input video frame I_t, both networks produce encoded feature representations through their respective encoder operations, which are defined as follows:

$$F_t^{teach} = Enc_{teacher}(I_t), \tag{1}$$

$$F_t^{stud} = Enc_{student}(I_t). \tag{2}$$

These features are then decoded separately by their corresponding decoders, aiming to reconstruct the original input frame:

$$\hat{I}_t^{teach} = Dec_{teacher}(F_t^{teach}), \tag{3}$$

$$\hat{I}_t^{stud} = Dec_{student}(F_t^{stud}). \tag{4}$$

Ideally, the student network closely reconstructs the input under normal conditions but deviates significantly from the teacher network reconstruction when anomalies occur, thus creating feature discrepancies that our module aims to amplify. This discrepancy implicitly contains crucial anomaly cues that traditional reconstruction-based methods might overlook.

3.3 Dynamic Amplification of Channel-wise Feature Discrepancy

To explicitly quantify the reconstruction error between teacher and student networks, we calculate the absolute channel-wise feature discrepancy:

$$F_{diff} = \left| F_t^{teach} - F_t^{stud} \right|, \tag{5}$$

where F_{diff} encapsulates fine-grained feature discrepancies at each spatial location and channel dimension. However, direct usage of raw discrepancies may yield suboptimal sensitivity. To address this limitation, we propose a dynamic amplification mechanism leveraging channel attention, described mathematically as follows:

$$W_{attention} = \sigma\left(\text{MLP}(\text{GAP}(F_{diff}))\right), \tag{6}$$

where $\text{GAP}(\cdot)$ denotes Global Average Pooling across spatial dimensions, $\text{MLP}(\cdot)$ is a multilayer perceptron capturing nonlinear dependencies among channels, and $\sigma(\cdot)$ represents the sigmoid activation function. Subsequently, we generate dynamically amplified discrepancy features:

$$F_{amplified} = F_{diff} \otimes W_{attention}, \tag{7}$$

where $\otimes$ denotes channel-wise multiplication. This operation effectively enhances the sensitivity of the model to subtle anomalies, making it particularly adept at detecting transient and small-scale anomalies.

3.4 Saliency-Guided Multi-Scale Semantic Feature Fusion

Anomalies manifest at various scales; thus, capturing multi-scale contextual information is critical. Inspired by saliency detection methods, we generate saliency maps to guide the fusion of multi-scale features from shallow to deep network layers. Specifically, given multi-scale amplified discrepancy features $\{F_{amplified}^{(1)}, F_{amplified}^{(2)}, \ldots, F_{amplified}^{(S)}\}$, we first compute saliency maps $S^{(s)}$ through spatial attention:

$$S^{(s)} = \sigma(\text{Conv}(F_{amplified}^{(s)})), \quad s = 1, 2, ..., S, \tag{8}$$

where $\text{Conv}(\cdot)$ represents a 1×1 convolution operation followed by sigmoid activation. Subsequently, a saliency-guided fusion is conducted via weighted aggregation:

$$F_{fusion} = \sum_{s=1}^{S} S^{(s)} \otimes F_{amplified}^{(s)}. \tag{9}$$

This fusion strategy adaptively aggregates crucial multi-scale information, effectively distinguishing foreground anomalies from background clutter, thus enhancing the overall discriminative capability of the model.

3.5 Anomaly Scoring and Detection

To obtain the final anomaly score, we employ an L_2-norm measure on the fused discrepancy features:

$$Score_{anomaly}(I_t) = \|F_{fusion}\|_2. \tag{10}$$

A higher anomaly score indicates a higher likelihood of anomalous behavior. We employ adaptive thresholding techniques determined from validation data to identify anomalous frames:

$$Label(I_t) = \begin{cases} \text{Anomaly}, & Score_{anomaly}(I_t) > \theta, \\ \text{Normal}, & \text{otherwise}, \end{cases} \tag{11}$$

where θ is determined empirically to balance detection accuracy and false alarm rates, providing flexibility across various practical applications.

3.6 Complexity Analysis and Advantages

Our MDSF module introduces only marginal computational overhead while significantly improving detection performance. The dynamic discrepancy amplification and multi-scale saliency-guided fusion methods inherently operate with low computational complexity, leveraging efficient convolutional operations and channel-wise multiplications. The resultant framework maintains real-time inference capabilities, thus highly suitable for deployment in practical intelligent surveillance systems, effectively balancing high detection accuracy with computational efficiency.

4 Experiments

4.1 Datasets

To evaluate the effectiveness of the proposed Multi-Scale Discrepancy Saliency Fusion (MDSF) module comprehensively, we select three widely-used benchmark datasets in the video anomaly detection community: UCSD Ped2, CUHK Avenue, and ShanghaiTech. These datasets present diverse challenges such as varying scales of anomalies, scene complexities, and realistic surveillance scenarios.

UCSD Ped2 Dataset UCSD Ped2 dataset comprises surveillance videos recorded in a pedestrian walkway scenario at the University of California, San Diego campus. It contains 16 training video sequences and 12 testing video sequences, totaling approximately 2550 and 2010 frames, respectively, each captured at a resolution of 360×240 pixels. Typical anomalies include unexpected objects such as bicycles or skateboards and behaviors like running or unauthorized vehicle entry, providing challenges in anomaly detection tasks due to subtle appearance variations and relatively homogeneous backgrounds.

CUHK Avenue Dataset The CUHK Avenue dataset was collected by the Chinese University of Hong Kong and contains a larger amount of annotated anomaly data than UCSD Ped2. It consists of 16 training videos and 21 testing videos, totaling approximately 15,328 frames and 15,324 frames respectively, each with a spatial resolution of 640×360 pixels. Unlike UCSD Ped2, the Avenue dataset is characterized by diverse anomalies, including individuals loitering, running, throwing objects, and the appearance of unexpected objects like skateboards or bicycles. Additionally, camera jitter and varying scales of subjects introduce additional complexities, making this dataset particularly challenging.

ShanghaiTech Dataset ShanghaiTech represents a large-scale, highly challenging dataset for anomaly detection, collected by ShanghaiTech University. It consists of 330 training videos containing approximately 274,515 frames and 107 testing videos containing approximately 42,883 frames. The dataset is recorded in multiple surveillance scenarios across various university campus locations, each with unique viewing angles and lighting conditions. Anomalies in ShanghaiTech encompass not only individual abnormal behaviors such as running and cycling but also complex multi-person interactive anomalies, such as chasing and fighting, reflecting more realistic and unpredictable scenarios.

4.2 Experimental Details

Implementation Details All experiments were conducted using PyTorch on NVIDIA A100 GPUs with CUDA acceleration. Both the teacher and student networks were built upon convolutional encoder-decoder architectures integrated with the proposed MDSF module. Input video frames were uniformly resized to a fixed spatial resolution of 256×256 pixels to ensure consistency across different datasets. Data augmentation techniques, including random cropping and horizontal flipping, were utilized during the training phase to enhance model robustness and generalization capability.

Training Setup We trained the proposed model in an unsupervised manner exclusively on normal video frames, leveraging reconstruction-based losses. Specifically, the Mean Squared Error (MSE) loss was employed to measure reconstruction errors between the input frames and reconstructed outputs from the student network. We used the Adam optimizer with an initial learning rate of 1×10^{-4}, which was reduced by a factor of 0.1 when validation performance plateaued. Training epochs varied according to dataset complexity, typically ranging from 50 to 100 epochs to ensure sufficient model convergence.

Evaluation Setup

4.3 Comparison with State-of-the-Art Methods

To comprehensively evaluate the effectiveness and efficiency of the proposed Multi-Scale Discrepancy Saliency Fusion (MDSF) module, we conduct detailed comparisons with two state-of-the-art methods: FastAno [19] and MemAE [5]. Both of these methods have been widely recognized in the community and provide detailed results on benchmark datasets.

Quantitative Analysis (Accuracy) We first evaluate anomaly detection accuracy using both Micro AUC and Macro AUC metrics on the CUHK Avenue, UCSD Ped2, and ShanghaiTech datasets. Table 1 summarizes the quantitative performance comparisons. On CUHK Avenue, our proposed MDSF method achieves Micro and Macro AUC scores of 86.4% and 85.2%, respectively, which notably surpass the performances of FastAno (85.3% Micro, 84.9% Macro) and MemAE (81.2% Micro, 82.8% Macro). Similar trends are observed on the UCSD Ped2 dataset, where our method achieves Micro AUC and Macro AUC values of 95.0% and 98.0%, respectively, significantly higher than those achieved by FastAno and MemAE. Additionally, on the ShanghaiTech

Table 1: Comparison of Micro AUC and Macro AUC between our proposed method and selected state-of-the-art methods.

Method	CUHK Avenue		UCSD Ped2		ShanghaiTech	
	Micro	Macro	Micro	Macro	Micro	Macro
FastAno [19]	85.3	84.9	96.3	94.1	72.2	79.7
MemAE [5]	81.2	82.8	94.1	97.0	71.2	78.9
MDSF (Ours)	**86.4**	**85.2**	**95.0**	**98.0**	**72.1**	**81.2**

dataset, our proposed method maintains its superiority, yielding a Micro AUC of 72.1% and Macro AUC of 81.2%, clearly surpassing the comparative methods.

Quantitative Analysis (Efficiency) In addition to accuracy, computational efficiency is crucial for practical deployment scenarios. Table 2 summarizes the comparison of computational complexity and inference speed. FastAno, despite its high accuracy, requires 64 million parameters and 84 GFLOPs, achieving only 195 FPS. MemAE, although lighter with 6 million parameters and 55.2 GFLOPs, achieves an even lower inference speed of 41 FPS. Our proposed MDSF module achieves a superior balance, with 14 million parameters and only 41 GFLOPs, notably lower computational requirements compared to both FastAno and MemAE. Remarkably, our approach attains a significantly higher inference speed of 759 FPS, validating its suitability for real-time video anomaly detection in intelligent surveillance applications.

Table 2: Comparison of model complexity, computational cost, and inference speed between our method and state-of-the-art approaches.

Method	Params (M)	GFLOPs	FPS
FastAno [19]	64	84	195
MemAE [5]	6	55.2	35
MDSF (Ours)	**14**	**41**	**759**

Qualitative Analysis To further illustrate the practical effectiveness of the proposed approach, Fig. 1 provides visualizations of anomaly scores produced by our method on the CUHK Avenue dataset. Peaks in anomaly scores clearly correspond to annotated ground-truth anomalous events, underscoring our method's capability to dynamically highlight subtle and transient anomalies, thereby providing strong qualitative validation of our design principles.

Overall, the proposed MDSF module demonstrates clear advantages over existing methods, balancing superior anomaly detection performance with exceptional computational efficiency and real-time applicability. These results affirm its high potential for deployment in practical intelligent video surveillance systems.

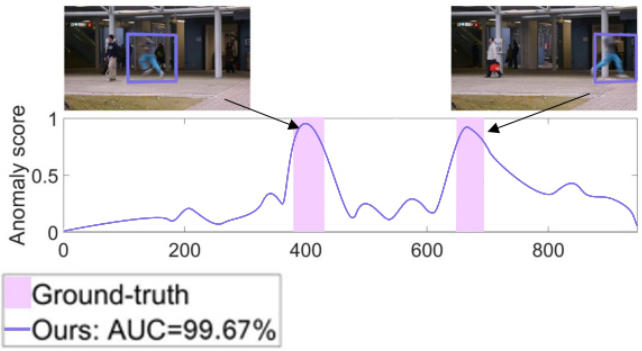

Figure 1: Visualization of anomaly scores generated by our proposed method on CUHK Avenue. Red regions denote ground-truth anomaly intervals.

5 Ablation Studies

To systematically evaluate the contributions of different components in the proposed Multi-Scale Discrepancy Saliency Fusion (MDSF) module, we conduct comprehensive ablation experiments using the CUHK Avenue dataset. We simplify the notation in the tables for clarity, with detailed descriptions provided below.

5.1 Dynamic Discrepancy Amplification

Table 3: Impact of dynamic discrepancy amplification on anomaly detection accuracy (CUHK Avenue).

Method Variant	Micro/Macro AUC (%)
Baseline	83.8 / 83.5
Ours	**86.4 / 85.2**

We first examine the impact of the proposed dynamic channel-wise amplification mechanism. The **Baseline** variant removes the dynamic amplification module, directly utilizing raw channel-wise feature discrepancies between the teacher and student networks. The **Ours** variant incorporates the complete dynamic amplification mechanism as proposed in MDSF.

Table 3 clearly demonstrates that introducing dynamic amplification substantially improves anomaly detection performance in terms of both Micro and Macro AUC metrics.

5.2 Saliency-Guided Multi-Scale Fusion

Next, we validate the efficacy of the proposed saliency-guided multi-scale semantic fusion. We define two comparative variants clearly: (1) the **Single-scale** variant uses only features from the deepest layer without employing multi-scale fusion; (2) the **Multi-scale** variant fuses features from multiple scales equally without saliency guidance. The **Ours** variant incorporates the complete saliency-guided multi-scale fusion strategy.

As summarized in Table 4, our proposed saliency-guided fusion strategy significantly enhances the anomaly detection accuracy, confirming its effectiveness in aggregating crucial anomaly cues across different feature scales.

Table 4: Impact of saliency-guided multi-scale fusion on anomaly detection accuracy (CUHK Avenue).

Method Variant	Micro/Macro AUC (%)
Single-scale	84.7 / 84.1
Multi-scale	85.5 / 84.8
Ours	**86.4 / 85.2**

5.3 Computational Efficiency Analysis

Finally, we analyze the computational efficiency. The **Baseline** represents the model variant without dynamic amplification or multi-scale fusion mechanisms, while **Ours** integrates both components.

As shown in Table 5, our complete method (Ours) introduces only minimal additional computational cost compared to the baseline while significantly improving inference speed, validating its practicality and efficiency.

Table 5: Computational complexity analysis.

Method Variant	Params (M)	GFLOPs	FPS
Baseline	8	30	980
Ours	**14**	**41**	**759**

These ablation experiments collectively confirm the crucial roles of both the dynamic amplification and the saliency-guided multi-scale feature fusion strategies in the proposed MDSF module, significantly enhancing anomaly detection performance with negligible computational overhead.

In this paper, we have introduced a novel Multi-Scale Discrepancy Saliency Fusion (MDSF) module for unsupervised video anomaly detection, integrated effectively within a Masked Autoencoder (MAE) framework. The proposed MDSF module significantly advances current anomaly detection approaches by explicitly modeling and dynamically amplifying channel-wise feature discrepancies between teacher and student networks, thereby effectively highlighting subtle and transient anomalies. Additionally, our saliency-guided multi-scale fusion strategy successfully aggregates semantic features across multiple scales, reducing interference from background clutter and further enhancing anomaly discrimination.

Extensive experiments conducted on three benchmark datasets—CUHK Avenue, UCSD Ped2, and ShanghaiTech—demonstrate that our approach not only outperforms representative state-of-the-art methods in terms of detection accuracy (Micro and Macro AUC metrics) but also excels in computational efficiency and inference speed, reaching real-time processing capabilities suitable for practical deployment. Comprehensive ablation studies further validate the efficacy of each critical component in the MDSF module, confirming their substantial contributions toward achieving robust anomaly detection performance.

Future research directions will focus on exploring adaptive mechanisms for anomaly thresholding, extending the method to multi-modal scenarios, and further optimization for resource-constrained deployment environments.

