Metric-Report Reviewer's Checklist

Six questions that turn a bare score back into a claim you can trust. Tick each before you act on a number.

Q1 · What metric, whose tool?

Name the metric and the implementation before reading the number.

- ☐ Metric named (PSNR dB / SSIM 0-1 / VMAF 0-100), not a bare "0.95".
- ☐ Tool and version named (libvmaf, VQMT, FFmpeg) — they differ.
- ☐ PSNR/SSIM compared only within one tool and one version.
- ☐ Remember: 0.95 SSIM is not 0.95 of anything else.

Q2 · Which VMAF model?

The model is part of a VMAF number; omitting it is the usual trick.

- ☐ Model named — default 1080p@3H, phone, or 4K@1.5H.
- ☐ Model matches the device the report claims to describe.
- ☐ A model-less VMAF means ±4-6 pts until you ask which one.
- ☐ Phone model reads several points higher than default.

Q3 · How was it pooled?

One headline number summarises thousands of per-frame scores.

- ☐ Pooling stated — mean, harmonic, median, min, or percentile.
- ☐ A low percentile (p5) or the min shown next to the mean.
- ☐ Mean hides a short disaster: 59 good seconds, 1 terrible.
- ☐ Ask "how bad was the worst stretch?", not just the average.

Q4 · What is the spread?

A VMAF score is a prediction; it carries a confidence interval.

- ☐ A 95% CI or stddev is reported, not just a point score.
- ☐ CI = score ± 1.96 × stddev (e.g. 75.44 ± 2.57).
- ☐ For a comparison, check whether the intervals overlap.
- ☐ A 1-point "win" inside the error bars is noise, not a result.

Q5 · Apples-to-apples?

A comparison is valid only with everything but the test held equal.

- ☐ Same source frames, reference, model, and pooling per column.
- ☐ Same resolution — low-res scored UPSCALED to the reference.
- ☐ A native low-res score is inflated by viewing distance.
- ☐ Check the resolution the metric actually ran at.

Q6 · Content or bad reference?

Every metric is blind on some content; full-reference needs an original.

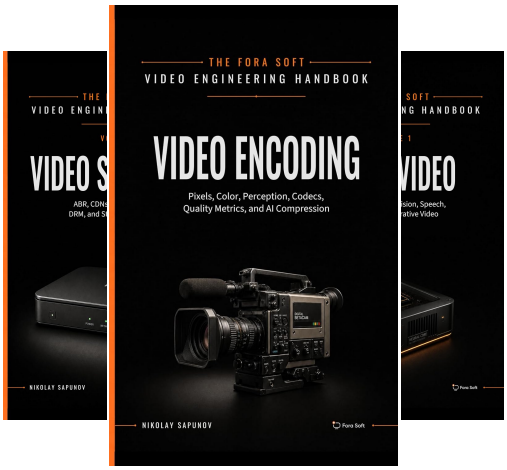
- ☐ Metric applicable? Banding / grain / dark / packet loss fool it.
- ☐ Reference is real, frame-aligned, and correctly color-tagged.
- ☐ Live or UGC has no original — full-reference is impossible.
- ☐ High score on hard content is blindness, not quality.

Worked example — is a one-point win real?

Encoder A: VMAF 93.0 (95% CI 91.5–94.5). Encoder B: VMAF 94.0 (95% CI 92.4–95.6). The intervals overlap almost completely, so the 1-point gap is inside the noise — the honest reading is "no measurable difference at this sample size," not "B wins."

BD-rate is not a quality score. “BD-rate −30%” means 30% fewer bits at equal quality (negative = a win) — and only with the metric, model, content set, and anchor named.

Tip: a video compared with itself scores ~98.7, not 100 — “not 100” is not a defect. Run the free VMAF report linter (vmaf-report-linter.py) to score these six questions automatically.



GO DEEPER · THE 7-VOLUME HANDBOOK

Fora Soft Video Engineering Handbook

Seven volumes, 9,000+ pages: encoding, streaming, audio and AI for video products — from the pixel to production.

By Nikolay Sapunov, CEO at Fora Soft · Kindle & paperback on Amazon

Browse the handbook on Amazon →

This checklist turns a bare quality score into a claim you can check: model, pooling, spread and reference. Our seven-volume handbook covers the stack behind those reports: Volume 1 for metrics and rate control, Volume 3 for QoE in production players, Volume 6 for AI models.

WHAT YOU WILL FIND INSIDE

- VMAF models, pooling and confidence intervals
- Rate control and ladders behind a metric report
- BD-rate and comparing encoders fairly
- QoE monitoring once the stream is live

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI
Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF
Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery
Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing
Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs
Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video
Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Got an encoder or vendor report you need to verify?

We will re-run the comparison on your content with named models, pooling and intervals, and send a plain verdict. 250+ video products shipped since 2005.

Book a 30-min call → **WhatsApp →** **Email us →**