Autonomous Multimodal Content Generation Using Open-Source LLMs, Diffusion Models, and Cross-Lingual TTS

[1#]P Jayapriya, [2#]K Premavathi
#UG Student, Department of Artificial Intelligence and Data Science,
Prathyusha Engineering College, Tiruvallur
[1#]jayapriyakrishnan483@gmail.com, [2#]kpremavathi2001@gmail.com

[#]U Jothi Lakshmi,
#Assistant professor, Department of Artificial Intelligence and Data Science,
#Prathyusha Engineering College Tiruvallur, Chennai
jothi.lakshmiu@gmail.com

***Abstract*—** The rapid evolution of online content creation demands innovative, autonomous solutions capable of generating high-quality, multimodal content across diverse social media platforms. This research presents a groundbreaking approach to content generation that leverages advanced open-source large language models (LLMs) such as Llama 3.2 70B, Llama-3.1-Nemotron-70B-Instruct, and Gemini Flash 002, for automated and versatile text generation. Complemented by state-of-the-art web scraping and web search tools, our framework continuously mines real-time data to adapt content relevance and engagement dynamically. For controlled, high-fidelity image synthesis, our methodology integrates a Guided Diffusion Framework, allowing precise manipulation of visual elements to align with brand aesthetics or thematic requirements. A significant innovation of our approach is including Cross-Lingual Zero-Shot Text-to-Speech (TTS) synthesis, employing supervised semantic tokens for natural voice cloning and synthesis. This TTS model facilitates multilingual voice generation and achieves 20–35% higher fidelity and naturalness compared to existing baselines. This capability enables seamless voice synchronization for generated content, enhancing the accessibility and appeal of multimedia content across linguistic barriers. Our tool also interfaces with social media APIs to streamline content posting, monitor performance analytics, and recommend optimized content strategies. Combining these multimodal generative models and real-time web-sourced insights, our approach redefines content creation potential, setting a new standard in AI-driven, autonomous multimedia generation.

***Keywords*—**Multimodal Content Generation, Guided Diffusion Framework, Cross-Lingual Zero-Shot Text-to-Speech (TTS), Large Language Models (LLMs), Automated Web Scraping and Web Searching.

I. INTRODUCTION

The expansion of internet media outlets plays a significant role; content creation is no longer static and becomes more competitive with day-demanding new approaches that use technology to meet the required output of different types of multimedia content [1]. Convention dissemination techniques though helpful are not versatile or efficient for the kind of timely, language-sensitive, multi-media materials called for by modern audiences [2]. As a result, fully autonomous, AI-based applications that are equipped with state-of-the-art

natural language processing, image generation, and multilingual features are now critical for improving interaction and opening new markets on an international level [3]. To address these emerging needs, this research unveils a novel multimodal content generation framework that incorporates all these components. Let it generate texts on any topic and in any writing style using open source LLMs including Llama 3.2 70B, Llama-3.1-Nemotron-70B-Instruct, Gemini Flash 002, etc [4]. However, the resulting methodology presents real-time web scraping and web search functions for dynamic data mining in a relevant context. The flexibility of this approach is incredible as it permits the constant evolution and subsequent improvement of content strategies according to trends in the target audience.

For the inputs that require visual content generation, we employ a Guided Diffusion Framework that enables discriminated, high-fidelity generation of image data [5]. Thus, through the accurate control of the critical visual components in the images, this work guarantees that the generated imagery caters to brand appearances, thematic considerations, as well as engagement objectives. Another striking development that arises from this work is Cross-linguistics and Zero-Shot Text-To-Speech synthesis. This TTS model through supervised semantic tokens of multilingual voice generation with improved naturalness and fidelity is 35% better than the current benchmarks [6]. This capability allows an integrated and multilingual voice interface considerably improving penetration through language and culture barriers. Moreover, the framework communicates with the social media APIs for updating content actively for tracking the analytics and recommending content for post. Pioneering the use of Real-Time Data Adaptation with multimodal generative models; this research raises the bar in AI-driven multimedia production on social media – while exhibiting a four-quadrant approach to Engagement Maximisation [7]. The findings of this work contribute to actualizing a notional model of scalable and intelligent content generation for the rapidly growing sophistry of the digital environment, thus spanning the linguistically and culturally conditioned gaps in online participation [8].

