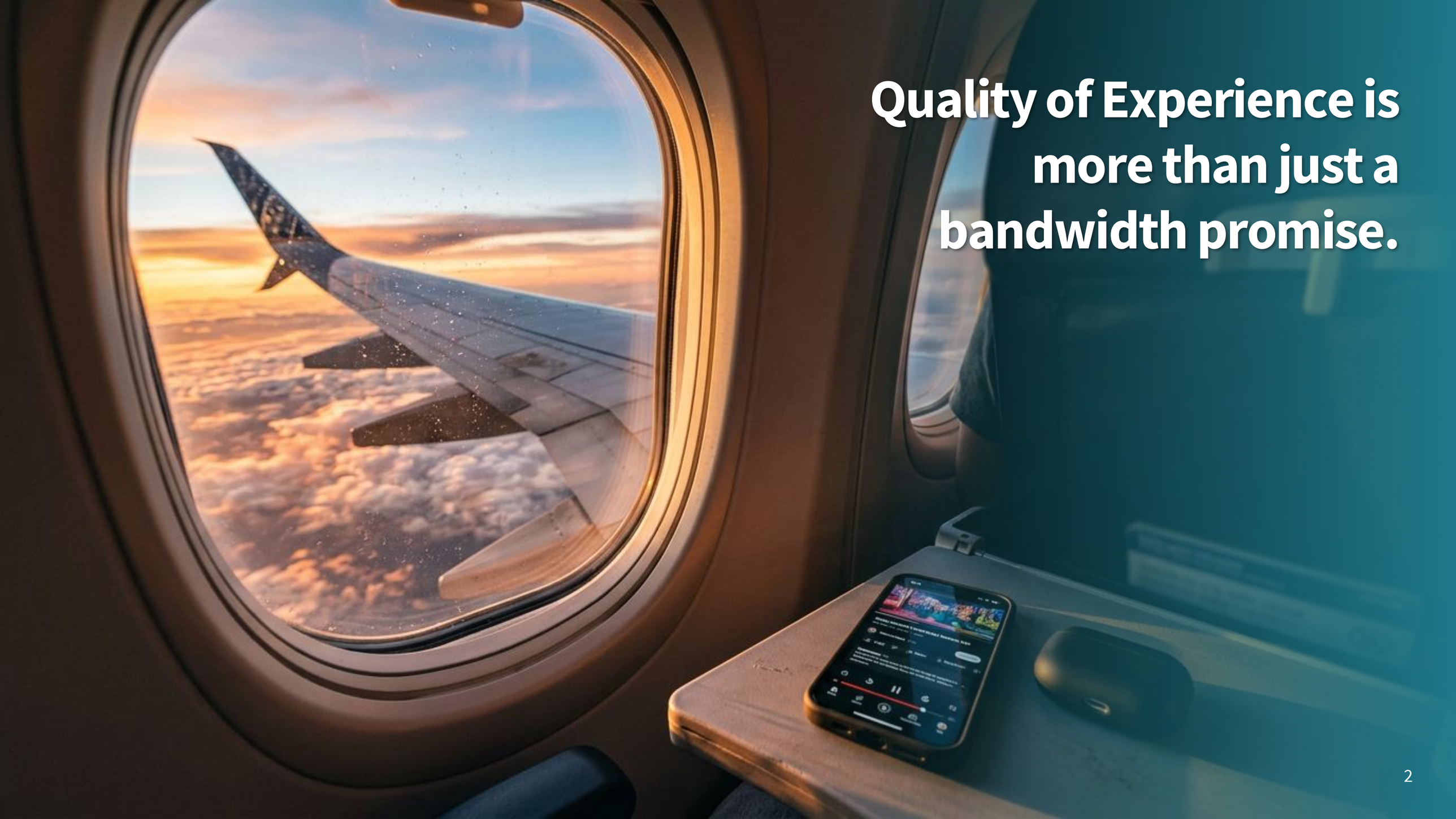
Dr. Werner Robitza — CEO

QoE Models in the Age of Satellite-based Video Streaming

Recent studies and discussion on the future of model developments



Quality of Experience is more than just a bandwidth promise.

