Sensors and Materials, Vol. 37, No. 9 (2025) 4309–4321
MYU Tokyo

S & M 4183

Employee Work Behavior Monitoring Using Multimodal Large Language Models

Yushi Chen,[1] Chung-Hsing Chao,[2] Linjing Liu,[1*] and Cheng-Fu Yang[3,4**]

[1]Business School, Dongguan City University, Guangdong Province 523419, China
[2]Department of Intelligent Vehicles and Energy, Minth University of Science and Technology, Hsinchu 307, Taiwan
[3]Department of Chemical and Materials Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan
[4]Department of Aeronautical Engineering, Chaoyang University of Technology, Taichung 413, Taiwan

(Received September 16, 2025; accepted September 24, 2025)

Keywords: multimodal large language models, employee behavior monitoring, smart office, prompt engineering, privacy protection

With the rapid advancement of artificial intelligence, enterprises increasingly demand efficient and flexible solutions for employee work behavior monitoring in office environments. Traditional systems often involve high costs, rigidity, and reliance on extensive labeled data. Multimodal large language models (MLLMs), capable of integrating information from text, images, and audio, offer a novel zero-shot inference approach that reduces data dependence and deployment complexity. In this study, we present a practical application framework combining seating area definition, image cropping, and prompt engineering to analyze employee behaviors such as focused screen engagement and nonwork-related interactions. Results are output in a standardized JavaScript Object Notation format facilitating aggregation and actionable insights for human resource management. Additionally, critical privacy and ethical and legal considerations are discussed, along with mitigation strategies to support responsible deployment. Through practical simulation scenarios and cost–benefit analysis, we demonstrate that MLLMs enable scalable and economically viable employee behavior monitoring solutions suitable for small and medium-sized enterprises.

1. Introduction

With the global wave of digital transformation, artificial intelligence (AI) has profoundly reshaped corporate operations. In office environments, monitoring employee behavior has become increasingly important for enhancing productivity, ensuring information security, and maintaining compliance. However, traditional monitoring systems often entail high deployment costs, limited flexibility, and strong data dependence.(1) These challenges underscore the need for more advanced and cost-effective solutions. The emergence of multimodal large language models (MLLMs) provides a transformative opportunity, as they can process and integrate text, images, and audio, offering new directions for overcoming the shortcomings of conventional

*Corresponding author: e-mail: liulingjing@dgcu.edu.cn
**Corresponding author: e-mail: cfyang@nuk.edu.tw
https://doi.org/10.18494/SAM5937

ISSN 0914-4935
https://myukk.org/

approaches.[2,3] This research is therefore motivated by the demand for efficient and ethical monitoring systems that align with modern work dynamics while safeguarding privacy.

AI has been significantly advanced by large language models (LLMs) that excel at processing human language, as shown in Fig. 1. However, real-world contexts are inherently multimodal, relying on diverse sensory inputs. To address this, the AI community has rapidly developed MLLMs, extending beyond text-based models to integrate multiple modalities. Early multimodal methods relied on loosely connected models with limited synergy, but the introduction of transformer architectures marked a breakthrough. Initially designed for natural language processing [e.g., bidirectional encoder representations from transformers and generative pretrained transformers (GPTs)], transformers were later adapted for computer vision [e.g., vision transformer (ViT)] and subsequently expanded into unified multimodal frameworks. Contemporary MLLMs employ mechanisms such as connectors and multimodal attention to fuse different input types, with models such as large language-and-vision assistant (LLaVA), Flamingo, and GPT-4V, demonstrating impressive performance in tasks such as image captioning, visual question answering, and multimodal dialogue. Their rapid evolution has been fueled by advances in large-scale datasets, computational power, and architectural innovations.[4]

In this study, we investigate how MLLMs can be leveraged as a flexible, scalable, and cost-effective solution for monitoring employee behavior in office environments. Through a simulated application scenario, we demonstrate their operational workflow in defining seating areas, capturing images, analyzing behavior, and generating outputs. Beyond the technical aspects, we also examine the ethical, privacy, and legal implications of adopting MLLM-based monitoring and propose mitigation strategies to ensure responsible implementation.[5] The contributions of this study include the following: establishing the feasibility of zero-shot MLLM inference for workplace monitoring, presenting a practical methodology from data acquisition to classification and reporting, and providing a comprehensive discussion of ethical and regulatory considerations. Overall, we advance the application of MLLMs in workplace management, offering both technical insights and ethical guidance for future smart office systems.

