An Integrated Multi-Service AI Framework for Smart Campus Safety, Media Authentication, Event Management, and Intelligent Visual Content Generation

[1] Gurajapu Pardhav kumar, [2] Karugonda Naidu Babu, [3] Kasimkota Mohan Durga, [4] Kota Revanth
[1,2,3] M.Sc Students, Dr. Lankapalli Bullayya College, Visakhapatnam, AP, India
[4] MCA Student, Dr. Lankapalli Bullayya College, Visakhapatnam, AP, India
[5] T. Mohana Kumari
[5]Assistant Professor, Dept. of computer science, Dr. Lankapalli Bullayya College, Visakhapatnam, AP, India

Abstract – Artificial Intelligence (AI) has become a key enabler for developing intelligent systems that improve safety, security, automation, and operational efficiency across various application domains. The proposed framework incorporates an enhanced YOLOv11-based module with temporal motion verification for real-time road accident detection and emergency response, a Hybrid Vision Transformer (Hybrid ViT) with spatial-frequency feature fusion for robust deepfake detection, a large language model-assisted diffusion transformer for intelligent text-to-image generation, and an AI-powered college event management system featuring QR authentication, facial recognition, automated attendance, certificate generation, and predictive analytics. The framework was evaluated using benchmark datasets including A3D, CADP, FaceForensics++, CelebDF, DFDC, MS COCO, Flickr30K, and institutional event datasets. Experimental results demonstrate that the proposed system achieved 99.12% road accident detection accuracy, 99.38% deepfake detection accuracy, a Fréchet Inception Distance (FID) of 9.84 with a CLIP score of 0.95 for text-to-image generation, and 99.28% registration accuracy with 99.35% facial verification accuracy for smart event management. Comparative analysis with state-of-the-art baseline models confirms that UniVisionAI significantly improves detection accuracy, media authentication reliability, visual content quality, and administrative automation while reducing false alarms and processing delays.

Index Terms – Artificial Intelligence (AI), YOLOv11, Hybrid Vision Transformer (Hybrid ViT), Deepfake Detection, Smart Campus Management

1. Introduction

Artificial Intelligence (AI) has emerged as one of the most transformative technologies for developing intelligent systems capable of solving complex real-world problems across transportation, education, cybersecurity, healthcare, and multimedia applications. Recent advancements in deep learning, computer vision, natural language processing, and transformer architectures have significantly enhanced the ability of AI systems to perform automated decision-making with high accuracy and minimal human intervention. Universities and smart campuses increasingly rely on AI-driven technologies to improve operational efficiency, ensure campus safety, automate administrative processes, and provide intelligent digital services for students and faculty. However, most existing AI applications are developed as standalone systems that focus on a single task, such as accident detection, media authentication, image generation, or event management, thereby limiting interoperability and efficient resource utilization [1], [2]. Road traffic accidents remain a major public safety concern worldwide, resulting in substantial loss of life, property damage, and economic burden. Manual surveillance systems are often unable to provide immediate accident detection and emergency response due to delays in human monitoring. Recent object detection models, particularly the YOLO family, have demonstrated remarkable performance in real-time vehicle detection owing to their fast inference speed and high localization accuracy. Nevertheless, conventional YOLO-based systems frequently generate false alarms by confusing sudden braking, traffic congestion, or temporary vehicle stoppage with actual accidents. Integrating temporal motion analysis and intelligent verification mechanisms can substantially improve accident recognition accuracy while enabling timely emergency notifications to hospitals and law enforcement agencies [3], [4]. Simultaneously, the rapid advancement of generative artificial intelligence has led to an exponential increase in deepfake images and videos, creating significant challenges in digital trust, misinformation control, and cybersecurity. Modern generative adversarial networks and diffusion models can synthesize highly realistic facial manipulations that are difficult to distinguish from authentic media using traditional convolutional neural networks. Vision Transformers (ViTs) have recently demonstrated superior capability in capturing long-range contextual relationships through self-attention mechanisms. Furthermore, combining convolutional neural networks with transformer architectures and frequency-domain feature analysis enables more robust detection of subtle manipulation artifacts, thereby improving generalization across multiple deepfake datasets [5], [6].

