Towards a Unified Learning-based Video Compression and Streaming Pipeline

Hannes Keunen
hannes.keunen@uhasselt.be
Hasselt University
Digital Future Lab
Diepenbeek, Belgium

Abstract

Video streaming accounts for a significant portion of global internet traffic, necessitating video delivery systems that efficiently utilize network resources while maximizing end-user Quality of Experience (QoE). Traditional techniques for video compression as well as adaptive bitrate (ABR) streaming rely on hand-crafted heuristics, but have recently been superseded by learned alternatives. However, these components are typically studied in isolation. This Ph.D. research investigates how learned video compression and streaming algorithms can be jointly optimized to improve over-the-top end-to-end QoE. The project focuses on neural video representations (NVRs) as a lightweight alternative to conventional codecs, analyzing their limitations in streaming scenarios and developing methods to reduce encoding complexity. In parallel, the research aims to build a systematic understanding of learning-based ABR streaming approaches and their design trade-offs. The overarching goal is to build a unified learning-based video compression and streaming pipeline optimized for QoE. Initial work includes a survey of NVR-based video compression methods and an ongoing study on accelerating NVR encoding using hypernetworks.

CCS Concepts

• **Information systems** → **Multimedia streaming**; • **Computing methodologies** → **Machine learning**; *Reinforcement learning*.

Keywords

Video streaming, Video compression, Adaptive bitrate streaming, Deep learning, Reinforcement learning

ACM Reference Format:
Hannes Keunen. 2026. Towards a Unified Learning-based Video Compression and Streaming Pipeline. In *ACM Multimedia Systems Conference 2026 (MMSys '26), April 4–8, 2026, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3793853.3798411

1 Introduction

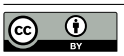
According to WifiTalents' 2025 Video Streaming Industry Statistics, over 80% of internet traffic worldwide is attributed to video streaming, while viewers spend 25% more time on platforms with adaptive streaming technology [20]. This highlights the importance of video delivery systems that can handle large amounts of high-quality video and actively strive to optimize the end-user's Quality of Experience (QoE). According to the same report, 4K video requires a median bandwidth of approximately 25Mbps. Advanced video compression standards can reduce the number of bits required for storing and retrieving video data, enabling higher-quality video to be transmitted over a limited network bandwidth. In platforms with adaptive bitrate (ABR) streaming, a video is typically split into temporal segments of a few seconds, and each segment is subsequently encoded at different bitrates. A rate adaptation algorithm then selects the most suitable bitrate for each segment under varying network conditions, thereby avoiding rebuffering and directly optimizing QoE.

Traditional video compression standards, with H.266/VVC [2] as the current state-of-the-art, rely on hand-crafted features and heuristics to exploit spatial and temporal correlations between nearby pixels. Similarly, ABR algorithms typically use heuristics based on measured throughput and/or buffer size [18, 21]. Despite the proven effectiveness of these methods, the reliance on heuristics and manually tuned parameters limits flexibility and adaptability to unseen circumstances. Recent advances have already shown that learned models can improve these two aspects of the video delivery pipeline. Neural video compression schemes replace hand-crafted codecs by deep neural networks [8, 15], while reinforcement learning (RL)–based ABR algorithms have demonstrated superior performance over heuristic approaches [1, 16, 19]. Yet both paradigms still face significant challenges, and little work has explored their integration. The overarching goal of this Ph.D. research is to investigate the extent to which learned video compression and learning-based streaming algorithms can jointly optimize QoE in realistic over-the-top (OTT) video consumption scenarios.

The remainder of this proposal is structured as follows. Sections 1.1 and 1.2 first outline the research challenges related to ongoing works in learned video compression and streaming, respectively. Section 1.3 identifies measuring and optimizing QoE as an underlying challenge related to both research areas. Based on these challenges, Section 2 summarizes the concrete objectives of this project. Section 3 provides a high-level plan of the steps that will be taken to achieve the objectives from Section 2.

1.1 Neural Video Compression

Neural video codecs replace traditional codecs with a compression model that is trained end-to-end [15]. These models are usually based on an autoencoder backbone and mimic the traditional video compression pipeline by explicitly modelling temporal redundancies. This can be trained end-to-end by jointly optimizing



MMSys '26, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2481-7/26/04
https://doi.org/10.1145/3793853.3798411

for low distortion and low bitrate. While recent works on these autoencoder-based methods have exceeded the compression performance of traditional codecs [8], this approach also has some downsides. Since the characteristics of videos can vary significantly across domains (e.g., live-action vs. animation, low vs. high motion, etc.), it is infeasible to train a single model that can capture the distribution of all possible videos. This is especially challenging if the decoder is deployed on a client device with limited computational power, such as a mobile phone.

