



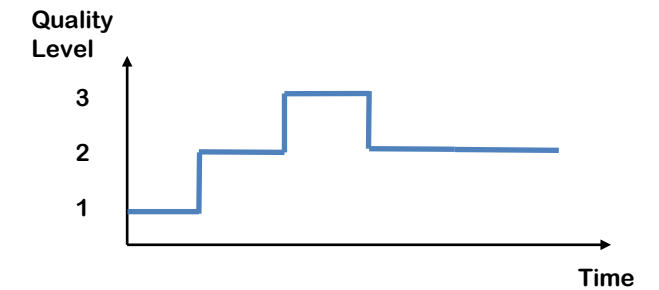
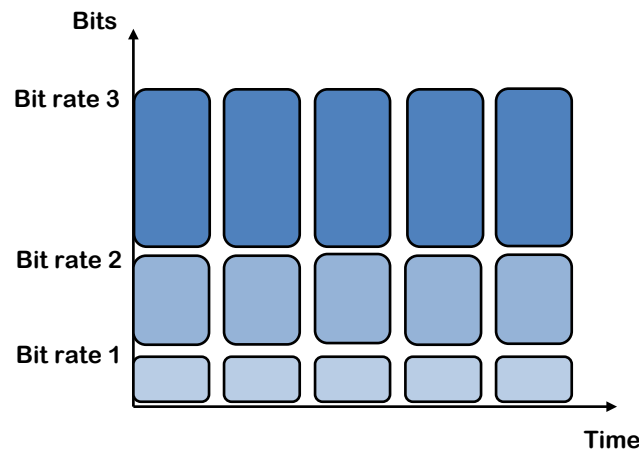
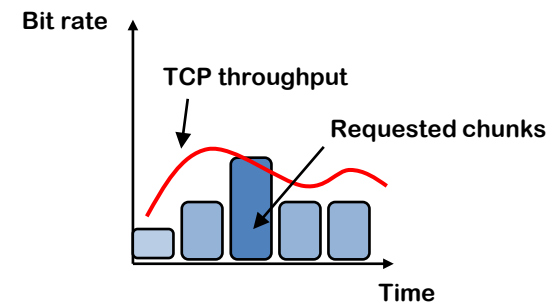
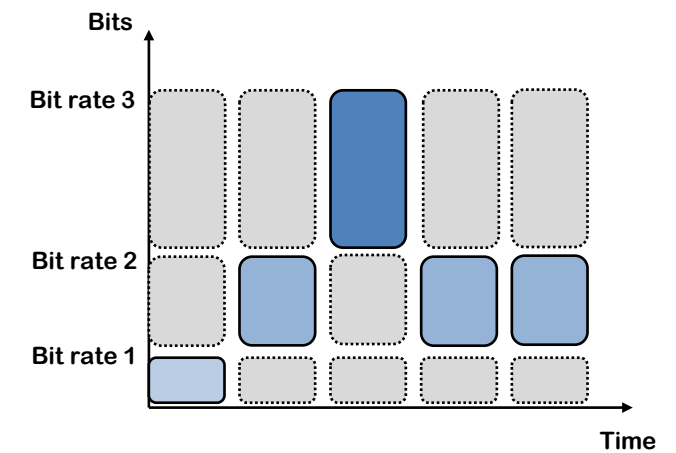
AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE
AGH UNIVERSITY OF KRAKOW

Segment-Level QoE Assessment Dataset For Adaptive Bitrate Video Streaming

Syed G. Uddin

HTTP Adaptive Streaming

- Video is available in several encodings and split into small segments/chunks.
- Client requests small chunks via HTTP
- When one segment is downloaded, the bit rate of the next segment is selected.
- Dimensions for adaptation:
 - Image quality (codec, quantization)
 - Spatial (resolution)
 - Temporal (frame rate, initial delay, stalling)



Motivation

- Importance of QoE in Video Streaming
- Objective measures not enough like Rebuffer, Bitrate, delay
- Objective measures correlation with user-perceived quality
- Subjective opinion is crucial
- Video Quality in Low-latency and mission critical systems

Research Questions

Q1: How does segment duration effect the subjective QoE in DASH streaming?

Q2: What is the association between objective metrics and subjective opinions in DASH based streaming?

Q3: How do ABR Algorithms impact user-percieved quality in low-latency streaming?

Q4: What is trade-off between low-latency and high subjective QoE across bandwidth scenarios?

Hypothesis

H1: There are significant differences exist between HAS and low-latency ABR algorithms subjective QoE

H2: The short-segment durations lead to quality fluctuations impacting subjective QoE

H3: The video quality is influenced by playback smoothness and stability in DASH streaming.

Research Framework

Dataset Generation

- Encoding Process
- Generate DASH
- MP4Box
- Deploy streaming server
- Test case scenarios

Objective Evaluation

- Analyze Bitrate, Rebuffer, Interruption
- VMAF, PSNR, SSIM
- Collect QoE metrics
- Compare ABR algorithms

