LARGE MULTIMODAL MODEL-BASED VIDEO ENCODING OPTIMIZATION

Z. Duanmu, M. Jiang

(zduanmu@imax.com, mjiang@imax.com)

IMAX Streaming and Consumer Technology (SCT), Canada

ABSTRACT

In the realm of video encoding, achieving the optimal balance between encoding efficiency and computational complexity remains a formidable challenge. This paper introduces a groundbreaking framework that utilizes a Large Multi-modal Model (LMM) to revolutionize the process of per-title video encoding optimization. By harnessing the predictive capabilities of LMMs, our framework estimates the encoding complexity of video content with unprecedented accuracy, enabling the dynamic selection of encoding configurations tailored to each video's unique characteristics. The proposed framework marks a significant departure from traditional per-title encoding methods, which often rely on expensive and time-consuming sampling in the rate-distortion space. Through a comprehensive set of experiments, we demonstrate that our LMM-based approach not only significantly reduces the computational complexity required for sampling-based per-title video encoding—by an astounding 13 times—but also maintains the same level of bitrate saving. These findings not only pave the way for more efficient and adaptive video encoding strategies but also highlight the potential of multi-modal models in enhancing multimedia processing tasks. The implications of this research extend beyond the immediate improvements in encoding efficiency, offering a glimpse into the future of multimedia content distribution and consumption in an increasingly video-centric digital landscape.

INTRODUCTION

Adaptive streaming [1] has become the cornerstone of modern video delivery, enabling content providers to offer a seamless viewing experience across a wide range of devices and network conditions. This technology dynamically adjusts video quality during playback, based on the user's bandwidth and device capabilities, utilizing a predefined set of bitrate-quality pairs known as a bitrate ladder. However, the traditional “one-size-fits-all" approach [2-4] to constructing these bitrate ladders often falls short. It fails to account for the unique characteristics of each video, leading to suboptimal use of bandwidth and a compromised viewing experience.

In response to these limitations, per-title encoding optimization [2] has emerged as a solution that tailors the encoding settings for each video title based on its content

complexity. This method promises to significantly enhance the viewer's experience by optimizing the balance between video quality and file size. However, per-title optimization is computationally expensive [5]. It involves analyzing each video to determine its optimal bitrate ladder, a process that requires extensive computational resources and time. This complexity limits the scalability of per-title encoding, making it a challenge for content providers with large libraries of video content.

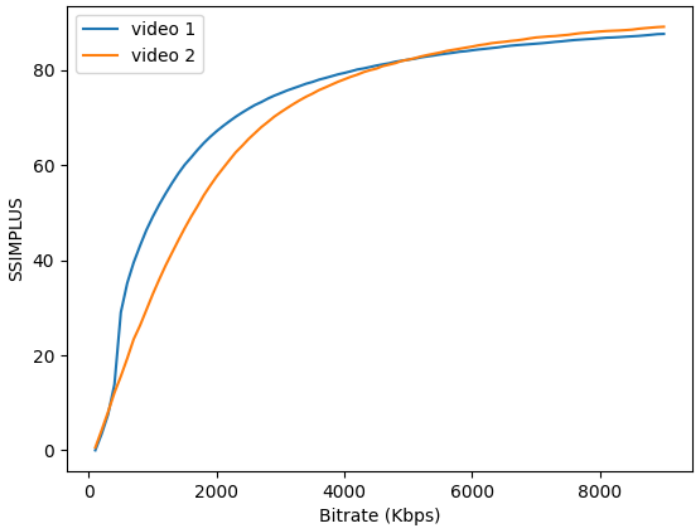

Figure 1: Rate-distortion curves of two videos with similar Spatial Information and Temporal Information

Recent efforts to streamline the per-title encoding process have explored the use of low-dimensional hand-crafted features such as Spatial Information and Temporal Information (SI/TI) [6] and regression models to predict the rate-distortion (RD) function [7-11], a key factor in determining optimal encoding settings. However, this approach suffers from two limitations. Figure 1 illustrates the R curves of two videos with similar SI/TI, from which we have two observations. First, low-level features are not a good representation of the encoding complexity, as they often overlook complex interplays of visual elements that significantly impact perceived quality. Second, multiple intersections between the two curves suggest that the encoding complexity lies in a high dimensional space. These limitations underscore the need for more sophisticated models that can better understand video content and predict encoding parameters.

