Generative AI for HTTP Adaptive Streaming

Emanuele Artioli
Alpen-Adria-Universität Klagenfurt, Christian Doppler Laboratory ATHENA
Klagenfurt, Austria

ABSTRACT

Video streaming stands as the cornerstone of telecommunication networks, constituting over 60% of mobile data traffic as of June 2023. The paramount challenge faced by video streaming service providers is ensuring high Quality of Experience (QoE) for users. In HTTP Adaptive Streaming (HAS), including DASH and HLS, video content is encoded at multiple quality versions, with an Adaptive Bitrate (ABR) algorithm dynamically selecting versions based on network conditions. Concurrently, Artificial Intelligence (AI) is revolutionizing the industry, particularly in content recommendation and personalization. Leveraging user data and advanced algorithms, AI enhances user engagement, satisfaction, and video quality through super-resolution and denoising techniques.

However, challenges persist, such as real-time processing on resource-constrained devices, the need for diverse training datasets, privacy concerns, and model interpretability. Despite these hurdles, the promise of Generative Artificial Intelligence emerges as a transformative force. Generative AI, capable of synthesizing new data based on learned patterns, holds vast potential in the video streaming landscape. In the context of video streaming, it can create realistic and immersive content, adapt in real time to individual preferences, and optimize video compression for seamless streaming in low-bandwidth conditions.

This research proposal outlines a comprehensive exploration at the intersection of advanced AI algorithms and digital entertainment, focusing on the potential of generative AI to elevate video quality, user interactivity, and the overall streaming experience. The objective is to integrate generative models into video streaming pipelines, unraveling novel avenues that promise a future of dynamic, personalized, and visually captivating streaming experiences for viewers.

KEYWORDS

Video Streaming, Generative AI

ACM Reference Format:
Emanuele Artioli. 2024. Generative AI for HTTP Adaptive Streaming. In *ACM Multimedia Systems Conference 2024 (MMSys '24), April 15–18, 2024, Bari, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3625468.3652912

MMSys '24, April 15–18, 2024, Bari, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0412-3/24/04.
https://doi.org/10.1145/3625468.3652912

1 INTRODUCTION

Video streaming is the most important service in telecommunication networks. According to the estimations provided in the *Ericsson Mobility Report*, [1], video streaming accounted for more than 60% of all mobile data traffic by June 2023. Given that such a vast portion of the Internet's traffic is associated with video content provisioning, video streaming service providers demand new techniques to provide high Quality of Experience (QoE) to their customers, *i.e.*, the users' satisfaction while watching video content [2]. In HTTP Adaptive Streaming (HAS) – including *Dynamic Adaptive Streaming over HTTP* (DASH) [3] and *HTTP Live Streaming* (HLS) [4] – a video content stored at the server is encoded at several quality versions, each identified by a specific bitrate. Each version is then divided into temporal segments of the same duration. This process generates a fixed number of segments for each quality *representation* of the video. The client sends HTTP requests to fetch the segments composing the video content sequentially. Based on the network conditions, the Adaptive Bitrate (ABR) algorithm implemented at the player selects the most suitable versions for the segments.

Artificial Intelligence (AI) is revolutionizing many industries, offering significant benefits and driving advancements in various areas. One notable area where AI excels is content recommendation and personalization [5]. By leveraging user data and applying sophisticated algorithms, video streaming platforms can deliver tailored content recommendations, enhancing user engagement and satisfaction. AI algorithms also enhance video quality by utilizing super-resolution, denoising, and frame interpolation, resulting in improved visual experiences for viewers [6, 7, 8]. Moreover, AI is employed in optimizing content delivery and minimizing rebuffering through adaptive bitrate algorithms that dynamically adjust video quality based on network conditions. However, several challenges lie ahead for AI in the video streaming industry. One challenge is the ever-increasing demand for real-time video processing, particularly on resource-constrained devices [9]. Developing efficient AI models and algorithms that can operate within the limitations of these devices is crucial. Another challenge is the need for large and diverse datasets for training AI models effectively. Access to high-quality, annotated datasets that capture the complexities of video streaming scenarios can be a hurdle [10]. Additionally, ensuring privacy and data security while utilizing user data for AI is a growing concern that must be addressed [11]. Furthermore, the interpretability and explainability of AI models in the video streaming context are essential for building trust and transparency [12]. Lastly, the rapid evolution of AI techniques necessitates continuous research and development efforts. The industry must stay updated with the latest advancements, explore new algorithms, and refine existing models to meet evolving user expectations and deliver seamless video streaming experiences. Despite these challenges, AI holds immense potential for the video streaming industry.

