

GIDLMM: Generated Images Detection Using Large Multimodal Model

Yu-ting Lu, Jiann-Liang Chen, Bagus Tri Atmaja, Shih-Ping CHIU

Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan
 m11207516@gapps.ntust.edu.tw, lchen@gmail.com, d11207806@mail.ntust.edu.tw, spchiu@mail.ntust.edu.tw

Abstract - The rapid advancement of Generative AI (GenAI) has produced highly realistic images, making it difficult to distinguish authentic content from synthetic fabrications and accelerating the spread of visual misinformation. Effectively identifying AI-generated content with transparent methods has become a critical challenge. This study proposes the Generated Images Detection Using Large Multimodal Model (GIDLMM), a dual-module framework for data-efficient and interpretable detection. The first module, an AI-generated image detector, builds upon a CLIP model fine-tuned with Low-Rank Adaptation (LoRA) for parameter efficiency. To train this module, we introduce a novel data generation process where a Large Multimodal Model (LMM) automatically creates rich semantic descriptions for a small dataset. For this process, we propose an efficient “AI Judges” mechanism, where multiple virtual judges score candidate descriptions in a single pass to ensure quality without the high computational cost of repeated queries. Trained on only 1,000 samples, our model significantly outperforms state-of-the-art methods that use 100k samples, achieving an improvement on commercial generator datasets. The second module provides generation trace interpretations; after an image is flagged as AI-generated, it produces a textual explanation identifying visual artifacts that informed the decision. This explainability helps users recognize suspicious regions, enhances trust, and improves the model's transparency and reliability.

Keywords - CLIP, LMM, AI-generated Image Detection, Low-Resource Data, LoRA

I. INTRODUCTION

The proliferation of generative artificial intelligence (GenAI) has enabled the creation of highly realistic images, blurring the line between authentic and synthetic content. While this technology offers benefits in creative fields and product advertising [1, 2], it also facilitates the spread of misinformation across digital platforms. Recent examples, such as widely circulated images of fictitious “Himalayan Pagoda Blossoms” claimed to bloom once every 400 years [3], illustrate how easily such content can be created and disseminated. The increasing sophistication of GenAI tools, including Generative Adversarial Networks (GANs) and Diffusion models [4,5], makes it difficult for human observers to distinguish real photographs from fabrications. This capability presents challenges, from undermining trust in digital media to enabling sophisticated criminal activities like the \$25 million deepfake fraud reported in 2024 [6].

The dual nature of GenAI—a tool for both innovation and deception—necessitates the development of reliable detection methods. Early detection models often focused on specific

generative artifacts, but as GenAI evolves, these methods struggle to keep pace. Malicious actors can exploit this technology to forge historical artifacts, manipulate narratives, or create fraudulent content, posing ethical and societal risks [7, 8]. Consequently, there is a need for robust systems that can not only identify AI-generated images but also provide transparency in their decision-making processes to build user trust.

To address this challenge, this study proposes a model named GIDLMM (Generated Images Detection Using Large Multimodal Model). The GIDLMM framework is designed with two primary functions. The first is an AI-generated image detection module that leverages a Contrastive Language-Image Pre-Training (CLIP) model [9] fine-tuned with Low-Rank Adaptation (LoRA) for parameter efficiency [10]. This module uses a Large Multimodal Model (LMM) to automatically generate rich semantic descriptions for training data, enhancing its detection capabilities even with a small dataset of only 1,000 samples.

The second function is a generation trace interpretation module. After an image is classified as AI-generated, this module provides a textual explanation identifying visual artifacts or inconsistencies that led to the classification. By integrating detection with explainability, this research aims to create a more transparent and trustworthy tool for combating visual misinformation. This approach offers a data-efficient and interpretable solution to the growing problem of distinguishing authentic content from AI-generated fabrications.

II. RELATED WORKS

Research in AI-generated image detection has primarily focused on adapting large-scale, pre-trained models, with CLIP [9] serving as a common foundation. Early approaches used CLIP as a feature extractor for traditional classifiers like SVMs [11] or performed classification directly within CLIP's feature space [12].