Against this backdrop, the promise of Large Multi-modal Models (LMMs) [12-14] offers a compelling solution. LMMs leverage advances in artificial intelligence to analyze video content across multiple modalities—combining visual, audio, and textual information—to understand content complexity comprehensively. This paper introduces a novel framework that utilizes LMMs for video encoding optimization, aiming to overcome the shortcomings of traditional per-title optimization methods. By harnessing the predictive power of LMMs, our proposed solution not only aims to reduce the computational expense associated with per-title encoding but also improves the accuracy of RD function prediction, leading to more efficient and viewer-centric adaptive streaming experiences. This approach signifies a paradigm shift in how video content is delivered, promising substantial improvements in streaming quality and resource utilization.

BACKGROUND

Per-title Encoding Optimization

Per-title encoding optimization [1] represents a targeted approach in the realm of video processing, designed to tailor encoding parameters specifically for each video based on its unique content characteristics. This method manipulates additional encoding dimensions such as spatial resolution, ensuring that each video is encoded in a way that delivers the highest perceptual quality within a fixed bitrate budget. However, despite its effectiveness in enhancing viewer experience, per-title encoding optimization is notoriously expensive and time-consuming [2]. The process involves extensive sampling and analysis within the RD space for each video title, requiring significant computational resources.

This intensive approach, while beneficial for achieving optimal encoding settings, places a substantial burden on resources, making it a challenging endeavour for content providers who must manage large libraries of video content. Recent advancements in per-title encoding optimization have led to two notable approaches: one using curve fitting to reconstruct RD functions [15-18] and another predicting these functions based on low-level video features with regression models [7-11]. While these methods offer more efficient alternatives to traditional exhaustive sampling, they come with limitations. The computational complexity of curve fitting techniques, for instance, increases exponentially with respect to the dimensionality of encoding configuration. Similarly, predicting RD functions using hand-crafted low-level features such as spatial information and temporal information may overlook the impact of higher-level content attributes, such as narrative elements, objects, and texture type, on encoding efficiency. Furthermore, hand-crafted features may fail to capture all the nuances of the data, leading to suboptimal performance in video encoding optimization. These limitations underscore the ongoing need for more sophisticated models that can holistically account for the multifaceted nature of video content in the encoding process.

Large Multimodal Model

Large Multi-modal Models (LMMs) [12-14] have emerged as a transformative force in video understanding, harnessing the power of integrating multiple data modalities—text, images, and audio—to achieve a comprehensive analysis of video content. Their success is largely attributed to their ability to discern intricate details and contextual nuances within videos, which traditional single-modality approaches might miss [12]. This capability enables LMMs to perform exceptionally well in various video understanding tasks, including video retrieval [19], content classification [20], and activity recognition [21], thereby setting new benchmarks in the field.

PROPOSED FRAMEWORK

We initiate our discussion by delineating the foundational assumptions that underpin our framework. The first of these is predicated on the notion that videos sharing congruent characteristics will elicit analogous verbal descriptions. This assumption is deeply rooted in the theoretical understanding that language functions as an information compression heuristic [22], a concept that is well-documented within the research field. In practical scenarios, this is exemplified by numerous subjective video quality assessment datasets [23-25], which often classify video content based on observable characteristics such as the level of motion, texture, and camera movement.

Our second assumption extends from the premise that videos with akin complexity profiles are likely to demonstrate comparable RD behaviours. This underlying hypothesis forms the

bedrock of regression-based RD models, where it is implicitly inferred that the intricacies of a video's content correlate directly with its RD function. The explicit recognition not only provides a more transparent foundation for the mathematical soundness, but also implies a more deliberate and methodical approach in the RD function modelling, as we will see in the subsequent section.