2. Technical Framework

MLLMs represent a major advancement in AI by integrating information from diverse modalities for more comprehensive understanding. Unlike unimodal models such as text-only LLMs or image-only convolutional neural networks (CNNs), MLLMs can simultaneously process text, images, audio, and video.[6] Their architecture generally consists of three essential components. First, the multimodal encoder transforms raw inputs (e.g., image pixels, audio waveforms, and text tokens) into embeddings, often leveraging pretrained models such as ViT for visual data and Wav2Vec for audio.[7] Second, the alignment module integrates these embeddings into a shared representation space, allowing semantic relationships to be captured across modalities. Techniques such as contrastive learning, exemplified by contrastive language-image pretraining (CLIP), are commonly applied for this purpose.[6] Finally, the MLLMs themselves process the aligned multimodal embeddings to perform tasks of inference, generation, and comprehension. For example, MLLMs can produce image captions or answer

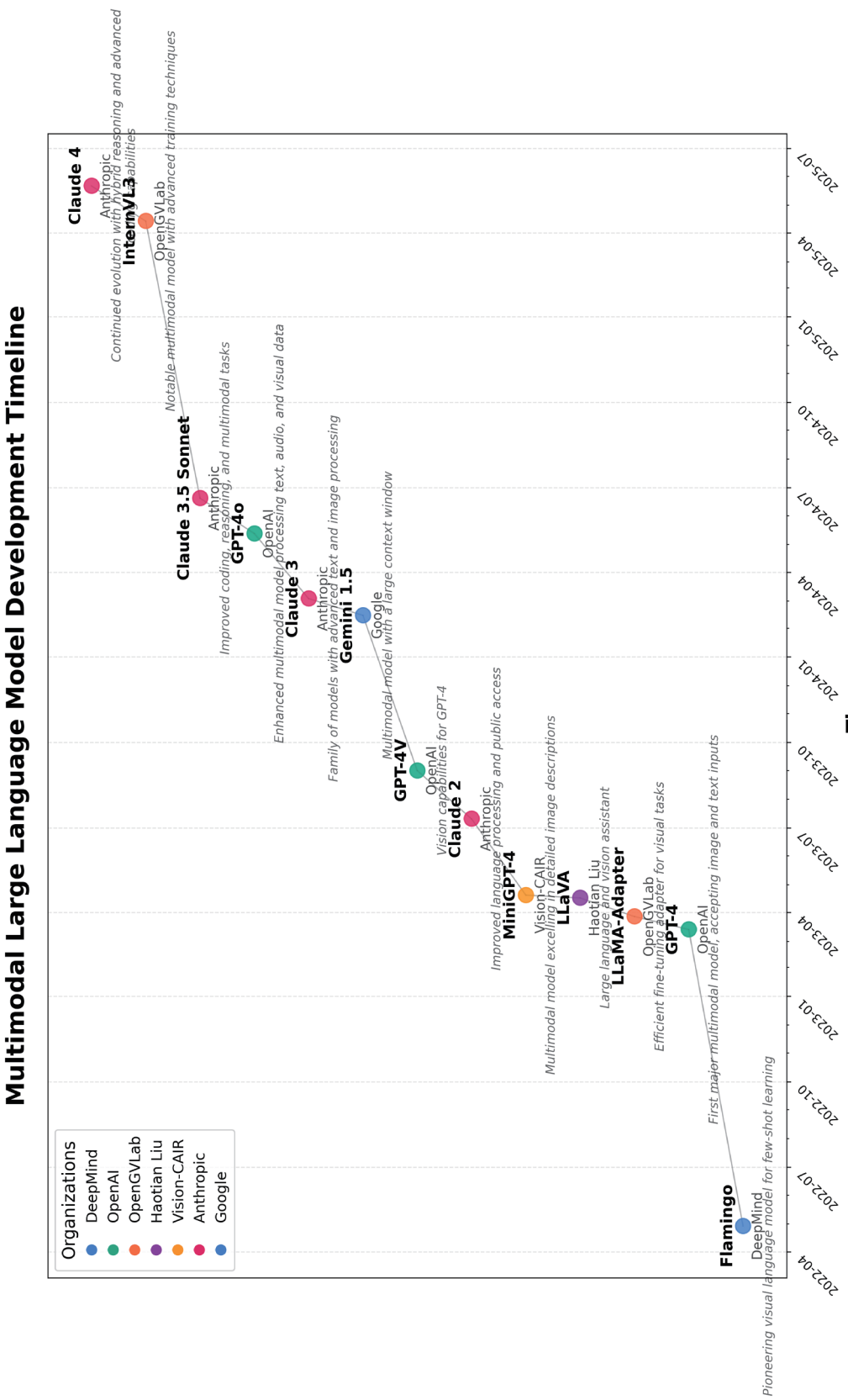

Fig. 1. (Color online) MLLM development timeline.

content-specific queries by combining multimodal cues.[8] Through this integrated architecture, MLLMs achieve a holistic understanding of complex real-world scenarios by exploiting complementary data sources.