One facet of AI, in particular, has emerged as a promising frontier in video streaming: Generative Artificial Intelligence. Generative AI encompasses a set of algorithms designed to generate new, synthetic data based on patterns learned from existing datasets [13].

In the context of video streaming, Generative AI offers a multitude of possibilities. It can create realistic and immersive visual content by leveraging deep learning techniques [6]. Moreover, these models can adapt in real time, tailoring content to individual preferences and enhancing user engagement and satisfaction. Additionally, generative AI can optimize video compression techniques, ensuring seamless streaming experiences even in low-bandwidth conditions [14]. By exploring the intersection of advanced AI algorithms and digital entertainment, we aim to unravel the potential of generative AI in enhancing video quality, user interactivity, and overall streaming experience. We plan to integrate generative models into video streaming pipelines and uncover novel avenues that will shape the future of entertainment, offering viewers a more dynamic, personalized, and visually captivating streaming experience.

2 RESEARCH QUESTIONS

In our research, we investigate the following research questions:

RQ-1. How can AI generate recommendations for user streaming engagement accurately? By leveraging the aforementioned AI advancements, in particular the combination of sequential machine-learning models and digital twins, we want to improve the modeling and prediction of user engagement. Literature regarding user sensitivities' impact on video streaming artifacts and anomalies is shallow, and we plan to offer a comprehensive model that can explore user differences and leverage them to predict user engagement. To do so, we will develop a digital twin-based model that learns from a vast dataset of real user video streaming sessions, models each user separately, and can act like the real user it models to provide insight into their decisions and suggestions on optimizing their experience.

RQ-2. How can AI techniques be leveraged in frame interpolation to efficiently reduce video data transmission? Generative AI allows us to remodel the main goals of video streaming with novel approaches. For example, codecs have revolutionized video transmission by greatly reducing the bandwidth required to stream content over the Internet without sacrificing final quality. An AI-based, intra-frame interpolation strategy could intelligently predict and send specific portions of frames, minimizing data size without spatial constraints, similar to HEVC spatial encoding but with unprecedented flexibility. Key areas of the frame would receive higher-quality rendering, optimizing bandwidth usage. Using synchronized models on both the server and client sides could validate real-time interpolations for accuracy before transmission, ensuring top-notch quality. Simultaneously, an inter-frame interpolation approach would prioritize frame transmission based on importance. By calculating frame significance concerning bandwidth estimations, critical frames will be sent in full, moderately important ones will employ advanced encoding methods, and the least crucial ones will be omitted to be accurately reconstructed by the client. With its innovative techniques and bidirectional AI-driven solutions, such a project is poised to greatly reduce the bandwidth necessary for video streaming.

RQ-3. How can AI techniques be leveraged in object interpolation to efficiently reduce video data transmission? Our research uses advanced AI methodologies like video transformers and generative adversarial networks to revolutionize object interpolation in video transmission. We employ generative AI models by generating segments from initial frames and movement data of identified objects. Moreover, we explore the feasibility of storing objects locally on clients and transmitting interaction data, constructing entire scenes from textual descriptions. To ensure generated video quality, we propose a Synthetic Video metric, a novel evaluation tool. This metric draws inspiration from established video QoE tools like PSNR [15] and ITU-T P.1203 [16], complemented by subjective tests where individuals assess synthetic videos. This comprehensive approach will enable us to assess the efficiency of data transmission and the perceptual quality of reconstructed scenes, addressing potential artifacts and peculiarities in synthetic videos.

3 RELATED WORK

The following studies offer an overview of the current landscape and possible new avenues:

RQ-1. The field of AI is currently experiencing rapid advancements, driven by the introduction of innovative technologies such as large language models [17, 6, 18], and digital twins [19, 20]. These technologies have already disrupted various domains and gained significant attention from diverse audiences. While efforts have been made to extend their applications to multiple industries, progress in the domain of video streaming has been relatively limited. Huang et al. [20] employed a basic digital twin approach to model personalized QoE for video streaming users. Wang et al. [21] addressed a similar problem by utilizing a Transformer model in conjunction with Deep Learning techniques to enhance QoE. Although these studies have shown promising results, it is important to note that this research direction is still in its early stages and lacks the refinement exhibited by these technologies in their original domains. Therefore, further endeavors are necessary to bridge the gap and bring video streaming up to par with the current advancements in ML.