More recent methods have shifted to parameter-efficient fine-tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) [10], to adapt CLIP more directly. This has been used to create specialized forgery awareness modules [13] or to incorporate fixed textual cues like “Camera” to guide classification [14]. A systematic comparison of various adaptation strategies, including prompt tuning and adapter

networks, provided strategies for effectively modifying CLIP for such tasks [15]. Beyond vision-language models, other strategies have targeted generative artifacts directly, such as by analyzing image gradients [16] or using reconstruction errors to distinguish real from synthetic images [17].

This study builds upon the PEFT approach by integrating LoRA for efficient fine-tuning. However, it is distinguished in two ways. First, instead of using simple labels, we employ LMMs to generate rich, descriptive text pairs for training, using prompting strategies inspired by methods like Self-Consistency [18]. Second, our framework combines the detection module with a generation trace interpretation module to provide explainability.

III. PROPOSED SYSTEM

This study proposes GIDLMM, a framework to detect AI-generated images and interpret the visual artifacts informing its classification. The model operates as a cohesive, four-stage pipeline as shown in the figure 1 below.

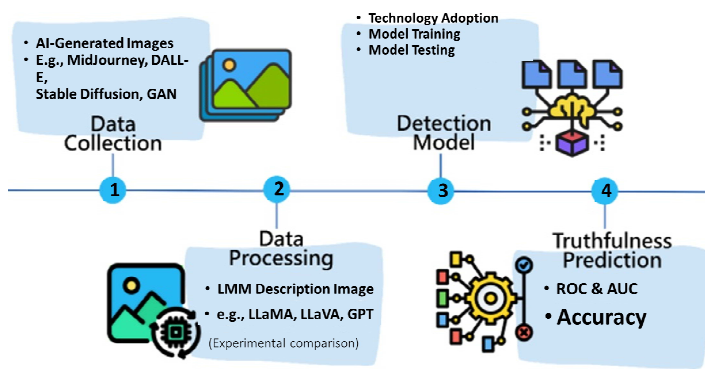


Figure 1. The four-stage architecture of the proposed GIDLMM framework, showing Data Collection, Data Processing, Detection Model, and Truthfulness Prediction.

A. Data Foundation and Semantic Enhancement

The model's performance relies on a high-quality dataset and enriched semantic data. This section details the dataset foundation and the data processing method used for training.

1) The GenImage Benchmark Dataset

This study uses the GenImage dataset [19], a large-scale benchmark constructed from the 1,000 object categories of ImageNet-1k. It contains real images and corresponding AI-generated versions from eight advanced generative models [22-28]. As detailed in Table 1, GenImage offers a balanced distribution of real and fake images, making it suitable for robust training and evaluation.

TABLE 1. GENIMAGE DATASET

Generation	Resolution	Train Dataset		Validation Dataset	
		Real	Fake	Real	Fake
BigGAN	128 x 128	162,000	161,996	6,000	6,000

GLIDE	256 x 256	162,000	161,999	6,000	6,000
VQDM	256 x 256	162,000	162,000	6,000	6,000
SD V1.4	512 x 512	162,000	161,997	6,000	6,000
SD V1.5	512 x 512	149,626	159,768	6,000	6,000
ADM	256 x 256	157,453	161,994	6,000	6,000
Midjourney	1024 x 1024	161,701	161,997	6,000	6,000
Wukong	512 x 512	160,739	162,000	6,000	6,000

2) LMM-Powered Image-Text Pair Generation

LMM-Powered Image-Text Pair Generation To improve the CLIP model's differentiation capabilities, we use LMMs to generate descriptive text for each image. To ensure factual consistency, a multi-LMM voting system verifies the accuracy of each description. We explored three prompt design strategies:

- A **Simple Descriptive** prompt, which instructs the LMM to describe either unnatural artifacts or the scene's content.
- A **Self-Consistency** method [18], where the LMM is queried multiple times for the same image, and the best description is selected through a subsequent vote.
- An **"AI Judges"** mechanism, our proposed efficient alternative, which prompts the LMM to generate and score multiple candidate descriptions from the perspective of different virtual judges in a single query. This method effectively balances quality and efficiency.

