WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

VIDEO QA SYSTEM USING GEN AI

Pavithra S[1], Rahul B[2], Rohith S[3], Uma shankar S[4], Vishal R[5]

[1]Assistant professor,Department of computer Science and Engineering United Institute of Technology(An Autonomous Institution) Coimbatore,India
pavishanthi1@gmail.com
[4]Student,Department of computer Science and Engineering United Institute of Technology(An Autonomous Institution) Coimbatore,India
shan132003@gmail.com
[2]Student,Department of computer Science and Engineering United Institute of Technology(An Autonomous Institution) Coimbatore,India
rahul637977@gmail.com
[5]Student,Department of computer Science and Engineering United Institute of Technology(An Autonomous Institution) Coimbatore,India
vishalmaadha54@gmail.com
[3]Student,Department of computer Science and Engineering United Institute of Technology(An Autonomous Institution) Coimbatore,India
Rohithrio481@gmail.com

ABSTRACT—

This paper examines the transformative impact of Generative AI and Large Language Models (LLMs) on video technology, focusing on video generation, understanding, and streaming. It highlights how these technologies enable realistic content creation, improve video comprehension, and support adaptive, user-centric streaming. The study outlines current progress, identifies challenges, and explores future directions, emphasizing their growing significance across multimedia, AI, and networking domains.

PROBLEM STATEMENT

his study explores the integration of Generative AI and Large Language Models (LLMs) in video generation, understanding, and streaming. It highlights their potential to enhance realism, interactivity, and personalization in video content, with implications for education, user experience, and analytics. The paper also emphasizes the role of LLMs in creating efficient, adaptive streaming services while addressing ethical concerns and guiding future research and responsible AI development in the video technology landscape.

Index Terms—Generative AI, Large Language Models (LLMs), Video Generation, Video Understanding, Video Streaming, GPT.

INTRODUCTION

Recent advances in video technology have revolutionized how video content is created, analysed, and delivered. With the integration of Generative AI and Large Language Models (LLMs), new possibilities are emerging in generating lifelike videos, understanding complex scenes, and enhancing streaming quality. Techniques like GANs have enabled realistic video synthesis, though challenges remain in maintaining consistency and control. In video understanding, LLMs show promise in tasks such as captioning and action recognition, moving beyond traditional feature-based methods. Additionally, LLMs can improve video streaming by enabling adaptive, context-aware content delivery, addressing issues like bandwidth limitations and varying user preferences.

EXISTING SYSTEM

- Manual Transcription : Traditional methods require human effort to transcribe video/audio content.
- Limited Retrieval : Users must manually search through transcripts or captions for relevant information.
- Inefficient Query Processing : Most systems do not provide direct question-answering capabilities on video content.
- Lack of Automation : Existing solutions often lack AI-driven indexing and retrieval, making searches time-consuming.

PROPOSED SYSTEM

- Automated Processing : Converts video to audio and transcribes speech into text using AI models like Whisper.
- Efficient Indexing : Uses Chroma DB to store and retrieve processed text efficiently.
- AI-Powered Query Handling : Allows users to ask questions and get precise answers using Large Language Models (LLMs).

- Continuous File Processing : Supports multiple file uploads, enabling users to build an indexed repository for future queries.

METHODOLOGIES

The proposed Video QA system leverages Generative AI and Large Language Models (LLMs) to enable intelligent interaction with video content. The methodology begins with converting video input into audio and transcribing it into text using advanced speech recognition models such as Whisper. The transcribed data is then indexed and stored efficiently using vector databases like Chroma DB to support fast retrieval. LLMs are employed to process and understand both textual and visual elements, enabling the system to comprehend user queries and generate accurate responses. The system supports continuous ingestion of multiple video files, allowing users to build a searchable knowledge repository over time. Additionally, generative capabilities are utilized to perform intermediate reasoning, summarize video scenes, and improve understanding of complex or temporally structured queries. By incorporating zero-shot and few-shot learning techniques, the system demonstrates adaptability across diverse domains with minimal supervision, enhancing its scalability and robustness.