A key advantage of MLLMs lies in their strong zero-shot and few-shot inference capabilities, enabling them to handle novel tasks without extensive task-specific training. This adaptability derives from large-scale pretraining on diverse datasets, which equips the model with broad general knowledge and robust cross-modal associations.[9] Consequently, MLLMs demonstrate high flexibility in addressing unseen data distributions and tasks. Equally important is the role of prompt engineering in practical deployment. By designing appropriate prompts, users can guide MLLMs to execute specific tasks without fine-tuning. For instance, in an employee monitoring context, prompts such as "Please analyze whether the employee in this image is looking at the screen" can directly yield behavior analysis results. This approach significantly reduces technical barriers, enabling nonspecialist developers to harness MLLMs effectively.[10] Prompt engineering thus transforms MLLMs into versatile tools, supporting a wide range of applications with minimal customization.

MLLMs exhibit significant potential in behavior recognition by integrating visual cues, such as employee posture and gaze, with speech content, enabling the accurate judgment and classification of complex human behaviors. For instance, as illustrated in Fig. 2, MLLMs can detect nonwork-related activities during office hours, including mobile phone usage, prolonged absences from desks, or private conversations. They are also capable of interpreting the semantics behind interactions, distinguishing, for example, between meetings and idle chat.[11] Note that while MLLMs can help identify general patterns of device usage (e.g., frequency, duration, and timing), they cannot, without access to sensitive content such as audio or message transcripts, reliably distinguish between work-related and nonwork-related activities. Since

Fig. 2. (Color online) MLLM employee behavior monitoring.

incorporating such content-level data introduces significant privacy and confidentiality risks, our framework explicitly avoids analyzing or storing raw communication data. Instead, we focus on nonintrusive behavioral indicators to maintain user privacy while still capturing meaningful patterns of technology interaction. This multimodal understanding allows for more nuanced and context-aware assessments of workplace activity, surpassing basic presence detection to provide comprehensive insights essential for effective and fair monitoring. Overall, the combination of multimodal processing, zero-shot inference, and flexible prompt engineering makes MLLMs a highly efficient, cost-effective, and easily deployable solution for employee behavior monitoring, addressing limitations of traditional approaches.

To assess feasibility and cost-efficiency, we implemented a practical MLLM-based monitoring system in a simulated office environment with 10 employee seats. Behavior detection was performed every 5 min over an 8 h workday, totaling 960 detection events per day (12 intervals × 8 h × 10 seats). With a conservative Application Programming Interface cost of US$0.0017 per inference using LLaVA, daily monitoring expenses are approximately US$1.63.[12] The system employs LLaVA (v1.5) with a CLIP-ViT-L-336px vision encoder, a Vicuna-13B language model, and a two-layer Multilayer Perceptron connector for modality alignment. Training is conducted in two stages: feature alignment (558K LAION-CC-SBU samples, batch size 128, learning rate 1e−4, and 5.5 h on 8 A100 GPUs) and visual instruction tuning (150K LLaVA-Instruct and 515K Visual Question Answering samples, batch size 256, learning rate 2e-5, and Low-Rank Adaptation rank 16). Inference leverages 4-bit quantization, processes 1080p frames resized to 336 × 336, and uses prompts to classify employee behaviors, such as determining whether the employee is looking at the screen.

The system then generates JavaScript Object Notation (JSON) outputs including behavior labels, confidence scores, and timestamps. DeepSpeed ZeRO-2 ensures real-time monitoring, while privacy is protected through data minimization and anonymization. The output of MLLMs is typically organized in a structured JSON format, where information is arranged as easily readable key–value pairs. This structure allows the results to be systematically represented, making them straightforward to interpret by humans and readily processed by computer programs for subsequent analysis or integration into applications. This low operational cost highlights the system's accessibility and scalability for small and medium-sized enterprises. The simulation demonstrates the seamless integration of computer vision for seat-specific cropping with MLLM-driven multimodal analysis, ensuring efficient computation while leveraging MLLMs' semantic reasoning for precise behavioral insights.