B. Model Architecture and Training

The GIDLMM framework comprises an AI-Generated Image Detection Module and a Generation Trace Interpretation Module. The overall operational flow is illustrated in Figure 2.

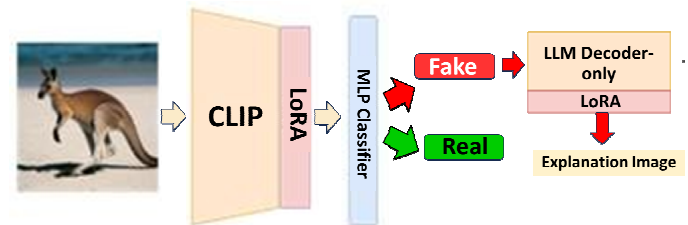


Figure 2. Operational Flow of the GIDLMM Model.

1) AI-Generated Image Detection Module

The detection module uses a CLIP-ViT model [9] fine-tuned with Low-Rank Adaptation (LoRA) [10]. LoRA is applied to the MLP components within the ViT Transformer blocks (Figure 3), allowing for efficient adaptation by updating only a small fraction of the model's parameters (<0.32%).

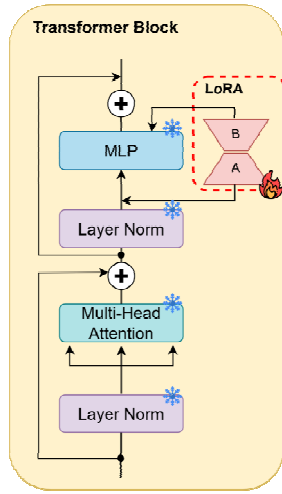


Figure 3. Transformer Block with Integrated LoRA Module [10]

During training, an image and its corresponding text description are passed through the LoRA-equipped image encoder and the frozen text encoder, respectively. The resulting features are used to compute a contrastive loss. The extracted visual features are then passed to a classifier consisting of three fully connected layers and a sigmoid function for final binary classification.

2) Generation Trace Interpretation Module

The second module provides explainability and is activated only after an image is classified as AI-generated. It takes the visual features (f_i) from the CLIP encoder and passes them through an MLP to project the embeddings into a space dimensionally compatible with a language model. The training process for this module is shown in Figure 4.

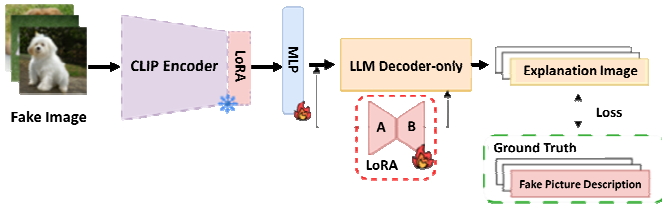


Figure 4. Training Architecture of the Generation Trace Interpretation Module

C. Inference Process and Evaluation Protocol

This section outlines the inference procedure for new images and the metrics used for quantitative performance evaluation.

1) Inference Process

During inference, the framework operates in an image-only input mode. An input image I is first processed by the trained LoRA-equipped CLIP image encoder to extract its visual feature embedding f_i . This vector is passed to the classifier, which outputs a prediction score between 0 and 1. Using a threshold of 0.5, the image is classified as AI-generated if the score is higher, and as real otherwise. The architecture for this stage is depicted in Figure 5.

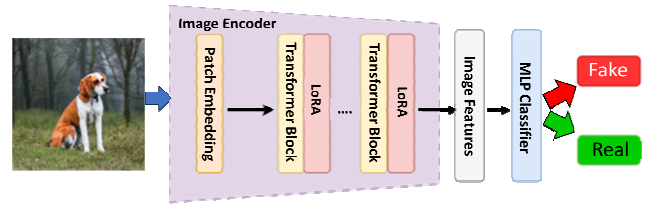


Figure 5. Inference Architecture of the AI-Generated Image Detection Module

If an image is classified as AI-generated, the same visual features f_i are relayed to the Generation Trace Interpretation Module. The features are processed by the alignment MLP and fed to the LoRA-equipped LLM to generate a natural language explanation of the visual artifacts that informed the decision, as shown in Figure 6.