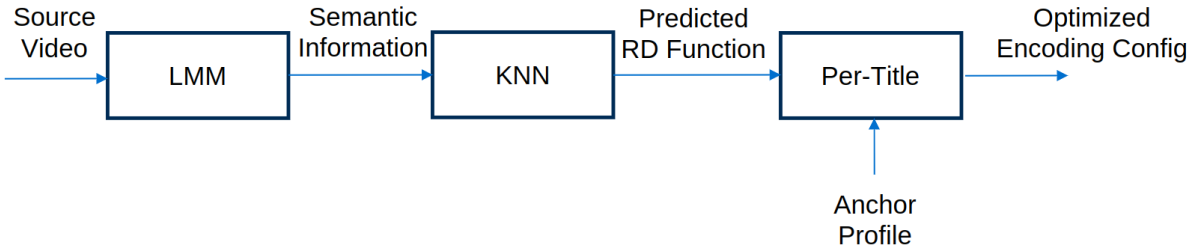

Figure 2. Proposed LMM-based video encoding optimization system

Building upon our foundational assumptions, we introduce an innovative framework, as depicted in Figure 2. At the genesis of this process stands the source video, which is ingested by a Large Multi-modal Model (LMM). The LMM is engineered to distil content characteristics from the video, translating complex visual and auditory data into a structured semantic representation. This semantic information, echoing our first assumption, encapsulates the notion that videos with shared characteristics can be uniformly described, compressing intricate content details into a descriptive language that mirrors human categorization.

Leveraging our second assumption—that videos with similar complexities will exhibit analogous RD functions—the semantic information is then fed into a k-nearest neighbours (KNN) algorithm. The KNN serves as a predictive tool that, using the distilled semantic descriptors, forecasts the RD function specific to the video. This function aims to map the relationship between bitrate and perceived quality, serving as a pivotal factor in the subsequent optimization steps.

The predicted RD function is then utilized in the Per-Title encoding stage. Here, the RD prediction is harmonized with an anchor profile, a predefined baseline of encoding parameters, to fine-tune the encoding process to satisfy the requirement from content distributors. This tailored encoding is crucial for transcoding the source video into the optimized encoding configuration, which is fitted to the video's unique content characteristics.

This framework embodies a strategic blend of linguistic theory and video analytics, transcending traditional exhaustive search and feature engineering and paving the way for a new era of content-specific encoding strategies that are both efficient and viewer-centric.

EXPERIMENTAL RESULTS

Experiment Setup

Dataset

We use the Waterloo Generalized Rate-Distortion Dataset (WaterlooGRD) [18] in our experiment. The dataset contains 1000 pristine semantically coherent videos. All the sequences are downsampled to FHD (1920x1080 pixels), converted to 4:2:0 chroma subsampling, and temporally cropped to 10 seconds. The source videos are then encoded at 90 bitrate levels at each of the following spatial resolutions 1920×1080, 1280×720,

720×480, 512×384, 384×288, and 320×240 according to the list of Netflix certified devices [2]. We evaluate the quality of each video representation using SSIMPLUS [26] due to its demonstrated effectiveness in perceptual video quality prediction [19]. In the end, we obtained 1000 generalized RD functions. The dataset is further segregated into three non-overlapping subsets: 490 for training, 210 for validation, and 300 for testing.

Competing Models

Our evaluation framework pits the proposed method against two predominant approaches in the realm of RD function reconstruction: sampling-based and feature-based methods. Within the sampling-based category, we benchmark against methods such as the piece-wise cubic Hermite interpolating polynomial (PCHIP), reciprocal regression [16], and logarithmic regression [15]. To boost the performance of these sampling-based models, our experiment utilizes an information-theoretic approach to sampling [18], designed to produce a sequence of samples that strategically reduces the uncertainty inherent in the RD function.