In an alternative line of work, neural video representations (NVRs) represent a video as an implicit neural representation, a neural network that is overfitted to a single data sample, or in this case, a single video. The video is then implicitly encoded in the parameters of the neural network. NeRV [4] first applied this paradigm to videos, where video frames are decoded through a simple forward pass with only a frame index as input. Subsequent works have introduced more parameter-efficient model architectures [10, 14, 23] and more powerful input embeddings [3, 12]. This method can also be optimized for rate and distortion jointly by imposing a rate penalty on the network parameters [7, 23]. Recent works on video compression with NVRs have surpassed the compression performance of traditional video codecs [11]. In contrast to autoencoder-based neural video codecs, NVRs do not rely on large-scale datasets for training and are more lightweight for deployment on streaming clients with limited computational power. On the other hand, NVRs suffer from longer encoding times due to the need to train a new model for each individual video, which limits their practical applicability.

1.2 Learned Adaptive Bitrate Streaming

As mentioned in the introduction, existing works for learned bitrate adaptation are based on RL, where an agent is trained to maximize some reward function. For the specific case of ABR streaming, the reward function should reflect the end user's QoE. Pensieve [16] formulates this as a trade-off between high bitrate, low rebuffering time, and a low number of oscillations between bitrates. Other works use perceptual video quality metrics such as VMAF, based on the observation that perceptual improvements tend to saturate at higher bitrates [1, 13, 19]. For mobile clients, energy usage may also play a role [19].

A second important decision is how the current state is presented to the RL agent. The state usually includes playback buffer occupancy and network-related metrics, as well as some information about the next downloadable segments [16]. Dan et al. [6] extend this with visual sensitivity information extracted from each chunk, allowing the agent to prioritize perceptually important content when making rate adaptation decisions.

Finally, there is the choice of which specific RL techniques to use. The shape of the action space plays an important role in this decision. If each bitrate is represented as a discrete action, the agent has to be retrained for each new bitrate ladder [19]. When the bitrate is regarded as continuous, each action still has to be mapped to a specific bitrate in the current bitrate ladder. Additionally, because network conditions can vary significantly, the agent must be able to adapt quickly to various unseen circumstances. Bentaleb et al. [1] propose a meta-learning approach to achieve this, while Zhang et al. [22] utilize online learning combined with a safe fallback to heuristic methods in extreme scenarios.

1.3 Measuring and Optimizing Quality of Experience

A common challenge for both compression and streaming consists of designing suitable metrics that accurately reflect end users' Quality of Experience (QoE). Existing learned video compression methods are typically trained and evaluated using distortion objectives such as mean squared error (MSE), or signal fidelity metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [4, 8]. While these measures quantify pixel-level reconstruction accuracy, they do not necessarily correlate with human perception.

As mentioned in Section 1.2, several works related to learning-based ABR streaming already use VMAF rather than bitrate to optimize perceptual video quality during training. VMAF is a perceptual metric that correlates better with human judgment than PSNR or SSIM [13]. However, its behavior in the context of neural video representations (NVRs) warrants further investigation, as NVR-based compression may produce artifacts that differ from those of traditional codecs. It is therefore not immediately self-evident that VMAF remains equally reliable as an optimization target. Beyond VMAF, ITU-T Rec. P.1203 [17] provides a standardized framework for estimating end-to-end QoE in HTTP adaptive streaming by combining video quality, rebuffering events, and playback interruptions into a single model. However, P.1203 is primarily designed as an evaluation metric and is not directly differentiable, which complicates its use as a training objective.

2 Objectives

Based on Sections 1.1 to 1.3, the research objectives of this Ph.D. project can be summarized as follows:

- **RO1: NVR-based video compression.** Investigate and identify the strengths and weaknesses of NVRs for video compression. Based on the identified weaknesses, develop techniques to improve the QoE provided by NVRs in the context of OTT video streaming. Relevant metrics that are directly related to QoE include bitrate, perceptual reconstruction quality, and the speed and computational complexity of encoding and decoding.
- **RO2: Learning-based adaptive bitrate streaming.** Analyze state-of-the-art learning-based ABR streaming algorithms, with a focus on RL approaches. This objective aims to identify how design choices such as state representations, action spaces, reward functions, perceptual quality metrics, and adaptation mechanisms affect robustness, efficiency, and QoE under dynamic network conditions.
- **RO3: A unified learning-based video compression and streaming pipeline.** Develop a unified video delivery system that integrates learned video compression and learning-based ABR streaming. Building on the insights from **RO2** and the NVR techniques from **RO1**, the goal is to jointly optimize compression and streaming decisions for end-to-end QoE in realistic network and device settings.

3 Methods and Contributions

The project can be divided into three chronological phases that correspond with the research objectives outlined in Section 2.