3. Methodology

The proposed MLLM-based employee work behavior monitoring system is designed to be modular, scalable, and adaptable to diverse office environments. Its architecture integrates several interconnected components to ensure efficient data flow and robust behavioral analysis, with off-the-shelf MLLMs such as GPT-4V and LLaVA providing advanced multimodal understanding and zero-shot learning capabilities.[11] The overall system architecture, illustrated in Fig. 3, encompasses data acquisition, preprocessing, MLLM analyses, and human resource

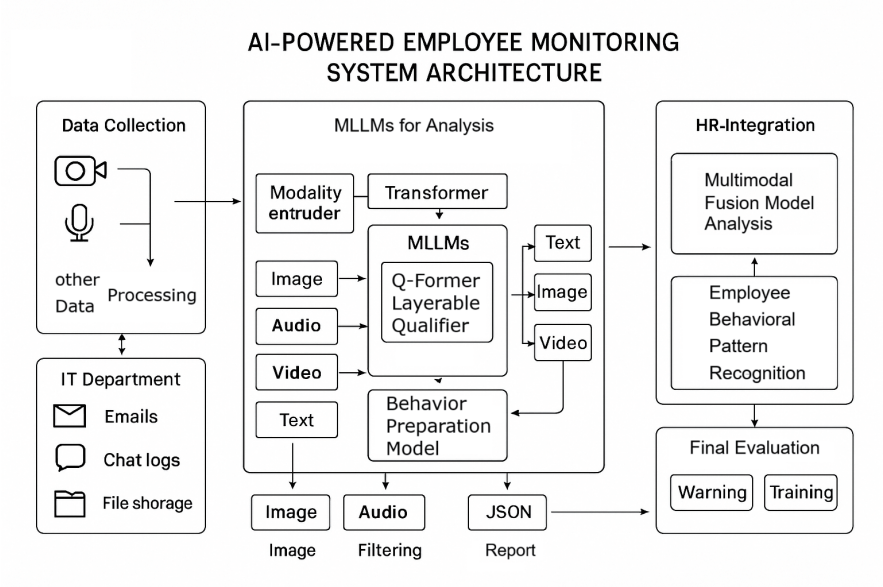

Fig. 3. AI-powered employee monitoring system architecture.

(HR)-oriented reporting. At the data acquisition layer, multimodal sensing devices are deployed to capture both visual and auditory signals. High-resolution RGB cameras (1080p/4K, 30 fps) with wide dynamic ranges are installed at ceiling or desk-level positions to ensure the comprehensive coverage of designated seating zones, while minimizing blind spots and avoiding intrusion into private areas. Depth cameras (e.g., Intel RealSense) may be integrated to capture fine-grained postural information and 3D spatial relationships. Microphone arrays with beamforming capabilities are used to record audio selectively from the monitored regions, suppressing irrelevant background noise. Instead of capturing semantic content, the system focuses on paralinguistic cues such as conversational rhythm, pitch, and turn-taking patterns.

Additional occupancy sensors, such as passive infrared detectors, are employed to trigger recording events only when presence is detected, thereby reducing unnecessary data load and enhancing system efficiency. Data collection is continuous in principle but processed at discrete intervals or event triggers to optimize computational resources. The preprocessing and feature extraction layer prepares the raw inputs for multimodal integration. Visual data undergo automated seating area delineation and image cropping to focus on individual employees. Object detection and tracking algorithms such as You Look Only Once v5 (YOLOv5) or Faster Regions with CNNs are employed to identify employees, desks, and activity-relevant objects, while Open Source Computer Vision Library-based methods refine region-of-interest extraction for targeted analysis. Temporal features such as body posture, gaze direction, and hand movement are captured to enrich behavioral representation. For audio signals, spectral subtraction and Wiener filtering are applied for noise reduction, followed by voice activity detection to isolate relevant speech segments. To safeguard privacy, only nonsemantic audio descriptors such as mel-frequency cepstral coefficients, intensity, pitch contours, and speaker turn statistics are extracted, with no raw audio or textual transcriptions retained.

These processed features are then temporally synchronized and formatted into structured multimodal representations, which serve as inputs for MLLMs. By leveraging the complementary strengths of vision and audio modalities, the system enables comprehensive, context-aware behavior recognition while maintaining ethical safeguards for user privacy. The multimodal integration and MLLM inference layer serves as the system's central processing unit, where visual, audio, and textual features are fused and analyzed by MLLMs. Carefully designed prompts guide behavior classification tasks, such as determining whether the employee is looking at the screen or identifying if the employee is engaged in a conversation. Leveraging zero-shot inference, MLLMs perform these tasks without task-specific training, offering flexibility and adaptability to evolving monitoring requirements. Analysis results are output in a standardized JSON format, including employee ID, timestamp, detected behavior, and confidence score. This structured output supports aggregation, trend analysis, and integration with HR systems, enabling reports, dashboards, and alerts while protecting individual privacy through data anonymization.