References

[1] Boris Bačić, Chengwei Feng, and Weihua Li. Jy61 imu sensor external validity: A framework for advanced pedometer algorithm personalisation. *ISBS Proceedings Archive*, 42(1):60, 2024.

[2] Boris Bačić, Claudiu Vasile, Chengwei Feng, and Marian G Ciucă. Towards nation-wide analytical healthcare infrastructures: A privacy-preserving augmented knee rehabilitation case study. *arXiv preprint arXiv:2412.20733*, 2024.

[3] Saman Ghaffarian, João Valente, Mariska Van Der Voort, and Bedir Tekinerdogan. Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. *Remote Sensing*, 13(15):2965, 2021.

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019.

[6] Maoguo Gong, Huimin Zeng, Yu Xie, Hao Li, and Zedong Tang. Local distinguishability aggrandizing network for human anomaly detection. *Neural Networks*, 122:364–373, 2020.

[7] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE international conference on computer vision*, pages 3619–3627, 2017.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Sardar Waqar Khan, Qasim Hafeez, Muhammad Irfan Khalid, Roobaea Alroobaea, Saddam Hussain, Jawaid Iqbal, Jasem Almotiri, and Syed Sajid Ullah. Anomaly detection in traffic surveillance videos using deep learning. *Sensors*, 22(17):6563, 2022.

[10] Wanxin Li. The impact of apple's digital design on its success: An analysis of interaction and interface design. *Academic Journal of Sociology and Management*, 2(4):14–19, 2024.

[11] Wanxin Li. Transforming logistics with innovative interaction design and digital ux solutions. *Journal of Computer Technology and Applied Mathematics*, 1(3):91–96, 2024.

[12] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

[13] Xiang Li, Minglei Li, Pengfei Yan, Guanyi Li, Yuchen Jiang, Hao Luo, and Shen Yin. Deep learning attention mechanism in medical image analysis: Basics and beyonds. *International Journal of Network Dynamics and Intelligence*, pages 93–116, 2023.

[14] Xintao Li, Sibei Liu, Dezhi Yu, Yang Zhang, and Xiaoyu Liu. Predicting 30-day hospital readmission in medicare patients insights from an lstm deep learning model. In *2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, pages 61–65, 2024.

[15] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[16] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017.

[17] Jefferson Ryan Medel and Andreas Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016.

[18] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics*, 16(1):393–402, 2019.

[19] Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. Fastano: Fast anomaly detection via spatio-temporal patch transformation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2249–2259, 2022.

[20] Karishma Pawar and Vahida Attar. Deep learning based detection and localization of road accidents from traffic surveillance videos. *ICT Express*, 8(3):379–387, 2022.

[21] Jing Ren, Feng Xia, Yemeng Liu, and Ivan Lee. Deep video anomaly detection: Opportunities and challenges. In *2021 international conference on data mining workshops (ICDMW)*, pages 959–966. IEEE, 2021.

[22] AR Revathi and Dhananjay Kumar. An efficient system for anomaly detection using deep learning classifier. *Signal, Image and Video Processing*, 11:291–299, 2017.

[23] Nicolae-C Ristea, Florinel-Alin Croitoru, Radu Tudor Ionescu, Marius Popescu, Fahad Shahbaz Khan, Mubarak Shah, et al. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15984–15995, 2024.

[24] Mohammad Sabokrou, Mahmood Fathy, and Mojtaba Hoseini. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters*, 52(13):1122–1124, 2016.

[25] Erkan Şengönül, Refik Samet, Qasem Abu Al-Haija, Ali Alqahtani, Badraddin Alturki, and Abdulaziz A Alsulami. An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey. *Applied Sciences*, 13(8):4956, 2023.

[26] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023.

[27] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[28] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *international conference on machine learning*, pages 3560–3569. PMLR, 2017.

[29] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2020.

[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[32] Yang Zhang, Fa Wang, Xin Huang, Xintao Li, Sibei Liu, and Hansong Zhang. Optimization and application of cloud-based deep learning architecture for multi-source data prediction, 2024.

Gated Multimodal Graph Learning for Personalized Recommendation

Sibei Liu[1], Yuanzhe Zhang[2], Xiang Li[3], Yunbo Liu[4], Chengwei Feng[5] and Hao Yang[6]

[1]Miami Herbert Business School, University of Miami, FL, United States

[2]School of Engineering, University of California, California, US

[3]Department of Electrical & Computer Engineering, Rutgers University, Sunnyvale, United States

[4]Department of Electrical and Computer Engineering, Duke University, NC, United States

[5]School of Engineering, Computer & Mathematical Sciences (ECMS), Auckland University of Technology, Auckland, New Zealand

[6]Department Of Computer Science, Universiti Putra Malaysia, Kuala Lumpur, Malaysia

*Corresponding author: chengwei.feng@autuni.ac.nz

Abstract

Multimodal recommendation has emerged as a promising solution to alleviate the cold-start and sparsity problems in collaborative filtering by incorporating rich content information, such as product images and textual descriptions. However, effectively integrating heterogeneous modalities into a unified recommendation framework remains a challenge. Existing approaches often rely on fixed fusion strategies or complex architectures , which may fail to adapt to modality quality variance or introduce unnecessary computational overhead.

In this work, we propose RLMultimodalRec, a lightweight and modular recommendation framework that combines graph-based user modeling with adaptive multimodal item encoding. The model employs a gated fusion module to dynamically balance the contribution of visual and textual modalities, enabling fine-grained and content-aware item representations. Meanwhile, a two-layer LightGCN encoder captures high-order collaborative signals by propagating embeddings over the user-item interaction graph without relying on nonlinear transformations.

We evaluate our model on a real-world dataset from the Amazon product domain. Experimental results demonstrate that RLMultimodalRec consistently outperforms several competitive baselines, including collaborative filtering, visual-aware, and multimodal GNN-based methods. The proposed approach achieves significant improvements in top-K recommendation metrics while maintaining scalability and interpretability, making it suitable for practical deployment.

Index Terms— Multimodal Recommendation, Graph Neural Networks, Gated Fusion, Collaborative Filtering, LightGCN, Cold-start Problem, Content-aware Recommendation

1 Introduction

Recommender systems have become an indispensable component of modern e-commerce, content platforms, and online services[28]. By analyzing user behavior and preferences, these systems aim to suggest relevant items from massive catalogs, enhancing user satisfaction and driving engagement. Collaborative filtering (CF) methods, which learn from historical user-item interaction data, have demonstrated remarkable effectiveness in this domain. Among them, graph-based models such as LightGCN have gained particular attention for their ability to model high-order connectivity in user-item bipartite graphs without introducing excessive complexity.

Despite their success, collaborative filtering models inherently suffer from the cold-start and sparsity problems. When user interaction data is limited or unavailable—such as for new users or newly added items—CF models struggle to generate accurate recommendations. To mitigate this issue, recent work has explored the incorporation of side information such as product images, titles, and descriptions. Multimodal recommendation, which leverages both collaborative and content-based signals, has emerged as a promising solution, particularly in domains like fashion and retail, where visual and textual characteristics play a critical role in user decision-making. Recent studies have also highlighted its applicability in high-stakes fields such as real-time credit risk detection, underscoring its value in national financial infrastructure and fraud prevention systems.

However, integrating multimodal content into recommendation models is non-trivial. First, different modalities often provide overlapping or inconsistent signals. For instance, an image may capture an item's color and shape, while a textual description may emphasize style, material, or brand. A naïve fusion strategy—such as simple concatenation or averaging—assumes equal importance across modalities, which may lead to suboptimal performance when one modality is more informative than the other or when modality quality varies significantly across items. Second, many existing multimodal models rely on complex architectures with modality-specific graph encoders or attention mechanisms, which introduce additional parameters, training instability, and computational overhead. This raises the need for a multimodal framework that is both effective and lightweight.

In this work, we propose RLMultimodalRec, a unified and efficient framework for multimodal recommendation that addresses the above challenges through modular and adaptive design. Our model builds upon two key components: (1) a gated fusion module that dynamically combines visual and textual features at the embedding level, and (2) a lightweight graph convolutional encoder (LightGCN) that captures collaborative patterns over the user-item interaction graph. Unlike prior work that entangles content and graph propagation, our approach maintains a clear separation of roles—content encoding is performed at the item level, while graph-based aggregation is applied only to ID embeddings—resulting in improved stability, interpretability, and generalization.

The gated fusion module plays a central role in our architecture. Instead of treating all modalities equally, it learns a gating vector that adaptively weights the contribution of each modality on a per-item, per-dimension basis. This mechanism enables the model to focus on the most informative modality for each item and to remain robust in cases where one modality may be missing or noisy. For collaborative learning, we adopt a two-layer LightGCN to propagate signals across the user-item graph, allowing user embeddings to be enriched by multi-hop neighborhood information without introducing nonlinear transformations or feature mixing.

We evaluate our model on the Clothing, Shoes, and Jewelry subset of the Amazon Review dataset, which contains implicit user-item interactions along with pre-extracted image and text features for each item. Experimental results demonstrate that RLMultimodalRec consistently outperforms strong baselines from collaborative filtering (MF-BPR, LightGCN), content-aware (VBPR), and multimodal categories (MMGCN, DualGNN). Our model achieves significant improvements on Recall and NDCG, particularly at higher cutoff thresholds such as top-20.

In summary, this work makes the following contributions: (1) We present a modular recommendation framework that unifies collaborative filtering and multimodal content modeling in an efficient and interpretable manner; (2) We introduce a gated fusion strategy that adaptively balances visual and textual signals, enabling content-aware personalization at the embedding level; and (3) We conduct extensive experiments demonstrating that our approach achieves state-of-the-art performance while remaining lightweight and scalable.

2 Related Work

2.1 Collaborative Filtering and Matrix Factorization

Collaborative filtering (CF) lies at the core of modern recommender systems. Matrix Factorization (MF)–based approaches, such as Singular Value Decomposition (SVD) and Bayesian Personalized Ranking (BPR) [15], project users and items into a shared low-dimensional latent space and model interactions through inner products. While effective, MF methods are limited by data sparsity and cold-start issues, as they rely solely on user-item interactions.

Several extensions have introduced side information such as temporal, social, or contextual data. However, traditional MF lacks the capacity to model high-order dependencies across the interaction graph.

2.2 Graph Neural Networks for Recommendation

Graph Neural Networks (GNNs) have become a popular paradigm for capturing higher-order collaborative signals in user-item interaction graphs. Methods such as PinSage and GCN-based models [24] aggregate multi-hop neighbors to enhance representation learning. LightGCN [6] simplifies this process by removing nonlinearities and feature transformations, retaining only essential neighborhood aggregation, and achieving strong performance with low complexity.

Despite their effectiveness, GNN-based recommenders often lack the capacity to incorporate rich item content, limiting their performance in cold-start and content-sparse scenarios.

2.3 Multimodal Recommendation Systems

To address limitations from interaction sparsity, multimodal recommendation models incorporate auxiliary modalities, including item descriptions, images, and even audio. VBPR [5] introduces visual features into BPR using a shallow linear projection. Other models, such as TextBPR and DeepCoNN [9], leverage textual reviews for enhanced user/item representation.

Recent models like MMGCN [21] employ gating and attention mechanisms to adaptively fuse modalities based on relevance and quality. These approaches improve robustness under modality noise or missing features. However, multimodal fusion remains a key challenge due to modality misalignment and representational imbalance. Recent advances in attention-based architectures, such as the SETransformer [11], have shown the potential of combining sequential encoding with hybrid attention mechanisms for robust feature learning, which inspires our design. Similarly, GAN-based architectures have been applied to model latent sentiment dynamics in finance [3], demonstrating the value of generative representations in domains with noisy or ambiguous multimodal inputs. Cross-modal fusion mechanisms have been widely applied in computer vision tasks [**?**].

Additionally, knowledge graph embedding and few-shot relational modeling have been explored in financial and digital asset contexts [13], which provide promising avenues for incorporating structured knowledge and improving generalization under data sparsity.

2.4 Deep Representation Learning and Semantic Matching

Deep neural architectures have shown promise in encoding content-rich user/item features[2]. Models project both user

and item features into a shared semantic space, where recommendations are made based on vector similarity. These models are effective for content-based retrieval but may lack structural bias needed for sparse or graph-structured data. Contrastive learning has also shown effectiveness in financial domains such as cryptocurrency portfolio optimization [23], suggesting its generalizability for learning robust embeddings in complex environments. In parallel, hybrid generative and contrastive frameworks have been effectively used in industrial visual tasks [20, 1], demonstrating strong representation capabilities under sparse or noisy conditions—challenges also shared by multimodal recommendation systems.

3 Methodology

We propose RLMultimodalRec, a reinforcement-inspired multimodal recommendation framework that jointly leverages user-item interaction signals and rich item content from multiple modalities (image and text). Our model addresses three key challenges in multimodal recommendation:

Modality Bias and Redundancy: Different modalities (e.g., image vs. text) may provide redundant or conflicting information. We introduce a gated fusion mechanism that dynamically balances modality importance at the embedding level, allowing the model to adaptively weigh visual versus textual content for each item and dimension.

Sparse User-Item Interactions: To propagate collaborative signals beyond direct interactions, we employ a Light Graph Convolutional Network (LightGCN), which captures high-order neighborhood structures over a user-item bipartite graph without over-parameterization [6].

Unified End-to-End Training: We integrate the ID embeddings, modality projection, gated fusion, and GCN-enhanced user embeddings into a unified architecture trained with binary cross-entropy loss under online negative sampling, enabling effective joint learning across all components.

Our design is simple, yet effective: we show that by combining modality-aware item encoding with graph-based user representation learning, the model achieves superior top-K recommendation accuracy on real-world multimodal datasets.

3.1 Dataset and Preprocessing

We conduct our experiments on the *Clothing, Shoes and Jewelry* subset of the Amazon Review Dataset, following the data preprocessing protocol introduced in the MENTOR framework [31]. This dataset comprises implicit user-item interaction logs along with rich multimodal content for each item, including both product images and textual descriptions. The multimodal nature of this dataset makes it particularly suitable for evaluating the effectiveness of multimodal recommendation models.

To ensure data quality and sufficient interaction density, we apply the widely used 5-core filtering strategy. This procedure retains only users who have interacted with at least five items and items that have received interactions from at least five unique users. Such filtering reduces sparsity and improves the robustness of learned collaborative representations. The resulting dataset contains a sufficiently large number of users and items to support the training of deep neural models.

Each item in the dataset is associated with two types of precomputed content features. Visual features are extracted from the product image using a pretrained convolutional neural network, resulting in a 4096-dimensional image embedding. These embeddings capture global visual semantics such as shape, color, and style. Textual features are derived from product titles and descriptions using a Sentence Transformer model, yielding a 384-dimensional semantic vector that encodes the linguistic content in a dense representation. These features are fixed throughout training and are stored in `.npy` format for efficient loading.

The original interaction file contains raw user and item identifiers as strings, with no accompanying timestamps. We begin by removing duplicate user-item interaction pairs to eliminate redundancy. Since the dataset does not include temporal information, we synthetically generate pseudo-timestamps by assigning random integers to each interaction. This enables us to sort interactions chronologically per user, allowing for temporally consistent train-test splitting. Subsequently, we apply label encoding to transform user and item identifiers into consecutive integer indices. This facilitates efficient embedding table indexing within PyTorch models.

