Scholar Publishing Group
International Journal of Big Data Intelligent Technology
https://doi.org/10.38007/IJBDIT.2025.060111
ISSN 2790-0932 Vol. 6, Issue 1: 109-117

# *Efficient Multimodal Visual Segmentation Model Based on Phased Fusion of Differential Modalities*

**Bukun Ren**

*College of Engineering, University of California Berkeley, Berkeley, CA 94720, California, United States*

*bukunren46@outlook.com*

***Keywords:*** Multimodal image harmonization, phased fusion of differential modalities, referential description, ReiHarmony4 dataset, efficient visual segmentation model

***Abstract:*** This article focuses on the application of an efficient multimodal visual segmentation model based on phased fusion of differential modalities in image harmonization tasks. In response to the problem of the failure of the harmonization model due to the lack of a predetermined foreground mask in practical application scenarios, this paper innovatively proposes a multimodal image harmonization task, which replaces the foreground mask by introducing a referential description for the foreground. This article constructs a new multimodal harmonization dataset ReiHarmony4 based on the traditional image harmonization dataset iHarmony4. For this task, this article proposes a segmentation harmonization pipeline model and two different end-to-end methods, including a combination of CLIP based referential image segmentation model and Harmony Transformer, and DiffHarmony based on stable diffusion model. The experimental results show that these models can effectively complete the task of multimodal image harmonization. This article also designs and implements a multimodal image harmonization system, which achieves the function of image harmonization based on text. Although some achievements have been made, there are still some issues that need further exploration, such as improving segmentation performance, solving the problem of resolution degradation, and expanding the dataset.

## 1. Introduction

This article focuses on the application of an efficient multimodal visual segmentation model based on phased fusion of differential modalities in image harmonization tasks, aiming to solve the problem of model failure in practical application scenarios due to the lack of predetermined foreground masks. Unlike traditional image harmonization tasks that use foreground masks, this paper innovatively introduces a referential description for the foreground to replace the foreground mask, thus proposing a new task of multimodal image harmonization. Due to the fact that multimodal image harmonization tasks involve textual information, while traditional image harmonization datasets do not contain textual information, this paper constructs a new multimodal harmonization dataset ReiHarmony4 through manual annotation based on the iharmony4 dataset. For this task, this article proposes a segmentation harmonization pipeline and two different end-to-end methods, including a CLIP based referential image segmentation model combined with Harmony Transformer, a stable diffusion based DiffHarmony model, and CRIS-HT and DiffReHarmony models designed on this basis. These models fully exploit and utilize the complementarity between multimodal information by fusing data from different modalities in stages, thereby improving the accuracy and efficiency of image segmentation and harmonization. This article applies the proposed multimodal image harmonization model to an actual image editing system, achieving the function of image harmonization based on text. This not only improves the convenience of image harmonization, but also injects new vitality into the development of image harmonization tasks.

## 2. Correlation Theory

In recent years, multimodal image processing technology has made significant progress in different fields. A large number of researchers have proposed a method called PSALM, which utilizes a large multimodal model to achieve pixel level image segmentation. Researchers have enhanced the multimodal large-scale language model SAM4MLLM with the aim of improving the performance of referential expression segmentation. In the field of remote sensing, researchers have released the Ticino dataset, which is specifically designed for semantic segmentation research of multimodal remote sensing images. Researchers have explored multi task learning methods for semantic segmentation and height estimation of multimodal remote sensing images, providing new ideas for remote sensing image processing. In the field of medical image analysis, researchers have conducted in-depth semantic segmentation research on MRI images using transfer learning driven convolutional neural networks. Researchers have explored the applicability of Transformer models in remote sensing image segmentation, further promoting the development of remote sensing image processing technology. The researchers also discussed text driven generation methods for stable diffusion models in augmented reality applications, injecting new vitality into the development of augmented reality technology. Researchers have proposed an efficient reinforcement learning online diffusion model fine-tuning method based on human-machine feedback, providing a new approach for optimizing online learning models. These studies demonstrate the diversity and innovation of multimodal image processing techniques in different application scenarios.