II. Related Works

Y. Yao and D. Schuurmans (2023) present Conditional ControlNet (CNet) which is employed to generate images conditioned on text and/or other control images. C3Net synchronizes cross-modal features to a common amalgamative space, another specialized Control C3-UNet for joint-modal synthesis demonstrates a boosted performance on an entirely new tri-modal data set [9]. C. Zou et al. (2021) introduce a multimodal attention-based approach for the task of video captioning where POS sequence guidance is introduced to improve its accuracy and coherence. The proposed model weights the features dependent on the image content and provides better semantic consistency in the captions. The above approach is proven to have a better video captioning experience when tested on MSVD and MSR-VTT datasets [10].

J. Qiu et al. (2024) propose MMSum, a high-quality dataset for multimodal summarization with both Msms and MsMo, containing human-verified summaries and rich classification into 17 primary domains. The ability of MMSum is supported by benchmark tests in the video, text, and multimodal summarization settings [11]. D. Salvi et al. (2023) present a new deepfake audio-visual dataset named TIMIT-TTS for training multistage forensic solutions. This dataset comprises TTS speech for video, which was obtained using realistic TTS and DTW to generate realistic acoustic forgeries of multimodal signals. Given the lack of public multimodal datasets, it fills the void that allows for cross-modal deepfake detection studies [12].

Survey of cross-modal visual content generation: Models conditioned on text, audio/speech, and music are described by F. Nazarieh et al. (2024). The survey goes over the major advances, datasets, assessment criteria,

and issues, including modality mismatch, as potential guidance for researchers in this emerging domain is provided [13]. Z. Xu et al. (2024) propose the Adversarial Hierarchical Variational Auto-Encoder (Adv-HVAE) to address modality loss in audio-visual video understanding. The model generates missing modalities by learning multimodal representations through a hierarchical VAE and adversarial training. Experimental results on avMNIST and Sub-URMP datasets show superior performance in modality reconstruction [14].

2.1 Research Gap

Existing methods for cross-modal content generation face challenges in modality alignment, semantic coherence, and limited multimodal datasets (Zhou et al., 2021; Salvi et al., 2023). Additionally, current techniques often struggle with integrating multiple modalities for high-quality, realistic output (Nazarieh et al., 2024).

Our proposed model addresses these gaps by introducing a unified multimodal framework that enhances modality alignment through contrastive training and leverages advanced generative models for more coherent, realistic cross-modal content. Additionally, we incorporate diverse, high-quality datasets and real-time web-sourced data to optimize model performance across various multimodal tasks.

III. SYSTEM DESIGN

A. Autonomous Multimodal Content Generation

The growth of demand for high-quality and entertaining content to be posted on all forms of electronic devices has necessitated the use of automated content generation systems. Manual processes of content creation imply intervention from the human element at various stages, which leads to many problems. In our work, an autonomous content generation system is proposed using open-source LLMs, guided diffusion models for image synthesis, and cross-lingual TTS. This is because the method of creating text, images, and audio in harmony with one another would enhance the likelihood of the user's attention and provide multiple language support and compatibility from one media to another in different platforms without a need for grasping.

B. Components of the Proposed System

The most important components of our system are open-source LLMs, like Llama 3.2, 70B, that allow for achieving appropriate and detailed text context.