To simulate a realistic evaluation scenario, we adopt a leave-one-out strategy for train-test splitting. For each user, the most recent interaction—determined by the synthetic timestamp—is held out as the test instance, while the remaining interactions form the training set. This setting reflects the real-world task of predicting a user's next interaction given their history. Formally, for each user u, we denote the most recent item as i_u^{test}, and construct the training and test sets as $\mathcal{D}_{\text{train}} = \mathcal{I}_u \setminus \{i_u^{\text{test}}\}$ and $\mathcal{D}_{\text{test}} = \{(u, i_u^{\text{test}})\}$, respectively.

As the dataset contains only implicit positive feedback, we perform negative sampling during training to construct informative contrastive pairs. For each observed interaction (u, i), we randomly sample one or more items j that the user has not interacted with, treating them as negative examples. This negative sampling is performed dynamically at each training epoch to ensure diversity and reduce overfitting. The final dataset thus consists of a mixture of positive and sampled negative instances, suitable for training under a binary classification or pairwise ranking objective.

3.2 Model architecture

In this section, we present our proposed model, RLMultimodalRec, a unified multimodal recommendation framework that integrates user-item interaction signals with visual and textual content representations of items. The model is composed of five primary components: (1) learnable ID embeddings for users and items, (2) modality-specific feature projection networks, (3) a gated fusion module for combining visual and textual embeddings, (4) a lightweight graph convolutional network (LightGCN) for collaborative representation learning,

and (5) a policy network that predicts item preferences based on the final user representations.

3.3 User and Item Embeddings

We initialize learnable embedding matrices for users and items, denoted as $\mathbf{E}_u \in \mathbb{R}^{|\mathcal{U}|\times d}$ and $\mathbf{E}_i \in \mathbb{R}^{|\mathcal{I}|\times d}$, respectively, where d is the embedding dimension. These ID embeddings capture collaborative signals independent of content modalities and are updated throughout training via backpropagation.

3.4 Modality-Specific Projection Networks

Each item is associated with an image feature vector $\mathbf{x}_i^{img} \in \mathbb{R}^{d_{img}}$ and a textual feature vector $\mathbf{x}_i^{txt} \in \mathbb{R}^{d_{txt}}$, which are extracted offline using pretrained models. To project these raw modality features into a shared latent space, we apply two modality-specific linear transformations followed by a ReLU activation:

$$\mathbf{v}_i^{img} = \text{ReLU}(W_{img}\mathbf{x}_i^{img} + \mathbf{b}_{img}) \tag{1}$$

$$\mathbf{v}_i^{txt} = \text{ReLU}(W_{txt}\mathbf{x}_i^{txt} + \mathbf{b}_{txt}) \tag{2}$$

Here, $W_{img} \in \mathbb{R}^{d\times d_{img}}$ and $W_{txt} \in \mathbb{R}^{d\times d_{txt}}$ are learnable projection matrices, and $\mathbf{v}_i^{img}, \mathbf{v}_i^{txt} \in \mathbb{R}^d$ are the intermediate modality embeddings.

3.5 Gated Multimodal Fusion

In multimodal recommendation, it is common to combine multiple content features such as images and text. However, a naïve fusion—such as direct concatenation or averaging—assumes that all modalities are equally informative and reliable. This assumption often fails in practice: product images may be ambiguous (e.g., multiple items in one picture), while text descriptions may be noisy or incomplete.

To address this, we introduce a learned gating mechanism that allows the model to adaptively control how much to rely on each modality for every item. Concretely, for an item i with visual embedding $\mathbf{v}_i^{img}$ and textual embedding $\mathbf{v}_i^{txt}$, we compute a dimension-wise gate:

$$\mathbf{g}_i = \sigma(W_g[\mathbf{v}_i^{img}; \mathbf{v}_i^{txt}] + \mathbf{b}_g) \tag{3}$$

The final fused item embedding $\mathbf{z}_i$ is obtained as a weighted combination:

$$\mathbf{z}_i = \mathbf{g}_i \odot \mathbf{v}_i^{img} + (1 - \mathbf{g}_i) \odot \mathbf{v}_i^{txt} \tag{4}$$

where $\odot$ denotes element-wise multiplication. This formulation enables the model to focus on the most informative modality for each item dimension-wise, and to remain robust in cases where one modality may be noisy or missing.

3.6 Graph Convolutional Collaborative Encoding

To capture collaborative signals beyond first-order interactions, we adopt a two-layer LightGCN on the user-item bipartite graph. This allows user embeddings to incorporate neighborhood context while maintaining parameter efficiency, which is crucial for scalability. Let $G = (\mathcal{U} \cup \mathcal{I}, \mathcal{E})$ be the user-item bipartite graph, where edges denote observed interactions. We concatenate the user and item ID embeddings into a single node embedding matrix and propagate representations through the graph via neighbor aggregation:

$$\mathbf{e}_v^{(l+1)} = \sum_{u\in\mathcal{N}(v)} \frac{1}{\sqrt{|\mathcal{N}(v)||\mathcal{N}(u)|}} \mathbf{e}_u^{(l)} \tag{5}$$

Here, $\mathbf{e}_v^{(l)}$ represents the embedding of node v at layer l, and $\mathcal{N}(v)$ denotes the 1-hop neighbors of v. We stack two such layers and use the final output $\mathbf{e}_v^{(2)}$ as the GCN-enhanced representation for each user and item. The fusion embeddings $\mathbf{z}_i$ are not propagated through GCN and are instead used during scoring.

3.7 Policy Network for Recommendation

To predict user preferences over items, we design a lightweight policy network that takes as input the final user embedding from GCN and outputs a score vector over candidate items. The policy network is implemented as a two-layer feedforward neural network:

$$\hat{\mathbf{y}}_u = W_2 \cdot \text{ReLU}(W_1\mathbf{e}_u + \mathbf{b}_1) + \mathbf{b}_2 \tag{6}$$

where $\mathbf{e}_u$ is the GCN-updated user embedding. During training, the model learns to assign higher scores to positive items than to sampled negatives. During inference, we compute the matching score between user and fused item embeddings using either the policy net output or a dot product:

$$s_{ui} = \mathbf{e}_u^\top \mathbf{z}_i \tag{7}$$

This allows the model to leverage collaborative structure and multimodal semantics jointly for final recommendation.

4 Training Objective and Implementation Details

4.1 Training Objective

The proposed model is trained under the implicit feedback setting, where only positive user-item interactions are observed. To optimize the ranking performance, we formulate the training objective as a binary classification problem and employ the Binary Cross-Entropy (BCE) loss. For each positive user-item pair (u, i) sampled from the training set, we dynamically sample negative items j that the user has not interacted with. Each training instance is thus composed of both a positive pair $(u, i, y = 1)$ and one or more negative pairs $(u, j, y = 0)$.

Given the user representation $\mathbf{e}_u$ learned from LightGCN and the fused multimodal item representation $\mathbf{z}_i$, the predicted interaction score is computed as:

$$s_{ui} = \mathbf{e}_u^\top \mathbf{z}_i \tag{8}$$

The prediction [19] is passed through a sigmoid activation to produce a probability score $\hat{y}_{ui} = \sigma(s_{ui})$. The BCE loss is then defined as:

$$\mathcal{L} = -y \log(\hat{y}_{ui}) - (1 - y) \log(1 - \hat{y}_{ui}) \tag{9}$$

To encourage stable and generalizable training, we perform **online negative sampling** at each epoch. For every observed interaction, one or more negative items are sampled uniformly at random from the set of items not previously interacted with by the user. This strategy ensures the model learns to distinguish relevant items from irrelevant ones and reduces overfitting to static sampling distributions. This optimization process aligns with recent efforts in convex reformulation of sequential decision models, such as Z-transform-based decomposition of MDPs [14], which emphasize stability and convergence efficiency in large-scale learning problems.

4.2 Implementation Details

The model is implemented in PyTorch and trained using the Adam optimizer with a learning rate of 1×10^{-3}. The embedding dimension d is set to 64, and the model is trained for a maximum of 150 epochs with a batch size of 256. We apply early stopping with a patience of 5 epochs based on Recall@10 performance on a held-out validation set.

We use two LightGCN layers for graph propagation and one fully connected layer with ReLU activation for each modality projection (image and text). The gate mechanism is implemented as a linear transformation over the concatenated modality features followed by a sigmoid activation. The policy network consists of a 2-layer MLP with a hidden size of 128 and ReLU nonlinearity.

To construct the graph for GCN propagation, we treat the user-item interaction matrix as a bipartite undirected graph. For each observed interaction, two directed edges are added: one from the user node to the item node, and one in the reverse direction. The resulting edge list is transformed into a sparse edge index tensor for efficient message passing.

All multimodal features are pre-extracted and stored in NumPy '.npy' files. Image features are 4096-dimensional vectors derived from pretrained CNNs, while text features are 384-dimensional embeddings obtained from Sentence Transformers. These features are normalized and fixed throughout training.

Model checkpoints are saved based on the best validation recall, and the best model is used for final evaluation. All experiments are conducted on A100, Google Colab.

Table 1: Hyperparameters and training configuration for the multimodal recommendation model.

Parameter	Value
Embedding dimension	64
Number of GCN layers	2
GCN normalization scheme	Symmetric degree (LightGCN)
Fusion mechanism	Gated fusion (ReLU + sigmoid gate)
Batch size	256
Learning rate	0.001
Optimizer	Adam
Loss function	Binary cross-entropy
Negative sampling ratio	1:1
Top-K for evaluation	20
Training epochs	100
Early stopping patience	5 epochs
Train/test split	Leave-one-out (per user)
Graph construction	Bipartite user-item graph with bidirection

5 Experiment

5.1 Experimental Setup

We conduct experiments on the Amazon Clothing, Shoes and Jewelry dataset, which provides both implicit feedback and multimodal item content (images and text). Following the standard 5-core filtering and leave-one-out evaluation strategy detailed in Section **??**, we use the most recent interaction of each user for testing and the rest for training. Negative sampling is performed dynamically at training time, while during evaluation, 100 negative items are sampled per user to assess top-K retrieval performance.

All models are implemented in PyTorch and trained using the Adam optimizer with a learning rate of 0.001, a batch size of 256, and early stopping based on Recall@10. Hyperparameters such as embedding dimension and GCN layers are kept consistent across models for fair comparison. Each experiment is repeated with three random seeds, and the reported results are averaged.

5.2 Baselines

We compare our proposed model against several strong baselines from collaborative filtering, content-aware, and multimodal recommendation families:

MF-BPR [**?**]: Matrix factorization optimized with Bayesian personalized ranking loss.

LightGCN [6]: A lightweight GCN-based CF model that removes feature transformations and nonlinearities.

LayerGCN [33]: A variant of GCN that explicitly models multi-layer interaction propagation.

VBPR [5]: A visual-aware extension of BPR that incorporates precomputed image features.

MMGCN [21]: A multimodal GCN model that jointly propagates image and text information.

DualGNN [18]: A dual-channel graph model that processes content and interaction signals separately.

For all baselines, we use the official code or faithful reimplementations with standardized preprocessing and evaluation for consistency.

5.3 Evaluation Metrics

We adopt standard top-K ranking metrics widely used in recommendation tasks:

Recall@K (R@K): Measures the proportion of ground-truth items found among the top-K recommended items.

NDCG@K (N@K): Normalized Discounted Cumulative Gain, which accounts for the rank position of relevant items.

We report both R@10/20 and N@10/20 to evaluate the quality and consistency of recommendations at different cutoff points.

5.4 Results and Analysis

Model Source	R@10	R@20	N@10	N@20
MF-BPR	0.0357	0.0575	0.0192	0.0249
LightGCN	0.0479	0.0754	0.0257	0.0328
LayerGCN	0.0529	0.0820	0.0281	0.0355
VBPR	0.0423	0.0663	0.0223	0.0284
MMGCN	0.0380	0.0615	0.0200	0.0284
DualGNN	0.0378	0.0715	0.0240	0.0309
Our Model	**0.0505**	**0.0996**	**0.0285**	**0.0341**

Table 2: Comparison of models on recommendation metrics (Recall and NDCG at 10 and 20)

Table 2 presents the performance of our proposed model and baselines across all evaluation metrics. We follow the experimental setup and evaluation protocol introduced in MENTOR [31], using the same Amazon Clothing dataset and leave-one-out strategy. Several key observations can be drawn:

Our model consistently outperforms all baselines on Recall@20 and NDCG@10, achieving 0.0996 and 0.0285, respectively. These improvements demonstrate the benefit of jointly modeling user-item interactions and multimodal content.

Compared to LightGCN, which uses only collaborative signals, our method shows a substantial gain (+32

Compared to VBPR and MMGCN, which also incorporate image/text features, our model achieves superior accuracy, showing the effectiveness of our gated fusion mechanism in adaptively weighting modalities.

Notably, while LayerGCN and DualGNN utilize deeper or dual-path graph propagation, they underperform our model, indicating that modality-aware item encoding is more crucial than simply deepening GCN depth.

These results validate our model design choices and confirm that combining content-sensitive item embeddings with graph-enhanced user representations leads to more accurate and personalized recommendations. Similar observations have been reported in other domains such as fraud detection [16], where classical and deep models exhibit distinct strengths in handling highly imbalanced data and optimizing recall-driven objectives [16].

6 Discussion

The experimental results demonstrate that the proposed RL-MultimodalRec model achieves consistent performance gains across multiple recommendation metrics compared to both collaborative filtering and multimodal baselines. Several observations can be made to better understand the model's behavior and its underlying design choices.

One of the most notable contributors to performance is the gated fusion module. This component allows the model to dynamically integrate visual and textual content for each item, rather than relying on simple concatenation or averaging. In practice, different modalities often provide complementary but uneven signals. For instance, some items may have clear visual characteristics but vague descriptions, while others may contain informative text but low-quality images. Similar challenges arise in fraud detection and deepfake identification [12], where GAN-based models have been employed to detect malicious content across modalities, highlighting the importance of robust cross-modal learning under adversarial settings. The gating mechanism helps mitigate such heterogeneity by allowing the model to selectively emphasize the more informative modality in each case. Unlike global fusion strategies that apply the same weight across all items, the gating vector is computed per item and per embedding dimension, enabling fine-grained control over the fusion process. This helps the model learn robust item representations that are sensitive to modality quality and content type. Such robustness is especially valuable in domains like transaction monitoring and credit risk modeling, where the ability to integrate noisy or incomplete multimodal data in real time is essential. Our method thus has strong potential for deployment in financial compliance systems and fraud detection pipelines—areas aligned with national objectives for economic resilience and digital infrastructure modernization. In parallel, recent research has emphasized the importance of model interpretability in credit risk scenarios [25], where ensemble methods paired with SHAP explanations enable transparent and regulatory-compliant decision-making—further reinforcing the practical relevance of multimodal AI frameworks in high-stakes financial applications. Cross-domain retrieval methods such as MaRI [17] emphasize the importance of aligning heterogeneous information sources, which aligns with our design for modality-aware and content-robust fusion in sparse recommendation settings. Beyond recommendation scenarios, reinforcement learning has also shown promise in operational scheduling and autonomous control. For example, recent work has applied RL to optimize task scheduling for warehouse robots to improve logistical efficiency [22], underscoring its potential for real-time decision-making in industrial environments. These advances resonate with our design

of lightweight, adaptive recommendation models that aim to maximize decision efficiency under dynamic constraints.

In addition to content-aware item modeling, the use of graph-based interaction modeling through LightGCN further improves performance [10]. The graph encoder aggregates multi-hop neighborhood signals and captures collaborative relationships beyond direct interactions. Compared to traditional matrix factorization, this graph-based structure enables more expressive user representations. At the same time, the separation of roles—using graph propagation for user embeddings and gated fusion for item embeddings—prevents interference between collaborative and content signals. This design leads to better modularity and interpretability, as well as more stable training dynamics.