Subjective Evaluation

- User Evaluation
- ACR/MOS
- Regression ANALYSIS
- ANOVA

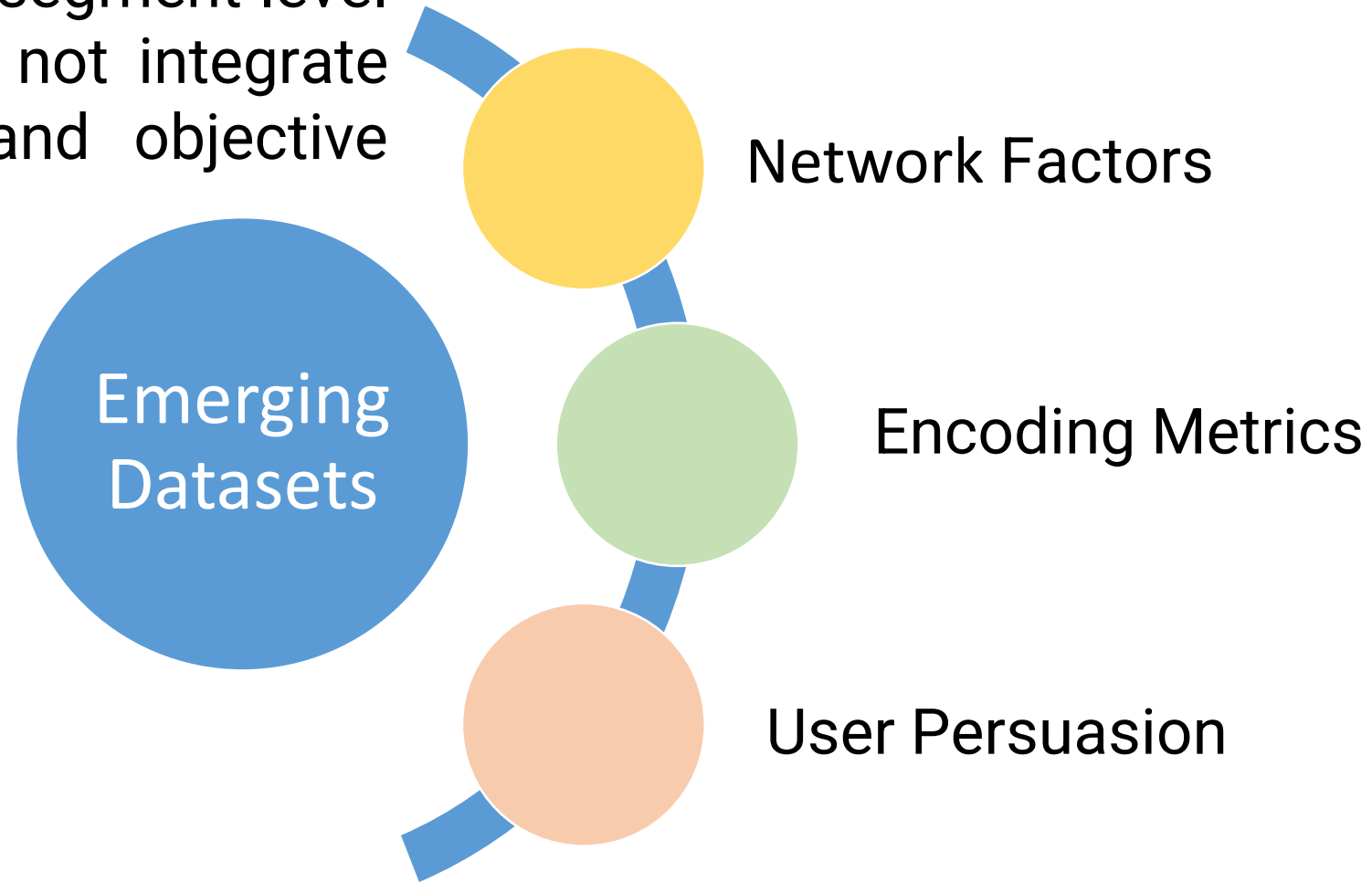
FRAMEWORK IMPLEMENTATION STATUS

Existing Video Datasets

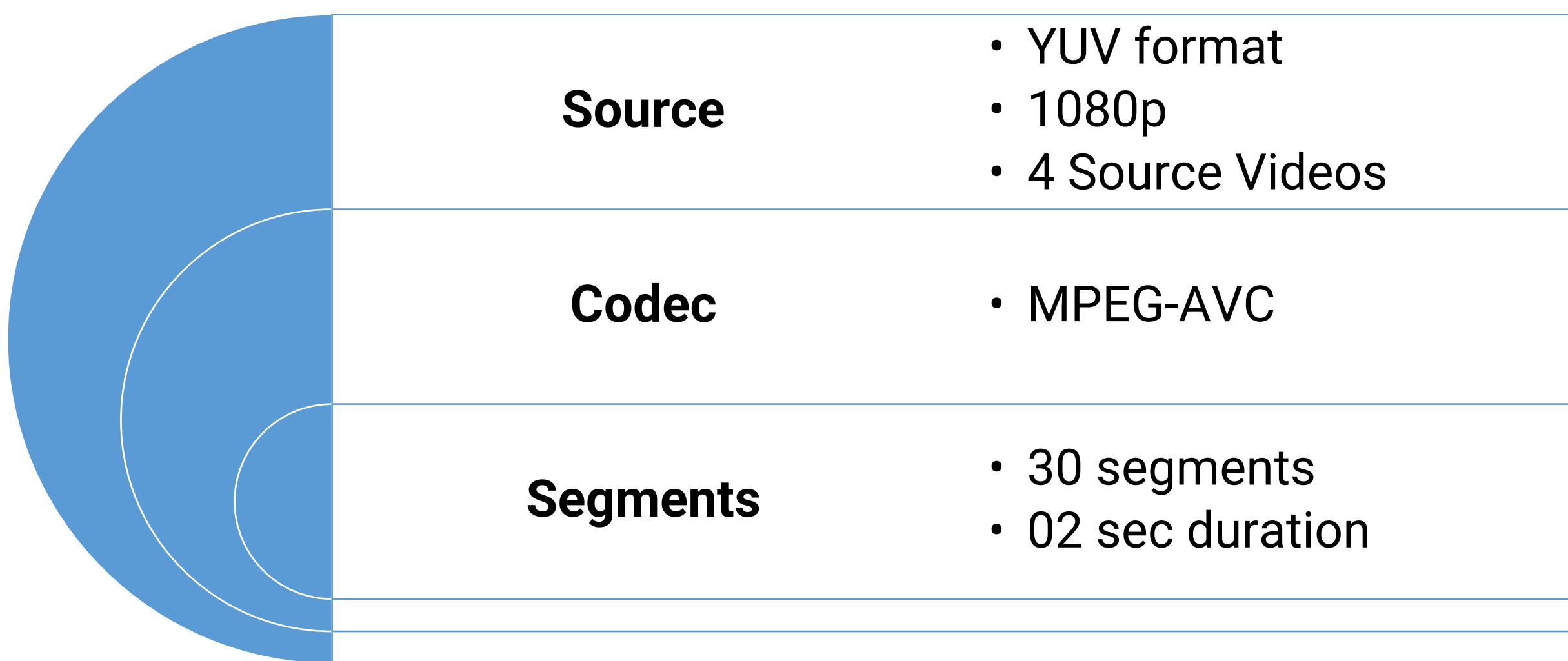
Dataset	Owner	Type of Data	QoE Link	Strengths	Limitations
LIVE Video Quality DB	Univ. of Texas at Austin	Impaired videos + MOS	objective metrics (SSIM, VMAF)	Standard dataset	Not DASH-segmented, no ABR adaptation
Netflix Video Dataset	Netflix	220 videos, multiple resolutions/bitrates	VMAF development, bitrate ladder tuning	Diverse content, high-quality encodings	No subjective QoE ratings
ITU-T P.1203 Dataset	ITU-T Study Group 12	Test videos with stalling/switching + MOS	Validation of ITU-T P.1203 QoE model	Playback impairments (stalling, switches)	Limited content diversity, lab-only setup
Waterloo SQoE-I/II	Univ. of Waterloo	Streaming sessions + subjective ratings	Evaluating ABR algorithms, QoE modeling	Segment-level logs, playback impairments	Moderate dataset size, synthetic conditions
Qualinet Databases	Qualinet (COST Action IC1003)	Curated videos + subjective ratings	Benchmarking video quality/QoE	Large collection, standardized methodology	Not segment-based, heterogeneous dataset
Oboe Dataset	Carnegie Mellon Univ.	Network traces + ABR logs + QoE	ML-based ABR evaluation	Real-world traces, ABR decision logs included	Not focused on content diversity