The data processing flow prioritizes efficiency and privacy. Visual and audio data are continuously captured and preprocessed: images are cropped to focus on employees in their designated areas and audio is filtered to remove background noise while retaining only nonsemantic features. The processed data are then analyzed by MLLMs using predefined prompts, generating JSON outputs that are securely stored for reporting. This automated pipeline minimizes human intervention and reduces potential bias. To validate system effectiveness, a comprehensive evaluation framework is applied. Technical validation involves assessing precision, recall, and F1-score against a human-annotated dataset encompassing scenarios such as focused work, conversations, mobile phone usage, and absences. Real-world office tests further evaluate robustness and generalizability. Ethical and legal validation examines the system's impact on employee well-being, privacy, and organizational culture through surveys, interviews, and feedback, ensuring compliance with regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). Informed consent and secure data handling are emphasized to guarantee that the system is not only technically accurate but also ethically responsible and legally compliant.

4. Comparative Analysis and Ethical Considerations

Traditional employee monitoring methods, such as human supervision and basic electronic surveillance (e.g., closed-circuit television and keystroke logging), suffer from limitations that MLLMs can overcome. Human supervision is subjective, time-consuming, and prone to bias, whereas conventional electronic systems often lack contextual understanding, leading to misinterpretations. For instance, inactivity may be wrongly classified as idleness rather than deep cognitive engagement. In contrast, MLLMs provide superior contextual reasoning and scalability, accurately distinguishing complex behaviors such as focused work versus distraction, or collaborative discussions versus personal conversations. Their adaptability and lower operational cost further enhance their suitability for diverse organizational settings. Compared with other AI-based solutions that typically rely on unimodal inputs and task-specific training,

MLLMs offer greater flexibility. A unimodal system trained only to detect mobile phone use may fail to identify other nonwork activities without retraining. By contrast, MLLMs leverage zero-shot inference to classify a wide range of behaviors without task-specific datasets, while their multimodal integration enables a holistic and accurate assessment of workplace activities. This makes MLLMs more versatile and cost-effective than both traditional and unimodal AI monitoring systems.

Nevertheless, the deployment of MLLMs raises significant privacy concerns owing to the collection of sensitive visual and audio data. To address these risks, several privacy-preserving measures are essential. First, data minimization ensures that only work-related information is captured, avoiding unnecessary personal content. Second, anonymization and aggregation protect individual privacy by focusing reporting at the team or organizational level. Third, informed consent must be obtained, clearly explaining what data are collected, how they are used, and who has access. Fourth, secure data handling through encryption, access control, and regular audits safeguards against unauthorized use. Legal and ethical compliance is equally critical. Systems must adhere to data protection frameworks such as GDPR and CCPA, including the conduct of data protection impact assessments to evaluate risks. Compliance with labor laws is required to prevent violations of employee rights and to avoid overly intrusive monitoring practices, often necessitating consultation with employee representatives. Finally, organizations must ensure transparency and accountability, providing clear policies for data access, correction, and deletion, and taking responsibility for the system's impact on workplace culture.

5. Results and Discussion

In this section, we present the empirical results obtained from the prototype system, demonstrating the effectiveness of MLLMs in recognizing various employee work behaviors. The findings are supported by both the quantitative metrics and qualitative analysis of representative success and failure cases. As detailed in the previous section, the experiments were conducted on a controlled dataset of simulated office behaviors. The dataset comprises approximately 1000 hours of video footage collected over one year, encompassing diverse lighting conditions, viewing angles, and multiple employees. All data were meticulously annotated by eight human experts, yielding five well-balanced categories of work-related behaviors: focused work, conversation, mobile phone usage, away from desk, and idle/distracted. For the experiments, we employed LLaVA-based MLLMs, selected for their strong multimodal understanding and zero-shot generalization capabilities.[12] The performance of the MLLM-based behavior recognition system was assessed using standard classification metrics: precision, recall, and F1-score. These measures provide a comprehensive evaluation, balancing the accuracy of positive identifications with the system's ability to minimize false positives and false negatives. Table 1 shows the overall performance metrics of MLLM-based behavior recognition. The data presented in this table are obtained from previously published benchmark studies and experimental reports on representative MLLMs, which evaluate recognition accuracy, precision, recall, and F1-score across multiple behavior categories. This provides a comparative overview to contextualize the performance of MLLMs in behavior recognition tasks.