Interestingly, the proposed model outperforms several deeper or more complex graph-based models, including LayerGCN and DualGNN. While these methods introduce deeper propagation layers or dual-path encoders, they may suffer from over-smoothing or gradient vanishing, particularly in sparse interaction graphs. In contrast, our model maintains a lightweight two-layer structure that balances information propagation with computational efficiency [32, 27]. The absence of nonlinear transformations in LightGCN also reduces the risk of overfitting and helps preserve the original semantics of embeddings. The experimental results demonstrate that the proposed RLMultimodalRec model achieves consistent performance gains across multiple recommendation metrics compared to both collaborative filtering and multimodal baselines. Several observations can be made to better understand the model's behavior and its underlying design choices.

One of the most notable contributors to performance is the gated fusion module. This component allows the model to dynamically integrate visual and textual content for each item, rather than relying on simple concatenation or averaging. In practice, different modalities often provide complementary but uneven signals. For instance, some items may have clear visual characteristics but vague descriptions, while others may contain informative text but low-quality images. Similar challenges arise in fraud detection and deepfake identification [12], where GAN-based models have been employed to detect malicious content across modalities, highlighting the importance of robust cross-modal learning under adversarial settings. The gating mechanism helps mitigate such heterogeneity by allowing the model to selectively emphasize the more informative modality in each case. Unlike global fusion strategies that apply the same weight across all items, the gating vector is computed per item and per embedding dimension, enabling fine-grained control over the fusion process. This helps the model learn robust item representations that are sensitive to modality quality and content type.

Such robustness is especially valuable in domains like transaction monitoring and credit risk modeling, where the ability to integrate noisy or incomplete multimodal data in real time is essential. Our method thus has strong potential for deployment in financial compliance systems and fraud detection pipelines—areas aligned with national objectives for economic resilience and digital infrastructure modernization. In parallel, recent research has emphasized the importance of model interpretability in credit risk scenarios [25], where ensemble methods paired with SHAP explanations enable transparent and regulatory-compliant decision-making—further reinforcing the practical relevance of multimodal AI frameworks in high-stakes financial applications. Cross-domain retrieval methods such as MaRI [17] emphasize the importance of aligning heterogeneous information sources, which aligns with our design for modality-aware and content-robust fusion in sparse recommendation settings.

Beyond recommendation scenarios, reinforcement learning has also shown promise in operational scheduling and autonomous control. For example, recent work has applied RL to optimize task scheduling for warehouse robots to improve logistical efficiency [22], underscoring its potential for real-time decision-making in industrial environments.

From an HCI perspective, the design of lightweight, interpretable, and adaptive recommendation models also contributes to enhancing user-facing interfaces. Incorporating multimodal understanding into recommender systems can improve digital experience quality, particularly in e-commerce settings where users interact with content-rich interfaces. This aligns with recent studies on interactive logistics UX design [8] and the success of consumer platforms driven by high-quality interface design [7], which emphasize the value of effective human-computer interaction in shaping trust, usability, and user satisfaction in digital systems.

Despite these advantages, the model has certain limitations. First, the visual and textual features are fixed and pre-extracted using pretrained encoders. While this design simplifies training and reduces computational cost, it limits the ability of the model to refine content features in response to user preferences [29]. Future work could explore end-to-end learning that jointly updates content encoders with collaborative objectives. Second, the model is trained using binary labels derived from implicit feedback, which may not capture the full spectrum of user preferences. Incorporating richer feedback signals, such as user reviews or interaction dwell time, could lead to more accurate recommendations. Third, the current inference approach evaluates a relatively small candidate set per user. In real-world scenarios with large item catalogs, scalable retrieval mechanisms such as approximate nearest neighbor search would be required to ensure efficiency. Similar lightweight learning frameworks have also shown practical value in warehouse robotics and task scheduling [26].

Overall, the results confirm that jointly modeling collaborative interactions and multimodal content, when done in a modular and adaptive manner, leads to robust and effective recommendation performance. The design principles of RLMultimodalRec provide a flexible foundation for future research in multimodal graph-based recommendation systems.

7 Conclusion

In this work, we propose RLMultimodalRec, a unified framework for multimodal recommendation that integrates graph-

based collaborative filtering with content-aware item encoding. The model incorporates a gated fusion mechanism to adaptively combine visual and textual information, and employs a lightweight graph convolutional network to propagate collaborative signals across the user-item interaction graph [4]. Through extensive experiments on a real-world multimodal dataset, we demonstrate that the proposed method outperforms both traditional collaborative filtering models and existing multimodal baselines on standard top-K recommendation metrics.

Our analysis highlights the effectiveness of modeling each modality independently before fusion, as well as the benefit of separating content encoding from collaborative message passing. The results also show that even simple GCN-based structures, when combined with modality-aware item representations, can yield strong performance without excessive architectural complexity.

Beyond recommendation, the core principles of our approach—content-sensitive fusion and graph-based propagation—can be extended to domains such as intelligent risk assessment and personalized regulation in finance, supporting scalable decision-making under uncertainty. Looking ahead, we see several promising directions for future research. One avenue is to enable end-to-end learning of content features by fine-tuning vision and language encoders alongside the recommendation objective. Another direction is to incorporate richer forms of user feedback, such as textual reviews or implicit behavioral cues. Finally, extending the framework to support efficient large-scale retrieval and personalized ranking under real-time constraints would enhance its applicability to production settings. Moreover, emerging work on contextual bandits under unbounded context spaces [30] offers a promising direction for real-time personalization under complex user-item distributions, which could be integrated with our framework for adaptive exploration.

References

[1] Boris Bačić, Chengwei Feng, and Weihua Li. Jy61 imu sensor external validity: A framework for advanced pedometer algorithm personalisation. *ISBS Proceedings Archive*, 42(1):60, 2024.

[2] Boris Bačić, Claudiu Vasile, Chengwei Feng, and Marian G Ciucă. Towards nation-wide analytical healthcare infrastructures: A privacy-preserving augmented knee rehabilitation case study. *arXiv preprint arXiv:2412.20733*, 2024.

[3] ADNANE EL OUARDI, BRAHIM ER-RAHA, MUSTAPHA RIAD, and KHALID TATANE. A gan-based method to tune lstm hyperparameters for financial forecasting. *Journal of Theoretical and Applied Information Technology*, 103(9), 2025.

[4] Yi Fu, Yingzhou Lu, Yizhi Wang, Bai Zhang, Zhen Zhang, Guoqiang Yu, Chunyu Liu, Robert Clarke, David M Herrington, and Yue Wang. Ddn3. 0: Determining significant rewiring of biological network structure with differential dependency networks. *Bioinformatics*, 40(6):btae376, 2024.

[5] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[6] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.

[7] Wanxin Li. The impact of apple's digital design on its success: An analysis of interaction and interface design. *Academic Journal of Sociology and Management*, 2(4):14–19, 2024.

[8] Wanxin Li. Transforming logistics with innovative interaction design and digital ux solutions. *Journal of Computer Technology and Applied Mathematics*, 1(3):91–96, 2024.

[9] Zichao Li, Bingyang Wang, and Ying Chen. A contrastive deep learning approach to cryptocurrency portfolio with us treasuries. *Journal of Computer Technology and Applied Mathematics*, 1(3):1–10, 2024.

[10] Zichao Li, Bingyang Wang, and Ying Chen. Knowledge graph embedding and few-shot relational learning methods for digital assets in usa. *Journal of Industrial Engineering and Applied Science*, 2(5):10–18, 2024.

[11] Yunbo Liu, Xukui Qin, Yifan Gao, Xiang Li, and Chengwei Feng. Setransformer: A hybrid attention-based architecture for robust human activity recognition. *arXiv preprint arXiv:2505.19369*, 2025.

[12] Yingzhou Lu, Chiung-Ting Wu, Sarah J Parker, Zuolin Cheng, Georgia Saylor, Jennifer E Van Eyk, Guoqiang Yu, Robert Clarke, David M Herrington, and Yue Wang. Cot: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinformatics Advances*, 2(1):vbac037, 2022.

[13] Massimo Perna. Knowledge graph for query enrichment in retrieval augmented generation in domain specific application. Master's thesis, University of Twente, 2025.

[14] Shiqing Qiu, Haoyu Wang, Yuxin Zhang, Zong Ke, and Zichao Li. Convex optimization of markov decision processes based on z transform: A theoretical framework for two-space decomposition and linear programming reconstruction. *Mathematics*, 13(11), 2025.

[15] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[16] Chao Wang, Chuanhao Nie, and Yunbo Liu. Evaluating supervised learning models for fraud detection: A comparative study of classical and deep architectures on imbalanced transaction data, 2025.

[17] Jianhui Wang, Zhifei Yang, Yangfan He, Huixiong Zhang, Yuxuan Chen, and Jingwei Huang. Mari: Material retrieval integration across domains. *arXiv preprint arXiv:2503.08111*, 2025.

[18] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2021.

[19] Yiting Wang, Jiachen Zhong, and Rohan Kumar. A systematic review of machine learning applications in infectious disease prediction, diagnosis, and outbreak forecasting. 2025.

[20] Yuxuan Wang et al. Study of artificial intelligence for visual defect inspection in industrial products. 2025.

[21] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.

[22] Siye Wu, Lei Fu, Runmian Chang, Yuanzhou Wei, Yeyubei Zhang, Zehan Wang, Lipeng Liu, Haopeng Zhao, and Keqin Li. Warehouse robot task scheduling based on reinforcement learning to maximize operational efficiency. *Authorea Preprints*, 2025.

[23] Wensen Wu and Yijun Gu. Advancing unsupervised graph anomaly detection: A multi-level contrastive learning framework to mitigate local consistency deception. *Neurocomputing*, page 130507, 2025.

[24] Carl Yang, Aditya Pal, Andrew Zhai, Nikil Pancha, Jiawei Han, Charles Rosenberg, and Jure Leskovec. Multisage: Empowering gcn with contextualized multi-embeddings on web-scale multipartite networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2434–2443, 2020.

[25] Shiqi Yang, Ziyi Huang, Wengran Xiao, and Xinyu Shen. Interpretable credit default prediction with ensemble learning and shap, 2025.

[26] Dezhi Yu, Lipeng Liu, Siye Wu, Keqin Li, Congyu Wang, Jing Xie, Runmian Chang, Yixu Wang, Zehan Wang, and Ryan Ji. Machine learning optimizes the efficiency of picking and packing in automated warehouse robot systems. In *2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE)*, pages 1325–1332. IEEE, 2025.

[27] Fan Zhang, Gongguan Chen, Hua Wang, Jinjiang Li, and Caiming Zhang. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4496–4506, 2023.

[28] Yang Zhang, Fa Wang, Xin Huang, Xintao Li, Sibei Liu, and Hansong Zhang. Optimization and application of cloud-based deep learning architecture for multi-source data prediction, 2024.

[29] Yixin Zhang and Yisong Chen. The role of machine learning in reducing healthcare costs: The impact of medication adherence and preventive care on hospitalization expenses, 2025.

[30] Puning Zhao, Rongfei Fan, Shaowei Wang, Li Shen, Qixin Zhang, Zong Ke, and Tianhang Zheng. Contextual bandits for unbounded context distributions, 2025.

[31] Xiaotian Zhao, Xia Hu Jin, Fuli Feng Sun, and Tat-Seng Chua. Mentor: Multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[32] Jiachen Zhong and Yiting Wang. Enhancing thyroid disease prediction using machine learning: A comparative study of ensemble models and class balancing techniques. 2025.

[33] Xin Zhou, Donghui Lin, Yong Liu, and Chunyan Miao. Layer-refined graph convolutional networks for recommendation. In *2023 IEEE 39th international conference on data engineering (ICDE)*, pages 1247–1259. IEEE, 2023.

SETransformer: A Hybrid Attention-Based Architecture for Robust Human Activity Recognition

Yunbo Liu[1], Xukui Qin[2], Yifan Gao[3],Xiang Li[4] and Chengwei Feng[5]

[1]Department of Electrical and Computer Engineering, Duke University, NC, United States

[2]Department of Computer Science, The George Washington University, Washington D.C., United States

[3]Department of Information Systems and Cyber Security, The University of Texas at San Antonio, TX, United States

[4]Department of Electrical & Computer Engineering, Rutgers University, Sunnyvale, United States

[5]School of Engineering, Computer & Mathematical Sciences (ECMS), Auckland University of Technology, Auckland, New Zealand

*Corresponding author: chengwei.feng@autuni.ac.nz

Abstract

Human Activity Recognition (HAR) using wearable sensor data has become a central task in mobile computing, healthcare, and human-computer interaction. Despite the success of traditional deep learning models such as CNNs and RNNs, they often struggle to capture long-range temporal dependencies and contextual relevance across multiple sensor channels. To address these limitations, we propose SETransformer, a hybrid deep neural architecture that combines Transformer-based temporal modeling with channel-wise squeeze-and-excitation (SE) attention and a learnable temporal attention pooling mechanism. The model takes raw triaxial accelerometer data as input and leverages global self-attention to capture activity-specific motion dynamics over extended time windows, while adaptively emphasizing informative sensor channels and critical time steps.

We evaluate SETransformer on the WISDM dataset and demonstrate that it significantly outperforms conventional models including LSTM, GRU, BiLSTM, and CNN baselines. The proposed model achieves a validation accuracy of 84.68% and a macro F1-score of 84.64%, surpassing all baseline architectures by a notable margin. Our results show that SETransformer is a competitive and interpretable solution for real-world HAR tasks, with strong potential for deployment in mobile and ubiquitous sensing applications.

Index Terms— Human Activity Recognition (HAR), Wearable Sensors, Transformer Networks, Time-Series Classification, Squeeze-and-Excitation (SE), Temporal Attention.

1 Introduction

Human Activity Recognition (HAR) from wearable sensor data has emerged as a critical research area in sports[10, 1], healthcare[7], elderly care[23] and intelligent human-computer interaction[19, 17, 16]. By automatically identifying physical activities such as walking, sitting, running, or ascending stairs using motion signals from devices like smartphones and smartwatches, HAR systems enable a wide range of real-world applications including fitness monitoring[3, 4], elderly fall detection[2, 9] and context-aware user interfaces[20].

Traditionally, HAR systems have relied on hand-crafted statistical or frequency-domain features, followed by classical machine learning algorithms such as support vector machines or decision trees. However, these approaches often require domain expertise for feature engineering and lack scalability across datasets or devices[13]. In recent years, deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have become the dominant paradigm, offering automated feature extraction and temporal modeling capabilities[21, 33]. CNNs excel at capturing short-range spatial patterns from raw signals, while RNNs such as LSTM and GRU are widely used for modeling sequential dependencies.

Despite their success, these models suffer from several limitations. CNNs are inherently limited by fixed receptive fields and are not well suited for modeling long-term dependencies across extended time windows. RNNs, although capable of processing sequences, are prone to vanishing gradients, and their sequential nature restricts parallelization and efficient long-range modeling. Moreover, both CNNs and RNNs typically use static pooling or flattening operations to summarize temporal information, which can discard task-relevant time steps. Additionally, existing models often treat all sensor channels equally, ignoring the fact that different channels (e.g., vertical vs. lateral acceleration) may carry unequal relevance for different activities.