RQ-2/3. In adherence to Moore's Law, computational capabilities have experienced exponential growth in recent times. Notably, mobile devices have emerged as key contributors to this trend by incorporating dedicated ML chips in recent years, significantly enhancing their capacity for AI applications [22]. Exploiting these advanced chips, numerous studies have successfully transposed applications that were traditionally confined to potent desktops or servers onto mobile devices [23, 18]. Despite these strides, these investigations are still in their nascent stages, warranting additional endeavors, particularly in the realm of frame interpolation. This involves predicting portions of frames [24] or entire frames [25, 26], a crucial aspect that demands further attention. By capitalizing on the inherent redundancies present in video frames and leveraging knowledge about the shapes and likely movement patterns of objects, AI implemented on devices can adeptly reconstruct videos from significantly reduced data feeds. This not only enhances computational efficiency but also markedly diminishes the reliance on network bandwidth.

4 PROPOSED METHODOLOGY

The process that will be followed when researching is based on design science research principles enunciated in [27], and is composed of the following steps:

(1) *Problem identification and motivation:* In the first step, we choose the research question to tackle. We identify the scope of content encoding, provisioning, content delivery, or end-to-end solution. We then justify the importance of our problem for the research audience to accept the results.
(2) *Objectives of a solution:* After determining the problem, we produce an achievable goal for that model. In the context of ML for adaptive video streaming, this typically means identifying the target metric of interest that the system will try to optimize for.
(3) *Design and development:* At this stage, we compare alternatives and determine the optimal model or paradigm for the identified goal, based on ML best practices and current state of the art. We proceed to design the development process, from data gathering, to processing, to model implementation and interpretation of its results.
(4) *Demonstration:* The model will then be implemented in a HAS pipeline, and extensive experiments on a wide range of test data and/or context conditions will be conducted.
(5) *Evaluation:* The proposed solutions will be compared with other state-of-the-art approaches to show their performance.
(6) *Communication:* Finally, we will offer insight and observations about the proposed architecture, its results both as a stand-alone and in comparison with state of the art, especially regarding the tradeoffs among different alternatives, and offer guidance towards the next steps in the research.

5 ONGOING AND FUTURE WORK

To address RQ-1, specifically regarding the digital twin technology, we propose DIGITWISE: Digital Twin-based Modeling of Adaptive Video Streaming Engagement, which has been accepted to the MMSys'24 conference. This paper focused on the modeling of users' sensitivities with respect to video streaming artifacts and issues, such as stalls and quality changes, to predict the percentage of video that will be watched after only the first few minutes of session data. Results show that this model can consistently predict user engagement with performance on par with state-of-the-art models while allowing for flexibility and opening the door to various applications, among which the possibility for content providers to change streaming parameters like video codec or bitrate ladder and simulate with precision the impact of the change on a selected user group's viewing time. DIGITWISE is planned to be expanded with an online version that offers the same key benefits while being integrated into the streaming player and updating its predictions in real-time via time-sensitive models such as Transformers [17] or LSTMs [28]. Furthermore, this model will expand the digital twin modeling capabilities, by allowing the model to predict user actions such as pausing and seeking.

With regards to RQ-2, we plan to develop an AI-driven intra-frame interpolation model, which predicts and transmits specific frame portions without spatial constraints, similar to HEVC spatial encoding but with increased flexibility. This targeted transmission optimizes bandwidth utilization by prioritizing critical frame areas for higher-quality rendering. Additionally, synchronized models on both server and client ends validate real-time interpolations before transmission, ensuring video quality. Concurrently, an inter-frame interpolation method prioritizes frame transmission based on their significance, considering bandwidth estimations. This nuanced approach enables the transmission of crucial frames intact, implements advanced encoding for moderately important frames, and omits the least important frames for accurate client-side reconstruction. Such an approach has the potential to greatly reduce bandwidth requirements for video streaming while maintaining content integrity and quality.