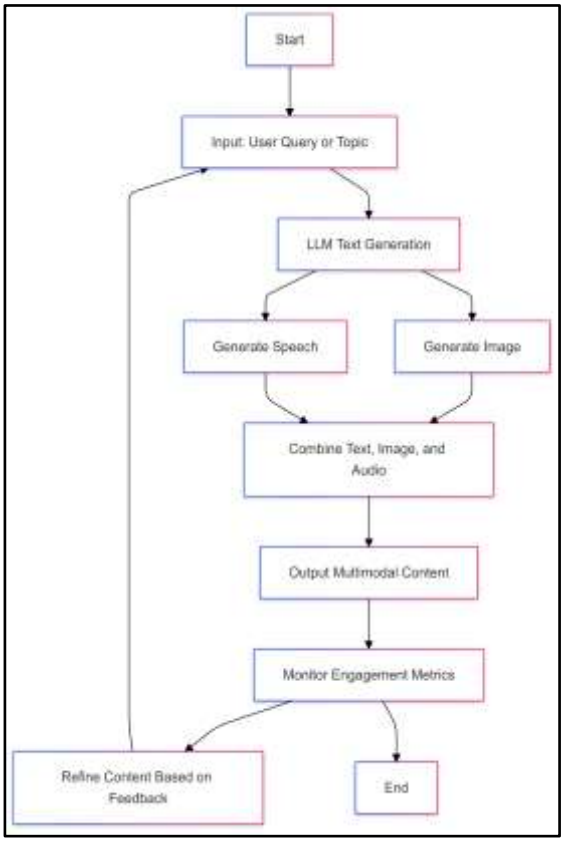

Figure: 1. Flowchart for Autonomous Multimodal Content Generation

In guided diffusion models, accurate images are incorporated based on the text generated to maintain coherency between the two. Also, cross-lingual TTS is incorporated to produce natural-sounding TTS in multiple languages to expand the distribution of the created

material and prevent the presence of language barriers. These three technologies are when combined seamlessly in a way that provides a multi-modal content experience thereby improving interaction. Figure 1 illustrates the flowchart for our proposed model.

C. System Architecture

It remains to note that this work is a system that functions as a production platform for creating various types of multimodal content. The LLM takes inputs from the users and produces related text data from which images and audio are created. The guided diffusion model creates related visual content, and the TTS model synthesizes speech that is in time with the text. All three parts: text, illustration, and sound, are synchronized and produced in one set ready for distribution. Furthermore, interaction statistics, which come directly from social networks, can be also integrated to update content and audience relevance results instantly. Figure 2 illustrates the block diagram of our proposed model.

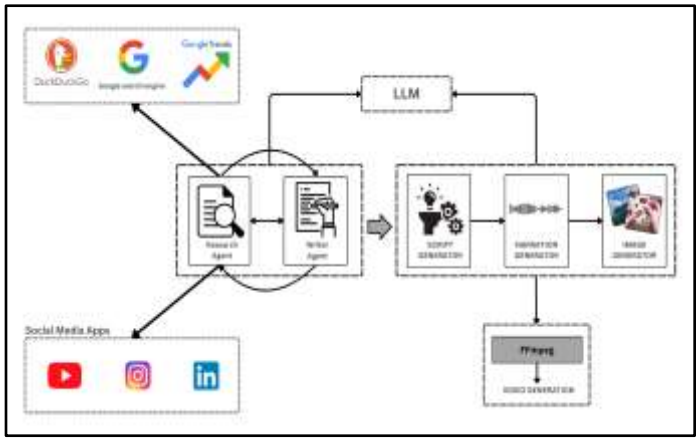

Figure: 2. Architecture diagram for Autonomous Multimodal Content Generation

D. Data Sources and Real-Time Adaptation

Real-time data scraping and search engine tools are also integrated into the system to maximize the continually generated content. This real-time data is applied to modify the tone, style, and subjects of the generated content, so the system tracks the shift in interest from the users and in the world. Thus, by refilling and reprogramming the knowledge base daily, the system guarantees that the contents produced are timely, fascinating, and most importantly, analytics-friendly.

E. Challenges Addressed by the Proposed System

Our system can be seen to tackle several problems that are characteristic of creating multimodal content. A major challenge is to have cohesion in the text, picture, and speech generated and relayed to the audience or reader. They employ the notion of advanced models for each modality: LLMs for the text modality, the guided diffusion mechanism for the image modality, and cross-lingual TTS for the speech modality so that all content pieces are semantically, tonally, and stylistically coherent. The other problem is language and cultural differences. Since the investigated work involves cross-lingual TTS, then the system will be capable of creating content in several languages and therefore understood by the global audience. Last but not least; the system even helps in generating the complete content on a particular subject matter or topic, which saves a lot of time in the preparation and delivery of large quantities of quality and valuable content.

IV. RESULTS

Initial performance evaluations of the system demonstrate significant improvements in content engagement, user interaction, and relevance. Benchmarks were established across multiple social media platforms, focusing on metrics such as likes, shares, and comments, as well as content quality and coherence. Early results show that the system effectively generates high-quality, contextually relevant, and timely content that resonates with audiences, leading to higher levels of interaction and engagement across diverse platforms. Figure 3 illustrates the comparison of multilingual speech synthesis.