The first phase focuses on **RO1: NVR-based video compression**. A major step towards this objective is presented in [9], which is under review as part of this research project. This work provides a systematic overview of existing NVR-based approaches to video compression and identifies several challenges that limit their applicability in practical video streaming scenarios. As already mentioned in Section 1.1, one of these challenges is related to encoding complexity. Chen et al. [5] approach this challenge by using a hypernetwork to predict NVR weights. Ongoing work in this Ph.D. project explores this idea further by treating the hypernetwork as an initialization method for an NVR-based encoder. The resulting NVR can then be used as is, or further trained independently of the hypernetwork to improve compression performance at the cost of a longer encoding time. This represents a concrete step towards making NVR-based video compression viable for adaptive streaming applications.

The second phase of the project addresses **RO2: learning-based adaptive bitrate streaming**. This phase will focus on a systematic analysis of existing RL–based ABR approaches. Rather than actively advancing the state of the art in RL-based video streaming, the goal of this phase is to gain an understanding of which learning-based techniques are most suitable for integration with learned video compression methods. The outcome of this phase is expected to be a set of design guidelines and evaluation criteria that inform the unified system developed in the final phase.

The final phase of the project targets **RO3: a unified learning-based video streaming pipeline**. This phase aims to integrate the insights gained from RO1 and RO2 into an end-to-end system, where learned video compression and adaptive bitrate streaming are jointly optimized for QoE. Concretely, this phase will investigate how content-dependent properties of learned compression methods, such as rate-distortion characteristics, decoding complexity, and encoding latency, can be exposed to and exploited by learning-based ABR algorithms. Rather than treating compression and streaming as independent modules, the objective is to enable tighter coupling between both components, allowing streaming decisions to account for the unique behavior of learned codecs. The resulting system will be evaluated in simulated streaming environments and compared against state-of-the-art modular pipelines.

4 Conclusion

This proposal outlines a Ph.D. research project aimed at advancing learning-based video streaming by jointly investigating learned video compression and adaptive bitrate streaming. After reviewing the limitations of traditional heuristics in both domains, the proposal identifies neural video representations as a method for video compression and reinforcement learning as a paradigm for adaptive streaming. It also identifies measuring and optimizing Quality of Experience (QoE) as a common challenge across both topics. The research objectives focus on improving the practicality of NVR-based video compression, developing a systematic design taxonomy of learning-based ABR algorithms, and ultimately integrating both components into a unified video delivery pipeline optimized for end-to-end QoE. This Ph.D. project is currently in its early to mid-stage, with substantial progress on learned video compression and an initial direction for future research on learning-based streaming and its integration.

Acknowledgments

Hannes Keunen (BOF24OWB27) is a Ph.D. candidate at Hasselt University supported by the Special Research Fund (BOF) and supervised by Prof. Dr. Jori Liesenborgs and Prof. Dr. Maarten Wijnants.

References

[1] Abdelhak Bentaleb, May Lim, Mehmet N. Akcay, Ali C. Begen, and Roger Zimmermann. 2024. Bitrate Adaptation and Guidance With Meta Reinforcement Learning. *IEEE Transactions on Mobile Computing* 23, 11 (November 2024), 10378–10392. doi:10.1109/TMC.2024.3376560

[2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. 2021. Overview of the Versatile Video Coding (VVC) Standard and its Applications. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 10 (2021), 3736–3764. doi:10.1109/TCSVT.2021.3101953

[3] Hao Chen, Matthew Gwilliam, Ser-Nam Lim, and Abhinav Shrivastava. 2023. HNeRV: A Hybrid Neural Representation for Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2023)*. Vancouver, BC, Canada, 10270–10279. https://doi.org/10.1109/CVPR52729.2023.00990

[4] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser-Nam Lim, and Abhinav Shrivastava. 2021. NeRV: Neural Representations for Videos. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. 21557–21568. Virtual Conference.

[5] Hao Chen, Saining Xie, Ser-Nam Lim, and Abhinav Shrivastava. 2024. Fast Encoding and Decoding for Implicit Video Representation. In *Proceedings of the 18th European Conference on Computer Vision (ECCV 2024)*. Milan, Italy, 402–418. doi:10.1007/978-3-031-72933-1_23

[6] Meng Dan, Jin Ye, Wenchao Jiang, and Yuanchao Shan. 2021. Visual Sensitivity Aware Rate Adaptation for Video Streaming via Deep Reinforcement Learning. In *2021 IEEE 23rd International Conference on High Performance Computing and Communications; 7th International Conference on Data Science and Systems; 19th International Conference on Smart City; 7th International Conference on Dependability in Sensor, Cloud and Big Data Systems and Application (HPCC/DSS/SmartCity/DependSys)*. IEEE, Haikou, Hainan, China, 141–148. doi:10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00045

