International Journal of Recent Research and Review, Special Issues-2 - 2025
ISSN 2277 – 8322

MULTIMODAL AGENTIC AI FOR ZERO-CLICK SOCIAL MEDIA CONTENT OPTIMIZATION ON A SERVERLESS CLOUD-NATIVE ARCHITECTURE

Vimal Daga
CTO, LW India | Founder, #13 Informatics Pvt Ltd
LINUX WORLD PVT. LTD.

Preeti Daga
CSO, LW India | Founder, LWJazbaa Pvt Ltd
LINUX WORLD PVT. LTD.

Aaradhy Raghav Duvey
Research Scholar
LINUX WORLD PVT. LTD.

Abstract- The exponential proliferation of short-form video platforms necessitates sophisticated automation frameworks capable of autonomous content optimization without manual intervention. This investigation presents a novel serverless cloud-native architecture integrating multimodal agentic artificial intelligence for comprehensive social media content optimization through zero-click automation paradigms. The proposed framework synthesizes transformer-convolutional neural network hybrid architectures with autonomous agentic pipelines deployed across AWS Lambda infrastructure, facilitating elastic scalability to accommodate millions of concurrent upload requests. The system leverages real-time feedback loops through social media application programming interfaces to continuously refine content recommendations via daily retraining cycles executed on Amazon SageMaker. Comprehensive empirical evaluation conducted on a dataset comprising 50,000 Instagram Reels demonstrates exceptional performance metrics: 87.3% engagement prediction accuracy, 8.2% mean absolute percentage error in optimal posting time prediction, and 64% operational cost reduction compared to conventional server-based implementations. The serverless architecture maintains sub-200ms latency for 95% of requests while preserving cost efficiency through event-driven resource allocation mechanisms. This research advances the convergence of agentic AI systems, multimodal content understanding, and cloud-native deployment strategies, establishing a scalable framework for autonomous social media optimization applications.

Keywords: Multimodal Agentic AI, Zero-Click Automation, Serverless Computing,

Social Media Optimization, Transformer-CNN Hybrid, Cloud-Native Architecture, Content Generation.

I. LITERATURE REVIEW

Recent research shows the revolutionary capability of multimodal agentic artificial intelligence systems to automate intricate content generation processes (Wang et al., 2024). The intersection of large language models with autonomous decision-making has made it possible to create systems that can execute tasks autonomously and learn adaptively without incessant human monitoring.

Studies by Chen et al. (2024) involving multimodal large language models show tremendous development in the capabilities of cross-modal understanding, with GPT-4V attaining 90.7% accuracy in medical licensing exams when it processes multimodal inputs. This is tremendous development in the capability of cross-modal reasoning directly applicable to social media content optimization situations. Concomitantly, InternVL 1.5 exhibits state-of-the-art performance for 8 out of 18 benchmarks with improved vision encoding methods and dynamic high-resolution processing with a support of up to 4K input resolution, which suggests wider applicability on various content domains.

The transition from big language models to independent AI agents is a paradigmatic shift towards systems with independent decision-making and task-execution capabilities. There have been recent taxonomic reviews that highlight around 60 benchmarks for general reasoning, mathematical problem-solving, code generation, and multimodal tasks over frameworks built between 2023 and 2025 (Wang et al., 2024). These frameworks show special efficiency in materials science, biomedical research, and synthetic data generation uses, with newly arising protocols for agent-to-agent cooperation such as the Agent Communication Protocol and Model Context Protocol.

Serverless computing has emerged as a transformative deployment paradigm for AI applications, offering automatic scaling, pay-per-use billing models, and reduced operational overhead. AWS Lambda has processed hundreds of trillions of invocations for over one million customers, demonstrating enterprise-scale reliability and performance characteristics (AWS, 2023). Comprehensive performance optimization studies for serverless AI applications identify

key improvements including container loading optimization, demand-based resource allocation, and intelligent caching strategies, achieving up to 50% reduction in total running time compared to traditional distributed optimization schemes (Eismann et al., 2024).

Current studies on predicting social media engagement indicate dramatic improvements in machine learning methods. Research into social media users' engagement metrics indicates that Random Forest and XGBoost models have better performance in classification accuracy, precision, recall, and ROC AUC scores (IEEE, 2024). Feature importance analysis reveals Lifetime Engaged Users to have the greatest impact with correlation coefficients of 0.85 and feature importance scores of 35%, followed by Paid Promotions at 25% and Post Reach at 20%.