Layers of QoS/QoE in Monitoring and Modeling

Network-level KPIs

Packet latencies

Jitter

Packet loss

Throughput

Application-level KPIs

Initial loading delay (video startup time)

Played video bitrate, resolution, …

Stalling position, frequency and duration

Website paint events (first contentful paint, largest contentful paint, …)

DNS resolution times

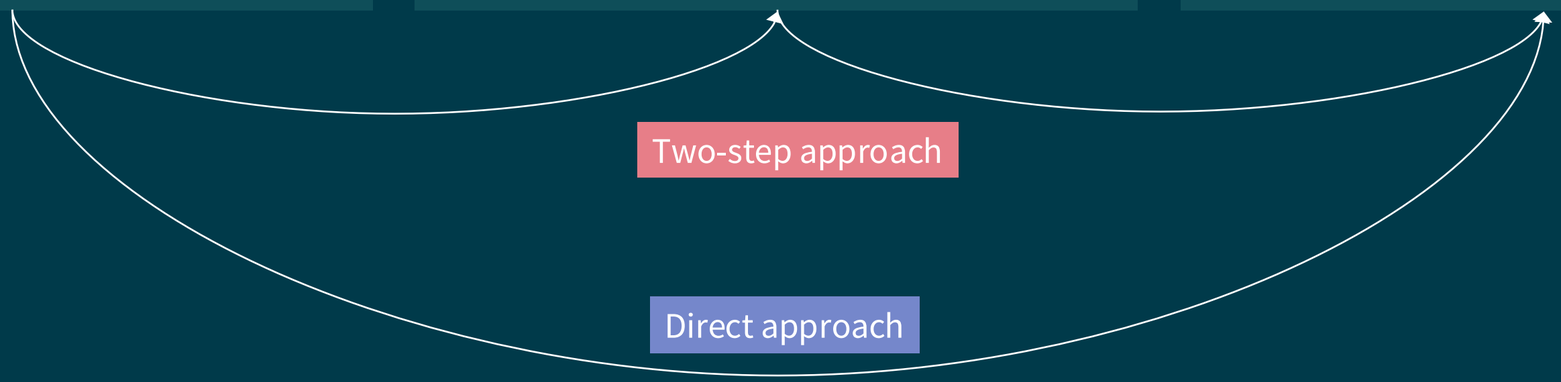
QoE (user-reported or modeled)

Video quality score (per-second)

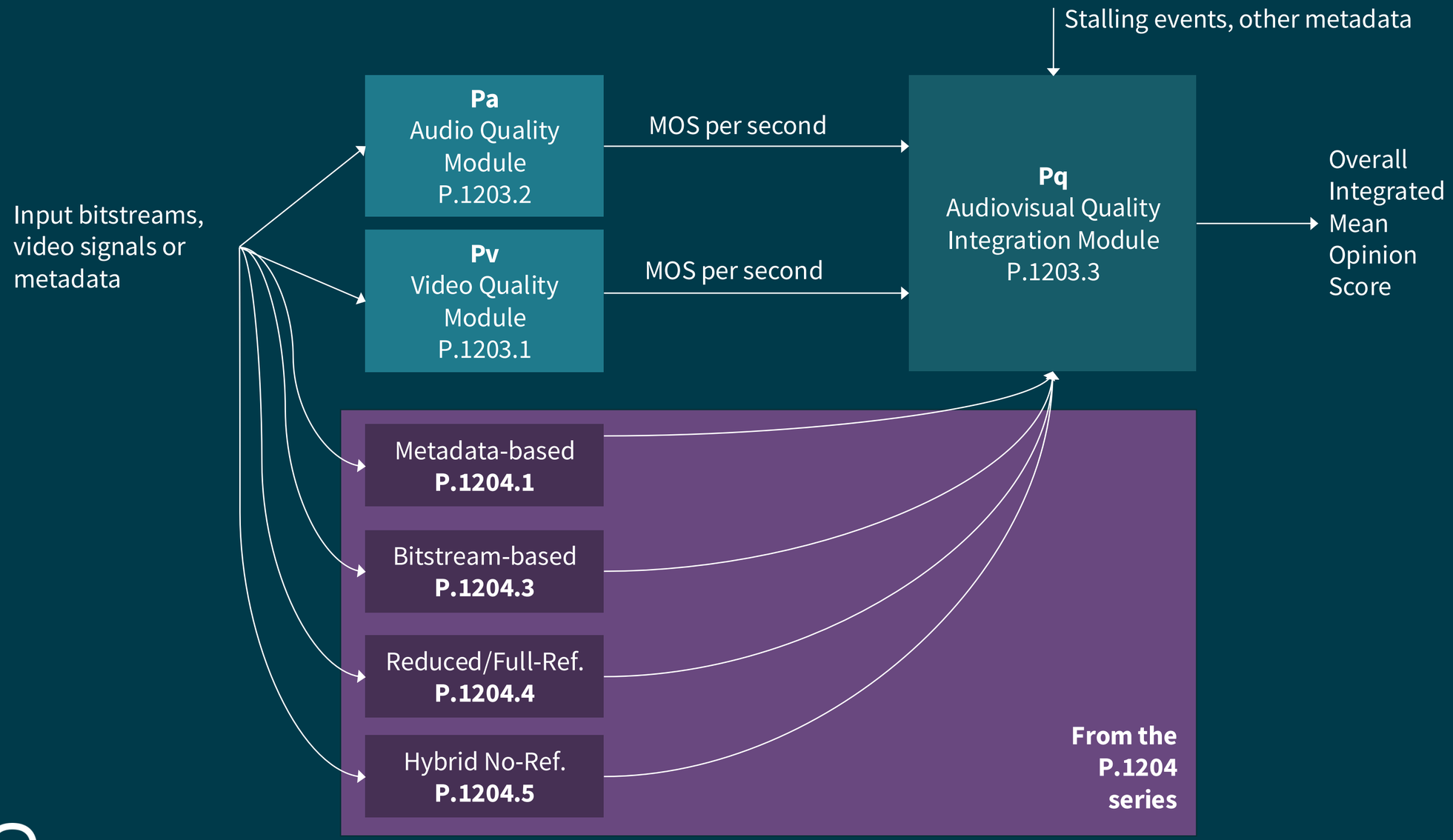
Overall QoE score (e.g. Mean Opinion Score according to a model)