Finally, for RQ-3, we plan on leveraging advanced AI techniques, including video transformers [29] and generative adversarial networks [30], to redefine object interpolation in video transmission. Our approach involves utilizing generative AI models to create segments based on initial frames and object movement data. Additionally, we investigate the viability of storing objects locally on client devices and transmitting interaction data, constructing entire scenes from textual descriptions. To evaluate the quality of the generated videos, we would introduce a novel assessment tool called the Synthetic Video metric. Inspired by established video quality evaluation methods like PSNR and ITU-T P.1203 [16], this metric is further enhanced through subjective tests, where individuals assess synthetic videos. By adopting this comprehensive methodology, we aim to assess both the efficiency of data transmission and the perceptual quality of reconstructed scenes. This approach allows us to identify and address potential artifacts and peculiarities in synthetic videos, paving the way for improved object interpolation techniques in video streaming applications.

6 TIME PLAN AND TARGET PUBLICATION VENUES

Table 1 shows the time plan for the thesis and Table 2 the possible target venues for future publications together with the corresponding quality factors [1][2][3].

REFERENCES

[1] Ericsson. [n. d.] Ericsson Mobility Report. [Online] Available: https://www.ericsson.com/49dd9d/assets/local/reports-papers/mobility-report/documents/2023/ericsson-mobility-report-june-2023.pdf. Accessed: 18 Dec 2023. ().

[2] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2015. A survey on quality of experience of http adaptive streaming. *IEEE Communications Surveys Tutorials*, 17, 469–492. DOI: 10.1109/COMST.2014.2360940.

[3] DASH Industry Forum (DASH-IF). [n. d.] dash.js JavaScript Reference Client. [Online] Available: https://reference.dashif.org/dash.js/. Accessed: 18 Dec 2023. ().

[4] Pantos Roger and May William. [n. d.] HTTP Live Streaming. RFC 8216. [Online] Available: https://www.rfc-editor.org/info/rfc8216. Accessed: 18 Dec 2023. ().

[5] Min Xu, Jesse S. Jin, and Suhuai Luo. 2008. Personalized Video Adaptation Based on Video Content Analysis. In *Proceedings of the 9th International Workshop on Multimedia Data Mining: Held in Conjunction with the ACM SIGKDD*, 26–35. DOI: 10.1145/1509212.1509216.

[6] Rombach Robin, Blattmann Andreas, Lorenz Dominik, Esser Patrick, and Ommer Björn. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695. DOI: 10.48550/arXiv.2112.10752.

[7] Tian Chunwei, Fei Lunke, Zheng Wenxian, Xu Yong, Zuo Wangmeng, and Lin Chia-Wen. 2020. Deep Learning on Image Denoising: An Overview. *Neural Networks*, 131, 251–275. DOI: 10.1016/j.neunet.2020.07.025.

[8] Ke Li, Bahetiyaer Bare, and Bo Yan. 2017. An Efficient Deep Convolutional Neural Networks Model For Compressed Image Deblocking. In *IEEE International Conference on Multimedia and Expo (ICME)*, 1320–1325. DOI: 10.1109/ICME.2017.8019416.

[1] http://www.conferenceranks.com/ (date of last access: 18 Dec 2023)
[2] http://www.guide2research.com/ (date of last access: 18 Dec 2023)
[3] https://www.scimagojr.com/journalrank.php?type=j (date of last access: 18 Dec 2023)

Table 1: time plan

	2023				2024				2025				2026			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Literature Review	■			■				■		■						
DIGITWISE		■	■													
Intra-frame Interpolation				■	■											
DIGITWISE [extended version]						■	■									
Inter-frame Interpolation								■	■							
Object Interpolation										■	■					
Dissertation												■	■			

Table 2: Possible target venues for future publications

Target publication venue	Type	Rank
ACM Multimedia Systems Conference (MMSys)	Conference	B [1]
ACM Multimedia Conference (MM)	Conference	A1 [1]
International Conference on Quality of Multimedia Experiences (QoMEX)	Conference	Top [2]
Symposium on Networked Systems, Design and Implementation (NSDI)	Conference	A1 [1]
ACM Transactions on Multimedia Computing, Communications, and Applications (ACM TOMM)	Journal	Q1 [3]
IEEE Transactions on Multimedia (IEEE TMM)	Journal	Q1 [3]

[9] Maxim Claeys, Steven Latré, Jeroen Famaey, and Filip De Turck. 2014. Design and Evaluation of a Self-Learning HTTP Adaptive Video Streaming Client. *IEEE Communications Letters*, 18, 4, 716–719. DOI: 10.1109/LCOMM.2014.020414.132649.