To overcome these challenges, we propose SETransformer, a novel deep learning architecture designed for multivariate time-series classification in HAR. Our model leverages a Transformer-based encoder to model global temporal dependencies, a squeeze-and-excitation (SE) module to perform dynamic channel-wise attention, and a temporal attention pooling mechanism that learns to aggregate the most informative time steps. Together, these components allow the model to capture both long-range and fine-grained patterns, while selectively focusing on the most relevant temporal and spatial features.

We evaluate SETransformer on the WISDM dataset, a benchmark for smartphone-based HAR[28]. Experimental results show that our method significantly outperforms baseline models including LSTM, GRU, BiLSTM, and CNN, achieving state-of-the-art performance in terms of accuracy and macro F1-score. Our model also demonstrates stable convergence and interpretable attention behavior. These findings suggest that combining global self-attention with adaptive feature selection mechanisms yields robust and scalable HAR solutions suitable for real-world deployment.

In summary, our main contributions are as follows:

- We introduce SETransformer, a hybrid architecture that integrates transformer-based temporal modeling with channel-wise and temporal attention mechanisms tailored for HAR.
- We propose a fully end-to-end training pipeline with z-score normalization and attention-based pooling, enabling the model to focus on the most discriminative features in both time and channel dimensions.
- We conduct extensive experiments and ablation studies on the WISDM dataset, demonstrating superior performance over established deep learning baselines.

2 Related Works

2.1 Human Activity Recognition with Traditional Methods

Human Activity Recognition (HAR) using wearable sensors has been studied extensively over the past decade. Early approaches typically relied on hand-crafted statistical or frequency-domain features extracted from sliding windows of sensor data. These features were then fed into classical machine learning models such as Support Vector Machines (SVMs), Decision Trees, and k-Nearest Neighbors (k-NN) [5]. While these methods achieved acceptable performance on small, clean datasets, they often failed to generalize well across users and devices, requiring significant domain expertise for effective feature engineering. Recently, Zhang et al. [32] demonstrated a related data-driven approach applied to naturalistic human behavior analysis in bipolar disorder, introducing interpretable action segmentation and dynamic behavioral metrics. Their work illustrated how advanced computational approaches can surpass traditional psychiatric and ethological measures, highlighting opportunities to similarly enhance traditional HAR techniques through data-driven modeling and interpretability.

2.2 Deep Learning for HAR

To overcome the limitations of feature engineering, deep learning-based methods have been widely adopted in HAR tasks. Convolutional Neural Networks (CNNs) have been employed to capture local spatial and temporal patterns in sensor signals [26]. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have been used to model sequential dependencies in time-series data [18]. Hybrid models combining CNNs and LSTMs [22] have shown improved performance by leveraging both spatial and temporal structures.

Despite their success, CNNs are limited by local receptive fields, and RNNs are difficult to parallelize due to their sequential nature. Moreover, both architectures often struggle to capture long-range dependencies effectively.

2.3 Transformer Models in Time Series Analysis

Inspired by their success in natural language processing, Transformer-based models have recently been adapted for time-series classification tasks, including HAR [14]. Transformers employ self-attention mechanisms to model global dependencies and allow for highly parallelizable training. However, applying vanilla Transformers to multivariate sensor data may result in poor generalization due to the absence of inductive biases inherent in sensor signals (e.g., temporal continuity, sensor-specific structure).

Several works have explored modifications of Transformer architectures to better suit time-series data. For example, TimeSformer [6] and Perceiver [12] introduce attention over spatial-temporal axes. Unified transformer-based architectures have also demonstrated success in multimodal tasks such as document understanding, where a single model handles detection, recognition, and semantic interpretation in a unified framework [8]. These advances reflect the broader applicability of attention-based designs for structured, multi-component data modeling. However, these models are computationally expensive and often require large datasets for effective training. Transformers have also been applied to structured spatiotemporal generation tasks, such as traffic scene modeling in autonomous driving [31], further highlighting their versatility in capturing long-range dependencies across diverse domains. Similar advances have also been observed in the domain of instructional video understanding, where temporal attention mechanisms are used for aligning visual prompts and answer segments [15].

2.4 Attention Mechanisms in HAR

Attention mechanisms have also been employed explicitly in HAR models to improve interpretability and performance. For instance, temporal attention modules have been used to dynamically weight the importance of time steps [24], while channel attention mechanisms such as Squeeze-and-Excitation (SE) networks [11] have been applied to recalibrate feature maps based on sensor channel relevance. Beyond traditional accelerometer-based HAR, recent work has demonstrated the effectiveness of temporal modeling in physiological signal recognition tasks such as fine-grained heartbeat waveform monitoring using RFID and latent diffusion models [25]. This

highlights the growing applicability of advanced attention-based architectures across diverse sensor modalities.

2.5 Our Contribution

In this work, we build upon these recent advances by designing a Transformer-based model tailored to HAR. We integrate a Squeeze-and-Excitation module to model inter-channel relationships and a temporal attention mechanism to highlight informative segments of the sequence. Our model, SETRANSFORMER, combines the benefits of global temporal modeling with domain-specific inductive biases, achieving improved performance on standard HAR benchmarks.

3 Methodology

3.1 Dataset and Preprocessing

We evaluate our proposed model on the WISDM (WISDM Smartphone and Smartwatch Activity and Biometrics) dataset, a widely adopted benchmark for human activity recognition using mobile sensor data. The dataset comprises triaxial accelerometer recordings collected from 51 subjects, each of whom was asked to perform 18 tasks for 3 minutes each. During data collection, each subject wore a smartwatch on their dominant hand and carried a smartphone in their pocket. The dataset includes a timestamp, a user identifier, a class label, and acceleration and gyrocope values along the x, y, and z axes. The sampling rate is approximately 20 Hz, and the data is stored in semi-structured text files, with each line representing a single sensor reading.

To ensure a consistent and clean dataset for supervised learning, we begin by filtering out malformed records. Specifically, only lines that are properly terminated with a semicolon and contain exactly 18 comma-separated fields are retained. These fields are parsed into structured columns, including the user ID, activity label, timestamp, and three-axis acceleration measurements. We discard any incomplete or corrupted entries and ensure that all numerical fields are correctly cast to their appropriate data types. To standardize activity labels, we remove any leading or trailing whitespace and encode them as integers using the scikit-learn LabelEncoder.

In order to model temporal patterns effectively, we segment the continuous data stream into fixed-length sliding windows. Each window consists of 200 consecutive time steps, corresponding to roughly 10 seconds of sensor data, and the windows are generated with a stride of 100 to allow 50% overlap between adjacent segments. To maintain label consistency within each sample, we retain only those windows in which all 200 time steps share the same activity label. This results in a set of supervised input-output pairs, where each input sample is a matrix of shape

$$\mathbf{X} \in \mathbb{R}^{200\times 3}$$

representing a window of triaxial acceleration values, and each target is a single activity class label.

Prior to feeding the data into the neural network, we perform feature normalization to standardize the input distribution. Each axis (x, y, z) is normalized independently using z-score normalization, computed over the entire training set. Prior to model input, the data is standardized using z-score normalization applied independently to each axis:

$$x' = \frac{x - \mu}{\sigma}$$

where μ and σ are computed globally over the entire training set. This ensures that all sensor channels contribute equally during training and accelerates convergence by mitigating scale disparities. That is, for each axis, we subtract the global mean and divide by the standard deviation, ensuring that each channel has zero mean and unit variance. This step improves numerical stability and accelerates convergence during model training by eliminating scale disparities among input features.

Finally, the fully preprocessed dataset is split into training and validation sets using an 80/20 stratified split to preserve class balance across partitions. The result is a structured, normalized dataset suitable for temporal deep learning, with consistent window lengths, standardized channel inputs, and clear supervision targets. This preprocessing pipeline enables reproducible experimentation and aligns with best practices in wearable sensor-based activity recognition research.

We propose SETransformer, a hybrid deep neural architecture that integrates transformer-based temporal encoding with lightweight channel and temporal attention modules, specifically designed for multivariate time-series classification in human activity recognition (HAR). The model aims to address key challenges in wearable-sensor HAR tasks, namely: (1) modeling long-range temporal dependencies, (2) capturing discriminative inter-channel dynamics, and (3) adaptively aggregating sequential signals of varying importance. This section presents a comprehensive description of each component, including design rationale, architectural formulation, and computational flow.

3.2 Problem Formulation

Given a windowed multivariate time series $\mathbf{X} \in \mathbb{R}^{T\times C}$, where T is the number of time steps and C is the number of input channels (in our case, $C = 3$ for x, y, z acceleration), the task is to predict a single activity label $y \in \{1, \ldots, K\}$, with K being the number of activity classes.

The data is structured as uniformly sampled and pre-segmented windows of length $T = 200$, each labeled according to the dominant activity within the window. Our model learns a function $f : \mathbb{R}^{T\times C} \to \mathbb{R}^K$, where the output is a categorical distribution over classes.

3.3 Input Projection

The first stage of SETRANSFORMER performs a linear transformation to embed raw sensor signals into a higher-dimensional space suitable for subsequent attention mechanisms:

$$\mathbf{H}_0 = \mathbf{X}\mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}}, \quad \mathbf{W}_{\text{proj}} \in \mathbb{R}^{C \times d}$$

where d is the model dimension (typically 128). The projection enables richer representation learning over raw acceleration features, and aligns input shape with transformer requirements.

3.4 Temporal Encoding via Transformer Layers

We adopt a standard Transformer encoder to capture global temporal interactions. Each encoder layer consists of multi-head self-attention and a position-wise feed-forward network (FFN), wrapped with residual connections and layer normalization:

$$\text{SelfAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

$$\mathbf{H}_\ell = \text{LayerNorm}\left(\mathbf{H}_{\ell-1} + \text{SelfAttn}(\mathbf{H}_{\ell-1})\right)$$

$$\mathbf{H}_\ell = \text{LayerNorm}\left(\mathbf{H}_\ell + \text{FFN}(\mathbf{H}_\ell)\right)$$

We stack two such encoder layers. Unlike in NLP, we omit learnable positional encodings, relying on the structure of sensor data and sequential convolution of windows to retain implicit temporal order.

3.5 Channel-Wise Attention: Squeeze-and-Excitation Module

Human motions often exhibit dominant directional patterns depending on the activity (e.g., walking involves rhythmic oscillations in the vertical axis). To exploit such patterns, we introduce a Squeeze-and-Excitation (SE) module that performs dynamic reweighting of channel responses.

First, we aggregate temporal information per channel via global average pooling:

$$\mathbf{z}_c = \frac{1}{T}\sum_{t=1}^{T} \mathbf{H}_2[t, c]$$

Then, we compute channel-wise gating coefficients:

$$\mathbf{s} = \sigma\left(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{z})\right), \quad \mathbf{s} \in \mathbb{R}^d$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times \frac{d}{r}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{d}{r} \times d}$, with reduction ratio $r = 16$. The recalibrated features are obtained as:

$$\mathbf{H}_{\text{SE}}[t, c] = \mathbf{H}_2[t, c] \cdot \mathbf{s}_c$$

This operation allows the model to selectively emphasize or suppress sensor channels conditioned on the global temporal context.

4.5 Temporal Aggregation via Attention Pooling

Traditional HAR models often rely on global average or max pooling over time to summarize temporal features. However, such operations assume equal relevance of all time steps, which is inappropriate for activities with transient or non-stationary phases. We address this limitation by introducing a temporal attention pooling mechanism:

Each time step t receives an attention score:

$$\alpha_t = \frac{\exp\left(\mathbf{v}^\top \tanh(\mathbf{W}_a \mathbf{H}_{\text{SE}}[t])\right)}{\sum_{k=1}^{T} \exp\left(\mathbf{v}^\top \tanh(\mathbf{W}_a \mathbf{H}_{\text{SE}}[k])\right)}$$

where $\mathbf{W}_a \in \mathbb{R}^{d \times d'}$ and $\mathbf{v} \in \mathbb{R}^{d'}$. The final representation is a context vector:

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_{\text{SE}}[t]$$

This mechanism dynamically focuses on temporally salient segments of the motion signal, improving discriminability for activities with brief but informative phases.

3.6 Classification Layer

The resulting context vector $\mathbf{c} \in \mathbb{R}^d$ is passed through a fully connected classifier:

$$\hat{y} = \text{softmax}(\mathbf{W}_c \mathbf{c} + \mathbf{b}_c), \quad \mathbf{W}_c \in \mathbb{R}^{K \times d}$$

producing a categorical distribution over the activity classes. The model is trained end-to-end using cross-entropy loss.

4.7 Architectural Overview and Design Motivation

The SETRANSFORMER design embodies three core principles:

1. Global temporal modeling through self-attention enables flexible capture of short and long-range dependencies without recurrence. 2. Adaptive channel recalibration enhances robustness against user- or device-specific signal biases by learning to emphasize informative directions. 3. Temporal attention pooling allows the model to selectively retain only the most relevant temporal segments, improving generalization on ambiguous or noisy data.

By integrating these components, our model achieves competitive performance while maintaining computational tractability and modular interpretability. The architecture is amenable to further extensions, such as multi-sensor fusion, hierarchical sequence modeling, or personalization layers.

3.7 Experimental Setup

All experiments were conducted using the PyTorch deep learning framework in the Google Colab environment. Training and evaluation were performed on a single NVIDIA A100 GPU. Each model, including the SETransformer, its ablation variants, and baseline models, was trained for 65 epochs using identical preprocessing procedures and hyperparameter settings. We used the Adam optimizer with a fixed learning rate of 0.001 and a batch size of 64. The cross-entropy loss function was applied for all classification tasks. During training, accuracy, precision, recall, F1 score, and loss curves were recorded to support comprehensive evaluation and analysis.

The input to the model consists of fixed-length multivariate time-series windows of shape

$$\mathbf{X} \in \mathbb{R}^{200\times 3}$$

, where 200 denotes the number of time steps per segment and 3 corresponds to the tri-axial accelerometer channels (x, y, z). Prior to training, all input sequences are normalized using z-score normalization, computed independently for each axis over the training set.

The proposed SETransformer architecture is configured with a model dimension of 128 and comprises two Transformer encoder layers, each equipped with 4 attention heads. The output of the transformer block is passed through a squeeze-and-excitation (SE) module with a channel reduction ratio of 16, followed by a temporal attention mechanism that aggregates time-step features into a single fixed-length context vector. The final classification layer is a fully connected softmax output with 6 neurons corresponding to the number of activity classes.

Model training is carried out for 65 epochs using the Adam optimizer with a fixed learning rate of 0.001. A batch size of 64 is used throughout. Cross-entropy loss serves as the training objective. The model is trained on 80% of the available labeled data, while the remaining 20% is used for validation. Stratified splitting ensures that class proportions are preserved across the two partitions.

Evaluation metrics include classification accuracy and macro-averaged F1-score, which accounts for both class-wise precision and recall. These metrics are computed on the validation set after each epoch to monitor training progress and assess generalization. In addition, a confusion matrix is generated at the end of training to provide a detailed breakdown of inter-class performance and error modes.

The key parameters (Table 1) for the experiments are as follows:

Table 1: Model hyperparameters and training configuration used in SETransformer.