Acceptability/annoyance

(Estimated) abandonment rates



ITU-T Rec. P.1203/P.1204 Model Overview



Performance of P.1204.x Models against PSNR/SSIM/VMAF

Model	RMSE	Pearson	Spearman
PSNR	0.716	0.630	0.615
SSIM	0.648	0.609	0.704
VMAF	0.611	0.761	0.773
P.1204.3 (Bitstream)	**0.422**	**0.899**	**0.883**
P.1204.4 (Reduced-Reference)	0.441	0.889	0.872
P.1204.5 (Hybrid No-Reference)	0.448	0.885	0.880

Based on P.1204 validation databases, see Raake et al., Multi-Model Standard for Bitstream-, Pixel-Based and Hybrid Video Quality Assessment of UHD/4K: ITU-T P.1204". In: IEEE Access 8 (2020)

Validation of ITU-T Rec. P.1203 on Satellite-based Conditions

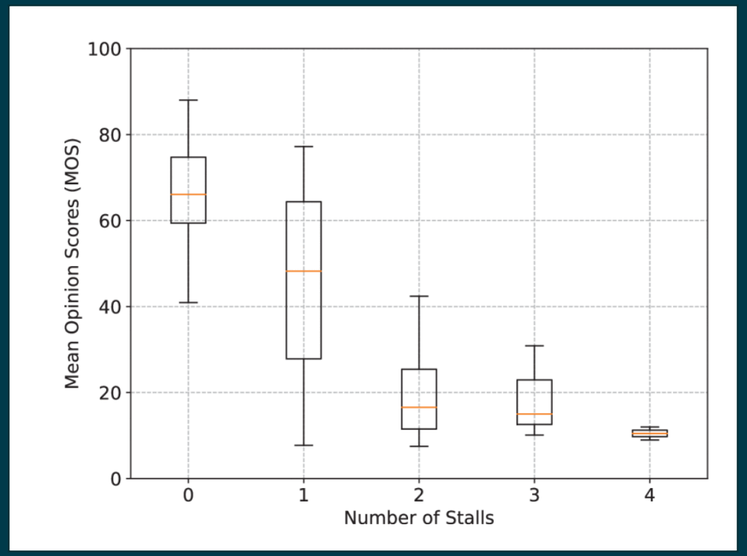
Peer-reviewed study by B. Chen et al., IEEE Access, 2025 — “Satellite Streaming Video QoE Prediction: A Real-World Subjective Database and Network-Level Prediction Models” (UT Austin, Viasat, AVEQ, Ilmenau University of Technology).