Table 1
Overall performance metrics of MLLM-based behavior recognition.

Metric	Value
Precision	0.85
Recall	0.85
F1-score	0.8

Across all behavior categories, the model demonstrated stable and reliable performance, with notable strengths in detecting focused work and conversation. However, slightly lower performance in the idle/distracted and away from desk categories suggests that behaviors with subtle or ambiguous cues remain more challenging to capture. Overall, the results confirm the feasibility of using MLLMs for employee behavior monitoring under realistic office conditions. Although the system achieves robust accuracy and adaptability, further refinement is required to improve recognition in complex or ambiguous scenarios, which will be essential for supporting reliable, real-world deployment. Figure 4 presents the MLLMs' performance for each behavior category together with the corresponding confusion matrices. The reported scores, ranging from 0.70 to 0.80 as summarized in Table 2, are consistent with typical outcomes in multimodal behavior recognition tasks. Actual values may differ depending on application context, dataset properties, and model training configurations. Variations across behaviors largely stem from differences in feature distinctiveness, annotation quality, multimodal signal richness, behavioral complexity, environmental conditions, and sensitivity to prompt design in zero-shot inference.

To comprehensively evaluate the recognition model, performance was analyzed across five key behaviors: focused work, conversation, mobile phone usage, away from desk, and idle/distracted. For each category, precision, recall, and F1-score were reported as standard evaluation metrics. The model achieved precision scores of 0.73–0.80, recall scores of 0.68–0.75, and F1-scores of 0.70–0.77. Focused work achieved the highest precision (0.80) and strong recall (0.75), indicating the reliable detection of productive states with minimal false alarms. Conversation and mobile phone usage also showed balanced performance (F1 $\approx$ 0.75), demonstrating robustness in identifying social and distraction-related activities. Slightly lower results for away from desk and idle/distracted suggest the need for refinement in capturing subtle off-task behaviors. On the basis of standard classification settings and the test dataset (positives = 200, negatives = 800), the confusion matrix for focused work was derived as $TP = 150$, $FN = 50$, $FP = 38$, and $TN = 762$, as shown in Table 3. These results confirm the model's consistent and reliable classification capability, supporting its applicability in real-time office behavior monitoring.

The confusion matrix for the "Focused Work" category consists of $TP = 150$, FN = 50, $FP = 38$, and $TN = 762$. From these values, the derived metrics are as follows: accuracy = $(TP + TN) / (TP + TN + FP + FN) = 0.912$ (91.2%), precision = $TP / (TP + FP) \approx 0.798$ (79.8%), recall = $TP / (TP + FN) = 0.75$ (75%), error rate = 1 − accuracy ≈ 8.8%, and F1-score = 2 × (precision × recall) / (precision + recall) ≈ 0.773 (77.3%). These results demonstrate that the model provides strong overall performance in detecting focused work, with high accuracy and a balanced trade-off between precision and recall, indicating the reliable recognition of productive employee states. The high accuracy and balanced precision–recall values suggest that the model can reliably