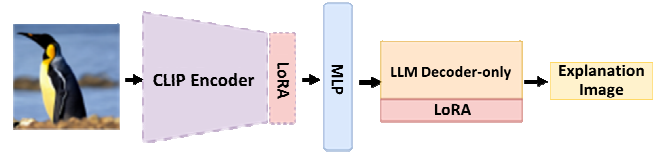


Figure 6. Inference Architecture of the AI-Generated Image Detection Module

2) Evaluation Metrics

The performance of the AI-Generated Image Detection Module is assessed using two standard metrics for binary classification:

Accuracy (Acc): Measures the overall proportion of correct predictions (both real and AI-generated) out of the total samples.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Area Under the Curve (AUC): Evaluates the model's ability to distinguish between classes across all classification thresholds. The AUC is the area under the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR)

$$TPR = \frac{TP}{TP + FN} \times 100\% \quad (6)$$

$$FPR = \frac{FP}{FP + TN} \times 100\% \quad (7)$$

IV. PERFORMANCE ANALYSIS AND RESULTS

This section presents a performance analysis of the proposed GIDLMM model. It begins by detailing the experimental environment and parameters, followed by an evaluation of the AI-generated image detection module across various configurations. The model is then compared against a state-of-the-art method to validate its effectiveness and data efficiency. Finally, the qualitative output of the generation trace interpretation module is presented.

A. Experimental Setup

Experiments were conducted on a system with an Intel i9-12900 CPU and an NVIDIA RTX 4070 Ti SUPER GPU. The hardware and training parameters for both modules are summarized in Table 2 and Table 3.

TABLE 2. KEY HARDWARE AND SOFTWARE SPECIFICATIONS

Device Name	Specification
CPU	Intel i9-12900
GPU	NVIDIA GeForce RTX 4070 Ti SUPER (VRAM 16 GB)
RAM	32 GB
Storage	2 TB SSD

TABLE 3. TRAINING PARAMETERS FOR GIDLMM MODULES

Parameter	Detection	Interpretation
Batch Size	8	2
Optimizer	SGD	AdamW
Loss Function	CrossEntropyLoss, BCELoss	CrossEntropyLoss
Learning Rate	1.00E-04	1.00E-04
LoRA Rank	8 (configurable)	8
LoRA Alpha	16	16
LoRA Dropout	0.05	0.05

B. Data Generation and Model Configuration Analysis

A series of studies confirmed the robustness of our design. Model performance was consistent regardless of the LMM used for data generation, with all configurations achieving around 89% accuracy (Table 4). The simple descriptive prompt design was most effective, likely due to CLIP's text input length limitations (Table 5). A comparison of CLIP variants showed that models with smaller patch sizes (ViT-B/16, ViT-L/14) were better suited for detecting subtle artifacts (Table 6).

TABLE 4. ACCURACY (%) AND AUC (%) RESULTS USING DIFFERENT LMMs

LMM for Data Processing	Average Accuracy	Average AUC
ChatGPT	89.47	94.09
LLaMA 3.2	89.59	96.43
Gemma 3	88.24	95.26
LLaVA	89.56	96.49

TABLE 5. ACCURACY COMPARISON OF DIFFERENT PROMPT DESIGNS

Prompt Design	Average Accuracy (%)
Simple Prompt	89.47
AI Judges	89.25
Self-Consistency	87.51

TABLE 6. ACCURACY RESULTS FOR DIFFERENT CLIP VARIANTS (SELF-CONSISTENCY PROMPTS)

CLIP Model	Average Accuracy (%)
ViT-B/32	81.56
ViT-B/16	89.60
ViT-L/14	89.25

C. Comparative Analysis with State-of-the-Art Methods

We benchmarked GIDLMM against the state-of-the-art CLIPping model [15], focusing on data efficiency. GIDLMM was trained on only 1k samples, while CLIPping was trained on 1k, 10k, and 100k samples.

As shown in Table 7, GIDLMM demonstrated superior data efficiency. It surpassed the CLIPping methods on the Diffusion and Commercial Tools series, even when they were trained on 100 times more data. On the Commercial Tools dataset, GIDLMM achieved an accuracy of 94.87%, an improvement of 13.85% over the best-performing CLIPping model.