OVERVIEW

Generative AI and Large Language Models (LLMs) are poised to play pivotal roles throughout the video lifecycle, including generation, understanding, and streaming. Bridging the domains of AI, multimedia, and networking, these technologies have rapidly evolved—from text-to-image models to text-to-video generation within a year. Recent advancements now even enable prompt-based 3D video creation, suggesting a potential shift away from traditional video generation methods. In video understanding, applications such as scene segmentation, activity monitoring, and captioning are gaining traction, driven by the enhanced multimodal capabilities of models like GPT-4 and Video-ChatGPT. For streaming, LLMs offer opportunities to optimize delivery by interpreting scene semantics and dynamically adjusting encoding. Moreover, in areas like 3D video and XR environments, LLMs can predict user focus to support intelligent pre-fetching, improving efficiency and user experience.

TECHNOLOGIES

A. Generative AI for Video Content Creation

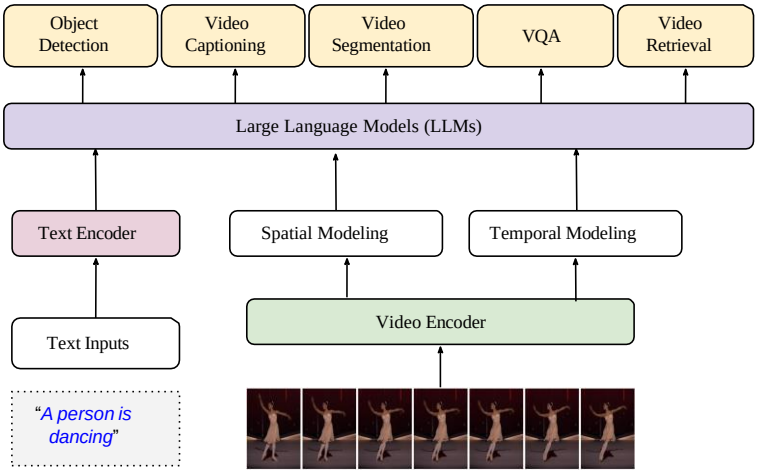
Generative AI is significantly transforming video content creation by enabling the automatic synthesis of realistic and high-quality videos. Core deep learning models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), autoregressive models, and diffusion models are employed to learn data distributions and generate content that reflects real-world dynamics. GANs utilize a generator-discriminator architecture to produce consistent video frames by modeling appearance and motion separately. VAEs focus on learning probabilistic latent spaces, with models like Stochastic Video Generation (SVG) capturing multi-scale temporal patterns. Autoregressive models, such as Video Pixel Networks (VPN), generate frames sequentially by incorporating temporal dependencies through LSTMs. Diffusion models approach frame generation as a denoising process; while Video Diffusion Models (VDM) and Imagen-Video achieve high-resolution results, they are computationally intensive. Recent efforts, like Video LDM, address this by using latent, motion-aware representations to reduce processing overhead.

B. LLMs for Video Scene Understanding

Under standing video scenes involves identifying and interpreting objects, actions, and events within a sequence. This task is inherently complex due to temporal variability and visual diversity. Large Language Models (LLMs), trained on extensive textual datasets, have shown strong capabilities in bridging vision and language. Their generative strengths enable them to interpret video scenes and provide natural language descriptions, supporting tasks such as object detection, activity recognition, and event analysis.