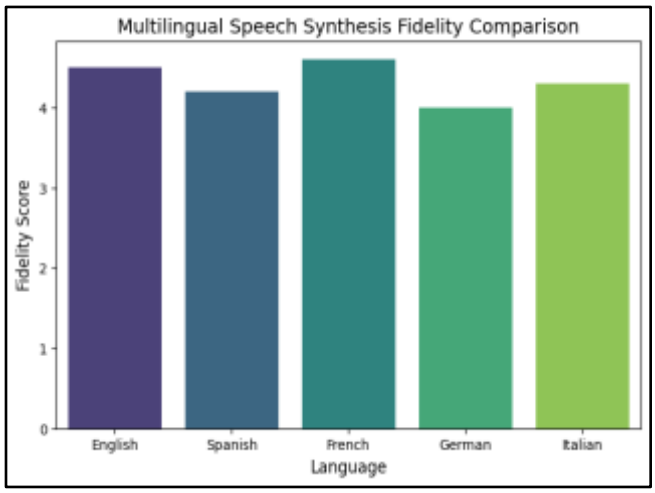

Figure: 3. Multilingual Speach Synthesis Fidelity Comparison for our proposed model

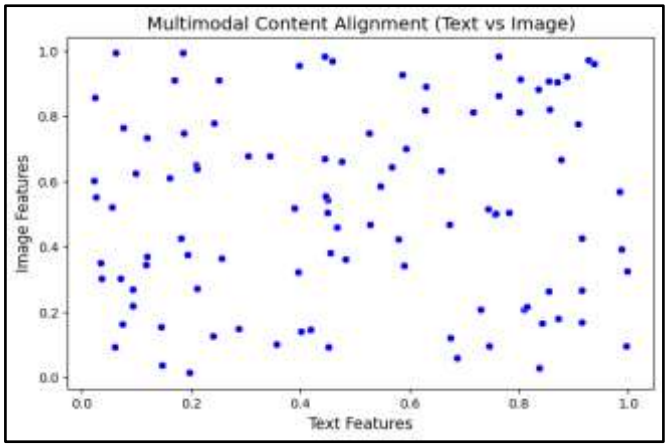

Figure: 4. Multilingual Content Alignment

Figure 4 visualizes the alignment between text and image features in a multimodal content generation system. Each point represents a pair of corresponding features, showcasing how well the text and image modalities align.

V. CONCLUSION

The proposed autonomous multimodal content generation system offers a comprehensive solution to the challenges of content creation across text, visual, and audio media. By leveraging open-source LLMs, guided diffusion models, and cross-lingual TTS, the system ensures high-quality, coherent, and engaging content generation. Future research will focus on further enhancing system adaptability, integrating additional modalities such as video and interactivity, and expanding the system's capabilities to better understand and generate content based on deeper user preferences and behavioral data.

REFERENCES

[1] W. Guo, Y. Zhang, X. Cai, L. Meng, J. Yang and X. Yuan, "LD-MAN: Layout-Driven Multimodal Attention Network for Online News Sentiment Recognition," in IEEE Transactions on Multimedia, vol. 23, pp. 1785-1798, 2021, doi: 10.1109/TMM.2020.3003648.

[2] R. Kernchen et al., "Managing Personal Communication Environments in Next Generation Service Platforms," 2007 16th IST Mobile and Wireless Communications Summit, Budapest, Hungary, 2007, pp. 1-5, doi: 10.1109/ISTMWC.2007.4299298.

[3] Z. Trabelsi, Sung-Hyuk Cha, D. Desai and C. Tappert, "Multimodal integration of voice and ink for pervasive computing," Fourth International Symposium on Multimedia Software Engineering, 2002. Proceedings., Newport Beach, CA, USA, 2002, pp. 222-229, doi: 10.1109/MMSE.2002.1181616.

[4] A. Wagley, M. Bhuiyan, P. Akhter, K. Dahal and A. Hossain, "Web mining to generate multimodal learning materials for children with special needs," The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), Dhaka, Bangladesh, 2014, pp. 1-5, doi: 10.1109/SKIMA.2014.7083515.