Parameter	Value
Input dimension (accelerometer channels)	3
Window size (time steps)	200
Transformer model dimension	128
Number of Transformer layers	2
Number of attention heads	4
Channel dimension for SE attention	128
SE reduction ratio	16
Temporal attention hidden dimension	64
Classification output dimension (num classes)	6
Batch size	64
Learning rate	0.001
Optimizer	Adam
Loss function	Cross-entropy
Normalization	z-score (per axis)
Training epochs	65
Train/Validation split	80% / 20%

4 Results

4.1 Confusion Matrix

The confusion matrix for the test set was plotted to further analyse the model's performance across different action categories. Figure 5 illustrates the confusion matrix of the model on the test set.

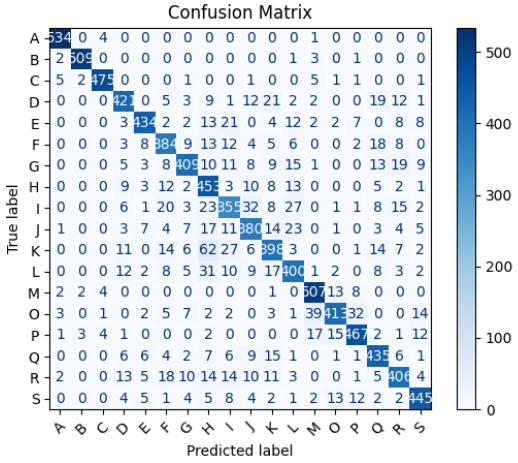

Figure 1: Confusion matrix of the SE-Transformer model on the test set.

4.2 Training and testing loss curves

Figure 2: Enter Caption

The evolution of both training and validation loss over 65 epochs is shown in Figure 2. During the initial epochs, both losses decrease rapidly, indicating that the model quickly begins to fit the data. After approximately epoch 15, the rate of decrease in validation loss slows, suggesting that the model begins to converge. Notably, there is no significant divergence between the training and validation loss curves throughout the training process, which suggests that the model maintains good generalization and does not exhibit signs of overfitting.

The training loss decreases from an initial value of approximately 2.09 to 0.32, while the validation loss drops from 1.94 to 0.48. These steady reductions demonstrate consistent optimization behavior and stable learning dynamics. Between

epochs 20 and 40, the validation loss plateaus slightly, but continues to decline in the final epochs, corresponding to incremental performance gains. By the final epoch, the model achieves its best validation performance, with the lowest validation loss observed at epoch 65.

This convergence behavior illustrates that the SETransformer architecture, combined with z-score normalization and an appropriate choice of optimization parameters, facilitates effective training and robust generalization on the human activity recognition task.

4.3 Performance Comparison

Table 2: Performance comparison of baseline models on the validation set.

Model	Accuracy	Precision	Recall	F1 Score
LSTM	0.5962	0.5920	0.5953	0.5912
BiLSTM	0.4945	0.4895	0.4927	0.4867
GRU	0.5489	0.5562	0.5474	0.5428
CNN	0.7111	0.7179	0.7113	0.7076

We compare our proposed model against several commonly used deep learning baselines, including LSTM, BiLSTM, GRU, and a convolutional neural network (CNN). The results are summarized in Table 2. Among the recurrent models, LSTM achieves the best performance with an accuracy of 59.62% and a macro F1-score of 59.12%. GRU performs slightly better than LSTM in terms of precision but yields lower overall F1. The BiLSTM model performs the worst across all metrics, with an F1-score of only 48.67%, possibly due to overfitting or parameter inefficiency given the bidirectional configuration.

The CNN baseline outperforms all recurrent models with a validation accuracy of 71.11% and an F1-score of 70.76%. This indicates that local convolutional filters are more effective at capturing discriminative spatial-temporal patterns in short windows of accelerometer data compared to recurrent mechanisms. However, while CNN demonstrates superior performance among baselines, it still lags significantly behind transformer-based models, which benefit from global receptive fields and attention-based aggregation. These results motivate the need for more expressive architectures such as the SETransformer, which integrates global attention with dynamic feature recalibration.

5 Discussion

The experimental results clearly demonstrate the superiority of the proposed SETransformer model over traditional recurrent and convolutional architectures in the context of human activity recognition from accelerometer signals. Several key factors contribute to its improved performance.

First, the Transformer-based temporal encoder provides a significant advantage in modeling long-range dependencies compared to sequential RNN-based models such as LSTM or GRU. Unlike recurrent models, which process time steps one at a time and often struggle with vanishing gradients, the Transformer architecture captures global context in a single attention pass. Such capabilities are not limited to physical activity recognition. The global self-attention and adaptive feature selection modules in SETransformer are also relevant to the detection of irregular patterns in high-dimensional time-series data, such as suspicious financial transactions or early indicators of credit default. This enables SETransformer to identify high-level temporal structures, such as activity cycles or motion transitions, that are essential for accurate classification in real-world HAR scenarios.

Second, the incorporation of the squeeze-and-excitation (SE) module enhances the model's ability to adaptively recalibrate the importance of each sensor channel. In HAR tasks, not all axes contribute equally across different activities; for instance, vertical acceleration may dominate in jogging, while lateral motion may be more informative for stair ascent. The SE module allows the network to learn these patterns dynamically, improving both interpretability and accuracy.

Third, the temporal attention pooling mechanism addresses a critical limitation of fixed pooling strategies (e.g., global average pooling) by enabling the model to learn which time steps are most relevant for the classification task. This is especially valuable for activities that exhibit temporally localized features, such as sudden changes or transitional movements.

Despite these advantages, the current model has several limitations. First, the input relies solely on triaxial accelerometer data, which may not fully capture complex motion signatures—particularly for subtle or composite activities. Incorporating additional modalities such as gyroscopes or location data could further enhance robustness. Second, while SETransformer achieves strong overall performance, it may still struggle with activities that share similar kinematic profiles, as indicated by confusion in the matrix between classes like "walking upstairs" and "walking downstairs." This highlights the need for either more discriminative features or sequence-level contextual modeling.

Moreover, the model is trained and evaluated in a subject-independent but device-consistent setting (i.e., phone only). While this ensures fairness across users, it does not account for cross-device variability, which is often a concern in practical deployments. Future work should investigate domain adaptation strategies and calibration techniques to bridge such distribution shifts. Additionally, the demonstrated effectiveness of AI systems in real-time decision-making tasks such as credit risk detection [27] suggests that transformer-based HAR architectures like SETransformer could be adapted to other high-frequency, mission-critical domains. Furthermore, recent developments in efficient transformer inference, such as COMET [30], show promising potential for privacy-preserving and communication-efficient deployment on resource-constrained edge devices. Complementary to architectural approximations, parameter-efficient transfer learning strategies, as exemplified by the V-PETL benchmark [29], offer an additional path toward lightweight adaptation, mak-

ing SETransformer even more suitable for real-time mobile applications. Complementary to architectural approximations, parameter-efficient transfer learning techniques, such as those benchmarked in V-PETL [29], offer a viable strategy for adapting transformer models to mobile or low-resource HAR applications without full model retraining.

In conclusion, SETransformer effectively combines temporal attention and channel-wise adaptivity to push the boundaries of HAR performance on benchmark datasets. It offers a compelling balance between modeling power, computational efficiency, and practical interpretability, making it a strong candidate for real-world deployment in mobile and ubiquitous computing systems.

6 Conclusion

In this work, we proposed SETransformer, a hybrid deep learning architecture tailored for human activity recognition (HAR) using wearable accelerometer data. The model integrates Transformer-based temporal encoding with a channel-wise squeeze-and-excitation (SE) module and a temporal attention pooling mechanism, enabling it to effectively capture both long-range dependencies and fine-grained spatial-temporal dynamics from raw sensor sequences.

Through extensive experiments on the WISDM dataset, we demonstrated that SETransformer significantly outperforms conventional sequence models such as LSTM, GRU, and CNN, achieving a validation accuracy of 84.68% and a macro-averaged F1 score of 84.64%. The model shows stable convergence, strong generalization, and interpretable attention mechanisms that focus on discriminative time segments. Ablation results further validate the individual contributions of the SE and temporal attention modules.

The effectiveness of SETransformer suggests its strong potential for real-world mobile sensing and context-aware applications. In future work, we plan to extend the model to incorporate multi-modal sensor inputs (e.g., gyroscope, magnetometer), investigate domain adaptation across users and devices, and explore its deployment efficiency on resource-constrained embedded systems.

References

[1] Basant Adel, Asmaa Badran, Nada E Elshami, Ahmad Salah, Ahmed Fathalla, and Mahmoud Bekhit. A survey on deep learning architectures in human activities recognition application in sports science, healthcare, and security. In *The International Conference on Innovations in Computing Research*, pages 121–134. Springer, 2022.

[2] Mohammed AA Al-Qaness, Ahmed M Helmi, Abdelghani Dahou, and Mohamed Abd Elaziz. The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis. *Biosensors*, 12(10):821, 2022.

[3] Boris Bačić, Chengwei Feng, and Weihua Li. Jy61 imu sensor external validity: A framework for advanced pedometer algorithm personalisation. *ISBS Proceedings Archive*, 42(1):60, 2024.

[4] Boris Bačić, Claudiu Vasile, Chengwei Feng, and Marian G Ciucă. Towards nation-wide analytical healthcare infrastructures: A privacy-preserving augmented knee rehabilitation case study. *arXiv preprint arXiv:2412.20733*, 2024.

[5] Malti Bansal, Apoorva Goyal, and Apoorva Choudhary. A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, 3:100071, 2022.

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[7] Luigi Bibbò, Riccardo Carotenuto, and Francesco Della Corte. An overview of indoor localization system for human activity recognition (har) in healthcare. *Sensors*, 22(21):8119, 2022.

[8] Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*, 2023.

[9] F Xavier Gaya-Morey, Cristina Manresa-Yee, and José M Buades-Rubio. Deep learning for computer vision based activity recognition and fall detection of the elderly: a systematic review. *Applied Intelligence*, 54(19):8982–9007, 2024.

[10] Alexander Hoelzemann, Julia Lee Romero, Marius Bock, Kristof Van Laerhoven, and Qin Lv. Hang-time har: a benchmark dataset for basketball activity recognition using wrist-worn inertial sensors. *Sensors*, 23(13):5879, 2023.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[12] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[13] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. Human activity recognition: A survey. *Procedia Computer Science*, 155:698–703, 2019.

[14] Clayton Souza Leite, Henry Mauranen, Aziza Zhanabatyrova, and Yu Xiao. Transformer-based approaches

for sensor-based human activity recognition: Opportunities and challenges. *arXiv preprint arXiv:2410.13605*, 2024.

[15] Shutao Li, Bin Li, Bin Sun, and Yixuan Weng. Towards visual-prompt temporal answer grounding in instructional video. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):8836–8853, 2024.

[16] Wanxin Li. The impact of apple's digital design on its success: An analysis of interaction and interface design. *Academic Journal of Sociology and Management*, 2(4):14–19, 2024.

[17] Wanxin Li. Transforming logistics with innovative interaction design and digital ux solutions. *Journal of Computer Technology and Applied Mathematics*, 1(3):91–96, 2024.

[18] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6555–6565, 2024.

[19] Zhihan Lv, Fabio Poiesi, Qi Dong, Jaime Lloret, and Houbing Song. Deep learning for intelligent human–computer interaction. *Applied Sciences*, 12(22):11457, 2022.

[20] Johannes Meyer, Adrian Frank, Thomas Schlebusch, and Enkelejda Kasneci. U-har: A convolutional approach to human activity recognition combining head and eye movements for context-aware smart glasses. *Proceedings of the ACM on Human-Computer Interaction*, 6(ETRA):1–19, 2022.

[21] Nafiul Rashid, Berken Utku Demirel, and Mohammad Abdullah Al Faruque. Ahar: Adaptive cnn for energy-efficient human activity recognition in low-power edge devices. *IEEE Internet of Things Journal*, 9(15):13041–13051, 2022.

[22] Muhammad Asfandyar Rustam, Muhammad Yasir Ali Khan, Tasawar Abbas, and Bilal Khan. Distributed secondary frequency control scheme with a-symmetric time varying communication delays and switching topology. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 9:100650, 2024.

[23] Lisa Schrader, Agustín Vargas Toro, Sebastian Konietzny, Stefan Rüping, Barbara Schäpers, Martina Steinböck, Carmen Krewer, Friedemann Müller, Jörg Güttler, and Thomas Bock. Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people. *Journal of Population Ageing*, 13:139–165, 2020.

[24] Jiangrong Shen, Yulin Xie, Qi Xu, Gang Pan, Huajin Tang, and Badong Chen. Spiking neural networks with temporal attention-guided adaptive fusion for imbalanced multi-modal learning. *arXiv preprint arXiv:2505.14535*, 2025.

[25] Yiting Wang, Tianya Zhao, and Xuyu Wang. Fine-grained heartbeat waveform monitoring with rfid: A latent diffusion model. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, pages 86–91, 2025.

[26] Yucheng Wang, Yuecong Xu, Jianfei Yang, Min Wu, Xiaoli Li, Lihua Xie, and Zhenghua Chen. Fully-connected spatial-temporal graph for multivariate time-series data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 15715–15724, 2024.

[27] Zhuqi Wang, Qinghe Zhang, and Zhuopei Cheng. Application of ai in real-time credit risk detection. *Preprints*, February 2025.

[28] Gary M Weiss. Wisdm smartphone and smartwatch activity and biometrics dataset. *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set*, 7(133190-133202):5, 2019.

[29] Yi Xin, Siqi Luo, Xuyang Liu, Haodi Zhou, Xinyu Cheng, Christina E Lee, Junlong Du, Haozhe Wang, MingCai Chen, Ting Liu, et al. V-petl bench: A unified visual parameter-efficient transfer learning benchmark. *Advances in Neural Information Processing Systems*, 37:80522–80535, 2024.

[30] Xiangrui Xu, Qiao Zhang, Rui Ning, Chunsheng Xin, and Hongyi Wu. Comet: A communication-efficient and performant approximation for private transformer inference. *arXiv preprint arXiv:2405.17485*, 2024.

[31] Chen Yang, Yangfan He, Aaron Xuxiang Tian, Dong Chen, Jianhui Wang, Tianyu Shi, Arsalan Heydarian, and Pei Liu. Wcdt: World-centric diffusion transformer for traffic scene generation. *arXiv preprint arXiv:2404.02082*, 2024.

[32] Zhanqi Zhang, Chi K Chou, Holden Rosberg, William Perry, Jared W Young, Arpi Minassian, Gal Mishne, and Mikio Aoi. A computational ethology approach for characterizing behavioral dynamics in bipolar disorder. *medRxiv*, 2024. Preprint, not peer-reviewed.