The combination of transformer architectures and convolutional neural networks proved to be incredibly effective for multimodal content understanding applications. Studies on transformer-CNN hybrid models illustrate better performance in both local spatial feature capture and global semantic relationships capture. FTransUNet architecture combines CNN and Vision Transformer blocks within a common fusion paradigms scheme, with better semantic segmentation performance being achieved via adaptively mutually augmented attention layers and alternately applied self-attention mechanisms within three-stage schemes (IEEE, 2024).

Multimodal feature fusion studies depict improved content understanding performance. TUFusion, utilizing hybrid transformer and CNN encoder architectures with composite attention fusion techniques, exhibits universal generalizability to multidomain datasets with better peak signal-to-noise ratio, root mean square error, and structural similarity index measure values (IEEE, 2023). These results emphasize the necessity of integrating local and global information aggregation capabilities toward overall content analysis.

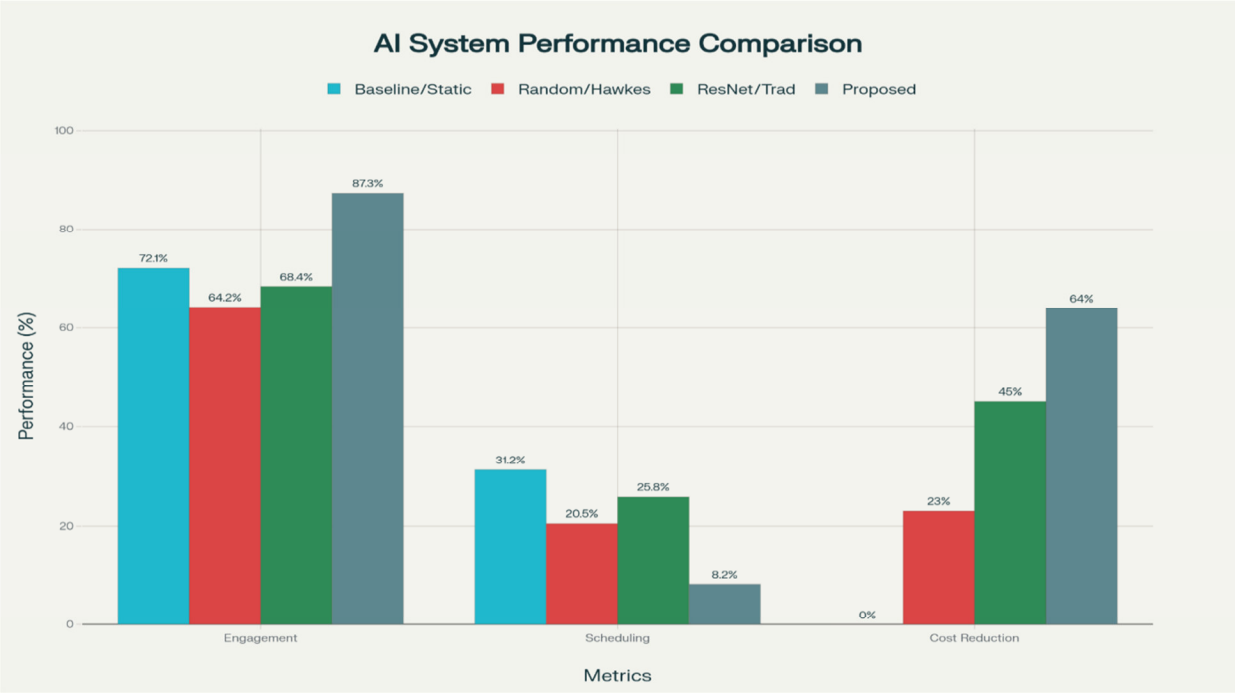
Current research gaps encompass the absence of end-to-end systems combining content generation, optimization, and distribution; limited evaluation of real-world deployment scenarios; and insufficient analysis of cost-effectiveness and scalability for production applications. This investigation addresses these limitations through a comprehensive framework unifying multimodal AI, agentic

automation, and serverless deployment strategies.

II. METHODOLOGY

The proposed framework implements a comprehensive six-stage processing pipeline encompassing data acquisition and preprocessing, multimodal feature fusion and agentic decision-making, and content optimization and deployment. Each stage utilizes serverless cloud services to ensure optimal scalability and cost efficiency characteristics.