Viasat dish on top of UT Austin lab with different network priorities

Videos streamed over real YouTube service under varying conditions

Video recordings + Network conditions + ITU-T Rec. P.1203/P.1204.1 scores

Subjective study correlating QoE models with user-reported scores



Validation of ITU-T Rec. P.1203 on Satellite-based Conditions

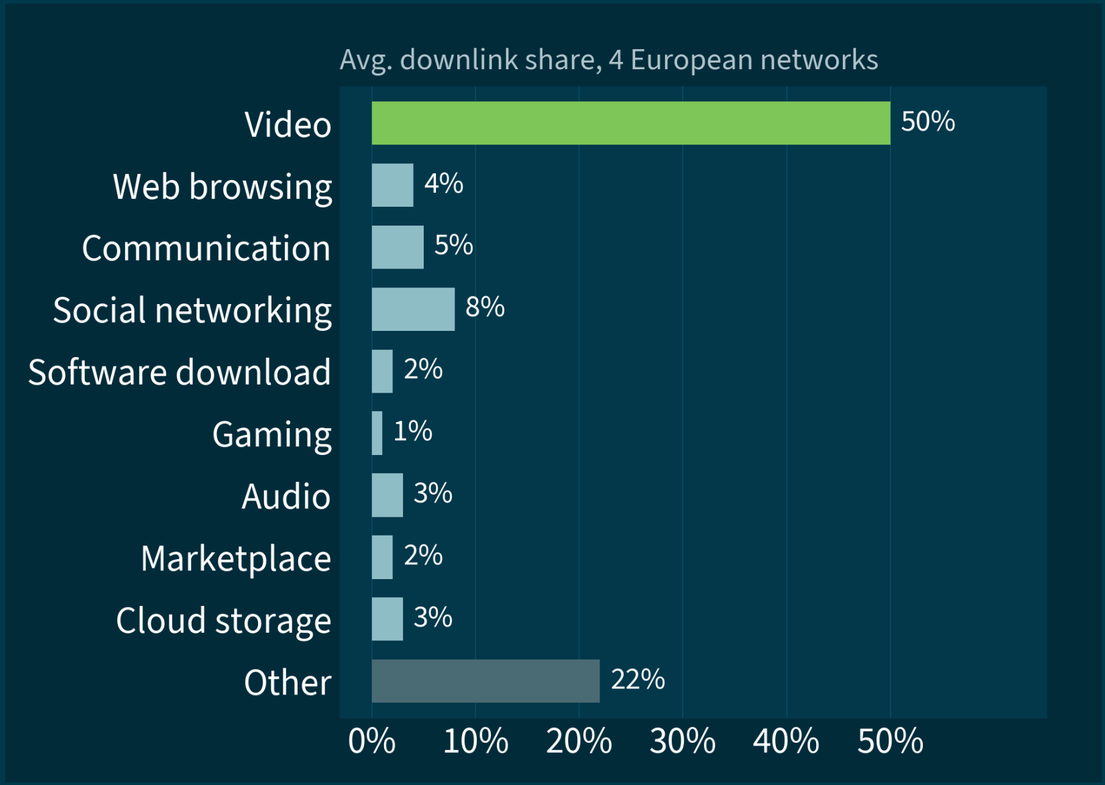
How well do modeled scores match with subjective results?