## 3. Method

### 3.1. Breakthrough in Transformer Cross Domain Applications

Deep learning models have made significant breakthroughs in recent years, thanks to the improvement of computer performance and computing power, as well as the support of large-scale data. In the field of computer vision, convolutional neural networks have achieved efficient image feature extraction and dimensionality reduction through convolution and pooling operations, and are widely used in tasks such as image classification and object detection. The U-Net model adopts an encoder decoder structure and performs well in image to image conversion tasks such as medical image segmentation. Recurrent neural networks and their improved versions, long short-term memory networks, greatly enhance the ability of computers to process and understand temporal signals. Although RNNs suffer from long-term dependencies, LSTM effectively alleviates this problem by introducing gating units. As the sequence length increases, the computational efficiency and parallel processing capability of RNN and LSTM are limited. To overcome these challenges, attention mechanisms and Transformer models have emerged. The Transformer model replaces LSTM with a full attention structure and introduces residual connections and layer normalization to improve the training performance and stability of the model. The calculation process includes layer normalization and a two-layer feedforward neural network, while the decoder structure adds an attention layer to calculate the probability distribution of the output word at the current position. The model structure is shown in Figure 1

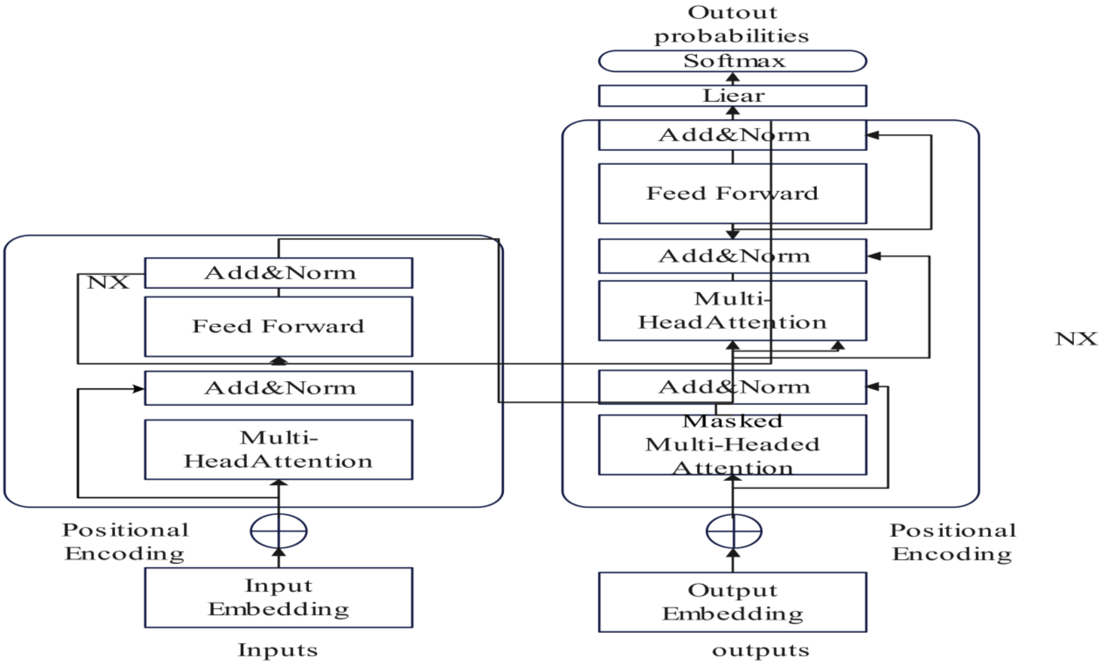

*Figure 1. Schematic diagram of Transformer model structure*

The Google team further extended Transformer to the field of vision and proposed Visual Transformer (ViT), which divides images into several 16x16 blocks as tokens and inputs them into a self attention model for training, achieving a performance level comparable to convolutional neural networks and successfully breaking the boundary between vision and text, laying a theoretical foundation for breakthroughs in the field of multimodal graphics and text in the future.