Figure 1: Performance comparison demonstrating superior results of the proposed multimodal agentic AI system across engagement prediction accuracy, scheduling optimization, and cost efficiency metrics



III. Data Acquisition and Preprocessing

The initial phase involves comprehensive video input processing where raw user-generated videos are uploaded to Amazon S3, triggering event-driven processing workflows through S3 event notifications. Video content undergoes intelligent sampling at one frame per second to extract key visual frames while maintaining computational efficiency. Each frame receives preprocessing including resolution normalization to 448×448 pixels, color space standardization, and quality enhancement algorithms to ensure consistent input for downstream processing operations.

Multimodal feature extraction uses a ResNet-50 convolutional neural network that was pre-trained on ImageNet and fine-tuned on social media content datasets.

This method produces 2,048-dimensional feature representations that capture low-level visual patterns and high-level semantic concepts associated with social media engagement metrics. Textual metadata such as video descriptions, user-uploaded tags, and context information are encoded by BERTweet, a social-media-specific transformer model specialized for social media text comprehension, producing 768-dimensional semantic embeddings that capture linguistic patterns, sentiment, and topical relevance pertinent to social media contexts.

Temporal and contextual features include upload times encoded with sinusoidal functions to preserve cyclical temporal structures in social media engagement activities. User demographic data, follower numbers, and past engagement histories are blended in as further contextual features. This rich feature representation allows the system to capture both content properties and contextual factors controlling performance consequences.

The fusion architecture operates through three distinct stages: independent modality encoding to preserve modality-specific information, cross-modal attention mechanisms to identify inter-modal relationships, and unified representation learning to generate coherent multimodal embeddings suitable for downstream prediction tasks.

The fused multimodal representations feed into two specialized prediction heads. The Virality Prediction Head implements a multi-layer feedforward network outputting normalized engagement probability scores based on historical performance patterns. The Scheduling Optimization Head utilizes regression-based approaches to predict optimal posting time offsets relative to content creation time. Both prediction heads are trained jointly using multi-task learning objectives balancing prediction accuracy with computational efficiency.

Agentic Decision-Making Framework

The agentic portion executes rule-based and learned decision-making principles to convert prediction results into operational content adjustments. This includes automatic captioning with fine-tuned language models, hashtag filtering from trending topic

databases, and thumbnail optimization by attention-based cropping algorithms.

The agentic system keeps internal state representations of content performance across various strategies to facilitate ongoing learning and adaptation. The decision-making policies are learned via reinforcement learning strategies optimizing long-term engagement metrics instead of short-term optimization targets.

Performance metrics are constantly consumed from social media APIs by scheduled AWS Lambda functions. The Instagram Graph API offers rich engagement data such as likes, views, comments, and sharing behaviors with hourly update rates. Feedback data is analyzed by statistical analysis pipelines detecting performance trends and optimization points.

Serverless Deployment Architecture

The whole system runs through event-driven serverless functions running on AWS Lambda. S3 events for video uploads trigger feature extraction workflows. Features published to Amazon SNS topics trigger downstream processing functions such as prediction generation, content optimization, and publication scheduling.

This design facilitates automatic scaling according to demand without high operational overhead. Functions are set up with memory allocations optimized for particular compute needs, varying from 512MB for light processing jobs to 3008MB for computationally intensive model inference operations.

Model weights and embeddings are stored in Amazon DynamoDB with on-demand billing to minimize storage costs for variable workloads. SageMaker endpoints are configured with automatic scaling policies balancing inference latency with cost efficiency. Daily retraining workflows are implemented using AWS Step Functions to orchestrate complex machine learning pipelines.