[10] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. 2011. Understanding the Impact of Video Quality on User Engagement. *ACM Computer Communication Review (SIGCOMM)*, 41, 4, 362–373. DOI: 10.1145/2043164.2018478.

[11] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Computer Survey*, 54, 2, 1–36. DOI: 10.1145/3436755.

[12] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, (NLPCC)*, 563–574. DOI: 10.1007/978-3-030-32236-6_51.

[13] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative ai and chatgpt: applications, challenges, and ai-human collaboration. In number 3. Vol. 25, 277–304. DOI: 10.1080/15228053.2023.2233814.

[14] Minh Nguyen, Ekrem Çetinkaya, Hermann Hellwagner, and Christian Timmerer. 2022. Super-Resolution Based Bitrate Adaptation for HTTP Adaptive Streaming for Mobile Devices. In *Proceedings of the 1st Mile-High Video Conference*, 70–76. DOI: 10.1145/3510450.3517322.

[15] Deepak S Turaga, Yingwei Chen, and Jorge Caviedes. 2004. No reference PSNR estimation for compressed pictures. In number 2. Vol. 19, 173–184. DOI: 10.1109/ICIP.2002.1038903.

[16] Alexander Raake, Marie-Neige Garcia, Werner Robitza, Peter List, Steve Göring, and Bernhard Feiten. 2017. A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6. DOI: 10.1109/QoMEX.2017.7965631.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30. DOI: 10.48550/arXiv.1706.03762.

[18] Andrey Ignatov, Andres Romero, Heewon Kim, and Radu Timofte. 2021. Real-Time Video Super-Resolution on Smartphones With Deep Learning, Mobile AI 2021 Challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2535–2544. DOI: 10.48550/arXiv.2105.08826.

[19] Michael W Grieves. 2019. Virtually Intelligent Product Systems: Digital and Physical Twins. In *Complex Systems Engineering: Theory and Practice*, 175–200. DOI: 10.2514/5.9781624105654.0175.0200.

[20] Xinyu Huang, Conghao Zhou, Wen Wu, Mushu Li, Huaqing Wu, and Xuemin Shen. 2022. Personalized QoE Enhancement for Adaptive Video Streaming: A Digital Twin-Assisted Scheme. In *IEEE Global Communications Conference (GLOBECOM)*, 4001–4006. DOI: 10.48550/arXiv.2205.04014.

[21] Shuoyao Wang, Suzhi Bi, and Ying-Jun Angela Zhang. 2021. Deep Reinforcement Learning with Communication Transformer for Adaptive Live Streaming in Wireless Edge Networks. In number 1. Vol. 40, 308–322. DOI: 10.1109/JSAC.2021.3126062.

[22] Vijay Janapa Reddi et al. 2022. MLPerf Mobile Inference Benchmark. In *Proceedings of Machine Learning and Systems*. Vol. 4, 352–369. DOI: 10.48550/arXiv.2012.02328.

[23] Wojciech Dudzik and Damian Kosowski. 2020. Kunster – AR Art Video Maker – Real time video neural style transfer on mobile devices. (2020). DOI: 10.48550/arXiv.2005.03415.

[24] Arief Bramanto Wicaksono Putra, Achmad Fanany Onnilita Gaffar, Muhammad Taufiq Sumadi, and Lisa Setiawati. 2022. Intra-frame based video compression using deep convolutional neural network (dcnn). *JOIV: International Journal on Informatics Visualization*, 6, 3, 650–659. DOI: 10.30630/joiv.6.3.1012.

[25] Simon Niklaus, Long Mai, and Feng Liu. 2017. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. (Oct. 2017). DOI: 10.48550/arXiv.1708.01692.

[26] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. 2019. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.48550/arXiv.1904.00830.

[27] K Peffers, T Tuunanen, C.E. Gengler, M Rossi, W Hui, V Virtanen, and J Bragge. 2006. The Design Science Research Process: A Model for Producing and Presenting Information Systems Research. In Claremont Graduate University, 83–106. DOI: 10.2753/MIS0742-1222240302.

[28] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9, 8, 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[29] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846. DOI: 10.48550/arXiv.2103.15691.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27. DOI: 10.48550/arXiv.1406.2661.