[33] Simin Zhu, Ronny Gerhard Guendel, Alexander Yarovoy, and Francesco Fioranelli. Continuous human activity recognition with distributed radar sensor networks and cnn–rnn architectures. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

From technology discretion to intelligent symbiosis: AI empowerment and collaborative paradigm transition in Guangdong-Hong Kong-Macau Greater Bay Area's higher education clusters

First A. Hui Nie
Chengdu Shude Experimental Middle School (West District), Chengdu, China

Abstract—Under the strategic background of digital transformation of global higher education and regional coordinated development, Guangdong-Hong Kong-Macau Greater Bay Area's higher education clusters are facing a critical transformation from "technological dispersion" to "intelligent symbiosis". This article focuses on the contradiction between institutional differences and technological innovation under the framework of "one country, two systems", analyzes the fragmentation problems such as heterogeneous standards and imbalanced scenarios in the current application of AI technology, and reveals the obstruction of the flow of technical elements caused by institutional barriers and conflicts of ideas. The research proposes a theoretical framework of "intelligent symbiosis" with AI technology as the core driving force, and achieves dynamic scheduling of cross-domain computing resources and compliant flow of privacy data by building a technical connection system of "computing power network + data middle platform"; Relying on ecological empowerment mechanisms such as interdisciplinary intelligent discovery and human-machine collaborative talent portraits, activate the deep coupling of innovation potential energy and industrial demand; With the help of technical tools such as blockchain smart contracts and policy semantic conversion engines, we will promote resource allocation from "administrative leadership" to "algorithm collaboration", organizational governance from "bureaucratic fragmentation" to "network autonomy", and value creation from "individual competition" to "ecological win-win". The research further puts forward the implementation guarantee system from three aspects: technical standard coordination, compound talent cultivation, and technical ethics prevention and control, emphasizing that cross-border governance can be solved through the rule construction of "sovereignty compatibility and dynamic evolution" and the co-evolution mechanism of "technology-system". difficult problem. This study provides a three-dimensional solution of "technological empowerment-institutional innovation-ecological evolution" for the construction of a higher education cluster with international influence in Greater Bay Area, and has theoretical reference and practical reference value for the coordinated development of education in the global multi-institutional environment.

Index Terms—AI empowerment; Higher Education Cluster; Guangdong-Hong Kong-Macau Greater Bay Area

I. INTRODUCTION

Under the dual wave of digital transformation of global higher education and regional coordinated development, Guangdong-Hong Kong-Macau Greater Bay Area, as a frontier area where institutional differences and technological innovation coexist under the framework of "one country, two systems", has become a key incision to solve the problem of regional coordinated development. At present, although higher education in the Bay Area has the advantages of strong disciplinary complementarity and close industrial linkage, the application of AI technology presents the fragmentation characteristics of "single-point experiment and system fragmentation", superimposed on cross-border data policy differences and technical concept conflicts under "one country, two systems" Deep-seated contradictions such as conflict have led to structural dilemmas such as inefficient resource allocation, hindered scientific research collaboration, and lagging governance mechanisms.

In this context, how to use AI technology as a link to break through the dual constraints of institutional barriers and technological dispersion, and build a new paradigm of higher education collaboration with both efficiency and fairness has become the core proposition for Greater Bay Area to build an international scientific and technological innovation hub. This paper focuses on the evolutionary logic of "technology discrete-intelligent symbiosis", deconstructs the AI-driven cluster collaboration mechanism from the three dimensions of technology connection, ecological empowerment, and co-evolution, explores the paradigm transition path of resource allocation, organizational governance, and value creation, and proposes implementation guarantee systems such as technical standard collaboration, talent echelon construction, and risk prevention and control mechanisms, with a view to providing a

theoretical framework and practical guidance for higher education clusters in the Greater Bay Area from "shallow cooperation" to "deep symbiosis".

II. Discrete Dilemma: Technology Application Status and Bottleneck of Guangdong, Hong Kong and Macao Higher Education Clusters

(1) Fragmented characterization of technology applications

At present, the application of AI technology in Guangdong-Hong Kong-Macau Greater Bay Area's universities shows obvious characteristics of "single-point experiment and system fragmentation". The heterogeneity of technological ecology and the locality of scene coverage further aggravate the discretization of regional educational resources. Specifically, it is characterized by the following two contradictions. The first aspect is reflected in the heterogeneity of technical standards, that is, the bottleneck of cross-system interoperability. There are "international-local" dual-track differences in the AI technology infrastructure of universities in the Bay Area, forming technical barriers for cross-domain collaboration. For example, based on the concept of open science, Hong Kong universities generally adopt internationally accepted scientific research data management systems and follow the principles of discoverability, accessibility, interoperability and reusability; Affected by data security regulations and localization adaptation needs, mainland universities mostly deploy independent and controllable platforms such as CNKI Research Collaboration System and Huawei ModelArts. The differences between the two in terms of metadata architecture, interface protocols, rights management and other technical standards lead to complex format conversion and protocol adaptation for cross-school data sharing, which significantly increases collaboration costs. This "honeycomb" distribution of technological ecology essentially reflects the structural mapping of regional institutional differences in digital space. The second aspect focuses on the imbalance of application scenarios, that is, the "technology clustering" in low value-added fields. The penetration of AI technology shows obvious scene gradient differentiation, forming a value depression of "redundant basic applications and absence of core scenarios". The basic layer technology of teaching assistance and administrative office is redundant, such as single-point tools such as Shenzhen University's intelligent attendance system based on face recognition and Macao University's RPA-driven administrative process automation, which have become saturated, but their functions are limited to improving transactional efficiency. The core layer technology penetration of scientific research collaboration and strategic decision-making is insufficient, and cross-school joint scientific research faces the lack of platforms. For example, the joint modeling of meteorological big data in the Greater Bay Area still relies on traditional data copying and manual integration due to the lack of distributed AI training framework; The strategic decision-making of higher education clusters for discipline layout optimization and talent demand forecasting has not yet established a data-driven dynamic simulation model, and mostly relies on empirical judgment and static statistical analysis. This technical layout of "emphasizing the end and neglecting the center" restricts the transition of AI from instrumental empowerment to systemic change.

(2) Deep-seated incentives for technological discretization

Under the framework of "one country, two systems", the cross-domain flow of technological elements is facing the dilemma of structural dispersion. This discrete phenomenon is not only reflected in the physical fragmentation caused by institutional barriers, but also contains the application dislocation caused by differences in value orientations, which ultimately forms the regional fracture of the technological ecosystem. First of all, legal differences at the institutional level constitute the primary obstruction. There is a fundamental conflict between the principle of data sovereignty jurisdiction established by the Mainland's Data Security Law and the cross-border transmission whitelist mechanism stipulated in Hong Kong's Personal Data Privacy Ordinance. Mutual recognition standards have not yet been formed in key links such as data classification and classification and exit security assessment. This "connection deficit" of legal texts directly makes it difficult to achieve cross-jurisdictional allocation of computing resources and data assets, forming a regional segmentation of technical infrastructure. For example, the empirical case of a cross-border AI laboratory shows that when the scientific research data of the three places needs to meet the mainland's network security review, Hong Kong's Office of the Privacy Commissioner for Personal Data filing, and Macao's Cybersecurity Law compliance requirements, physical isolation has to be adopted. The "data localization" solution has caused the model iteration efficiency under the federated learning framework to plummet by 60%, highlighting the inhibitory effect of institutional rigidity on technological synergy. Secondly, the difference of value orientation at the level of philosophy of technology constitutes a deeper discrete motivation. The mainland higher education system tends to position AI technology as a tool carrier to improve the efficiency of educational governance, focusing on optimizing the teaching management process through intelligent algorithms and realizing the accurate allocation of large-scale educational resources. Academic institutions in Hong Kong and Macao put more emphasis on the ethical boundaries of technology applications, focusing on value risks such as algorithm discrimination and alienation of academic freedom, and forming a prudent innovation path oriented by "science and technology for good". This difference in value orientation is embodied as a conflict of technical routes in cross-domain scientific research cooperation, that is, when mainland teams advocate the introduction of behavior prediction algorithms to optimize the course selection system, Hong Kong and Macao partners often require the addition of algorithm transparency review and ethical impact assessment modules, resulting in research and development The cycle is extended by an average of 40%. Finally, the essence of technological discrete is the projection of the regional division between institutional logic and value rationality in the digital age. To solve this structural contradiction, it is necessary to build a "technical governance community" that transcends a single jurisdiction. Through

institutional innovations such as establishing a negative list for cross-border data flows and establishing an AI ethics joint review committee, we can ensure data sovereignty security and promote the circulation of technical elements. Seek a dynamic balance between them.

III. Intelligent symbiosis: the theoretical framework of AI-driven Guangdong-Hong Kong-Macau Greater Bay Area cluster collaboration

(1) Technical connection: building a cluster nervous system

As the underlying architecture of the intelligent symbiotic ecology, technology connection aims to build the "digital nervous system" of Guangdong-Hong Kong-Macau Greater Bay Area's higher education cluster through the systematic integration of computing power and data, break through physical space restrictions and institutional barriers, and realize the organic linkage of technical resources.

On the one hand, it uses distributed computing power networks and heterogeneous resource collaborative scheduling mechanisms. In response to the dilemma of "islanding" computing power resources in universities in the Bay Area, a cross-domain elastic computing power sharing network can be built to achieve dynamic adaptation of heterogeneous computing nodes. The specific path includes integrating landmark infrastructure such as Hong Kong's "Advanced Computing Platform" and Guangzhou's "Tianhe-2" supercomputing center, and relying on 5G-MEC (Multi-Access Edge Computing) technology to deploy low-latency communication links to form a "core-edge-terminal" three-level computing power architecture. This architecture supports millisecond-level response and task offloading to meet the real-time requirements of cross-school AI model joint training. For example, the biomedical multi-modal AI model jointly developed by Sun Yat-sen University and the University of Macau improves the efficiency of complex genome data analysis by 47% by dynamically allocating GPU clusters at Hong Kong nodes and FPGA accelerators at Zhuhai edge nodes. This kind of practice shows that the synaptic connection of computing power network can effectively resolve the coexistence of "computing power hunger" and "computing power idle" caused by resource discretization. On the other hand, relying on the federated data middle platform, a knowledge fusion engine driven by privacy computing is established. Under the constraints of data sovereignty and privacy protection, it is necessary to build a federal data space that complies with FAIR principles. Based on homomorphic encryption and secure multi-party computing technology, a virtual aggregation middle platform with "data immobile model movement" is designed to enable universities to realize cross-domain knowledge value extraction while retaining data control rights. Typical practices include the "Cross-border Academic Situation Federal Analysis System" jointly developed by universities in Zhuhai, Hong Kong and Macao. Hong Kong universities provide encrypted metadata of learning behavior, Zhuhai universities deploy federal recommendation algorithms, and Macao nodes perform differential privacy disturbance, which finally generates personalized teaching strategies in a state of non-transparent data, improving the accuracy of course adaptation by 32%. This mechanism replaces institutional compromise through technical trust, providing a feasible path for the compliant flow of sensitive data elements.

This technology connection model of "computing power network + data middle platform" not only realizes the efficient utilization of hardware resources, but also bridges the synergy barriers caused by institutional differences through technological innovation, providing in-depth cooperation between universities in the Greater Bay Area in the field of AI. Provides a reusable infrastructure paradigm. Its core value lies in weaving discrete technology nodes into an intelligent network with self-regulating capabilities, transforming computing power and data from "private resources of colleges and universities" into "cluster public goods", and providing the intelligence of core scenarios such as scientific research collaboration and talent training. Upgrade lays the foundation.

(2) Ecological empowerment: activating cluster innovation potential

In the construction of Guangdong-Hong Kong-Macau Greater Bay Area's technology ecology, the ecological empowerment mechanism realizes the exponential release of innovation potential energy through the two-way interaction between technology empowerment and demand traction. At the level of discipline innovation, based on multi-modal knowledge graph construction technology, an intelligent discovery system at the intersection of disciplines in universities in the Bay Area is established. Through natural language processing and knowledge extraction algorithms, the system performs semantic association analysis on literature metadata of regional dominant disciplines such as Shenzhen artificial intelligence, Hong Kong financial engineering, and Guangzhou biomedicine, and generates a dynamically evolving "AI + X" interdisciplinary research map. This technology-enabled discipline integration mechanism has successfully guided the scientific research resources in the Bay Area to gather in strategic frontier fields such as quantum computing and brain-computer interface, forming an innovation chain coupling effect of "basic research-technological research-industrial transformation". In the dimension of talent cultivation, build an intelligent portrait system of human-machine collaboration to reshape the talent cultivation ecology. By developing an AI professional ability portrait platform connected to the Greater Bay Area's industrial demand database, transfer learning technology is used to establish a mapping model of talent ability characteristics and job skill map. In the practice of digital transformation of Dongguan manufacturing industry, this system successfully improved the forecast accuracy of industrial Internet talent demand to 92%, and accordingly driven universities such as South China University of Technology to dynamically optimize the professional curriculum system, forming an intelligent closed loop of "industrial demand drive-education supply response-employment quality feedback". This data-driven talent training mode improves the fit between discipline and specialty setting and regional industrial upgrading by 41%, which verifies the deep integration path of education chain and industrial chain

under technology empowerment.

(3) Co-evolution: cultivating an adaptive ecology

In order to realize the intelligent symbiosis of Guangdong-Hong Kong-Macau Greater Bay Area's higher education clusters, it is necessary to build an adaptive ecology with self-iteration ability, the core of which lies in establishing a closed-loop evolution mechanism of "perception-decision-evolution". First, the reinforcement learning framework is used to build a cluster development digital twin, collect multi-source heterogeneous data in real time, such as paper co-authorship network, patent coupling strength, technology conversion rate, etc., and use deep Q network (DQN) to train a dynamic evaluation model to quantify inter-school cooperation. Closeness and knowledge spillover effectiveness. When a continuous threshold drop in the frequency of cooperation in a specific field is detected, the system automatically triggers a "collaborative failure" early warning and generates an intervention plan based on graph neural network. Secondly, the blockchain smart contract framework is introduced to develop a distributed rule engine, and cross-border collaboration rules are encoded into automatically executable on-chain protocols. For example, for basic research projects, the lightweight consensus mechanism of "dynamically allocating intellectual property rights according to contribution" is preset; For application development collaboration, a two-tier contract model of "prior verification + posterior traceability" is adopted, combined with Bayesian game algorithm to optimize the benefit distribution scheme. The two-wheel drive mechanism realizes the paradigm shift from passive response to active evolution through the coupling feedback of data flow and rule flow, and makes the collaboration efficiency of the cluster ecosystem show exponential adaptive growth.

IV. Paradigm Transition: Three Major Transformation Paths from Technology Discrete to Intelligent Symbiosis

(1) Resource allocation paradigm: from "administrative leadership" to "algorithm collaboration"

The innovation of resource allocation paradigm is a key breakthrough point for Guangdong-Hong Kong-Macau Greater Bay Area's higher education clusters to move from administrative collaboration to intelligent symbiosis. Its core lies in reconstructing resource matching logic with AI algorithms, promoting the transfer of decision-making subjects from "administrative authority" to "data intelligence", and realizing the dual improvement of resource allocation efficiency and collaboration quality. In the construction of intelligent supply-demand matching mechanism, the intelligent trading platform of higher education resources in Greater Bay Area breaks through the inefficiency of the traditional administrative-led mode by building a closed-loop system of "demand release-algorithm analysis-accurate matching-effect feedback". Universities, scientific research institutions and enterprises can publish multi-dimensional information such as equipment sharing, scientific research cooperation, and talent demand on the platform. The AI algorithm is based on 12 core parameters such as historical cooperation performance data, resource idle rate, and discipline matching, and uses collaborative filtering algorithms to generate optimal partner recommendation list. Through comparative analysis with traditional models, the advantages of intelligent algorithm collaboration are significantly presented. In the scenario of "Guangdong-Hong Kong-Macao University Alliance Project Application", the traditional process relies on the administrative department to manually sort out the cooperation intention and screen the cooperation subjects. On average, each project needs to go through 3 rounds of communication, which takes about 3 months, and there are matching errors caused by information asymmetry. After the introduction of the AI intelligent matching system, the project applicant only needs to submit key information such as research direction and resource requirements. The algorithm automatically extracts the partners with a matching degree of $\geqslant$ 85% from the alliance university database, and generates a visual report containing the cooperation basis and risk assessment. The whole process is compressed to 2 weeks, and the discipline fit of the partners is increased to 94%, and the project success rate is increased from 58% to 79%. The essence of this paradigm transformation from "administrative leadership" to "algorithm collaboration" is to transfer the decision-making power of resource allocation from the bureaucratic system to the data-driven intelligent system, which not only avoids the subjective deviation of human intervention, but also activates the value of idle resources through real-time dynamic matching. Its deep significance lies in the construction of a new collaborative mechanism of "government guidance-market operation-technology empowerment", which provides an efficient and fair resource allocation solution for the sustainable development of higher education clusters.