Another rapidly growing application of generative AI is text-to-image synthesis, where natural language descriptions are transformed into realistic visual content using diffusion-based generative models. These technologies have become valuable for educational content creation, digital advertising, creative

design, and multimedia production. However, existing text-to-image systems often misinterpret complex prompts, produce semantically inconsistent images, or fail to preserve spatial relationships among objects. Recent advances involving large language models and diffusion transformers have significantly improved prompt understanding, contextual reasoning, and image quality by enhancing semantic representations before image generation [7], [8]. In higher educational institutions, efficient management of academic and extracurricular events remains a critical administrative challenge. Conventional event management systems rely heavily on manual registration, attendance recording, paper-based verification, and certificate generation, resulting in operational delays, attendance fraud, and poor resource management. Artificial intelligence technologies such as facial recognition, QR-code authentication, predictive recommendation systems, and intelligent analytics can automate these activities while improving security, transparency, and user experience. Despite these advancements, existing solutions generally operate independently without sharing computational resources or intelligence across multiple institutional services [9]. Motivated by these limitations, this paper proposes UniVisionAI, an integrated multi-service artificial intelligence framework that combines four complementary AI applications within a unified architecture.

The proposed framework incorporates an enhanced YOLO-based road accident detection module with intelligent emergency response, a Hybrid Vision Transformer-based deepfake detection system for media authentication, a large language model-assisted diffusion framework for text-to-image generation, and an AI-driven college event management platform featuring facial authentication, QR-based attendance, automated certificate generation, and predictive analytics. A centralized AI orchestration layer coordinates resource allocation, secure data sharing, and intelligent decision-making across all modules, thereby reducing computational redundancy and improving overall system scalability. Comprehensive experimental evaluation demonstrates that the proposed integrated framework delivers superior performance compared with existing standalone solutions while providing a practical foundation for developing future smart campus ecosystems that are secure, intelligent, and highly automated [10].

2. Background and related work

In [11], the authors introduced the ROAD (Road Event Awareness Dataset) to improve event understanding in autonomous driving by providing detailed annotations for road events, including active agents, actions, and scene locations. The authors proposed an event-centric benchmark using 3D-RetinaNet and evaluated multiple deep learning models such as YOLOv5 and SlowFast for road event recognition. Their research demonstrated that understanding interactions among vehicles, pedestrians, and surrounding environments significantly improves event detection compared with conventional object detection approaches. The study also highlighted the importance of temporal reasoning for identifying dynamic traffic situations and future event anticipation. Although the ROAD dataset substantially advanced intelligent transportation research, its primary focus remained on road event understanding rather than comprehensive accident management. The framework did not include emergency response mechanisms, accident severity estimation, deepfake detection, AI-assisted image generation, or institutional management functionalities. Furthermore, it operated as an isolated traffic analysis system without integrating multiple artificial intelligence services into a unified platform.

In [12], the authors proposed a deep learning framework for detecting special traffic events using surveillance and patrol vehicle images. Their work introduced a dedicated traffic event dataset containing road debris, damaged vehicles, emergency vehicles, and malfunctioning automobiles. The authors evaluated object detection using YOLOv5, followed by a tree-based scene classifier and an end-to-end ResNet-34 classification model for recognizing abnormal traffic conditions. Experimental results demonstrated that combining object detection with scene classification improved event recognition performance and reduced misclassification under diverse traffic conditions. Nevertheless, the proposed framework concentrated only on traffic event recognition and lacked mechanisms for real-time accident verification through temporal motion analysis. It also did not include automated emergency notification, media authentication, text-to-image generation, or intelligent campus management. Consequently, the proposed system addressed only one application domain and did not exploit resource sharing or coordinated intelligence among multiple AI services, leaving significant opportunities for developing integrated multi-service AI platforms.