[7] Carlos Gomes, Roberto Azevedo, and Christopher Schroers. 2023. Video Compression with Entropy-Constrained Neural Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2023)*. Vancouver, BC, Canada, 18497–18506. doi:10.1109/CVPR52729.2023.01774

[8] Zhaoyang Jia, Bin Li, Jiahao Li, Wenxuan Xie, Linfeng Qi, Houqiang Li, and Yan Lu. 2025. Towards Practical Real-Time Neural Video Compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2025)*. Nashville, TN, USA, 12543–12552. doi:10.1109/CVPR52734.2025.01170

[9] Hannes Keunen, Maarten Wijnants, and Jori Liesenborgs. 2026. A Survey of Implicit Neural Representations for Video Compression. *Multimedia Tools and Applications*. Manuscript submitted for review.

[10] Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. 2023. HiNeRV: Video Compression with Hierarchical Encoding-based Neural Representation. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. New Orleans, LA, USA.

[11] Ho Man Kwan, Ge Gao, Fan Zhang, Andrew Gower, and David Bull. 2024. NVRC: Neural Video Representation Compression. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. Vancouver, BC, Canada, 132440–132462.

[12] Joo Chan Lee, Daniel Rho, Jong Hwan Ko, and Eunbyung Park. 2023. FFNeRV: Flow-Guided Frame-Wise Neural Representations for Videos. In *Proceedings of the 31st ACM International Conference on Multimedia (MM 2023)*. Ottawa, ON, Canada, 7859–7870. doi:10.1145/3581783.3612444

[13] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. 2016. *Toward A Practical Perceptual Video Quality Metric*. Retrieved Dec 16, 2025 from https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[14] Zizhang Li, Mengmeng Wang, Huaijin Pi, Kechun Xu, Jianbiao Mei, and Yong Liu. 2022. E-NeRV: Expedite Neural Video Representation with Disentangled Spatial-Temporal Context. In *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)*. Tel Aviv, Israel, 267–284. doi:10.1007/978-3-031-19833-5_16

[15] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. 2019. DVC: An End-To-End Deep Video Compression Framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR 2019). Computer Vision Foundation / IEEE, Long Beach, CA, USA, 11006–11015. doi:10.1109/CVPR.2019.01126

[16] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM 2017)*. ACM, Los Angeles, CA, USA, 197–210. doi:10.1145/3098822.3098843

[17] Werner Robitza, Steve Göring, Alexander Raake, David Lindegren, Gunnar Heikkilä, Jörgen Gustafsson, Peter List, Bernhard Feiten, Ulf Wüstenhagen, Marie-Neige Garcia, Kazuhisa Yamagishi, and Simon Broom. 2018. HTTP adaptive streaming QoE estimation with ITU-T rec. P. 1203: open databases and software. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMsys 2018)*. Association for Computing Machinery, Amsterdam, Netherlands, 466–471. doi:10.1145/3204949.3208124

[18] Kevin Spiteri, Rahul Urgaonkar, and Ramesh K. Sitaraman. 2016. BOLA: Near-optimal bitrate adaptation for online videos. In *35th Annual IEEE International Conference on Computer Communications (INFOCOM 2016)*. IEEE, San Francisco, CA, USA, 1–9. doi:10.1109/INFOCOM.2016.7524428

[19] Bekir Oguzhan Turkkan, Ting Dai, Adithya Raman, Tevfik Kosar, Changyou Chen, Muhammed Fatih Bulut, Jaroslav Zola, and Daby Sow. 2024. GreenABR+: Generalized Energy-Aware Adaptive Bitrate Streaming. *TOMCCAP* 20, 9 (September 2024). doi:10.1145/3649898

[20] WifiTalents. 2025. *Video Streaming Industry Statistics*. Retrieved January 8, 2026 from https://wifitalents.com/video-streaming-industry-statistics/

[21] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. 2015. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. *Computer Communication Review* 45, 5 (2015), 325–338. doi:10.1145/2829988.2787486

[22] Huanhuan Zhang, Anfu Zhou, Jiamin Lu, Ruoxuan Ma, Yuhan Hu, Cong Li, Xinyu Zhang 0003, Huadong Ma, and XiaoJiang Chen. 2020. OnRL: improving mobile video telephony via online reinforcement learning. In *MobiCom '20: The 26th Annual International Conference on Mobile Computing and Networking*. ACM, London, United Kingdom. doi:10.1145/3372224.3419186

[23] X. Zhang, R. Yang, D. He, X. Ge, T. Xu, Y. Wang, H. Qin, and J. Zhang. 2024. Boosting Neural Representations for Videos with a Conditional Decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*. 2556–2566. doi:10.1109/CVPR52733.2024.00247

Received 16 January 2026; accepted 14 February 2026; revised 28 February 2026