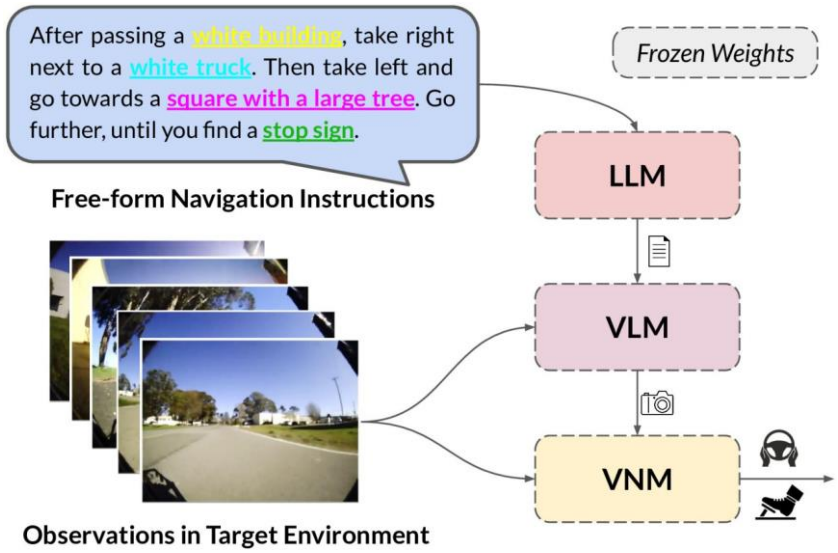
C. LLMs in Video Streaming

LLMs also present opportunities to enhance the video streaming pipeline, which includes video capture, encoding, transmission, decoding, and frame reconstruction. By leveraging their contextual understanding, LLMs can assist in optimizing encoding rates based on scene semantics, addressing format-specific challenges, and adapting content delivery to user behaviour. This can improve overall streaming efficiency and enable predictive capabilities such as pre-fetching in immersive environments like XR

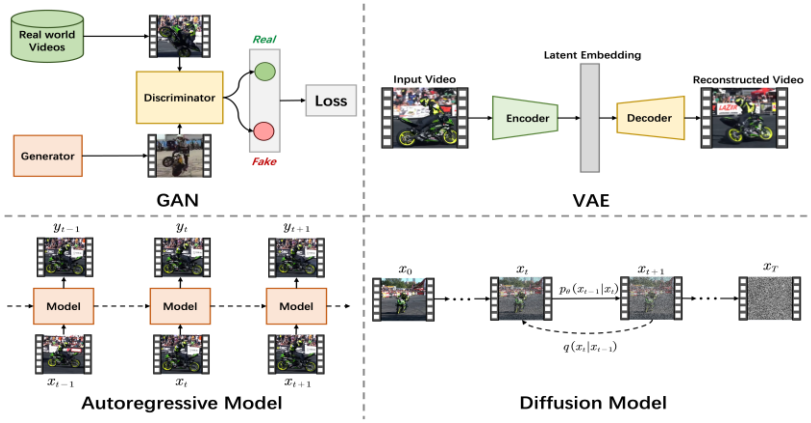


An overview of LLMs for video scene understanding task

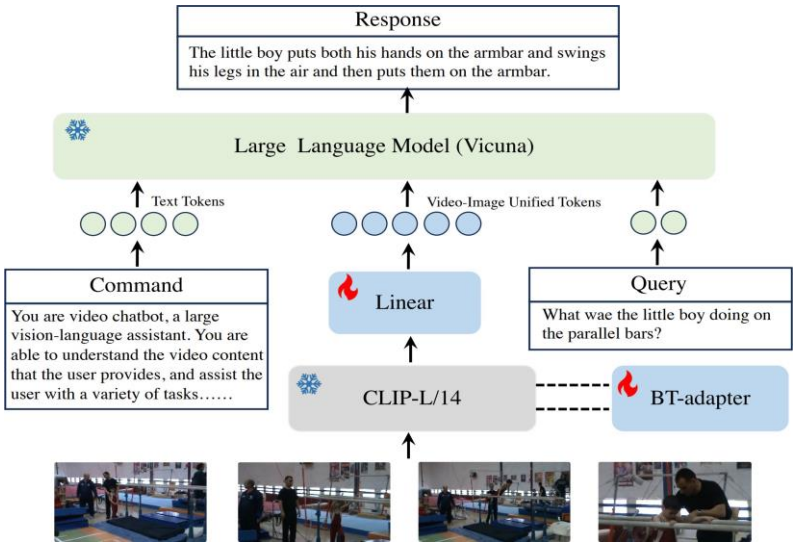
APPLICATIONS :