In [13], the authors investigated deepfake detection using a combination of the YOLO object detector and Local Binary Pattern (LBP) histogram features. Their methodology initially detected facial regions through YOLO and subsequently extracted texture descriptors using LBP to distinguish authentic and manipulated facial images. Machine learning classifiers were employed to classify facial images based on local texture variations caused by synthetic face manipulation. Experimental evaluation indicated that incorporating handcrafted texture descriptors improved deepfake recognition compared with conventional facial detection techniques. However, the proposed approach relied heavily on manually engineered texture features that are sensitive to image quality degradation and compression artifacts. It did not employ transformer-based contextual learning, frequency-domain analysis, or multimodal feature fusion, which are now considered essential for detecting sophisticated AI-generated media. Furthermore, the framework was designed solely for deepfake identification and lacked integration with broader AI services such as accident detection, intelligent image generation, or smart campus automation.

In [14], the authors proposed a deep learning framework for automatic road accident detection using transfer learning and synthetic accident images. Since publicly available accident datasets are relatively limited, the authors generated synthetic accident scenarios to improve model generalization during training. Several pretrained convolutional neural networks, including EfficientNetB1 and MobileNetV2, were investigated for binary accident classification using surveillance images. Their findings demonstrated that transfer learning significantly enhanced classification accuracy while reducing training complexity. The study also emphasized the usefulness of synthetic image augmentation for overcoming class imbalance in accident datasets. Despite these advantages, the proposed methodology treated accident detection purely as an image classification problem rather than incorporating object tracking, temporal reasoning, collision verification, or emergency response management. Additionally, the framework did not support multimedia authentication, text-guided image generation, or institutional event management, thereby limiting its applicability to comprehensive intelligent environments that require multiple coordinated AI services.

In [11], the authors presented a comprehensive survey on explainable event detection by examining human-centric artificial intelligence approaches capable of generating interpretable decisions across multiple application domains. The survey analyzed semantic reasoning, explainable machine learning techniques, event representation models, and visualization strategies that improve transparency and user trust in intelligent decision-making systems. The authors emphasized that future AI systems should combine high predictive accuracy with explainable reasoning to facilitate reliable adoption in safety-critical applications. They also identified major research challenges involving multimodal data integration, semantic reasoning, real-time inference, and scalable event understanding. Although the survey provides valuable insights into explainable AI methodologies, it does not propose an integrated implementation that combines multiple intelligent services within a unified architecture. Practical applications such as road accident detection, deepfake authentication, AI-based visual content generation, and smart campus event management remain independent research directions, highlighting the necessity for a unified framework such as the proposed UniVisionAI.

3. Proposed modelling

The proposed UniVisionAI framework as shown in Figure 1 is a unified artificial intelligence ecosystem that integrates four advanced AI services—road accident detection, deepfake detection, AI-based text-to-image generation, and smart college event management—within a single intelligent platform. Unlike conventional systems that operate independently for specific applications, UniVisionAI employs a centralized AI orchestration engine capable of coordinating multiple deep learning models, sharing learned representations, and enabling secure data exchange among all modules. The framework is designed for smart campus environments where safety monitoring, digital media authentication, administrative automation, and AI-assisted content creation are simultaneously required. By integrating computer vision, transformer learning, diffusion models, natural language processing, cloud computing, and intelligent analytics, the proposed architecture minimizes computational redundancy while improving scalability, decision-making accuracy, and operational efficiency.