TABLE 7. ACCURACY COMPARISON (%) WITH CLIPPING ADAPTER NETWORK AND LINEAR PROBING METHODS

Method	Samples	GAN	Diffusion	Commercial
CLIPping Linear (1k)	1k	56.28	68.95	78.7
CLIPping Adapter (1k)	1k	81.02	74.58	83.33
CLIPping Adapter (10k)	10k	85.32	81.61	91.78
CLIPping Adapter (100k)	100k	88.33	85.06	89.63
GIDLMM (r=8)	1k	74.49	89.06	94.87
GIDLMM (r=2)	1k	82.58	92.05	93.42

D. Qualitative Results of Generation Trace Interpretation

The interpretation module provided coherent explanations for the model's classifications. It successfully identified common visual artifacts, such as unnatural textures and inconsistent lighting, in AI-generated images (Table 8). An analysis of misclassifications showed that the model sometimes misinterpreted unusual but genuine visual features as synthetic artifacts, highlighting areas for future improvement (Table 9).

TABLE 8. EXAMPLE OUTPUTS FROM THE GENERATION TRACE INTERPRETATION MODULE

Generator	Generated Image	Model's Explanation
SD V1.4		The image appears fake due to unnatural textures, inconsistent lighting, and distorted features. The spider's body is overly smooth, with an unnatural sheen, and its legs appear too uniform in coloration.

GLIDE		The image appears to be AI-generated due to unnatural textures, inconsistent lighting, and distorted features. The monkey's fur lacks depth and dimension, with visible pixelation and an overall lack of realism.
-------	---	--

TABLE 9. CASES WHERE REAL IMAGES WERE MISCLASSIFIED AS AI-GENERATED IMAGES

Real Image	Model's Explanation for Misclassification
	The image exhibits unnatural textures, particularly on the leaves' edges, which appear jagged and lack the organic texture found in real-world images. This artifact is a common issue in AI-generated images.
	The image exhibits unnatural textures and lighting inconsistencies. The surface appears too uniform, lacking organic imperfections or subtle variations in color and tone, which are characteristic of real-world materials.

V. CONCLUSIONS

This study introduced the GIDLMM framework to address the need for transparent detection of realistic AI-generated images. The proposed dual-module architecture combines a data-efficient detector, using a LoRA-finetuned CLIP model trained on LMM-generated descriptions, with a module for explainable generation trace interpretation. Using significantly less training data, the model achieved high accuracy (~89%) and outperformed the state-of-the-art CLIPPING model by up to 13.85% on commercial datasets while providing coherent qualitative explanations. Future work will focus on incorporating more advanced models and developing a multi-turn, interactive interpretation system for more dynamic visual analysis.