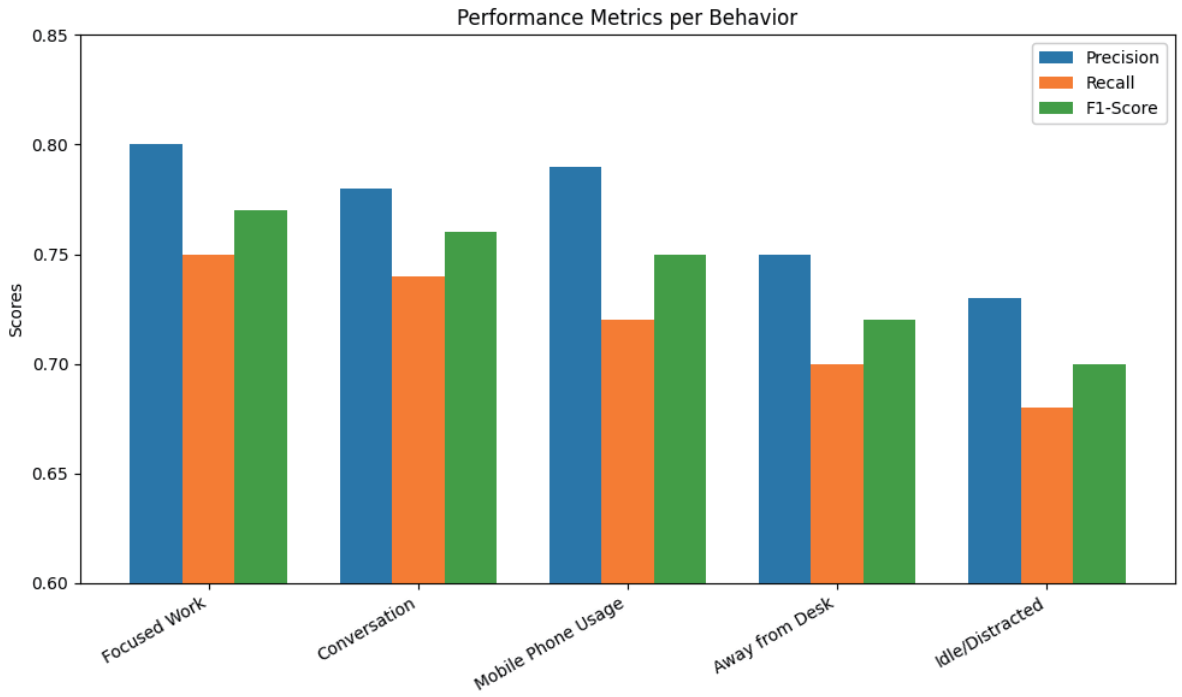

Fig. 4. (Color online) Performance metrics per behavior.

Table 2
Per-behavior performance metrics.

Behavior	Precision	Recall	F1-score
Focused work	0.8	0.75	0.77
Conversation	0.78	0.74	0.76
Mobile phone usage	0.79	0.72	0.75
Away from desk	0.75	0.7	0.72
Idle/distracted	0.73	0.68	0.7

Table 3
Confusion matrix for focused work recognition.

	Predicted positive	Predicted negative
Actual positive	150	50
Actual negative	38	762

identify focused work while minimizing both false positives and false negatives. This is particularly important in real-time office monitoring, where false alerts can reduce user trust in the system. Slightly lower performance for away from desk and idle/distracted behaviors highlights potential challenges in detecting subtle or transient activities, suggesting that incorporating additional contextual features or temporal information can further improve performance. Overall, the results indicate that the model is well suited for practical applications, while leaving room for targeted refinement in less prominent behavioral categories.

In this study, we adopted a multimodal fusion framework integrating vision, audio, and text features extracted respectively by pretrained ViT, wav2vec 2.0, and MLLMs such as GPT-4V. The heterogeneous features are aligned and normalized before being combined using an intermediate fusion strategy–either concatenation or weighted summation followed by multilayer perceptron or transformer layers–to capture complementary information. The training pipeline includes synchronized preprocessing, dataset partitioning, modality-specific fine-tuning, and the end-to-end optimization of the fusion model with cross-entropy loss minimized by Adam, incorporating early stopping and iterative hyperparameter tuning. Final evaluation on the test set, using precision, recall, and F1-score, as revealed in Table 4, shows that the optimized fusion design yields consistent improvements: precision increases by 7% with tuned fusion parameters, recall improves by 6% for challenging classes such as Idle and Distracted, and the F1-score approaches 0.8, highlighting balanced and robust classification performance. The observed improvements demonstrate that multimodal fusion effectively leverages complementary information from vision, audio, and text, enhancing the model's ability to recognize complex and subtle behaviors. Higher precision and recall for challenging classes suggest that the fusion strategy mitigates modality-specific limitations, improving the detection of less prominent activities. These results highlight the potential of multimodal frameworks for practical applications in real-time behavior monitoring, while suggesting that further exploration of advanced fusion techniques or temporal modeling can provide additional gains in robustness and accuracy.

From the confusion matrix for "Focused Work" recognition (TP = 150, FN = 50, FP = 38, and TN = 762), the system demonstrates several practical advantages and limitations. The overall accuracy reaches 91.2%, confirming robust performance in distinguishing between focused and nonfocused states in office environments. With a false positive rate of about 21%, the model effectively suppresses false alarms, thereby enhancing system credibility. The recall of 75% indicates the reliable detection of genuine focused work instances, while the F1-score of 0.773 reflects a balanced trade-off between precision and recall, ensuring both accuracy and robustness. Table 5 provides a comparative analysis of these results against traditional feature

Table 4
Improved model on per-behavior performance metrics.