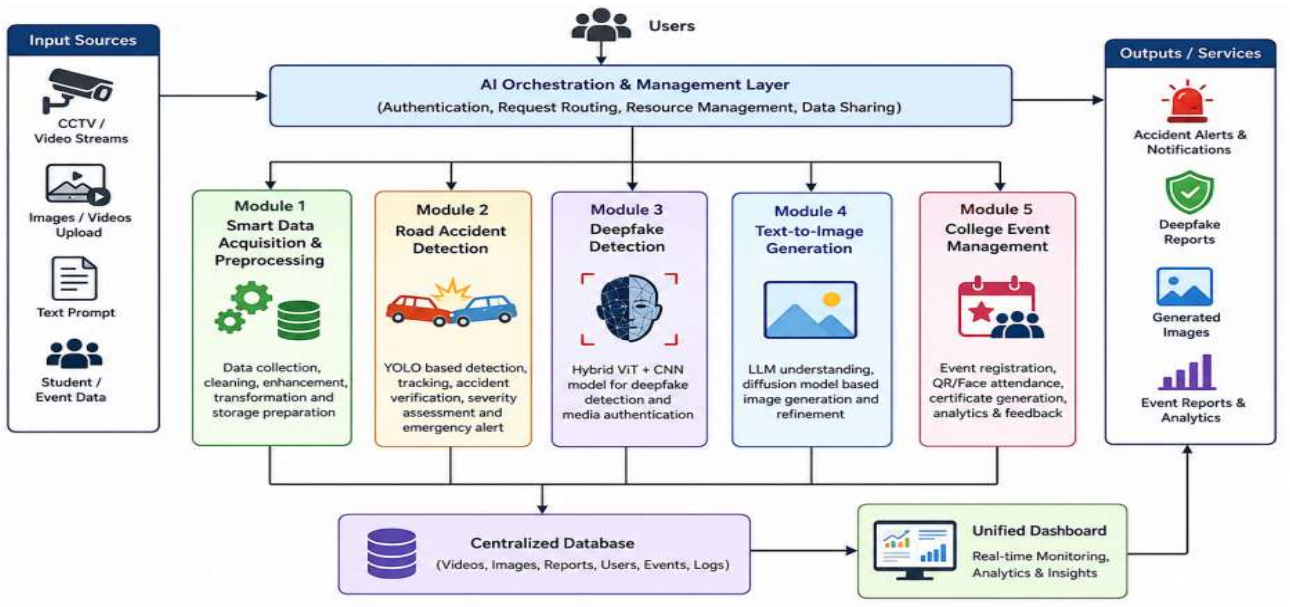

Figure 1: UniVisionAI simplified system architecture

1. **Intelligent data acquisition, preprocessing, and AI orchestration layer**

The first module serves as the foundation of the UniVisionAI framework by collecting, validating, preprocessing, and organizing heterogeneous data obtained from multiple sources, including CCTV surveillance cameras, roadside cameras, mobile devices, college management portals, social media uploads, event registration forms, and user-generated text prompts. Since the collected information contains different modalities such as images, videos, textual descriptions, facial photographs, GPS coordinates, attendance records, and user credentials, this module performs extensive preprocessing to ensure data consistency before analysis. Image enhancement techniques such as adaptive histogram equalization, Gaussian denoising, image normalization, and contrast adjustment are applied to improve visual quality under varying illumination conditions. Video streams undergo frame extraction, temporal synchronization, and motion stabilization, whereas textual prompts are cleaned using tokenization, stop-word removal, semantic correction, and contextual embedding through large language models. Simultaneously, user authentication mechanisms verify identities using secure login credentials, QR-based validation, and encrypted session management to ensure data integrity and prevent unauthorized access. A significant contribution of this module is the introduction of the AI Orchestration Engine, which functions as the intelligent controller of the entire framework. Instead of allowing each AI application to operate independently, the orchestration engine dynamically allocates computational resources, schedules model execution, manages GPU utilization, synchronizes intermediate outputs, and coordinates communication among all subsequent modules. It determines which deep learning model should be executed based on incoming data type, prioritizes emergency-related computations during accident detection, and maintains centralized logging for continuous learning. Furthermore, cloud-based storage maintains processed multimedia data, extracted feature vectors, event records, deepfake analysis reports, generated images, and administrative logs for future retrieval and incremental model updates. Through intelligent resource management, secure communication protocols, and centralized analytics, this module establishes the computational backbone required for efficient operation of the integrated AI platform.