Highlights

Datasets lack segment-level DASH logs or do not integrate both subjective and objective QoE measures



Dataset

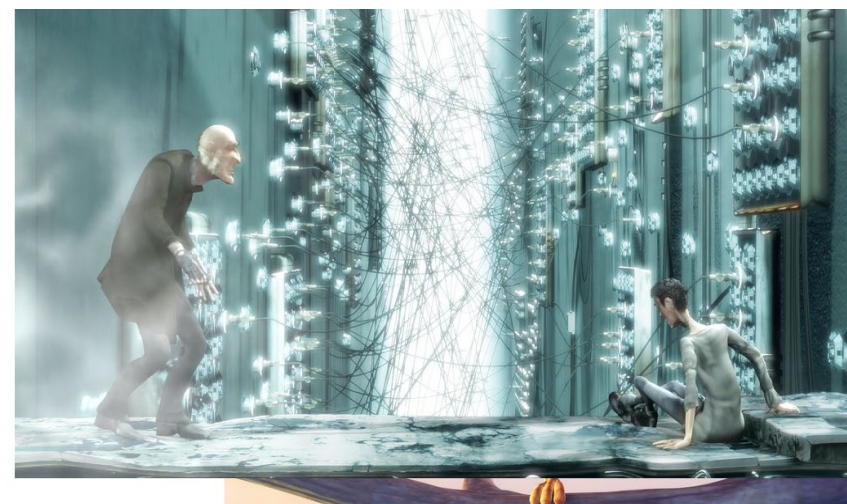


Contents

Big Buck Bunny (1920x1080 @ 24fps)



Elephant Dream (1920x1080 @ 24fps)



Tears of Steel (1920x1080 @ 24fps)

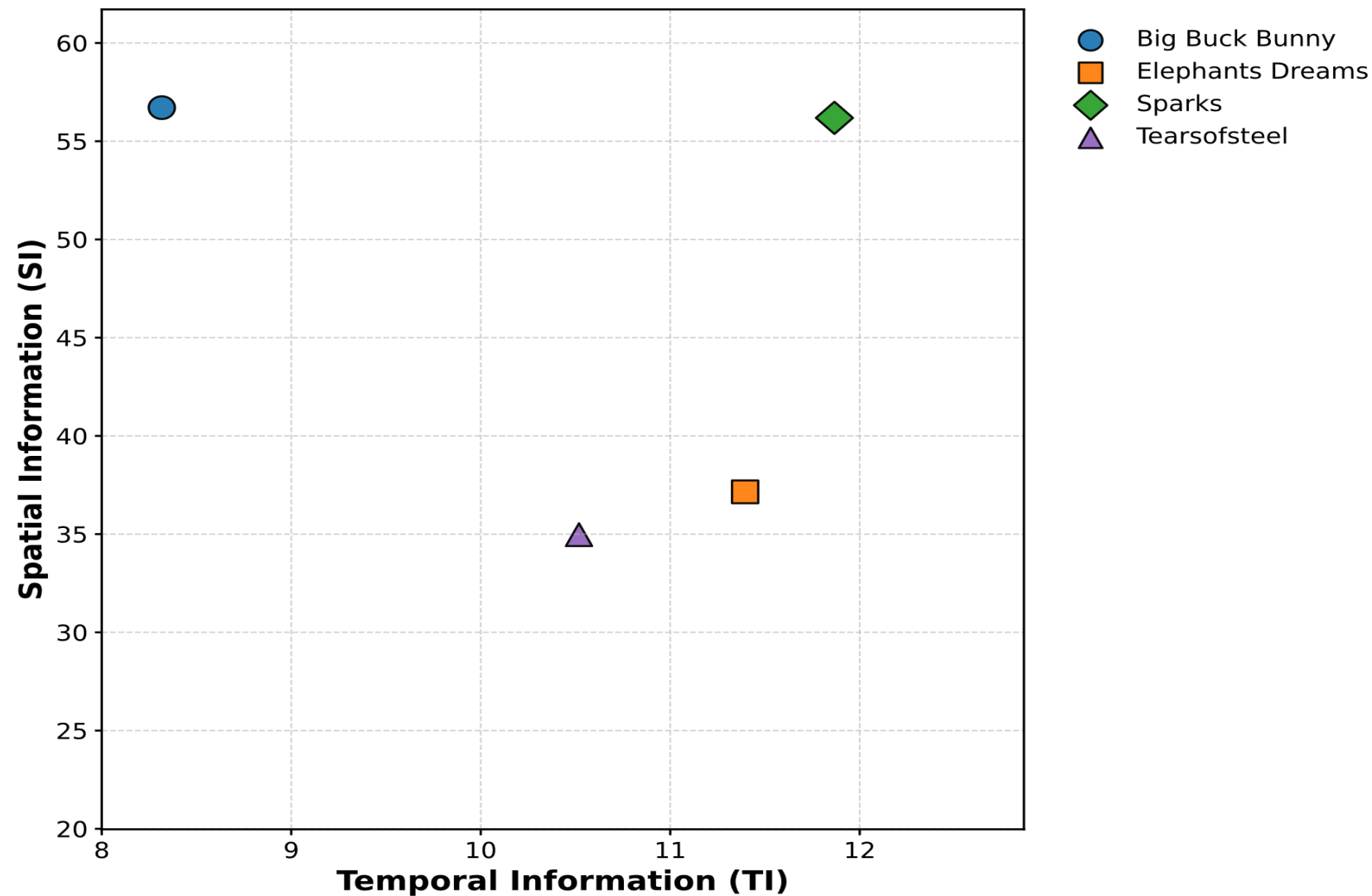


Sparks (1920x1080 @ 24fps)