In the arena of feature-based approaches, our comparison extends to methodologies like SI/TI [6], and the Video Complexity Analyzer (VCA) [7]. For these methods, we employ an array of multi-layer perceptrons, honing them through the gradient descent algorithm on a designated training dataset. The optimal architecture is then chosen based on its superior performance against a set of pre-determined criteria on the validation dataset. This meticulous training and selection process ensures that our feature-based approach is finely tuned for accurately modelling the RD function.

Implementation Details

In our study, we selected the CLIP model [12] for its demonstrated robustness, efficiency, and adaptability to serve as the LMM. However, it is noteworthy that other LMMs could also be integrated into this framework. We specifically employ CLIP's vision component to distil semantic features from test videos. These extracted features, when derived from identical segments, are subjected to average pooling to consolidate them into segment-level features. Consistent with the guidelines provided in [12], we employ cosine similarity as our metric for quantifying the likeness between feature sets. For the KNN classifier, we have determined that setting K to 3 yields the most favourable outcomes, as evidenced by the enhanced performance metrics observed within our validation dataset.

Evaluation Criteria

Our assessment of the RD function models encompasses two critical performance dimensions: the accuracy of the prediction and the bitrate savings achieved within a per-title encoding optimization framework. To gauge prediction accuracy, we calculate the Mean Absolute Error (MAE) by comparing the model-estimated RD functions against the ground truth for each piece of source content. Regarding bitrate savings, we utilize the predicted RD functions to inform and guide the per-title optimization process. This involves constructing an actual RD function reflective of the predicted convex hull at various bitrates and then calculating the Bjøntegaard Delta rate (BD-rate) [26] to quantify the bitrate efficiency relative to Apple's established bitrate ladder [3]. The performance is averaged across all content in the test set. The experimental procedure is repeated for 50 times with different training/validation split, and we report the median performance.

Performance in Prediction Accuracy

Table 1 details the performance of various competing models in predicting rate-distortion functions, as measured by the Mean Absolute Error (MAE). This measure quantifies the

average magnitude of errors in the predictions, with a lower MAE indicating higher predictive accuracy.

Sample #	0	18	24
PCHIP	N.A.	7.81 ± 0.05	5.12 ± 0.03
Reciprocal	N.A.	9.67 ± 0.08	5.83 ± 0.07
Logarithmic	N.A.	3.61 ± 0.04	**2.24 ± 0.04**
SI/TI	2.49 ± 0.05	2.49 ± 0.05	2.49 ± 0.05
VCA	2.56 ± 0.04	2.56 ± 0.04	2.56 ± 0.04
Proposed	**2.32 ± 0.07**	**2.32 ± 0.07**	2.32 ± 0.07

Table 1. Performance of Competing Models with Different Number of Samples on Rate-Distortion Functions in Terms of Mean Absolute Error

The competing models are evaluated with different number of samples in the rate-distortion space: 0, 18, and 24. Two key observations can be made from the table:

- The proposed model exhibits superior performance over all regression-based counterparts, underscoring the enhanced predictive power of LMM features. Most importantly, the improvement in performance presented by the proposed model is statistically significant.
- Remarkably, the proposed model achieves the best performance among all competing models even with 18 quality analysed encoding samples are available to the sampling-based models. At the same time, our model maintains parity with the best-performing sampling-based method, the Logarithmic model, even when 24 additional RD samples.

Overall, the table underscores the superiority of the proposed model in terms of prediction accuracy for RD functions, which is a pivotal aspect of optimizing video encoding processes.

Performance in Bitrate Saving

Table 2 provides a detailed comparative analysis of RD models in terms of bitrate saving in the context of per-title optimization. Alongside the competing models—PCHIP, Reciprocal, Logarithmic, SI/TI, VCA, and the proposed method—the table also introduces the 'Offline Optimal' result. This result represents an ideal scenario where each rate-distortion function in the dataset is known in advance, serving as a benchmark for the utmost bitrate saving achievable.