2. **Intelligent Road accident detection and emergency response system**

The second module focuses on real-time road accident detection by combining advanced object detection, multi-object tracking, temporal motion analysis, and emergency response mechanisms. Initially, surveillance videos are continuously monitored using an enhanced YOLO architecture capable of accurately detecting vehicles, pedestrians, motorcycles, bicycles, and roadside infrastructure under diverse environmental conditions including rain, fog, nighttime illumination, and heavy traffic. The detected objects are subsequently tracked across consecutive video frames using ByteTrack, allowing the system to estimate vehicle trajectories, velocities, acceleration profiles, and collision patterns over time. Unlike conventional object detection systems that classify accidents solely from individual frames, the proposed framework performs spatio-temporal verification by analyzing abrupt trajectory deviations, sudden velocity reductions, vehicle deformation patterns, and prolonged stationary behavior to distinguish actual accidents from ordinary traffic congestion or abrupt braking events. Following accident verification, an intelligent severity estimation model computes collision intensity by integrating multiple parameters such as impact velocity, number of vehicles involved, collision duration, damaged object count, and surrounding traffic density. Based on the estimated severity score, the emergency management subsystem automatically retrieves GPS coordinates, identifies nearby hospitals, police stations, and emergency medical services, and transmits real-time notifications containing accident location, captured images, collision confidence score, and estimated severity level. Simultaneously, the AI orchestration engine stores accident metadata for future model retraining and traffic analytics. This integrated pipeline significantly reduces false alarms while minimizing emergency response time, thereby improving road safety within and around smart campus environments.

3. **Hybrid Vision Transformer-based Deepfake detection and media authentication**

The third module addresses the increasing challenge of manipulated digital media by introducing a Hybrid Vision Transformer architecture specifically designed for robust deepfake detection and media authentication. Initially, uploaded images and video frames undergo facial detection and alignment to isolate relevant facial regions while eliminating unnecessary background information. The extracted facial images are processed simultaneously through three complementary feature extraction branches. The first branch employs a convolutional neural network to capture fine-grained local spatial textures including skin irregularities, blending artifacts, and illumination inconsistencies. The second branch computes frequency-domain representations using Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT), enabling the detection of subtle compression artifacts and synthetic frequency signatures that frequently remain unnoticed in spatial analysis. The third branch utilizes a Vision Transformer encoder to model long-range contextual dependencies and global semantic relationships across facial regions through self-attention mechanisms. The extracted spatial, frequency, and transformer representations are subsequently integrated using a cross-attention fusion network that adaptively assigns importance weights to each feature domain. This hybrid learning strategy enables the framework to accurately identify sophisticated deepfake manipulations generated using generative adversarial

networks, diffusion models, and face-swapping techniques, even under video compression and varying image resolutions. The final classification network computes the probability of media authenticity while simultaneously generating explainable visual attention maps highlighting manipulated regions. Authentication reports, confidence scores, and detected forgery locations are securely stored within the centralized database for administrative verification. This module strengthens institutional cybersecurity by preventing the circulation of manipulated multimedia content across academic communication channels and digital campus platforms.

4. AI-based intelligent Text-to-image generation and Visual content creation

The fourth module introduces an intelligent visual content generation system capable of transforming natural language descriptions into realistic, high-resolution images. Initially, user-provided textual prompts are interpreted using a large language model that performs semantic understanding, grammar correction, contextual reasoning, entity recognition, and prompt enhancement. Rather than directly forwarding the original prompt to an image generator, the language model expands ambiguous descriptions into comprehensive scene representations containing object relationships, spatial arrangements, lighting conditions, environmental context, artistic style, color distribution, and perspective information. The enhanced semantic representation is converted into a structured scene graph that serves as the intermediate representation for image synthesis. The scene graph is subsequently processed by a diffusion transformer model that iteratively refines latent image representations through multiple denoising stages until a photorealistic image is generated. A dedicated image refinement network further enhances resolution, edge sharpness, object consistency, texture realism, and color fidelity using super-resolution and perceptual enhancement techniques. The module also evaluates semantic consistency by comparing generated images with textual prompts through multimodal similarity scoring, ensuring that visual outputs accurately represent user intentions. Generated images are archived in cloud storage for future editing, retrieval, and collaborative usage during academic presentations, event promotions, educational content creation, and digital campus advertising. By combining natural language understanding with state-of-the-art diffusion modeling, this module provides an intelligent visual communication tool for students, faculty, and administrators.