REFERENCES

- [1] Adobe Communications Team, "Creative pros are leveraging Generative AI to do more and better work," 2024. [Online]. Available: <https://blog.adobe.com/en/publish/2024/02/02/creative-pros-generative-ai-usage>. (Visited 2025/01/04)
- [2] Amazon, "Amazon rolls out AI-powered image generation to help advertisers deliver a better ad experience for customers," 2024. [Online]. Available: <https://www.aboutamazon.com/news/innovation-at-amazon/amazon-ads-ai-powered-image-generator>. (Visited 2025/01/04)
- [3] Yahoo News, "【錯誤】網傳圖片「喜馬拉雅山上的寶塔花，每400年盛開一次」," 2024. [Online]. Available: <https://tw.news.yahoo.com/錯誤-網傳圖片-喜馬拉雅山上的寶塔花-每400年盛開一次-065918699.html>. (Visited 2025/01/04)
- [4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401-4410.
- [5] H. Jonathan, J. Ajay, and A. Pieter, "Denoising diffusion probabilistic models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840-6851, 2020.
- [6] F. Muhly, E. Chizzonic, and P. Leo, "AI-deepfake scams and the importance of a holistic communication security strategy," International Cybersecurity Law Review, vol. 6, pp. 53-61, 2025.
- [7] E. Ferrara, "GenAI against humanity: nefarious applications of generative artificial intelligence and large language models," Computational Social Science, vol. 7, pp. 549-569, 2024.
- [8] G. Bansal, A. Nawal, V. Chamola, and N. Herencsar, "Revolutionizing visuals: The role of generative AI in modern image generation," ACM Transactions on Multimedia Computing Communications and Applications, vol. 20, no. 11, 2024.
- [9] A. Radford, et al., "Learning Transferable Visual Models From Natural Language Supervision," in International Conference on Machine Learning, PMLR, 2021, pp. 8748-8763.
- [10] E. Hu, et al., "Lora: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.
- [11] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the Bar of AI-generated Image Detection with CLIP," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4356-4366.
- [12] U. Ojha, Y. Li, and Y. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24480-24489.
- [13] J. Xu, Y. Yang, H. Fang, H. Liu, and W. Zhang, "FAMSEC: A Few-shot-sample-based General AI-generated Image Detection Method," IEEE Signal Processing Letters, pp. 1-5, 2024.
- [14] C. Tan, R. Tao, H. Liu, G. Gu, B. Wu, Y. Zhao, and Y. Wei, "C2P-CLIP: Injecting Category Common Prompt in CLIP to Enhance Generalization in Deepfake Detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 7, pp. 7184-7192, 2025.
- [15] S. A. Khan and Duc-Tien Dang-Nguyen, "CLIPPING the Deception: Adapting Vision-Language Models for Universal Deepfake Detection," in Proceedings of the 2024 International Conference on Multimedia Retrieval, 2024, pp. 1006-1015.
- [16] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12105-12114.
- [17] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for Diffusion-Generated Image Detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22445-22455.
- [18] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," arXiv preprint arXiv:2203.11171, 2022.
- [19] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, and Y. Wang, "GenImage: A million-scale benchmark for detecting AI-generated images," in Proceedings of the Advances in Neural Information Processing Systems, 36th Conf., 2024.
- [20] Meta, "LLaMA," 2023. [Online]. Available: <https://www.llama.com/>. (Visited 2025/01/04)
- [21] A. Q. Nichol, et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," arXiv preprint arXiv:2112.10741, 2021.
- [22] S. Gu, et al., "Vector quantized diffusion model for text-to-image synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10696-10706.
- [23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684-10695.
- [24] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," Advances in Neural Information Processing Systems, vol. 34, pp. 8780-8794, 2021.
- [25] MidJourney, "Midjourney," 2023. [Online]. Available: <https://www.midjourney.com/home>. (Visited 2025/01/04)

- [26] J. Gu, et al., “Wukong: A 100 million large-scale Chinese cross-modal pre-training benchmark,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 26418-26431, 2022.



Yu-Ting Lu was born in Taiwan in 1999. He received the B.S. degree in 2022. She is currently pursuing an M.S. degree in electrical engineering with the National Taiwan University of Science and Technology, Taipei. Her main research interests include artificial intelligence and computer vision.



Jiann-Liang Chen, a Senior Member of IEEE, was born in Taiwan in December 1963. In 1989, he earned a Ph.D. in electrical engineering from the National Taiwan University in Taipei. Since August 1997, he has been a Professor and the Vice Dean of the Science and Engineering College at the Department of Computer Science and Information Engineering, National Dong Hwa University. He is currently a Distinguished Professor in the Department of

Electrical Engineering at the National Taiwan University of Science and Technology. His research focuses on cellular mobility management, cybersecurity, personal communication systems, and the Internet of Things (IoT).



Bagus Tri Atmaja, born on December 4th, 1996, is currently a doctoral candidate in Electronic Engineering at the National Taiwan University of Science and Technology (NTUST). Bagus completed his Master's Degree at NTUST in 2023, following a Bachelor's Degree from the Institut Teknologi Sepuluh Nopember (ITS), Indonesia, in 2019. His research interest focused on cybersecurity and IoT communication



Shih-Ping Chiu was born in Taipei, Taiwan, in 1984. She received the B.S. degree in 2007. She is a research assistant in electrical engineering from the National Taiwan University of Science and Technology, Taipei. Her main research interests include data analysis and the Internet of Things (IoT).