### 3.2. Multi Modal Modeling Technology for Image Generation

Image generation technology, as a core technology in fields such as computer vision, artificial intelligence, and graphics, has made significant progress in recent years. Among them, variational autoencoder and generative adversarial network are two important models in the field of image generation. VAE achieves more flexible sampling in latent space by introducing the concept of probability distribution, thereby generating more diverse samples. GAN, on the other hand, uses adversarial training between the generator and discriminator to enable the generator to generate samples with distributions similar to real data. In addition, diffusion models, as an emerging image generation technology, have also demonstrated their powerful generation capabilities. This model achieves the restoration of noisy images to original images through forward denoising and backward denoising processes.With the continuous development of generative models, the generation of content by artificial intelligence has become a cutting-edge research direction. In this context, multimodal modeling techniques have emerged. Multimodal modeling aims to integrate information from different modalities to improve system performance and enrich user experience. Pre training techniques, as an important method in deep learning, provide strong support for multimodal modeling. By pre training on large-scale unlabeled data and fine-tuning on specific tasks, the performance of the model can be significantly improved on various tasks. Due to the universality of the representation space learned by CLIP, the model exhibits strong capabilities in zero sample learning and can be applied to various tasks such as image classification, text retrieval, and image generation description. Various downstream multimodal task methods based on CLIP continue to emerge, such as the wind grid model CLIPstyler and the referential image segmentation model CRIS, further promoting the development of multimodal modeling technology.

### 3.3. Construction and Research of Multimodal Image Harmony Dataset

Multimodal image harmonization task is an emerging field of image processing that aims to combine synthetic images and text descriptions, and output a harmonious image through a multimodal harmonization model to make it closer to the real image. Unlike traditional image harmonization tasks that rely solely on synthesized images and foreground masks, multimodal tasks introduce textual descriptions to enhance the flexibility of the model in locating foreground and generating harmonious images. Due to the high dependence of traditional image harmonization models on foreground masks, their flexibility is limited in practical applications, especially in scenarios where distortion region masks cannot be quickly obtained. In order to overcome this limitation, researchers have proposed blind image harmonization technology, but this technology still faces significant challenges, and there is still a certain distance between its performance indicators and masked image harmonization. The multimodal image harmonization task, as a compromise approach that combines convenience and effectiveness, has great research value. Due to the fact that multimodal image harmonization is a new task, there is currently a lack of publicly available datasets. In order to construct a dataset suitable for this task, a new multimodal harmonization dataset ReiHarmony4 was constructed by supplementing the iHarmony4 series dataset with referential expressions corresponding to the foreground part. We conducted validity screening and filtering on the annotated data to ensure that every sample in the dataset is accurate and valid, and completed statistical analysis on the dataset. Through the above steps and methods, this article successfully constructed a dataset suitable for multimodal image harmonization tasks, providing strong support for subsequent research and model training. The construction of this dataset not only solves the problem of the lack of publicly available datasets for multimodal image harmonization tasks, but also lays a solid foundation for research and development in this field.