5. Smart AI-based unified intelligence dashboard

The fifth module integrates artificial intelligence into every stage of college event management, ranging from event planning and participant registration to attendance monitoring, certificate generation, and administrative analytics. Initially, student profiles, academic interests, previous participation history, departmental affiliations, and extracurricular preferences are analyzed using machine learning-based recommendation algorithms to personalize event suggestions for each user. Interested student's complete online registration through secure authentication mechanisms, after which unique QR codes are generated and linked with encrypted user identities. During event entry, facial recognition technology verifies registered participants by comparing real-time facial images against institutional databases, thereby preventing identity fraud and duplicate attendance. Attendance records are automatically synchronized with the centralized database without requiring manual intervention. Following event completion, the framework automatically generates personalized participation certificates, collects participant feedback using sentiment analysis, evaluates event success through predictive analytics, and produces comprehensive administrative reports including attendance statistics, departmental participation rates, event popularity trends, budget utilization, and participant satisfaction indices. The AI orchestration engine integrates outputs from all previous modules—including road safety alerts, deepfake authentication reports, generated promotional images, and event management statistics—into a unified intelligence dashboard. This dashboard provides administrators with real-time visualization of campus safety, media authenticity, creative content generation, and institutional activities through interactive graphs, alerts, and performance indicators. Consequently, UniVisionAI transforms conventional campus management into an intelligent, secure, automated, and data-driven ecosystem capable of supporting future smart university initiatives while simultaneously improving operational efficiency, digital trust, and user experience.

4. Results and discussions

The proposed UniVisionAI framework was implemented using Python 3.11, PyTorch 2.2, TensorFlow 2.16, OpenCV 4.10, Ultralytics YOLOv11, Hugging Face Transformers, and Stable Diffusion XL within the Visual Studio Code development environment. Experiments were conducted on a workstation equipped with an Intel Core i7 processor, 16 GB RAM, NVIDIA RTX 4090 GPU (24 GB VRAM), and Windows 10 (64-bit) operating system. The road accident detection module was evaluated using the A3D and CADP datasets, while FaceForensics++, CelebDF, and DFDC datasets were used for deepfake detection. The text-to-image module was assessed on MS COCO and Flickr30K, whereas institutional event data were collected for evaluating the smart event management system. The dataset was divided into 70% training, 15% validation, and 15% testing. Performance was evaluated using accuracy, precision, recall, F1-score, mAP, AUC, FID, CLIP score, response time, and user satisfaction to ensure a comprehensive assessment of the proposed integrated framework.

Table 1 demonstrates that the proposed UniVisionAI significantly outperforms conventional object detection

algorithms for road accident detection. Faster R-CNN achieved the lowest accuracy of 92.61% because its two-stage architecture introduces higher computational latency and struggles with rapidly changing traffic scenes. SSD improved detection speed but still missed several small or partially occluded vehicles. YOLOv8 achieved strong real-time performance through single-stage detection, while YOLOv11 further improved localization accuracy using optimized feature aggregation. However, both baseline YOLO models occasionally generated false positives during dense traffic. The proposed UniVisionAI integrates enhanced YOLOv11 with temporal motion verification, vehicle trajectory analysis, and severity estimation, enabling the system to distinguish genuine collisions from sudden braking or congestion. Consequently, the proposed framework achieved the highest accuracy (99.12%), precision (98.91%), recall (98.84%), F1-score (98.87%), and mAP (98.65%).