Sample #	0	18	24
PCHIP	N.A.	15.1% ± 0.04%	16.3% ± 0.02%
Reciprocal	N.A.	15.3% ± 0.04%	15.9% ± 0.04%
Logarithmic	N.A.	**18.6% ± 0.04%**	**20.2% ± 0.04%**
SI/TI	13.5% ± 0.06%	13.5% ± 0.06%	13.5% ± 0.06%
VCA	17.2% ± 0.05%	17.2% ± 0.05%	17.2% ± 0.05%
Proposed	**18.6% ± 0.07%**	**18.6% ± 0.07%**	18.6% ± 0.07%
Offline Optimal	28.4%	28.4%	28.4%

Table 2. Performance of Competing Models with Different Number of Samples on Rate-Distortion Functions in Terms of Bitrate Saving

The results align with the patterns identified in the previous table, confirming the general

trend observed earlier. Two key insights emerge from the analysis of the table. Firstly, the proposed method, even with zero RD samples, matches the bitrate saving performance of the state-of-the-art sampling-based algorithmic model at a sampling size of 18. This indicates that the proposed model, when integrated into a per-title optimization system, can drastically reduce computational demands while attaining equivalent levels of bitrate saving. Such efficiency suggests that the LMM method leverages its predictive capabilities to streamline the optimization process without compromising on performance outcomes.

Secondly, the proposed method demonstrates a statistically significant improvement over all regression-based models in terms of bitrate saving. This enhancement not only underscores the robustness and effectiveness of the proposed method but also highlights its superiority in optimizing video encoding parameters.

Computation Complexity

We evaluate the processing efficiency of various competing algorithms by examining their computational complexity. This is quantified by the average computation time in the context of per-title optimization. In the case of sampling-based approaches, the computation time encompasses the encoding at three different bitrate levels for each resolution, objective quality assessment tasks, and the fitting of RD curves. Conversely, for the regression-based method, the computation involves the feature extraction, and the estimation of the RD function. The assessment is conducted using 300 10-second 1080p videos as the source material, with an Amazon EC2 g5.2xlarge instance serving as the platform for benchmarking.

Models	Computation Time (s)
PCHIP	84.003 ± 1.88
Reciprocal	83.512 ± 1.83
Logarithmic	83.027 ± 1.81
SI/TI	4.613 ± 0.50
VCA	1.126 ± 0.41
Proposed	6.047 ± 0.72

Table 3. Computation Time in Seconds

Table 3 presents the computation times for various competing models, measured in seconds, and includes a margin of error for each measurement. From the data, several observations stand out:

- Feature-based approaches demonstrate significantly higher speed compared to sampling-based methods, with the VCA model being the quickest among them. This distinction underscores the efficiency of feature-based models in processing video content.
- The proposed method showcases the capability to operate in real-time, as indicated by its computation time. This attribute makes it a viable option for applications requiring immediate video processing and encoding decisions.
- When juxtaposed with the results from the previous table, it is evident that the proposed method not only matches the Logarithmic model in terms of bitrate saving but also drastically reduces the computation time by approximately 13 times. This efficiency gain highlights the proposed method's advantage in offering substantial bitrate savings with significantly lower computational demand.

Overall, the table illustrates the computational efficiency of the proposed method compared to traditional sampling and feature-based approaches, establishing its potential for real-time applications and substantial computational savings without compromising on performance.

DISCUSSION

Advantage of LMM

The integration of LMM into per-title encoding optimization presents a transformative approach to video processing, offering substantial advantages over traditional methods. At the heart of its benefit is the LMM's unparalleled ability to analyze and interpret complex video content at a granular level. Unlike conventional models that rely on basic features or extensive sampling, LMMs leverage deep learning to understand the nuances of video data, including visual elements, audio cues, and textual context. This comprehensive understanding enables the LMM-based framework to make more accurate predictions about the optimal encoding parameters for each video title. As a result, videos are encoded in a way that maximizes quality and efficiency, tailored to the specific characteristics of the content.