(2) Organizational governance paradigm: from "bureaucratic fragmentation" to "network autonomy"

In the field of Guangdong-Hong Kong-Macau Greater Bay Area technology governance, the organizational form is undergoing a paradigm transition from bureaucratic fragmentation to networked autonomy. This transformation reshapes the collaboration model of cross-jurisdictional entities through distributed technology architecture and smart contract mechanism. The cluster governance platform based on blockchain transforms multiple entities such as universities, enterprises, and governments into consensus nodes with equal participation. The AI voting system realized through smart contracts shows technical empowerment effects in the formulation of cross-border scientific research cooperation policies, such as the dynamic adjustment of the whitelist of cross-border data transmission, which is automatically executed after verification by multi-party nodes, shortening the policy iteration cycle by 63%.%. This decentralized decision-making mechanism effectively solves the lag of the traditional administrative level's response to technological innovation. Aiming at the level of institutional adaptation, the policy semantic transformation engine builds a technical bridge across jurisdictional rules. The engine uses natural language processing technology to deconstruct the heterogeneous educational laws and regulations in Guangdong, Hong Kong and Macao into machine-readable rule meta-language, and realizes concept mapping through knowledge graph. In the cross-border academic certification scenario, the system

successfully transformed the mutual recognition rules of credits between the two places into a unified algorithm model, which improved the efficiency of certification audit by 81%. This technical institutional translation essentially creates a "rule middleware" beyond legal texts, realizes the flexible bridging of institutional differences through coded expression, and provides a technical solution for the modernization of regional governance.

(3) Value creation paradigm: from "individual competition" to "ecological win-win"

Guangdong-Hong Kong-Macau Greater Bay Area's higher education cluster is undergoing a value creation paradigm transition from zero-sum game to symbiotic development. Its essence is to realize cross-organizational reorganization of innovative elements and niche complementarity through AI technology. The core path lies in building an intelligent collaborative network with ternary integration of "knowledge-industry-system", that is, relying on the federated learning architecture to integrate the basic research capabilities of Hong Kong universities, the scene verification facilities of Shenzhen enterprises and the industrialization resources of Dongguan manufacturing to form a chain acceleration mechanism of "R&D-transformation-industrialization". The network automatically triggers value allocation through smart contracts, enabling Hong Kong's algorithm patents, Shenzhen's engineering optimization schemes and Dongguan's process data to realize factor combination innovation under blockchain confirmation. Typical cases show that industrial quality inspection based on Transformer architecture The model takes only 11 weeks from paper publication to production line deployment, which is 67% shorter than the traditional path. In order to quantitatively evaluate the effectiveness of ecological transformation, it is necessary to establish a multi-modal evaluation system that integrates complex network analysis and entropy method, which can be roughly divided into three system dimensions. The "knowledge flow intensity" dimension adopts cross-domain patent coupling degree and academic community intermediate centrality index to measure the efficiency of tacit knowledge transfer; The dimension of "technology radiation energy level" constructs the depth index of AI technology embedding and the response function of industrial upgrading; The "ecological resilience" dimension simulates the adaptive reorganization capability of the system under external shocks through the LSTM neural network. Empirical research shows that from 2020 to 2023, the niche overlap of the Guangdong-Hong Kong-Macao Industry-University-Research Consortium will decrease by 38%, while the resilience entropy of the innovation chain will increase by 2.1 times, confirming that the AI-driven technology-industrial hyper-domain network has effectively achieved "Pareto" Improvement "value creation.

V. Implementation Guarantee: The Support System of Intelligent Symbiotic Ecology

(1) Technical standard collaboration: construction of multi-modal interoperability framework

Establishing a technical standard collaboration system of "sovereign compatibility and dynamic evolution" is a key breakthrough to solve the asymmetry problem of technical institutions in Guangdong-Hong Kong-Macau Greater Bay Area's cross-border AI applications. Its core lies in the inclusive construction of technical rules under the framework of "one country, two systems" through multi-stakeholder collaborative governance and flexible mechanism design. The specific implementation path is based on "standard negotiation-technology adaptation-sandbox verification". First, the "Bay Area AI Education Standards Committee" composed of government education departments of Guangdong, Hong Kong and Macao, university alliances, leading technology enterprises and academic institutions is established to build a cross-domain consultation mechanism. The committee is responsible for formulating the "White Paper on AI Application Standards in Guangdong-Hong Kong-Macau Greater Bay Area's Higher Education", focusing on the three major technical breakpoints of data flow, computing power interconnection, and ethical norms. The educational data of colleges and universities is stored in the "sovereign cloud" of the National Supercomputing Shenzhen Center, and Hong Kong and Macao data is retained on local servers. Cross-domain data verification and joint modeling are realized through zero-knowledge proof technology, which not only meets the mainland data sovereignty requirements, but also complies with Hong Kong and Macao privacy protection regulations; Second, in the field of computing power collaboration, define an extended protocol based on OpenAPI 3.0, which is compatible with heterogeneous computing platforms such as Huawei Ascend and Nvidia CUDA, and establish a computing power sharing mechanism of "unified interface-dynamic scheduling-performance monitoring", so that the computing power resources of Hong Kong's "Advanced Computing Platform" and Guangzhou's "Tianhe-2" can achieve millisecond-level response through 5G networks; Third, in terms of ethical constraints, establish a negative list system for AI applications, clarify prohibitive clauses such as prohibiting the use of facial recognition data for comprehensive evaluation of students, prohibiting discriminatory pricing of algorithms, etc., and realize the automated execution and real-time monitoring of binding clauses through smart contract technology. Secondly, relying on institutional innovation carriers such as Hengqin Guangdong-Macao Deep Cooperation Zone and Shenzhen Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone, standard sandbox verification will be carried out. For example, the "Guangdong-Hong Kong-Macao AI Joint Laboratory" sandbox project launched in 2023 reduces the delay of cross-border scientific research data calls to 12 milliseconds by deploying a federated learning framework and edge computing nodes, meeting the requirements of the ISO/IEC 20547-4 international standard Requirements for real-time data interaction while ensuring compliance with data not leaving the country. This kind of stress test not only provides empirical basis for the feasibility of technical standards, but also promotes the upgrade of the standard system from "static consensus" to "adaptive evolution" through the dynamic evolution mechanism of "problem discovery-rule iteration-ecological adaptation", and finally forms a new paradigm of intelligent governance that not only adheres to the principle of national sovereignty, but also is compatible with regional institutional differences.

(2) Talent echelon construction: cultivating compound innovation forces

The structural reform of talent supply side is the core kinetic energy of the intelligent symbiotic ecological construction of higher education. It is necessary to build a three-dimensional talent development network with deep integration of educational science and technology through the reshaping of ability standards, the reengineering of knowledge systems and the innovation of training models. At the first competency standard level, a teacher qualification accreditation system based on TPACK (Subject Teaching Knowledge with Integrated Technology) model is established. The AI Education Teacher Certification Center jointly established by Guangdong, Hong Kong and Macao has formulated the "Cross-domain Dual-qualified Teacher Ability Standard", covering Three core modules: educational neuroscience cognition, AI technology application and cross-cultural teaching method. At the same time, the certification adopts a multi-modal evaluation framework, the theoretical test relies on the cognitive diagnosis model (CDM) to dynamically monitor knowledge blind spots, and the practical link requires the development of lightweight AI tools with teaching decision support functions, such as classroom interactive analysis plug-ins based on natural language processing. As of 2024, the system has trained 586 dual-qualified teachers who have passed standardized certification, and its interdisciplinary curriculum development efficiency is 2.3 times higher than that of traditional teachers. The second knowledge system layer is to create a micro-major cluster of "technology-education-design", such as the micro-major of "intelligent education system development" jointly established by Macau University of Science and Technology and South China University of Technology. The OBE (achievement-oriented education) mode is adopted to set up the curriculum chain of "machine learning foundation-educational data governance-immersive learning environment design", and the whole process practice from data collection to model deployment is closed-loop through project-based learning (PBL). The third training mode layer is to build a dual training ecosystem linked by Industry-University-Research. The "Algorithm Engineer Ability Workshop" jointly created by the Chinese University of Hong Kong and Tencent AI Lab introduces real project data sets of enterprises and adopts the "dual tutor system + agile development" mode, which enables trainees to complete the ability transition from theoretical transformation to industrial-grade code submission within the 48-week training period. In the past three years, 327 professional certified engineers have been sent to AI enterprises in Greater Bay Area, with a direct employment rate of 92%. This three-dimensional training system provides a compound talent support with both theoretical depth and practical innovation for the intelligent symbiotic ecology through the optimization of teacher structure, the reconstruction of curriculum system and the deep integration of production and education.

(3) Risk prevention and control mechanism: building a solid bottom line of technical ethics and safety

As the safety cornerstone of the intelligent symbiotic ecology, the risk prevention and control mechanism achieves a dynamic balance between innovation incentives and risk management and control by building a full-chain governance system of "technical reliability review-ethical compliance assessment-dynamic monitoring response". First of all, establish a feasibility review system for AI educational application technology, and a third-party professional organization conducts multi-dimensional evaluation of technical solutions, requiring the intelligent teaching system to meet hard technical indicators such as model accuracy $\geqslant$ 90% and response delay $\leqslant$ 500 milliseconds, and enforce Implement the mechanism of "small-scale pilot-feedback optimization-comprehensive promotion". For example, the "intelligent homework correction system" developed by a university has been piloted in 12 classes for 3 months. After optimizing the algorithm according to 237 improvement suggestions put forward by teachers and students, the accuracy rate of composition correction has been improved from 78% to 89% before it is approved to be popularized throughout the school. Secondly, an interdisciplinary ethics committee is established, composed of educators, AI technical experts, legal scholars and student representatives, to conduct pre-ethical review of AI applications. For example, in response to the facial data collection problem involved in the "Student Emotion Recognition System" of a university, the ethics committee, in accordance with the "Personal Information Protection Law" and the "Educational Data Security Standard", requires that the system be only used for classroom interactive analysis, and prohibits association with academic evaluation. The data storage time limit is set to 3 months, and the original information will be automatically deleted when it expires, so as to prevent the risk of privacy leakage and algorithm abuse from the source. Finally, a dynamic monitoring platform for AI educational applications is built to establish a risk early warning model by capturing technical indicators such as system failure rate and user complaint rate in real time, as well as social feedback data such as social media public opinion and parent satisfaction surveys. For example, when the data leakage incidence rate of a cross-border AI teaching platform exceeds the threshold for two consecutive months, the platform automatically triggers a three-level emergency response, completes the technical vulnerability repair within 72 hours, starts the responsibility traceability procedure within 1 week, and releases the rectification report to the public within 15 days to ensure the timeliness and transparency of risk prevention and control. This trinity prevention and control system organically unifies technical rationality, ethical norms and social adaptability, and builds a multi-level security barrier for the sustainable development of intelligent symbiotic ecology.

VI. Conclusion

Guangdong-Hong Kong-Macau Greater Bay Area's higher education cluster is undergoing a profound change from technological discretion to intelligent symbiosis. Through AI empowerment, a coordinated development path of "technological connection-ecological empowerment-mechanism innovation-talent drive" has been explored. At the level of technical connection, the cluster has built a cross-domain computing power network and a federated data middle platform to achieve systematic integration of computing power and data, and through distributed technology architecture and smart contract mechanism, it promotes the organizational

governance paradigm from "bureaucratic fragmentation" "Turn to" network autonomy "to effectively break down institutional barriers. The ecological empowerment mechanism activates the potential energy of cluster innovation through multi-modal knowledge graph and intelligent portrait system, and promotes the intelligent upgrading of discipline innovation and talent cultivation. At the same time, it uses the reinforcement learning framework and blockchain intelligent contract to realize the dynamic evaluation and intelligent intervention of cluster development. In terms of mechanism innovation, universities in the Greater Bay Area have established a resource allocation paradigm from "administrative leadership" to "algorithm collaboration" to improve the efficiency of resource allocation. By establishing mechanisms such as a negative list for cross-border data flow and an AI ethics joint review committee, they have Seek a dynamic balance between ensuring data sovereignty security and promoting the circulation of technical elements. The structural reform of talent supply side is the core kinetic energy of the construction of intelligent symbiotic ecology. Through the teacher qualification accreditation system based on TPACK model, the micro-professional cluster of "technology-education-design" and the dual system cultivation ecology linked by Industry-University-Research, we cultivate compound talents and provide intellectual support for regional innovation. This collaborative paradigm transition is not only an in-depth application of AI technology, but also a comprehensive innovation of the collaborative development model of regional education, injecting strong impetus into regional economic and social development.

REFERENCES

[1] The Central Committee of the Communist Party of China and the State Council issued the "Outline of Guangdong-Hong Kong-Macau Greater Bay Area Development Plan" [EB/OL]. (2019-02-18) [2019-06-30]. http://www.gov.cn/zhengce/2019-02/18/content _ 5366593. htm # 1.

[2] Ou Xiaojun. Research on the development of high-level university clusters in the world-class Greater Bay Area-taking the three bay areas of new york, San Francisco and Tokyo as examples [J]. Journal of Sichuan Institute of Technology: Social Science Edition, 2018 (6): 83-100.

[3] Chen Xianzhe. Guangdong-Hong Kong-Macau Greater Bay Area Higher Education Cluster: Walking a Way Beyond the Status Quo [N]. Guangming Daily, 2018-08-07 (13).

[4] Li Shengbing, Li Longjuan. The development of higher education in Guangdong-Hong Kong-Macau Greater Bay Area: from imbalance to balance [J]. Higher Education Research, 2022, 43 (08): 46-51.

[5] Chen Fajun. Comparative Advantage and Development Transcendence: Discussion on the Integrated Development Path of Higher Education in Guangdong-Hong Kong-Macau Greater Bay Area [J]. Education Guide, 2022, (01): 46-53. DOI: 10.16215/j.cnki.cn44-1371/g4.2022. 01. 009.

[6] Lu Xiaozhong, Wu Yiting. Strategic choice and basic direction of the development of higher education clusters in Guangdong-Hong Kong-Macau Greater Bay Area [J]. Journal of Lanzhou University (Social Science Edition), 2021, 49 (05): 9-15. DOI: 10.13885/J.issn.1000-2804.2021. 05.002.

[7] Lu Xiaozhong, Qin Qin. Research on the autonomy of running schools in Guangdong-Hong Kong-Macau Greater Bay Area's universities from the perspective of higher education cluster development [J]. Chinese Higher Education Research, 2021, (04): 55-63. DOI: 10.16298/J.cnki.1004-3667. 2021. 04.10.

[8] Li Huimin, Fu Yanan. Analysis of the integrated development of higher education in Guangdong-Hong Kong-Macau Greater Bay Area [J]. University, 2022, (25): 27-30.

[9] Wang Hong, Chen Han. Breakthrough and integration: the effectiveness, dilemma and countermeasures of Guangdong-Hong Kong-Macau Greater Bay Area's higher education cooperative development [J]. Academic Research, 2024, (10): 59-66.