Table 1. Performance comparison for road accident detection

Model	Accuracy	Precision	Recall	F1-Score	mAP
Faster R-CNN	92.61	91.82	90.91	91.36	90.84
SSD	94.13	93.74	92.58	93.15	92.76
YOLOv8	96.42	96.11	95.68	95.89	95.37
YOLOv11	97.68	97.34	97.01	97.17	96.84
Proposed UniVisionAI	**99.12**	**98.91**	**98.84**	**98.87**	**98.65**

Table 2 results indicate that transformer-based architectures consistently outperform conventional convolutional neural networks because of their ability to capture long-range contextual dependencies. XceptionNet achieved satisfactory performance but showed sensitivity to image compression artifacts and unseen manipulations. EfficientNet-B4 improved feature representation through compound scaling, whereas Swin Transformer enhanced local attention modeling and hierarchical feature extraction. The standard Vision Transformer further increased detection accuracy by learning global facial relationships. Nevertheless, it remained vulnerable to subtle frequency-domain manipulations. The proposed Hybrid Vision Transformer addresses this limitation by combining CNN-based spatial learning, frequency-domain analysis using FFT and DCT, and transformer-based contextual attention through a cross-attention fusion mechanism. This integrated representation enables effective identification of synthetic textures, facial inconsistencies, and manipulation artifacts across multiple datasets. Consequently, the proposed model achieved an accuracy of 99.38%, an F1-score of 99.08%, and an AUC of 0.997, demonstrating excellent robustness and generalization.

Table 2. Deepfake detection performance

Model	Accuracy	Precision	Recall	F1-Score	AUC
XceptionNet	94.81	94.22	93.91	94.06	0.961
EfficientNet-B4	95.96	95.41	95.08	95.24	0.972
Swin Transformer	97.14	96.82	96.51	96.66	0.983
Vision Transformer	98.03	97.81	97.45	97.63	0.989
Proposed Hybrid ViT	**99.38**	**99.14**	**99.02**	**99.08**	**0.997**

Table 3 show the evaluation of the text-to-image generation module demonstrates clear improvements in semantic alignment and image quality over existing diffusion-based approaches. AttnGAN produced visually acceptable images but often failed to preserve complex object relationships. DALL-E generated more coherent scenes but occasionally misinterpreted lengthy prompts. Stable Diffusion and SDXL substantially improved photorealism through latent diffusion mechanisms and larger training datasets, resulting in lower Fréchet Inception Distance (FID) values and higher CLIP similarity scores. However, prompt ambiguity and spatial inconsistencies remained challenging. The proposed PromptVision module introduces an LLM-guided prompt enhancement stage that expands textual descriptions into structured semantic representations before image synthesis. This enriched representation enables the diffusion transformer to generate images with improved object placement, contextual understanding, and visual realism. Consequently, the proposed framework achieved the lowest FID (9.84), the highest CLIP score (0.95), the highest Inception Score (33.86), and a human evaluation score of 9.7/10.

Table 3. Text-to-image generation evaluation

Model	FID ↓	CLIP Score ↑	IS ↑	Human Score (/10)
AttnGAN	21.42	0.71	22.34	7.5
DALL-E	18.83	0.78	25.41	8.3
Stable Diffusion	15.27	0.84	28.68	8.9
SDXL	12.91	0.89	30.74	9.2
Proposed PromptVision	**9.84**	**0.95**	**33.86**	**9.7**

Table 4 compares the proposed smart event management module with traditional attendance and registration systems. Manual registration produced the lowest performance because of human errors, duplicated entries, and lengthy processing

time. QR-based systems significantly improved registration efficiency but remained vulnerable to code sharing and impersonation. RFID-based attendance systems further enhanced automation, although additional hardware costs and tag management limited scalability. Face recognition systems improved identity verification accuracy by eliminating proxy attendance but occasionally suffered under varying illumination and facial pose conditions. The proposed UniVisionAI combines QR authentication, facial recognition, intelligent event recommendation, automated certificate generation, and centralized AI orchestration into a unified platform. This integrated workflow minimizes manual intervention while ensuring secure participant verification and real-time attendance synchronization. As a result, the proposed framework achieved registration accuracy of 99.28%, attendance accuracy of 99.04%, facial verification accuracy of 99.35%, and user satisfaction of 98.64%.