A representative pipeline of video conversation based on LLMs

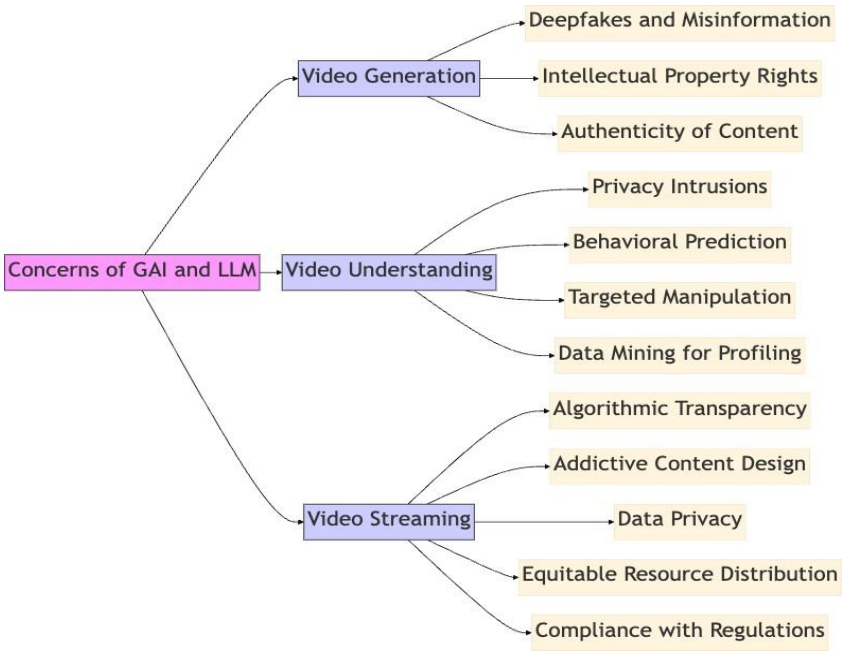


An overview of advanced AI-based video generation technologies



CONCERNS

A representative pipeline of video conversation based on LLMs



CONCLUSION

In this paper, we conduct a comprehensive examination of how generative artificial intelligence (Generative AI) and large language models (LLMs) are revolutionizing the video technology sector, focussing on video generation, understanding and streaming. The innovative technologies of this integration technologies results in high realistic digital creation, enhanced video understanding by extracting meaningful information from visual content, and more efficient and personalized streaming experiences, thus improving user interaction with videos and user preference-tailored experience provision. The paper navigates through current achievements, ongoing challenges, and future possibilities in applying Generative AI and LLMs to video-related tasks. It underscores the immense potential these technologies hold for advancing video technology across multimedia, networking, and AI communities. It also highlights the challenges and concerns that require further exploration. Observed from the reviewed works, we can see that, overall, advanced AI technologies like GAI and LLMs are making profound impacts on several key sectors of video-related research fields. The biggest advantage of AI-based methods is their automation capability with lower manual costs. However, it comes at the price of challenges uniquely faced by AI**,** such as lack of large-scale datasets, high computational cost, consistency issues, and concerns such as misinformation and security, etc. Therefore, academia and industry should be cautious during the rapid development to ensure a sustainable market.