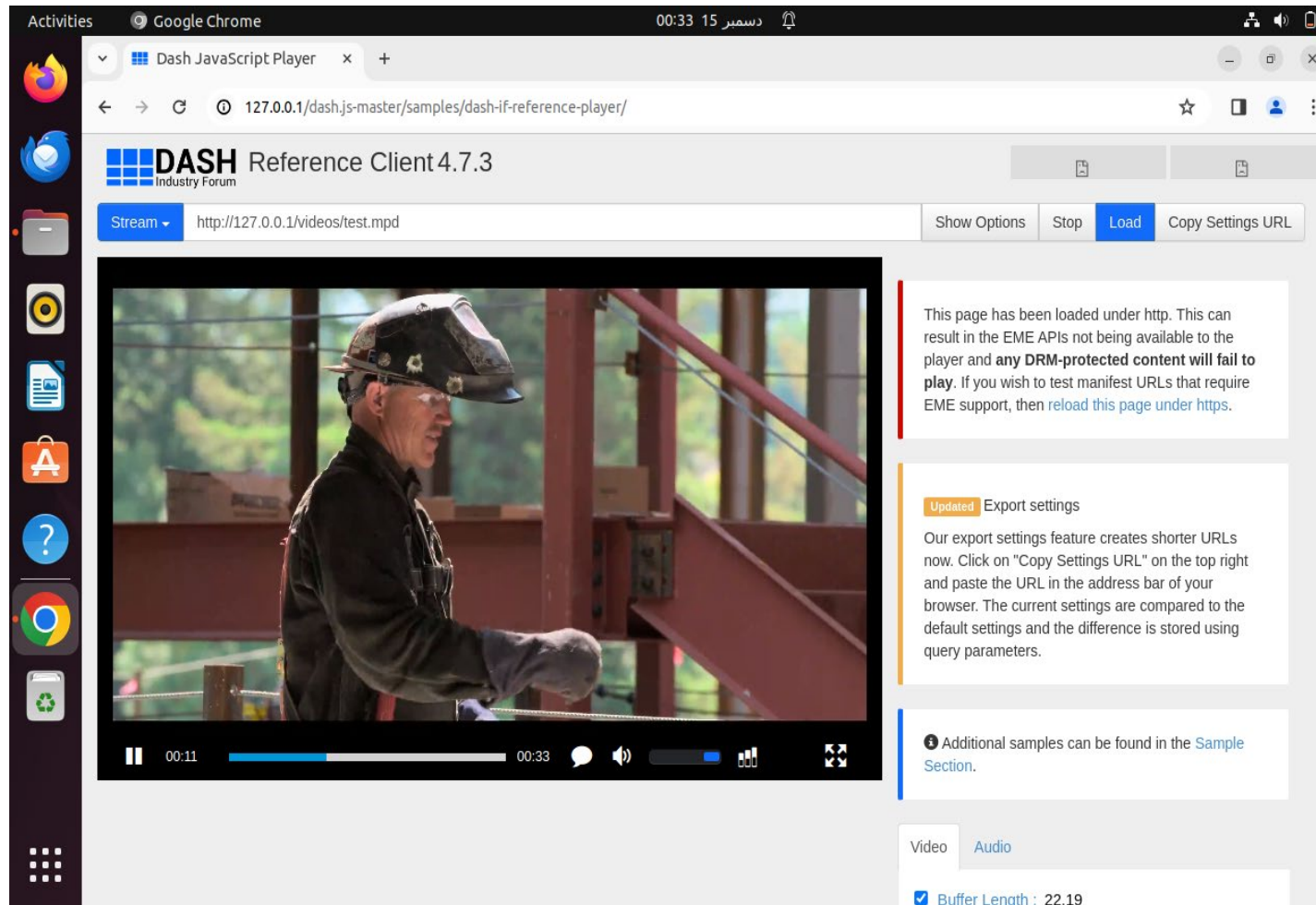


Temporal and Spatial Characteristics

SI vs TI of Videos



Evaluation Framework



- **dash.js**
- Reference client
- Playback of MPEG DASH
- Java script based
- Enables segment level logging: Bitrate, buffer level, stalls, dropped frames.
- Supports live and low-streaming modes (LL-DASH)
- Provide environment for playback testing and real-time adaptation analysis

Adaptive Bitrate Algorithms (ABR)