Moreover, the use of LMMs in per-title encoding optimization significantly reduces the computational overhead traditionally associated with video processing. By accurately predicting rate-distortion functions and encoding parameters from a deep, semantic understanding of the content, LMMs eliminate the need for brute-force sampling and testing across multiple bitrate and resolution settings.

Limitations and Challenges

A noteworthy observation is that while the proposed method's savings are impressive, they fall short of the 'Offline Optimal' result, which stands at a significant 28.4%. This gap indicates the scope of potential improvement and the ceiling of performance that could be aimed for in future iterations or enhancements of the model.

Another notable limitation of the current work is its performance relative to state-of-the-art sampling-based approaches, particularly when a high number of RD samples are available. In scenarios where extensive RD sample data can be utilized, sampling-based methods tend to outperform our feature-based approach, capturing nuances in video encoding optimization that our current model may overlook. This gap underscores the need for a more holistic framework that integrates the precision and depth of feature-based approaches, like the one presented here, with the comprehensive data utilization of sampling-based methods. Such a combined approach would ideally leverage the strengths of both methodologies, ensuring that the predictive accuracy and efficiency of the encoding optimization process are maximized across all scenarios.

CONCLUSION

Our work represents a significant step forward in the pursuit of more efficient and adaptive video encoding technologies. By harnessing the power of Large Multi-modal Models, we have not only achieved substantial improvements in encoding efficiency but have also laid the groundwork for future innovations in the field of multimedia processing. As the digital landscape continues to evolve, we are confident that the insights and methodologies presented in this paper will contribute to the development of more sustainable, efficient, and user-centric video content delivery solutions.

REFERENCES

[1] T. Stockhammer. Dynamic adaptive streaming over HTTP: Standards and design principles. In Proceedings of the ACM Multimedia Systems Conference, pages133–144, San Jose, CA, USA, Feb. 2011.

[2] A. Aaron, Z. Li, M. Manohara, D. J. Cock, and D. Ronca. (2015) Per-Title encode optimization. [Online]. Available: https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2.

[3] Apple. (2016) Best practices for creating and deploying HTTP live streaming media for iPhone and iPad. [Online]. Available: http://is.gd/LBOdpz.

[4] G. Michael, T. Christian, H. Hermann, C. Wael, N. Daniel, and B. Stefano. (2013) Combined bitrate suggestions for multirate streaming of industry solutions. [Online]. Available: http: //alicante.itec.aau.at/am1.html.

[5] J. Dahl. (2018) Instant per-title encoding. [Online]. Available: https://www.mux.com/blog/instant-per-title-encoding.

[6] ITU-T P. 910. 1999. Recommendation: Subjective video quality assessment methods for multimedia applications. [Online]. Available: https://www.itu.int/rec/T-REC-P. 910-199909-S.

[7] V. V. Menon, C. Feldmann, H. Amirpour, M. Ghanbari, and C. Timmerer. VCA: Video complexity analyzer. In Proceedings of the ACM Multimedia Systems Conference, New York, NY, USA, pp. 259–264. Aug. 2022.

[8] H. Amirpour, P. T. Rajendran, V. V. Menon, M. Ghanbari, and C. Timmerer. Light-weight video encoding complexity prediction using spatio temporal features. In Proceedings of the IEEE International Workshop on Multimedia Signal Processing, Shanghai, China, pp. 1-6. Sep. 2022.

[9] A. V. Katsenou, M. Afonso, D. Agrafiotis and D. R. Bull. Predicting video rate-distortion curves using textural features. In Picture Coding Symposium, Nuremberg, Germany, pp. 1-5, Dec. 2016.

[10] A. Telili, W. Hamidouche, S. A. Fezza, and L. Morin. Benchmarking learning-based bitrate ladder prediction methods for adaptive video streaming. In Picture Coding Symposium, San Jose, CA, USA, December, pp. 325-329, Dec. 2022.