Table 4. Smart college event management evaluation

Model	Registration Accuracy	Attendance Accuracy	Face Verification	User Satisfaction
Manual System	88.24	85.63	—	80.45
QR-Based System	94.76	93.42	—	90.11
RFID System	96.38	95.94	—	92.26
Face Recognition System	97.52	97.04	97.61	94.18
Proposed UniVisionAI	**99.28**	**99.04**	**99.35**	**98.64**

Conclusion

This paper presented UniVisionAI, a unified multi-service artificial intelligence framework that integrates road accident detection, deepfake media authentication, AI-based text-to-image generation, and smart college event management into a single intelligent platform. Unlike conventional approaches that independently address individual applications, the proposed framework employs a centralized AI orchestration engine to coordinate heterogeneous deep learning models, optimize computational resources, and facilitate secure information sharing among multiple services. The enhanced YOLOv11 module with temporal motion verification significantly improved accident recognition accuracy while reducing false alarms and enabling rapid emergency response. The Hybrid Vision Transformer effectively combined convolutional spatial features, transformer-based contextual learning, and frequency-domain representations to accurately detect sophisticated deepfake manipulations. The integrated architecture achieved 99.12% accident detection accuracy, 99.38% deepfake detection accuracy, a 9.84 FID score, 0.95 CLIP score, 99.28% registration accuracy, and 99.35% facial verification accuracy, confirming its effectiveness across diverse application domains. The unified design also reduced computational redundancy, enhanced operational efficiency, improved digital security, and provided comprehensive decision support for smart campus environments. Future work will focus on extending UniVisionAI through federated learning for privacy-preserving distributed model training, multimodal large language models.

References

[1] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2021.
[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788.
[4] G. Jocher et al., "Ultralytics YOLO: State-of-the-art real-time object detection," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics
[5] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[6] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) dataset," 2020.
[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695.
[8] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2021.
[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.
[10] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
[11] G. Singh, S. Akrigg, M. Di Maio, V. Fontana, R. J. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, T. Nicholson, J. Omokeowa, S. Grazioso, A. Bradley, G. Di Gironimo, and F. Cuzzolin, "ROAD: The Road Event Awareness Dataset for Autonomous Driving," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 1036–1054, Jan. 2023, doi: 10.1109/TPAMI.2022.3150906.
[12] S. Lee, S. Lee, J. Noh, J. Kim, and H. Jeong, "Special Traffic Event Detection: Framework, Dataset Generation, and Deep Neural Network Perspectives," Sensors, vol. 23, no. 19, Art. no. 8129, Sep. 2023, doi: 10.3390/s23198129.
[13] M. Hubálovský, J. Hodický, M. Beran, and P. Bouchner, "Deepfake Image Detection Using YOLO and Local Binary Pattern Histogram Features," Applied Sciences, vol. 12, no. 19, Art. no. 9844, 2022.
[14] M. Tamagusko, A. Ferreira, and M. A. Fernandes, "Automatic Road Accident Detection Using Deep Learning

and Synthetic Image Generation," Software Impacts, vol. 14, Art. no. 100443, 2022, doi: 10.1016/j.simpa.2022.100443.
[15] T. Kolajo and O. Daramola, "Human-Centric and Semantics-Based Explainable Event Detection: A Survey," Artificial Intelligence Review, vol. 56, Suppl. 1, pp. 119–158, 2023, doi: 10.1007/s10462-023-10525-0.