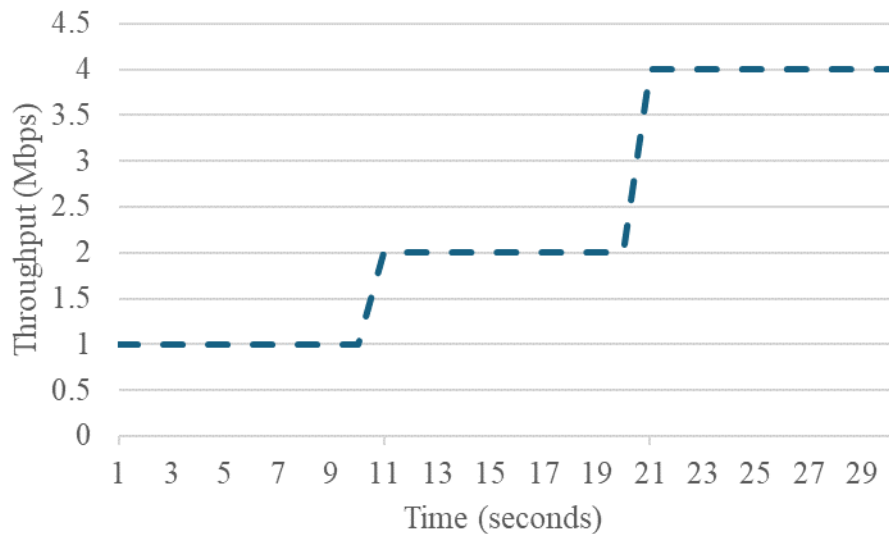
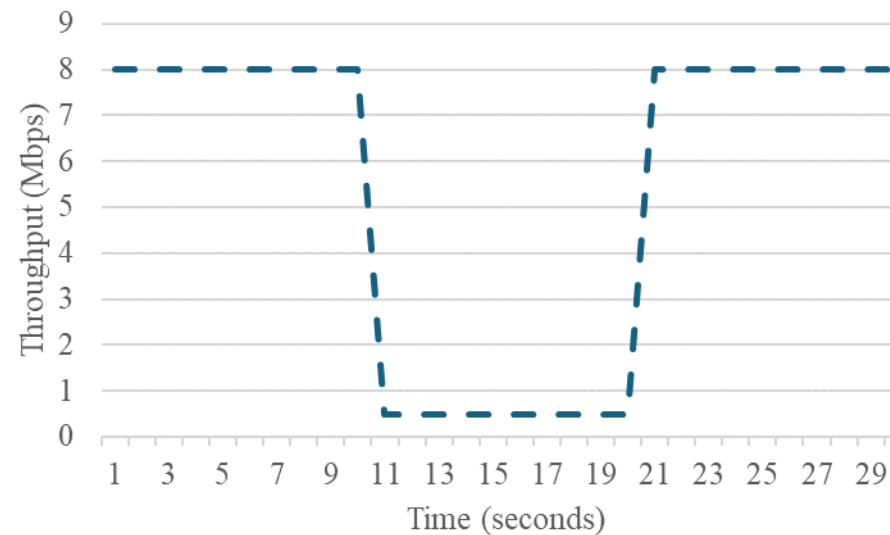
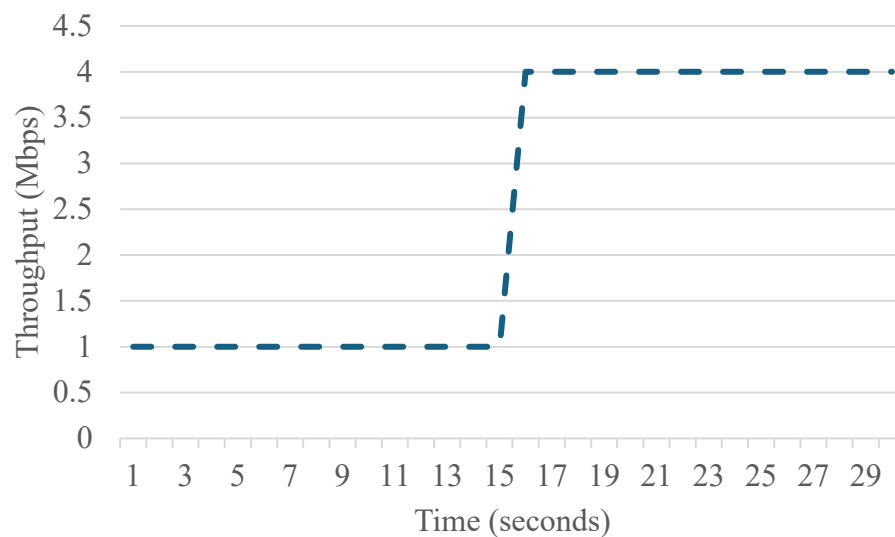
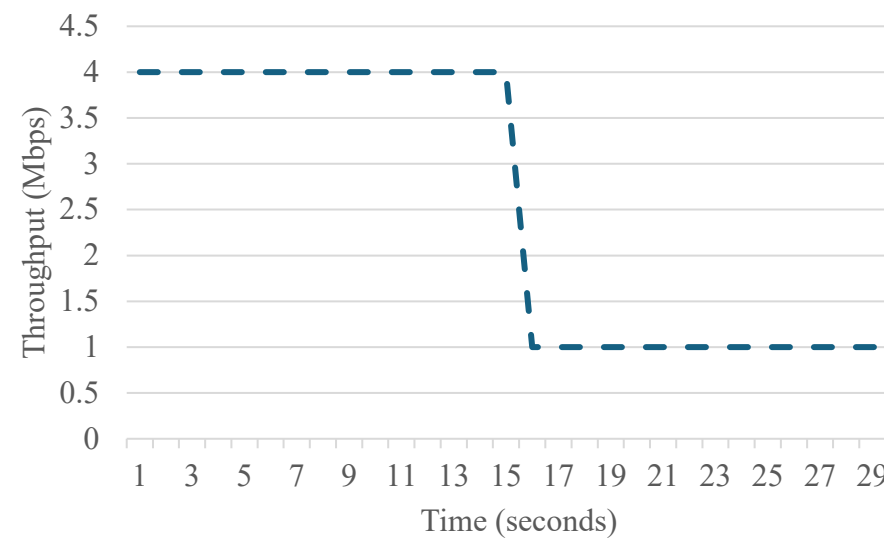
- **Dynamic**
 - A hybrid algorithm which use throughput estimation and buffer levels
 - **Throughput**
 - The algorithm makes a decision based on the throughput of the network.
 - **Buffer Occupancy based Lyapunov Algorithm (BOLA)**
 - Algorithm decides which bitrate to download based on the client buffer.
 - **Learn2Adapt Low Latency (L2A-LL)**
 - Low-latency algorithm works on online convex optimisation principle.
 - **Low on Latency (LOL+)**
 - This is a heuristic based algorithm that uses learning principles to optimise for the best QoE.
- HAS Algorithms
- LL Algorithms

[1] Spiteri, K.; Sitaraman, R.; Sparacio, D. From Theory to Practice: Improving Bitrate Adaptation in the DASH Reference Player. 558 ACMTrans. Multimedia Comput. Commun. Appl. 2019, 15. <https://doi.org/10.1145/3336497>.

[2] Lim, M.; Akcay, M.N.; Bentalab, A.; Begen, A.C.; Zimmermann, R. When they go high, we go low: low-latency live streaming in 598 dash.js with LoL. In Proceedings of the 11th ACM Multimedia Systems Conference; Association for Computing Machinery: New York, NY, USA, 2020; MMSys '20, p. 321–326. <https://doi.org/10.1145/3339825.3397043>.

[3] Bentalab, A.; N. Akcay, M.; Lim, M.; C. Begen, A.; Zimmermann, R. Catching the Moment With LoL+ in Twitch-Like Low-Latency 601 Live Streaming Platforms. IEEE Transactions on Multimedia 2022, 24, 2300–2314. <https://doi.org/10.1109/TMM.2021.3079288>.