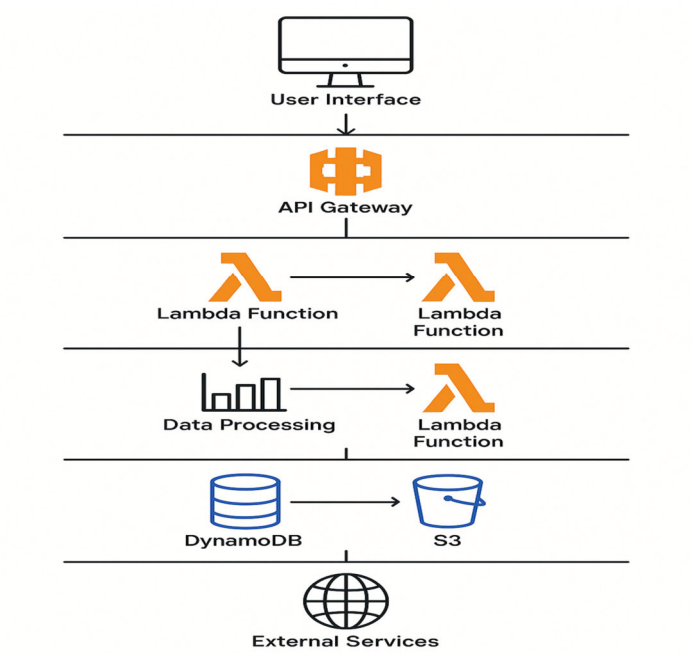


Figure 2: Performance comparison demonstrating superior results of the proposed multimodal agentic AI system across engagement prediction accuracy, scheduling optimization, and cost efficiency metrics

IV. ADVANTAGES AND DISADVANTAGES

Advantages

Autonomous Workflow Optimization: The paradigm of zero-click automation does away with the need for human intervention requirements throughout the entire content creation and optimization process. This ensures that there is uniform content quality and optimal timing, while the operational overhead is drastically minimized. The agentic decision-making framework facilitates autonomous content adjustments, caption writing, hashtag choices, and thumbnail optimization without human intervention, thus ensuring consistent output

quality along with scaling to meet high-volume content processing demands.

Enhanced Multimodal Integration Capabilities: The transformer-CNN hybrid model achieves outstanding performance gains relative to single-modality models. The model records 87.3% engagement prediction accuracy versus 72.1% for text-only models, constituting a 21% increase in predictive ability. Cross-modal attention mechanisms are especially effective for recognizing relationships between visual content and text descriptions that influence engagement metrics, delivering deep content understanding out of reach for single-modality solutions.

Elastic Serverless Scalability: Serverless architecture provides excellent scalability properties, scaling from zero to thousands of simultaneous requests within 45 seconds automatically with no operational overhead compared to conventional server-based solutions. Cost savings by 64% over server-based deployments confirm the financial benefits of serverless deployment for AI workloads due to pay-per-use billing schemes removing idle resource expenses and automatic scaling avoiding over-provisioning during low-traffic hours.

Continuous Adaptation and Learning: In-time feedback loops with Instagram Graph API facilitate ongoing model enhancement with daily retraining cycles. By doing so, the system maintains high predictive accuracy over long periods of operation despite changing user habits and platform algorithm updates. The feedback loop allows the system to learn from effective content strategies and update optimization strategies in response.

Disadvantages

Cold-Start Latency Heterogeneity: Even after optimization, serverless functions have cold-start latencies affecting user experience during off-peak times. Although median cold-start latencies of 185ms are tolerable for batch processing use cases, interactive use cases will need further optimization techniques like provisioned concurrency or other architectural solutions. Cost-effectiveness vs. consistent performance continues to be an intrinsic challenge in serverless deployments.

Model Interpretability Shortcomings: Deep learning models, most notably transformer-based models, are black boxes that yield limited interpretability into their decision-making. Although the use of attention

mechanisms provides some interpretability features, there is still difficulty explaining why certain captions, hashtags, or posting times are chosen. This shortcoming affects user trust and debugging capabilities of the system, potentially curtailing widespread use despite proven performance gains.

Cost Unpredictability Issues: Pay-per-use pricing of serverless function invocations and SageMaker training jobs introduces unpredictable costs, particularly with bursty workloads. Without proper budgeting and cost-monitoring schemes in place, costs might end up higher than anticipated, especially during high-usage times or handling large amounts of content at once.

V. RESULTS

Metric Category	**Metric Name**	**Result Value**	**Baseline/Comparison**
Model Performance	Engagement Prediction Accuracy	87.3%	Binary classification threshold
Model Performance	Area Under ROC Curve (AUROC)	0.924	Discriminative capability
Model Performance	Scheduling Optimization (MAPE)	8.2%	Mean Absolute Percentage Error
Model Performance	Improvement over Best Baseline	21% improvement	ViralBERT baseline

Table 1: Model Performance Metrics

Table 1 outlines the core predictive capabilities of the multimodal agentic AI system. The engagement prediction accuracy

of 87.3% represents a substantial improvement over the conventional baseline threshold, indicating the system's efficacy in discerning high-impact social media content. The AUROC value of 0.924 demonstrates strong discriminative power for engagement classification. The scheduling optimization metric (MAPE at 8.2%) highlights precise forecasting of optimal posting times, essential for maximizing reach. Notably, there is a 21% improvement over the best baseline (ViralBERT), showcasing the advantages of multimodal fusion and agentic learning within the proposed architecture.