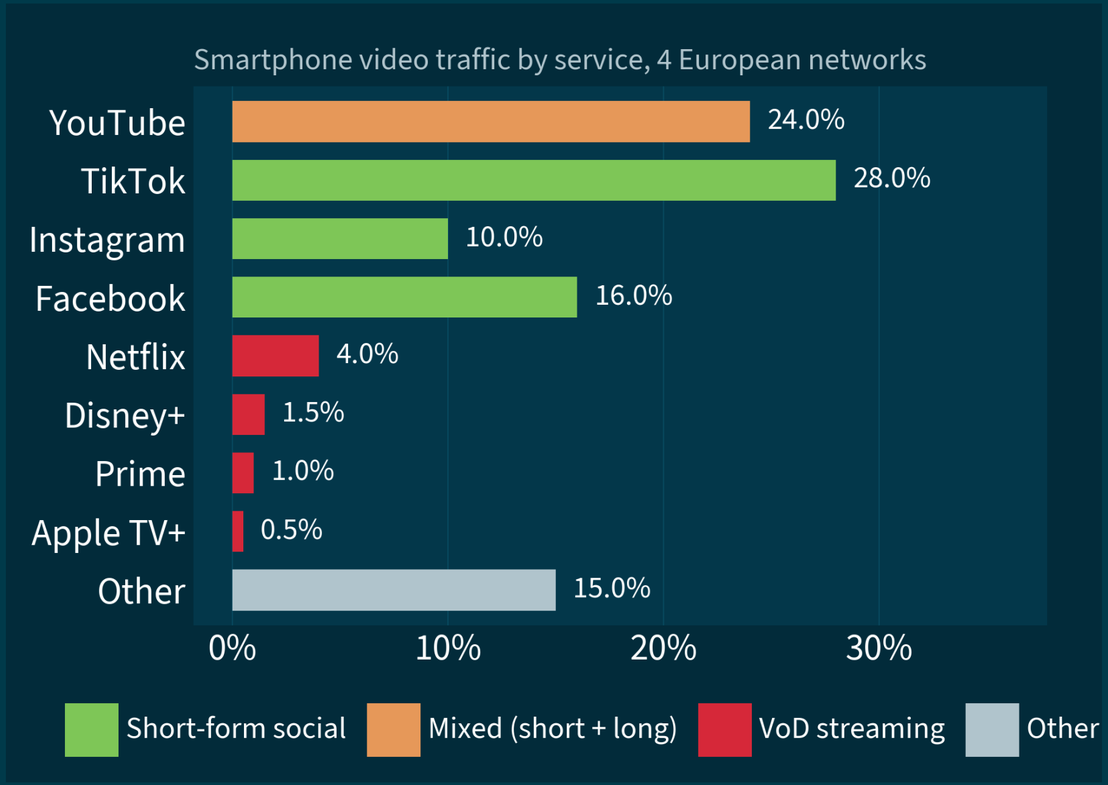
Model	SROCC	PLCC	RMSE
NIQE	0.78	0.80	49.58
BRISQUE	0.67	0.70	21.66
SQI	0.80	0.81	18.92
Video ATLAS	**0.89**	0.90	**11.60**
P.1203 + P.1204.1	0.87	**0.94**	19.15

→ P.1203 model is still accurate — even on previously unseen network conditions

Models are missing context & expectations.

Short Form Video on The Rise





* https://www.ericsson.com/en/reports-and-papers/mobility-report/articles/short-form-dominates-video-traffic

What Works Well — And What Doesn’t

Modularity/Adaptability
Versatile P.1203/P.1204 framework can adapt to new codecs/resolutions/services

Future-Proofness
High correlation with subjective scores established even in previously unseen conditions

Two-step modeling approach is future-proof in light of new network/service developments (satellite networks, transmission over QUIC, …)

Scale of deployment
Broad rollout of active P.1203-style measurement approaches may not be achievable vs crowdsourced collection of network data

Accuracy issues
Direct modeling approaches from network → QoE may not be accurate enough, and services keep changing over time

Model trust
Models in the industry are often proprietary, not validated

Call to Action for Characterization/New Model Developments

Immediate gaps in P.1203/P.1204 series:

1. P.1203.3 model should receive an update to consider **higher user expectations** wrt. shorter video loading times

2. Effects of short form video **interactivity** on QoE may not be reflected properly yet

3. The use of “long form” (1-5min) video QoE models like P.1203.3 for short sessions <30s should be characterized

Questions to ITU-T Experts and the Industry

1. How do we bridge the gap between **operators' needs** for continuous and broadly gathered, actionable QoE data and **regulatory and SLA-related needs** for transparency, accuracy and validity of the results?

2. How can ITU-T standardize the joint collection of acceptability, annoyance, and MOS to support different contexts of use (application use and *reporting* use)?

3. How should user profiles and expectation baselines be captured and task criticality be tested and modeled? Can such models be operationalized?

4. What role can crowd-based testing play in gathering satellite-specific QoE (not QoS!) ground truth at scale?

Thank you.