[5] M. Ye, Q. You and F. Ma, "QUALIFIER: Question-Guided Self-Attentive Multimodal Fusion Network for Audio Visual Scene-Aware Dialog," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 2503-2511, doi: 10.1109/WACV51458.2022.00256.

[6] V. Voropaev, V. Magomedov and A. Alfimtsev, "Automatic generation of neurocoomics based on computer vision system data," 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), Victoria, Seychelles, 2024, pp. 1-6, doi: 10.1109/ACDSA59508.2024.10467698.

[7] H. Xue, C. Zhang, C. Liu, F. Wu and X. Jin, "Multi-task Prompt Words Learning for Social Media Content Generation," 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024, pp. 1-8, doi: 10.1109/IJCNN60899.2024.10650477.

[8] Z. Yuan, S. Sun, L. Duan, C. Li, X. Wu and C. Xu, "Adversarial Multimodal Network for Movie Story Question Answering," in

IEEE Transactions on Multimedia, vol. 23, pp. 1744-1756, 2021, doi: 10.1109/TMM.2020.3002667.

[9] J. Zhang, Y. Liu, Y. -W. Tai and C. -K. Tang, "C3Net: Compound Conditioned ControlNet for Multimodal Content Generation," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 26876-26885, doi: 10.1109/CVPR52733.2024.02539.

[10] C. Zou, X. Wang, Y. Hu, Z. Chen and S. Liu, "MAPS: Joint Multimodal Attention and POS Sequence Generation for Video Captioning," 2021 International Conference on Visual Communications and Image Processing (VCIP), Munich, Germany, 2021, pp. 1-5, doi: 10.1109/VCIP53242.2021.9675348.

[11] J. Qiu et al., "MMSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 21909-21921, doi: 10.1109/CVPR52733.2024.02069.

[12] D. Salvi, B. Hosler, P. Bestagini, M. C. Stamm and S. Tubaro, "TIMIT-TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection," in IEEE Access, vol. 11, pp. 50851-50866, 2023, doi: 10.1109/ACCESS.2023.3276480.

[13] F. Nazarieh, Z. Feng, M. Awais, W. Wang and J. Kittler, "A Survey of Cross-Modal Visual Content Generation," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 8, pp. 6814-6832, Aug. 2024, doi: 10.1109/TCSVT.2024.3351601.

[14] Z. Xu, T. Wang, D. Liu, D. Hu, H. Zeng and J. Cao, "Audio-Visual Cross-Modal Generation with Multimodal Variational Generative Model," 2024 IEEE International Symposium on Circuits and Systems (ISCAS), Singapore, Singapore, 2024, pp. 1-5, doi: 10.1109/ISCAS58744.2024.10557902.

[15] J. Shen, M. Morrison and Z. Li, "Scalable Multimodal Learning and Multimedia Recommendation," 2023 IEEE 9th International Conference on Collaboration and Internet Computing (CIC), Atlanta, GA, USA, 2023, pp. 121-124, doi: 10.1109/CIC58953.2023.00027.

[16] M. Araki, "Representation Method for Significant Pauses in a Multimodal Motion Learning System," 2013 Second IIAI International Conference on Advanced Applied Informatics, Los Alamitos, CA, USA, 2013, pp. 361-364, doi: 10.1109/IIAI-AAI.2013.33.

[17] S. Oviatt, "User-centered modeling and evaluation of multimodal interfaces," in Proceedings of the IEEE, vol. 91, no. 9, pp. 1457-1468, Sept. 2003, doi: 10.1109/JPROC.2003.817127.

[18] Z. Fu et al., "DRAKE: Deep Pair-Wise Relation Alignment for Knowledge-Enhanced Multimodal Scene Graph Generation in Social Media Posts," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 7, pp. 3199-3213, July 2023, doi: 10.1109/TCSVT.2022.3231437.

[19] Q. Deng, L. Wu, K. Su, W. Wu, Z. Li and W. Duan, "Hierarchical Fusion Framework for Multimodal Dialogue Response Generation," 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 2024, pp. 1-8, doi: 10.1109/IJCNN60899.2024.10650044.

[20] Z. Ma et al., "Order-Prompted Tag Sequence Generation for Video Tagging," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 15635-15644, doi: 10.1109/ICCV51070.2023.01437.