FUTURE ENHANCEMENT

The integration of Generative AI into Video Question Answering (Video QA) systems presents promising avenues for future research and development. Current Video QA systems often rely on multimodal deep learning models to interpret visual and textual data. However, these systems face limitations in understanding complex temporal events, abstract reasoning, and multi-step interactions within videos. Generative AI, particularly large-scale pre-trained models such as vision-language transformers and multimodal foundation models, can significantly enhance Video QA by generating richer scene representations and contextual embeddings. Future systems are expected to incorporate generative models capable of producing intermediate reasoning steps, temporal summaries, or synthetic video-text pairs to improve answer accuracy and generalizability. Additionally, the fusion of generative video synthesis with QA frameworks can facilitate the generation of hypothetical scenarios, enabling counterfactual reasoning and deeper semantic understanding. With advancements in zero-shot and few-shot learning, these systems can also adapt more efficiently to diverse domains with minimal supervision, paving the way for more robust, interpretable, and human-like video understanding.

REFERENCES

1. Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A Joint Model for Video and Language Representation Learning. ICCV.
2. Luo, Y., et al. (2020). UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation.
3. Alayrac, J.-B., et al. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. DeepMind.
4. Li, J., et al. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.
5. Yang, J., et al. (2021). ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering.
6. Y. Ma *et al.*, "Dreamtalk: When expressive talking head generation meets diffusion probabilistic models," *arXiv preprint arXiv:2312.09767*, 2023.
7. B. Qin *et al.*, "Dancing avatar: Pose and text-guided human mo- tion videos synthesis with image diffusion model," *arXiv preprint arXiv:2308.07749*, 2023.
8. T. Wang *et al.*, "Disco: Disentangled control for realistic human dance generation," 2023.
9. Van den Oord *et al.*, "Conditional image generation with pixelcnn decoders," *Advances in neural information processing systems*, vol. 29, 2016.
    A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
10. O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Comp*
    A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learn- ing*. PMLR, 2021, pp. 8748–8763.
11. Y. Chang *et al.*, "A survey on evaluation of large language models,"
12. *arXiv preprint arXiv:2307.03109*, 2023.
13. G. Chen *et al.*, "Videollm: Modeling video sequence with large language models," *arXiv preprint arXiv:2305.13292*, 2023.
14. Y. Zhu *et al.*, "A comprehensive study of deep video action recogni- tion," *arXiv preprint arXiv:2012.06567*, 2020.
15. M. Bain, "Understanding video through the lens of language," Ph.D. dissertation, University of Oxford, 2023.
16. S. Wu *et al.*, "Next-gpt: Any-to-any multimodal llm," *arXiv preprint arXiv:2309.05519*, 2023.
17. N. Aldausari *et al.*, "Video generative adversarial networks: a review,"
18. *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–25, 2022.
19. G. Rafiq *et al.*, "Video description: A comprehensive survey of deep learning approaches," *Artificial Intelligence Review*, pp. 1–80, 2023.
    A. Singh *et al.*, "A comprehensive review on recent methods and challenges of video description," *arXiv preprint arXiv:2011.14752*, 2020.
20. N. Aafaq *et al.*, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6,
21. pp. 1–37, 2019.
22. S. K. Muhammad Maaz, Hanoona Rasheed *et al.*, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *ArXiv 2306.05424*, 2023.
23. Vondrick *et al.*, "Generating videos with scene dynamics," *Advances in neural information processing systems*, vol. 29, 2016.
24. E. Denton *et al.*, "Stochastic video generation with a learned prior," in *International conference on machine learning*. PMLR, 2018, pp. 1174–1183.
25. N. Kalchbrenner *et al.*, "Video pixel networks," in *International Con- ference on Machine Learning*. PMLR, 2017, pp. 1771–1779.
26. J. Ho *et al.*, "Video diffusion models," *arXiv preprint arXiv:2204.03458*, 2022.
27. ——, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.
28. U. Singer *et al.*, "Make-a-video: Text-to-video generation without text- video data," *arXiv preprint arXiv:2209.14792*, 2022.
29. Chan *et al.*, "Everybody dance now," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5933–5942.