Bandwidth Profiles



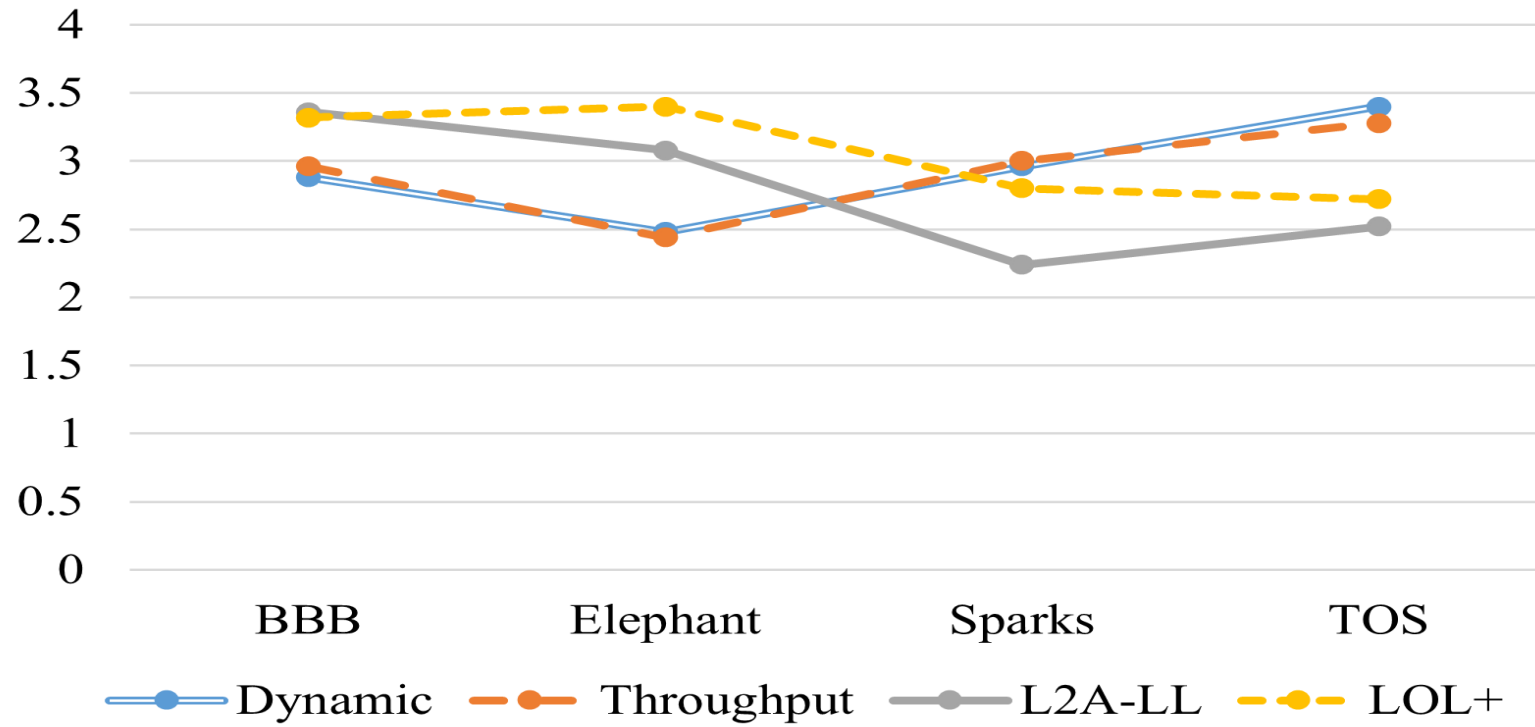
Study Description

- Crowdsource video quality study based on prolific framework
- 60 seconds sequences from BBB, ED, TOS and Sparks in 3 quality levels (resolutions, 854x480, ,1280x720, 1920x1080)
- 5 Adaptive Bitrate (ABR) algorithms selected for evaluation
- 70 participants take part in the crowdsourcing study
- Discarding of unreliable user ratings
 - Test subjects have to watch whole video sequence
 - User have to rate and answer video content question
 - (0.75 Pearson correlation) ITU-R recommendation BT.500-13
- Different quality profiles to identify influence of quality parameters
- 24 test conditions need to be rated by test participants

Methodology

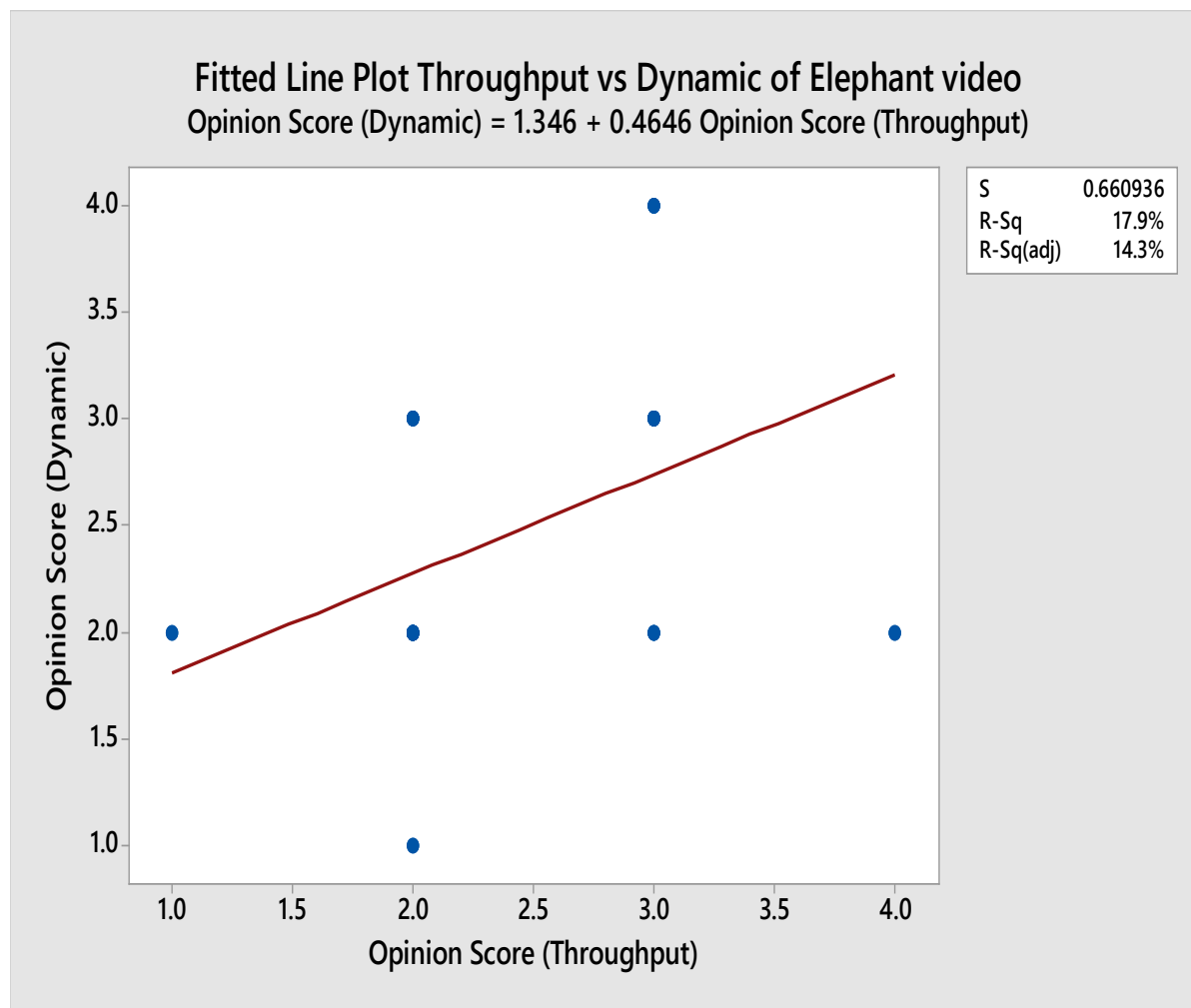
- Absolute Category Rating (ACR) scale
 - Single-stimulus, discrete rating scale (1=Bad, 5= Excellent)
 - Participants watch a video clip once and immediately rate the video
- ITU BT-500
 - One test session completes in 30 minutes
 - Alternative as many different contents as possible
- Details
 - Never the same content in consecutive presentations
 - Randomize to avoid possible effect of content presentation order
 - Added some sequences to stabilize observer's rating

Outcome Analysis



Mean Opinion scores of HAS Algorithms.