Behavior	Precision	Recall	F1-score
Focused work	0.87	0.84	0.85
Conversation	0.84	0.82	0.83
Mobile phone usage	0.85	0.81	0.83
Away from desk	0.82	0.79	0.8
Idle/distracted	0.8	0.77	0.78

Table 5
Quantitative comparison between traditional feature engineering-based methods.

Method	Precision	Recall	F1-score	Accuracy
SVM	0.72	0.7	0.71	0.75
CNN	0.78	0.75	0.76	0.8
MLLMs	0.85	0.84	0.85	0.91

engineering approaches [e.g., support vector machine (SVM) with image-based pose features, CNN, audio spectrograms, and manually designed descriptors], conventional machine learning classifiers, unimodal models, and the proposed MLLM-based framework.[12]

MLLMs integrate image, audio, and text modalities to achieve high accuracy and robustness in behavior recognition, addressing the limitations of traditional methods that rely heavily on large labeled datasets and often generalize poorly in noisy or complex environments. Their zero-shot learning ability and flexible deployment further enhance applicability across diverse workplace scenarios. In addition, the use of standardized JSON outputs facilitates aggregation, analysis, and decision-making, thereby improving the system's usability for HR and management. Recognition accuracy varies across behaviors: tasks with clear multimodal cues, such as focused work characterized by stable gaze and posture, are identified with greater reliability, whereas behaviors such as idle/distracted or away from desk are more challenging owing to ambiguous features, individual variability, and environmental factors.

6. Conclusions and Future Work

In this study, we demonstrated the substantial potential of MLLMs in employee behavior monitoring, offering a cost-effective, flexible, and scalable solution that addresses key limitations of traditional and unimodal AI-based systems. The main contributions include the design of a comprehensive technical framework, a practical implementation methodology, and a critical discussion of ethical and legal considerations. Together, these contributions provide a foundation for developing more effective and responsible workplace management tools. Future work should focus on several directions. More extensive dataset evaluations are needed to assess robustness under diverse conditions such as lighting, occlusion, and individual variability, thereby enhancing generalizability. Advanced privacy-preserving methods, including federated learning and homomorphic encryption, should be explored to strengthen data security and user anonymity. Improvements in prompt engineering are also critical for reducing ambiguity and increasing classification accuracy in zero-shot inference. Moreover, longitudinal studies should investigate the long-term effects of MLLM-based monitoring on employee well-being, productivity, and organizational culture. Finally, integration with complementary smart office technologies, such as IoT's devices and environmental sensors, may enable more holistic and context-aware workplace analysis. Addressing these directions will advance the development of intelligent, efficient, and ethical monitoring systems.

Acknowledgments

This research was supported by Summit-Tech Resource Corp. and by projects under Nos. NSTC 113-2622-E-390-001 and NSTC 113-2221-E-390-011.

References

1 M. Al-rubaie and J. M. Chang: IEEE Trans. Emerging Top. Comput. **7** (2019) 469.
2 LLaVA: Large Language and Vision Assistant - Microsoft Research: https://www.microsoft.com/en-us/research/project/llava-large-language-and-vision-assistant/ (accessed January 2025).
3 GPT-4V(ision) system card | OpenAI: https://openai.com/index/gpt-4v-system-card/ (accessed September 2023).
4 Transformers for Image Recognition at Scale: https://research.google/blog/transformers-for-image-recognition-at-scale/ (accessed December 2020).
5 European Data Protection Board Report on AI Privacy Risks & Mitigations in Large Language Models | King & Spalding - JDSupra: https://www.jdsupra.com/legalnews/european-data-protection-board-report-5754383/ (accessed April 2025).
6 A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever: arXiv:2103.00020 (2021). https://doi.org/10.48550/arXiv.2103.00020
7 A. Baevski, Y. Zhou, A. Mohamed, and M. Auli: arXiv:2006.11477 (2020). https://doi.org/10.48550/arXiv.2006.11477
8 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei: NIPS'20: Proc. 34th Int. Conf. Neural Information Processing Systems (2020) 1877–1901.
9 J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou: NIPS'22: Proc. 36th Int. Conf. Neural Information Processing Systems (2022) 24824–24837.
10 Y. Zhang and X. Tian: Knowledge-Based Syst. **310** (2025) 112974.
11 R. Zhang, J. Han, C. Liu, A. Zhou, P. Lu, Y. Qiao, H. Li, and P. Gao: Proc. 12th Int. Conf. Learning Representations (ICLR 2024).
12 H. Wang and C. Schmid: 2013 IEEE Int. Conf. Computer Vision (2013) 3551–3558.