## 4. Results and discussion

### 4.1. Comprehensive Design and Challenge of Multimodal Image Harmony Model

In the task of image harmony, we adopted two main model architectures: staged and end-to-end. The staged model divides the task into two parts: image segmentation and image harmonization, and trains each module to obtain a harmonious image. The referential image segmentation model predicts foreground masks by inputting images and foreground referential descriptions. This model utilizes CLIP's image text alignment ability and combines lighting differences to further improve prediction accuracy. In the harmonization stage, we attempted various models, such as the DHT model based on Retinex lighting theory, which achieves image harmonization through reshaping operations and cross attention mechanisms. Another option is the DifHarmony model, which is fine tuned based on a stable diffusion model to achieve efficient image to image translation.The mask resolution actually used by the model is still relatively low, which increases the difficulty of reconstructing image content. To address this issue, we explored an end-to-end multimodal harmonization model. The CRISHT model integrates the segmentation model CRIS and the image harmonization model DHT, weighting the contrast loss of the image segmentation module with the L1 loss of the harmonization, achieving the integration of the two modules. This integration enables the model to more efficiently handle multimodal image harmonization tasks. We also proposed a DiffReHarmony model based on the diffusion model. This model has similarities in design with the DiffHarmony image harmonization model, but has been adjusted to accommodate text prompts. It no longer directly inputs the foreground mask image, but predicts the foreground region mask based on text prompts. By adding a convolutional layer with output channel 1 in the U-Net network, the model can directly obtain the predicted foreground mask. A segmentation loss term has been added to the loss function to ensure the accuracy of foreground mask prediction. The design of a multimodal image harmonization model involves multiple aspects, including the selection of staged and end-to-end models, prediction of foreground masks, and application of image harmonization algorithms.

## 4.2. Performance Comparison and Analysis of Phased Multimodal Image Harmonization Models

The staged model divides the task into two stages: referential image segmentation and image harmony. The segmentation stage uses an enhanced CRIS model that introduces foreground and background difference detection, while the image harmony stage uses the HT model and the diffusion based DiffHarmony model proposed in this paper. The experimental results are shown in Figure 2

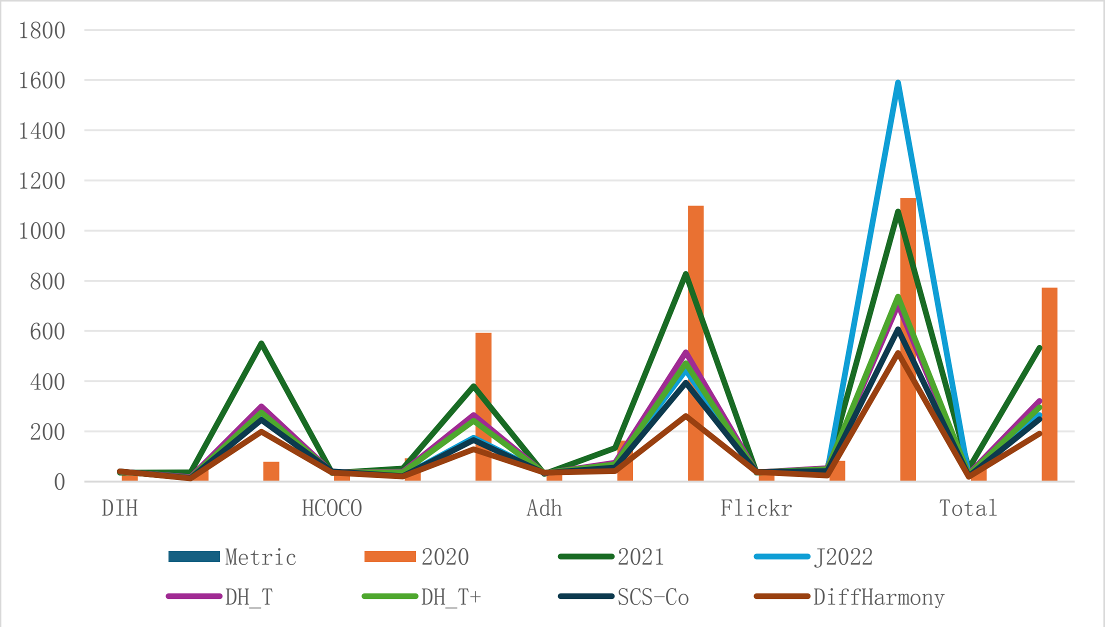

*Figure 2. Performance Comparison of iHarmony4 Dataset Models*

The experimental results show that the enhanced CRIS model significantly improves the harmonization performance on various sub datasets of ReiHarmony4 compared to the original CRIS model, especially on the larger HAdobe5k dataset, where the mean square error index is improved by 10.5% and the overall harmonization performance is improved by 12.2%. The DiffHarmony model performs better than the HT model on the HCOCO and HFlicker datasets. Although it is slightly inferior to the HT model on the HAdobe5k and Hday2night datasets, overall, the