Outcome Analysis



Regression Analysis

$$\text{Opinion Score (Dynamic)} = 1.346 + 0.46 \text{ Opinion Score (Throughput)}$$

$$S = 0.66 \quad R - Sq = 17.9 \% \quad R - Sq (adj) = 14.3 \%$$

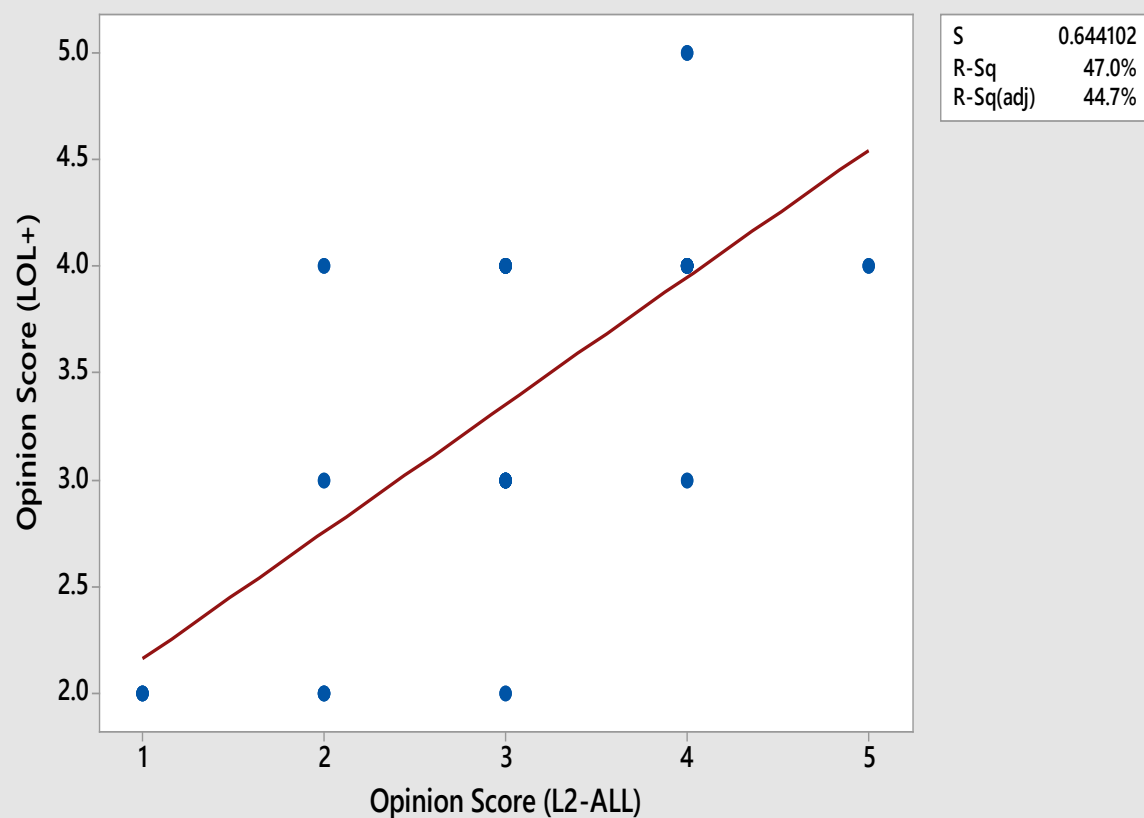
Source	DF	SS	MS	F	P
Regression	1	2.19	2.19	5.02	0.035
Error	23	10.04	0.43		
Total	24	12.24			

$F = 5.02, P < 0.05$ This is a statistically significant regression model.

- $P - \text{Value} = 0.035$ The relationship between Dynamic and Throughput scores is significant
- The positive slope (0.46) shows a moderate positive association

Outcome Analysis

Fitted Line Plot L2-All vs LOL + of Elephant video
Opinion Score (LOL+) = 1.565 + 0.5956 Opinion Score (L2-ALL)



- Regression Analysis

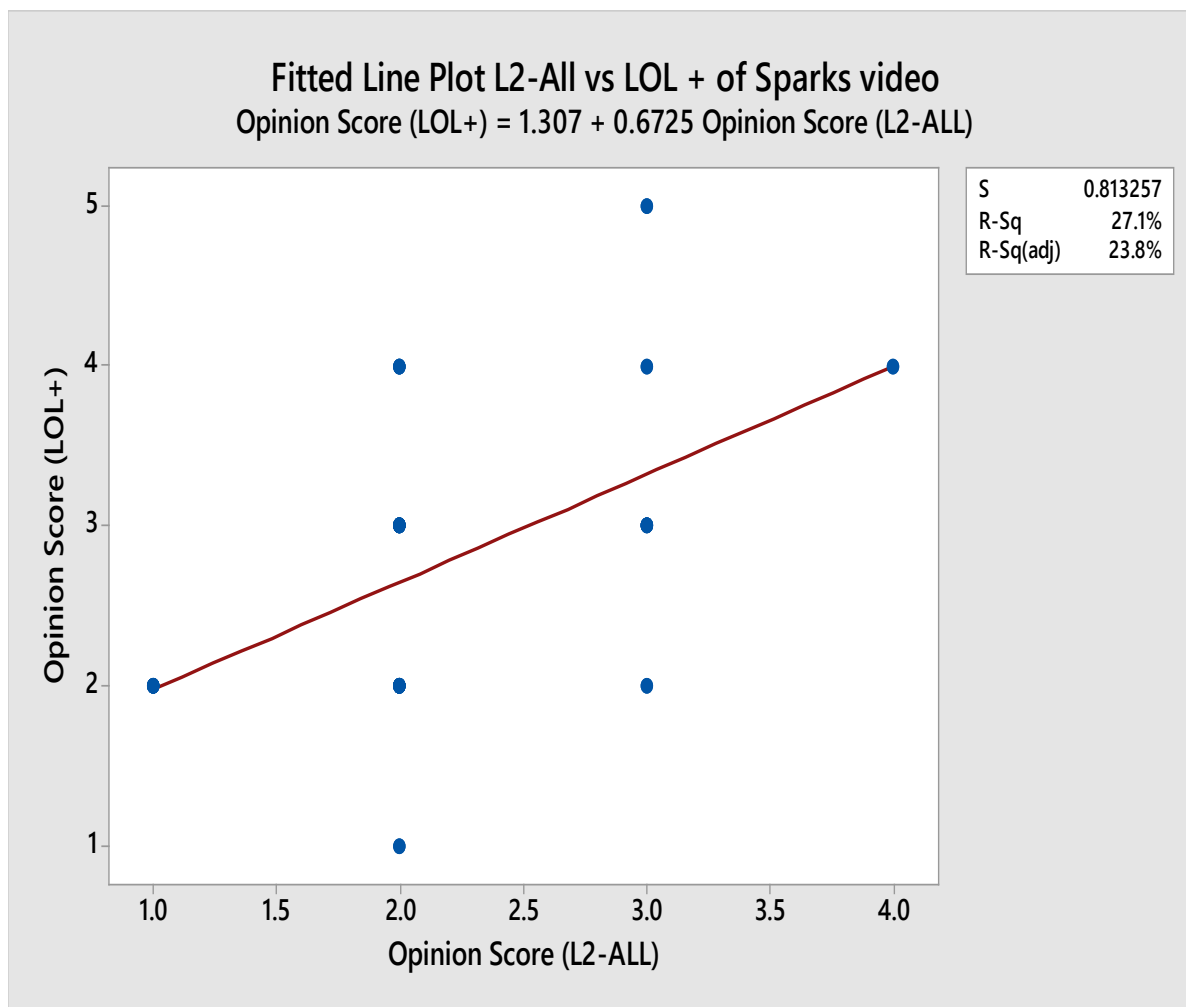
$$\text{Opinion Score (LOL +)} = 1.56 + 0.59 \text{ Opinion Score (L2A - LL)}$$