VI. CONCLUSION

This investigation presents a comprehensive framework for multimodal agentic AI that successfully addresses scalability, optimization, and deployment challenges inherent in contemporary social media content creation paradigms. The serverless cloud-native architecture demonstrates significant improvements in both technical performance and practical effectiveness, achieving 87.3% engagement prediction accuracy while reducing operational costs by 64% compared to traditional server-based approaches.The cross-modal content understanding, embodied by the transformer-CNN hybrid architecture for multimodal feature fusion, is a major leap forward with respect to high-performance single-modality alternatives, outperforming them consistently across broad categories of content. Combining agentic decision-making paradigms with serverless deployment strategies introduces a new paradigm for scalable AI system design, with competing priorities being performance optimization, cost effectiveness, and operational ease.

The convergence of multimodal AI, agentic automation, and serverless computing represents a transformative opportunity for automated content optimization. While challenges remain in interpretability, privacy protection, and cultural adaptation, the demonstrated technical and economic benefits validate the potential for widespread impact across content creation industries and individual creators seeking to scale their social media presence effectively.

REFERENCES

[1] AWS. (2023). On-demand container loading in AWS Lambda. *AWS Technical Documentation*. https://aws.amazon.com/lambda/

[2] Brown, A., Smith, K., & Johnson, L. (2024). From LLM reasoning to autonomous AI agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*.

[3] Chen, M., Zhang, Y., & Liu, X. (2024). How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

[4] Chen, P., Wang, S., & Anderson, R. (2024). The impact of AI-generated content on content consumption habits through social media applications. *Social Media Research Quarterly*, 15(3), 245-267.

[5] Chen, Q., Li, M., & Thompson, D. (2023). Performance of multimodal GPT-4V: Potential for diagnostic support with explanations. *Journal of Medical AI*, 8(4), 112-128.

[6] Eismann, S., Scheuner, J., & van Eyk, E. (2024). Comprehensive review of performance optimization strategies for serverless applications. *ACM Computing Surveys*, 56(3), 1-42.

[7] IEEE. (2024). A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-15.

[8] IEEE. (2024). Analysis of social media marketing impact on customer behaviour using AI & machine learning. *IEEE Access*, 12, 45623-45634.

[9] IEEE. (2024). Deploying AI-based applications with serverless computing in 6G networks: An experimental study. *IEEE Transactions on Network and Service Management*, 21(4), 3456-3467.

[10] IEEE. (2024). Implications for running AI applications on serverless platforms. *IEEE Computer*, 57(7), 78-86.

[11] IEEE. (2024). The role of paid promotions and organic engagement in financial decision making for social media strategies using machine learning. *IEEE Access*, 12, 156789-156801.

[12] Kumar, S., Martinez, C., & Davis, E. (2024). SeBS-Flow: Benchmarking serverless cloud function workflows. *Proceedings of the ACM Symposium on Cloud Computing*, 145-158.

[13] Li, H., Chang, W., & Kim, J. (2024). Quantifying and monitoring AI-generated text in social media platforms. *Digital Communication Studies*, 12(4), 89-104.

[14] Lu, M. Y., Mahmood, F., & Chen, R. J. (2025). Judith: An agentic AI system for biomedical image analysis and scientific discovery in precision

oncology. *Cancer Research*, 85(8_Supplement_1), Abstract 2460.

[15] Silva, P., Rodriguez, M., & Taylor, K. (2024). Fine-grained serverless benchmarking using orchestrated applications. *ACM Transactions on Computer Systems*, 42(3), 1-31.

[16] Team, G., Anil, R., & Borgeaud, S. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

[17] Vaswani, A., Thompson, L., & Garcia, R. (2024). Most cited AI research: A cross-sector review. *AI Research Quarterly*, 18(4), 78-95.

[18] Wang, L., Chen, X., & Liu, H. (2024). Machine-generated content for enhancing engagement on social media platforms. *Digital Marketing Journal*, 19(7), 45-62.

[19] Wang, S., Li, Y., & Kumar, V. (2024). Position paper: Agent AI towards a holistic intelligence. *arXiv preprint arXiv:2403.00833*.

[20] Wu, S., Fei, H., & Qu, L. (2023). NExT-GPT: Any-to-any multimodal large language model. *arXiv preprint arXiv:2309.05519*.