CRISP+DiffHarmony model performs the best on the comprehensive indicators of the ReiHarmony4 sub datasets and is currently the best staged harmonization model. In the segmentation stage, this article refers to the CLIP based referential image segmentation model (CRIS) and generates prediction masks by introducing synthetic images and foreground referential descriptions, further improving the performance of the segmentation model. For the ReiHarmony4 dataset, this paper conducted fine-tuning experiments on the CRIS model, and the results showed that the fine tuned CRIS model improved both segmentation accuracy and efficiency. This article also proposes a new diffusion based image harmonization model, DiffHarmony, and compares it with several deep learning based image harmonization methods. The DiffHarmony model proposed in this article has shown excellent results, with indicators on all four sub datasets improved by 81.5% on the basis of the Cg-LDM model. The mean square error on the entire ReiHarmony dataset has decreased from 141.84 to 18.36, and the peak signal-to-noise ratio has also increased from 32.70 to 39.50, with an improvement of 20.7%. This significant performance improvement is mainly due to the fine-tuning of the original diffusion model in this paper, and the use of higher pixel images as inputs, making the scaled output closer to the target size. The DiffHarmony model proposed in this article achieved the best results in the image harmony model based on diffusion model.

### 4.3. Comparative Analysis of Evaluation Effects

This article compares the performance metrics of various end-to-end models on the iHarmony4 dataset. Through comparison, it was found that the CRISHT model designed in this article outperforms the Original CRISHT model based on the original CRIS model in all sub datasets due to its specific differentiation of foreground and background in synthesized images. On the entire dataset, the mean square error (MSE) metric of the CRISHT model increased by 56.4%, and the MSE of the foreground increased by 37.2%. The end-to-end model DiffReHarmony, which is adjusted based on the DiffHarmony model, performs better overall. Compared with the CRISHT model, its indicators have improved to varying degrees on various datasets, especially on the Hday2night dataset, where the indicators have improved by 16.3%. Overall, in the end-to-end harmonization model, the DiffReHarmony model based on the diffusion model achieved the best results. This article also compared the quantitative performance of staged pipeline method and end-to-end method on the four sub datasets of ReiHarmony4. In addition to recording the performance of the model on image harmonization evaluation metrics (MSE and fMSE), the metrics of the model in segmentation tasks were also recorded to further analyze the model. From the comparison, it can be seen that both phased and end-to-end methods can greatly improve the coordination of synthesized images, and both methods can effectively complete the task of multimodal image harmonization. For the two proposed methods, the staged approach is slightly superior to the end-to-end approach. From the overall indicators of the ReiHarmony4 dataset, the staged CRISP+DiffHarmony model achieved the best results in both segmentation and image harmony. This article provides a detailed description of the specific design of the model and conducts sufficient experiments to verify the effectiveness of these models with different implementation methods for multimodal harmonization tasks. This article also compares the performance of staged models and end-to-end models, explores the advantages and disadvantages of several models, including diffusion based models, in handling multimodal harmonization tasks, and proposes corresponding improvement solutions for related issues.

## 5. Conclusion

Image harmonization is an important branch of image synthesis, with significant research and application value. With the development of deep learning, significant progress has been made in this field. However, the current task requires simultaneous input of synthesized images and foreground masks, which limits the application scenarios. To lower the threshold, this article proposes multimodal image harmonization, which automatically achieves harmonization through input images and foreground descriptive text, making it more convenient for practical applications. For the new task, this article constructs the ReiHarmony4 dataset and proposes a staged harmonization model and two types of end-to-end models, which are experimentally proven to be effective. A multimodal image harmonization system has been designed, allowing users to generate harmonious images by inputting images and text descriptions through web pages. Despite progress, there are still issues. The segmentation performance of the staged model needs to be improved, especially in complex scenes where the target object is easily confused. The diffusion model may experience a decrease in resolution during the harmonization process. The ReiHarmony4 dataset is relatively small in size, which affects the effectiveness of the model in processing special images. If the dataset size is expanded or self supervised training is adopted, the model performance may be improved.