[11] K. S. Durbha, H. Tmar, C. Stejerean, I. Katsavounidis, and A. C. Bovik. Bitrate ladder construction using visual information fidelity. arXiv preprint arXiv:2312.07780. Dec. 2023.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and L. Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of International Conference on Machine Learning, Virtual Event, pp. 8748-8763. Jun. 2021.

[13] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. VideoBERT: A joint model for video and language representation learning. In Proceedings of International Conference on Computer Vision, Long Beach, CA, USA, pp. 7464-7473. Jun. 2019.

[14] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems. Vancouver, BC, Canada, pp. 13-23. Dec. 2019.

[15] C. Chen, S. Inguva, A. Rankin, and A. Kokaram. A subjective study for the design of multi-resolution ABR video streams with the VP9 codec. In Electronic Imaging, pp.1–5. Nov. 2016.

[16] C. Kreuzberger, B. Rainer, H. Hellwagner, L. Toni, P. Frossard. A comparative study of DASH representation sets using real user characteristics. In Proceedings of International Workshop on Network and Operating Systems Support for Digital Audio and Video, pp. 1-6. May 2016.

[17] Z. Duanmu, W. Liu, Z. Li, and Z. Wang. Modelling generalized rate-distortion functions. IEEE Transactions on Image Processing. vol. 23, no. 29, pp. 7331-7344. Jun. 2020.

[18] Z. Duanmu, W. Liu, Z. Li, K. Ma, and Z. Wang. Characterizing generalized rate-distortion performance of video coding: An eigen analysis approach. IEEE Transactions on Image Processing. vol. 29, pp. 6180-6193. Apr. 2020.

[19] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, Y. Qiao. Unmasked teacher: Towards training-efficient video foundation models. arXiv preprint arXiv:2303.16058. Mar. 2023.

[20] Z. Wang, K. Kuan, M. Ravaut, G. Manek, S. Song, Y. Fang, S. Kim, N. Chen, L. F. D'Haro, L. A. Tuan, H. Zhu. Truly multi-modal youtube-8M video classification with video, audio, and text. arXiv preprint arXiv:1706.05461. Jun. 2017.

[21] S. S. Kalakonda, S. Maheshwari and R. K. Sarvadevabhatla, Action-GPT: Leveraging large-scale language models for improved and generalized action generation. In Proceedings of the IEEE International Conference on Multimedia and Expo, Brisbane, Australia, pp. 31-36. Jul. 2023.

[22] G. Delétang, A. Ruoss, P. A. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L. K. Wenliang, M. Aitchison, L. Orseau, and M. Hutter. Language modeling is compression. arXiv preprint arXiv:2309.10668. Sep. 2023.

[23] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, Z. Wang. A quality-of-experience index for streaming video. IEEE Journal of Selected Topics in Signal Processing, vol.11, no. 1, pp. 154-166. Sep. 2016

[24] S. Wang, A. Rehman, Z. Wang, S. Ma, W. Gao. SSIM-motivated rate-distortion optimization for video coding. IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 4, pp. 516-529. Sep. 2011.

[25] Z. Li, Z. Duanmu, W. Liu, Z. Wang, AVC, HEVC, VP9, AVS2 or AV1? — A comparative study of state-of-the-art video encoders on 4K videos. In Proceedings of the International Conference on Image Analysis and Recognition, Waterloo, ON, Canada, pp. 162-173. Aug. 2019.

[26] A. Rehman, K. Zeng, and Z. Wang, Display device-adapted video Quality-of-Experience assessment. in Proceedings of SPIE, San Francisco, CA, USA, pp. 939406.1-11. Mar. 2015.

[27] G. Bjøntegaard. Calculation of average PSNR differences between RD curves. Tech. Rep. VCEGM-33, ITU-TSG16/Q6, 13th VCEG Meeting. Telenor Satellite Services, Oslo, Norway. Apr. 2001.