$$S = 0.64 \quad R - Sq = 47\% \quad R - Sq (adj) = 44.7\%$$

Source	DF	SS	MS	F	P
Regression	1	8.458	8.458	20.39	0.000
Error	23	9.541	0.414		
Total	24	18.00			

- $P - \text{Value} = 0.00$ The relationship between L2-ALL and LOL+ scores is highly significant
- The positive slope (0.59) shows a moderate to strong positive association

Outcome Analysis



Regression Analysis

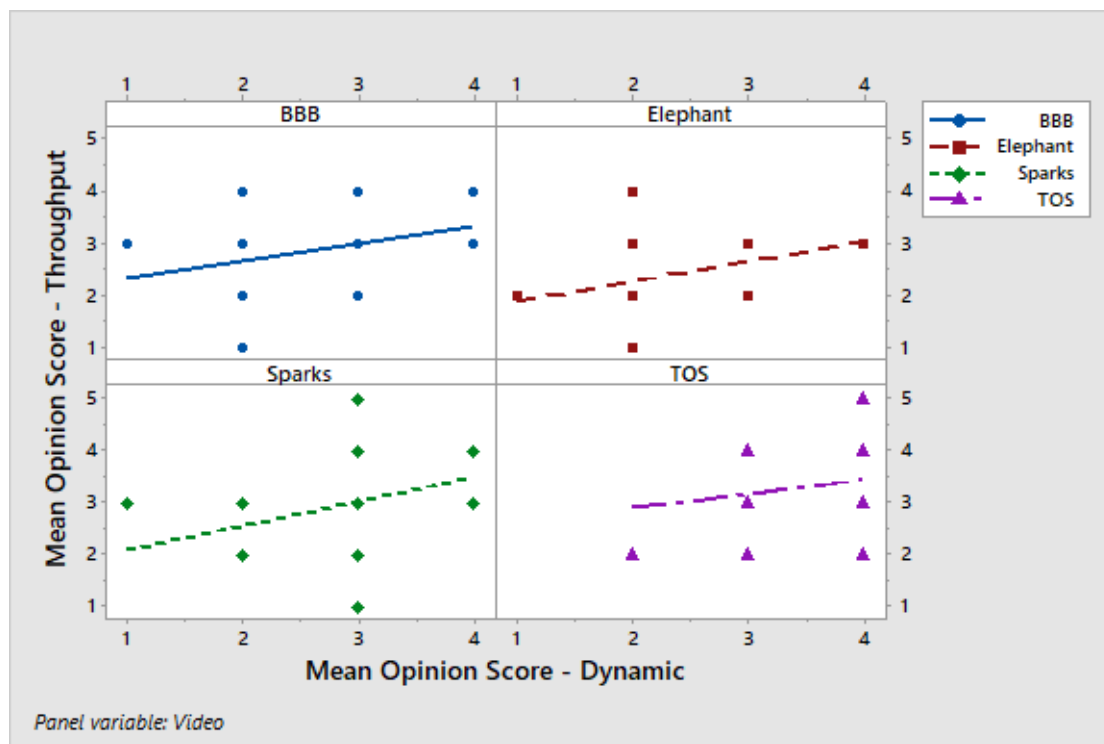
$$\text{Opinion Score (LOL+)} = 1.307 + 0.67 \text{ Opinion Score (L2A - LL)}$$

$$S = 0.81 \quad R - Sq = 27.1 \% \quad R - Sq(adj) = 23.8 \%$$

Source	DF	SS	MS	F	P
Regression	1	5.40	5.40	8.18	0.009
Error	22	14.55	0.66		
Total	23	19.95			

- $P - \text{Value} = 0.009$ The relationship between L2-ALL and LOL+ scores is highly significant
- The positive slope (0.67) shows a moderate to strong positive association

Outcome Analysis



HAS ABR Algorithms Comparison Scatterplot



Low-latency ABR Algorithms Comparison Scatterplot

Conclusion

- Adaptation strategy related parameters of HAS and LL-Algorithms investigated and evaluated using subjective methods performed in a crowdsourcing platform.
- Low-latency ABR algorithms effectively reduce end-to-end delay
- Latency reduction often comes at the cost of perceptual video quality
- Objective metrics alone are insufficient to capture user experience.
- Based on the results QoE model formulated for adaptive streaming in DASH and LL streaming domain.

Findings and Implications

- Low latency often degrades perceptual QoE
- Traditional DASH shows higher stability
- Objective QoE $\neq$ subjective perception
- High-motion content amplifies QoE loss
- Latency and QoE must be jointly optimized
- Stability should guide LL-ABR design
- Subjective evaluation is essential
- Content-aware ABR strategies are needed

Recent Project Showcase

Uddin, S.; Grega, M.; Rahman, W.u.; Leszczuk, M. Crowd-Sourced Subjective Assessment of Adaptive Bitrate Algorithms in Low-Latency MPEG-DASH Streaming. *Applied. Sciences*. **2025**, 15, 13092. <https://doi.org/10.3390/app152413092>