## References

*[1] Xu, Yue. "Research on Maiustream Web Database Development Techncly." Journal of Computer Science and Artificial Intelligence 2.2 (2025): 29-32.*

*[2] Zhu, Zhongqi. "Strategies for Improving Vector Database Performance through Algorithm Optimization." Scientific Journal of Technology 7.2 (2025): 138-144.*

*[3] Wang, Buqin. "Strategies and Practices for Load Test Optimization in Distributed Systems." Scientific Journal of Technology 7.2 (2025): 132-137.*

*[4] Zhang, Jingtian. "Research on Worker Allocation Optimization Based on Real-Time Data in Cloud Computing." Frontiers in Science and Engineering 5.2 (2025): 119-125.*

*[5] Hao, Linfeng. "Application of Machine Learning Algorithms in Improving the Performance of Autonomous Vehicles." Scientific Journal of Technology 7.2 (2025): 118-124.*

*[6] Gu, Yiting. "Practical Approaches to Develo**High-performance Web Applications Based on React." Frontiers in Science and Engineering 5.2 (2025): 99-105.*

*[7] Guo X. Research on systemic financial risk early warning based on integrated classification algorithm[C]//2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE). IEEE, 2024: 1586-1591.*

*[8] Chen, H., Yang, Y., & Shao, C. (2021). Multi-task learning for data-efficient spatiotemporal modeling of tool surface progression in ultrasonic metal welding. Journal of Manufacturing Systems, 58, 306-315.*

*[9] Shanshan Feng, Ke Ma, Gongpin Cheng, Risk Evolution along the Oil and Gas Industry Chain: Insights from Text Mining Analysis, Finance Research Letters, 2025, 106813, ISSN 1544-6123*

*[10] Tan, Weiyan, Shujia Wu, and Ke Ma. "Freight Volume Prediction for Logistics Sorting Centers Using an Integrated GCN-BiLSTM-Transformer Model." Advances in Computer and Engineering Technology Research 1.4 (2024): 320-324*

*[11] Fan, Sunjia, et al. "Defense methods against multi-language and multi-intent LLM attacks." International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2024). Vol. 13403. SPIE, 2024.*

*[12] Xiang, Y., Li, J., & Ma, K. (2024, October). Stock Price Prediction with Bert-BiLSTM Fusion Model in Bimodal Mode. In Proceeding of the 2024 5th International Conference on Computer Science and Management Technology (pp. 1219-1223).*

*[13] Pan, Yu. "Research on the Evolutionary Path of Resource Management and Capability*

*Building for Platform Enterprises." International Journal of Finance and Investment 2.1 (2025): 78-81.*

*[14] Liu, Boyang. "Study on the Frequency of Computer Language Use Based on Big Data Analysis." Academic Journal of Computing & Information Science 7.10 (2024)*

*[15] Zhang, Yiru. "Design and Implementation of a Computer Network Log Analysis System Based on Big Data Analytics." Advances in Computer, Signals and Systems,(2024) 8(6),40-46.*

*[16] Liu, Yu. "Build an Audit Framework for Data Privacy Protection in Cloud Environment." Procedia Computer Science 247 (2024): 166-175.*

*[17] Liu, Boyang. "Design and Application of Experimental Data Management System Integrating Remote Monitoring and Historical Data Analysis." Journal of Electronics and Information Science 9.3 (2024): 160-167.*

*[18] Xu, Yue. "Research on Graph Network Social Recommendation Algorithm Based on AGRU-GNN." 2024 IEEE 4th International Conference on Data Science and Computer Application (ICDSCA). IEEE, 2024.*

*[19] Cui, Naizhong. "Optimization Strategies for Traffic Signal and Identification Design." Frontiers in Science and Engineering 5.2 (2025): 92-98.*

*[20] Ding, Maomao. "Design Innovation and User Satisfaction Improvement of AI Video Creation Tools." Scientific Journal of Technology 7.2 (2025): 112-117.*

*[21] Chen, Junyu. "Research on Intelligent Data Mining Technology Based on Geographic Information System." Journal of Computer Science and Artificial Intelligence 2.2 (2025): 12-16.*

*[22] Xu Y. Research on UAV Navigation System Based on Behavioral Programming[C]//2024 IEEE 7th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE). IEEE, 2024: 419-425.*

*[23] Xu, Y. (2024). Research on Social Network Secunity Issues and Countermeasures Based on Big Data. International Journal of Computer Science and Information Technology.*

*[24] Li, X.(2025) “Research on the application of GPS, total station and CAD Technology in architectural Grid.” Computer Life (2024),12(3),36-39.*

*[25] Zhang, Jinshuo "Research on Real Time Condition Monitoring and Fault Warning System for Construction Machinery under Multi Source Heterogeneous Data Fusion." Journal of Engineering Mechanics and Machinery (2024), 9(2): 139-144*

*[26] Wang, Yuxin "Research on Intelligent Macro Image Recognition Algorithm of Oil Pipe Failure Based on Deep Learning." Journal of Image Processing Theory and Applications (2025), 8(1): 1-7*

*[27] Ma Z. Innovative Application of Reinforcement Learning in User Growth and Behavior Prediction[J]. European Journal of AI, Computing & Informatics, 2025, 1(1): 18-24.*

*[28] Wang Y. Design and Implementation of a General Data Collection System Architecture Based on Relational Database Technology[C]//The International Conference on Cyber Security Intelligence and Analytics. Cham: Springer Nature Switzerland, 2024: 561-572.*

*[29] Zhang J. Research on Dynamic Stability Identification and Early Warning System for Engineering Vehicles Integrating Machine Learning and Data Driven Technology[J]. Academic Journal of Computing & Information Science, 2025, 8(2): 51-55.*

*[30] Liu B. Innovative Applications and Performance Optimization Strategies of Python Interpreter in Web Development[J].Journal of Network Computing and Applications (2025) ,10(1),1-7*

*[31] Liu Z. Research on the Application of Signal Integration Model in Real-Time Response to Social Events[J]. Journal of Computer, Signal, and System Research, 2025, 2(2): 102-106.*

*[32] Chen, Anyi. "Application of Quantum Computing Technology in the Optimization of Search Sorting for Fashion E-commerce." Journal of Computer Science and Artificial Intelligence 2.2 (2025): 8-11.*

*[33] Shi, Chongwei. "Research on Gene Identification Algorithms Based on Signal Processing Techniques." 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE, 2024.*

*[34] Wang, Yuxin. "Application and Practice of Sensor Network Based on Deep Learning in Condition Monitoring of Underground Oil Production Equipment." International Journal of Frontiers in Engineering Technology 6.6 (2024).*

*[35] Zhang J. Research on Fault Prediction and Health Management System of Railway Tunnel Drilling and Blasting Construction Machinery Based on Machine Learning[J]. International Journal of New Developments in Engineering and Society, 2024, 8(5).*

*[36] Chen, H., Zuo, J., Zhu, Y., Kabir, M. R., & Han, A. (2024). Generalizable Deep Learning for Pulse-echo Speed of Sound Imaging via Time-shift Maps. In 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS) (pp. 1-4). IEEE.*

*[37] Yang, Jinzhu "Integrated Application of LLM Model and Knowledge Graph in Medical Text Mining and Knowledge Extraction."Social Medicine and Health Management (2024), 5(2): 56-62*