

Native resolution settings and predicted image quality in three hosted video generators

Jeffrey Seders

Independent researcher, Eagle, Colorado, United States

Directed Generative Production (DGP) Methodology

Manuscript draft for journal submission | 27 October 2026

Abstract

Output dimensions are readily observable specifications of generated video, but they do not establish perceptual quality. This exploratory study evaluates a low-cost workflow for comparing native video outputs from three hosted generative tools. Kling 4.0, Wan 3.0 Prime, and Seedance 2.5 were each used to generate a six-second video at a lower available resolution and at 1080p, using the same cowboy reference image and prompt. The prompt requested close-up, medium, and long shots separated by hard cuts. Nine frames per video were scored using the KonIQ-trained MUSIQ implementation in IQA-PyTorch, yielding 54 native-output observations across six generations. Mean scores at the lower setting and 1080p were 76.551 and 73.564 for Kling, 74.309 and 73.251 for Wan, and 70.615 and 64.818 for Seedance. Kling therefore had the highest observed mean at both settings, although the 1080p difference between Kling and Wan was only 0.313 points. Each tool's higher-resolution generation received a lower mean score. These descriptive findings do not identify a causal effect of resolution: only one generation was obtained per condition, seeds were not controlled, and compositions differed. Frame observations are nested within videos and are not independent generation replicates. The native generation expenditure was US\$6.8625; cloud rental for the wider scoring session consumed US\$0.35. The contribution is an auditable practitioner pilot and reporting procedure, rather than a validated model leaderboard or a direct measurement of resolved detail.

Keywords: generated video; image quality assessment; MUSIQ; native resolution; production decisions; exploratory study

1 Introduction

Directors choosing a generative video service face decisions about resolution, cost, framing, and continuity. Output dimensions describe the number of samples in a decoded image. They do not directly measure distinguishable fine structure, natural texture, or perceived quality. The present study asks what a practitioner can establish by applying a documented image-quality estimator to a small set of generated outputs.

The study arose within Directed Generative Production (DGP) Methodology, the author's developing production framework. Here, numerical image-quality assessment is treated separately from character identity consistency, environmental continuity, and compliance with requested actions or camera instructions. No validation of the wider DGP framework is claimed. A frame-quality score is one input to production judgment, rather than a composite measure of directorial success.

The research questions are descriptive. First, how do the three available tool configurations compare at the shared 1080p setting? Second, what differences occur between each tool's lower-resolution and 1080p generations? Third, how do scores vary across the three fixed time intervals corresponding to the requested shot sequence? The study also records expenditure to assess the practicality of independent testing. Upscaled derivatives were collected in the wider experiment but are excluded from the present analysis and reserved for a separate study.

2 Background and measurement scope

MUSIQ is a transformer-based image-quality assessment method designed to accommodate images of varying sizes and aspect ratios through multiscale representations [1]. IQA-PyTorch provides an accessible implementation and inference interface [2]. The present experiment uses its KonIQ-trained checkpoint, not a newly trained model. KonIQ-10k contains

naturally occurring photographic content and human quality ratings [3]. Performance on that source domain does not establish calibration for synthetic video frames.

Generated-video evaluation also includes temporal and semantic dimensions. VBench, for example, separates multiple dimensions rather than treating video quality as a single property [4]. The present study does not run VBench. It deliberately examines one frame-level estimator, allowing its outcome to remain distinct from motion smoothness, flicker, identity consistency, and spatial fidelity.

The operational outcome is therefore MUSIQ-predicted image quality under a specified implementation. Scores are reported in model points, not percentages, visual-acuity units, or independently validated thresholds of acceptability. The study does not interpret score differences as ratios of perceived quality, and does not equate higher scores with more resolved detail. It estimates neither a minimum perceptible difference nor a threshold for production suitability.

3 Materials and methods

3.1 Design and sampling unit

The design comprised three tool labels and two resolution settings per tool, with one generated video per condition. Lower settings were 720p for Kling and 480p for Wan and Seedance. The higher setting was 1080p for all three. The six videos are the experimental units; the nine observations extracted from each are subsamples. Differences between resolution settings compare separately generated videos, not the same latent generation rendered at two resolutions.

The test was conducted on 27 October 2026 through the author’s shared hosted generation interface. Tool names are the labels recorded in that interface: Kling 4.0, Wan 3.0 Prime, and Seedance 2.5. They are not independently verified backend model identifiers. Exact backend builds, sampling seeds, hidden prompt transformations, and any undocumented provider processing were not captured. Accordingly, conclusions attach to the obtained service outputs, not to independently reproduced model weights.

3.2 Reference and requested scene

An AI-generated still depicting a seated cowboy beside a timber building, with a bench, coiled rope, fencing, vegetation, and distant mountains, supplied the common scene reference. The retained reference has dimensions of 1672×941 pixels; it is not a 4K source. Its spatial information and the providers’ treatment of that reference may influence generated detail. The full generation prompt is reproduced in Appendix A.

The prompt requested six seconds at a 16:9 aspect ratio, with an eye-level close-up during seconds 0–2, a medium shot during seconds 2–4, and a long shot during seconds 4–6. The character was to remain seated with subtle breathing and blinking. Two instantaneous hard cuts were requested, with no movement inside shots. Deep focus, consistent daylight, and the absence of intentional grain or motion blur were requested to reduce obvious stylistic sources of score variation. These were instructions, not guarantees that the output met each constraint.

The generation record specifies the same reference and prompt across conditions. However, a complete per-run export of reference-slot assignments, seeds, and service settings was not retained. This limits exact reproduction of the generation process. The frozen output files permit reproduction of the subsequent scoring analysis.

3.3 Media verification

Decoded frame dimensions were checked against the nominal resolution settings. Kling’s lower output was 1280×720 ; Wan’s and Seedance’s lower outputs were 854×480 ; all higher outputs were 1920×1080 . Kling and Seedance outputs were nominally 24 frames per second, while Wan outputs were nominally 30 frames per second. Container durations were slightly greater than six seconds; analysis used only the specified sampling times through 5.5 seconds.

Kling and Wan native files were MP4 containers. Seedance native files were MOV containers. The retained Seedance 1080p file used HEVC, whereas the other five native files used H.264. Scoring was performed on decoded RGB frames. Container type itself is not interpreted as a quality indicator; encoding and service-side export differences remain part of the delivered-output

comparison.

The SHA-256 hashes in the scoring-session record were checked against the retained video bytes, and all matched. Wan filenames containing “1020p” were transcription errors; decoded dimensions confirmed 1920×1080 . Appendix B records hashes for identification independently of those filenames.

3.4 Frame extraction and scoring

For each video, frames were requested at 0.5, 1.0, 1.5, 2.5, 3.0, 3.5, 4.5, 5.0, and 5.5 seconds. The script obtained nominal frame rate from OpenCV and selected the nearest integer frame index using Python `round(time × frame rate)`. The 1.5-second offset in 30 fps videos can entail rounding a half-frame index. The resulting discrepancy is at most approximately half a frame; all requested times and selected indices were logged.

OpenCV decoded each selected frame, converted BGR channels to RGB, and passed it as a floating-point tensor with shape $1 \times 3 \times \text{height} \times \text{width}$ and values divided by 255. No external resizing, crop, sharpening, denoising, or color correction was applied by the scoring script. MUSIQ’s own preprocessing remained active; supplying native-sized tensors does not imply an absence of internal transformations or scale sensitivity.

Inference used `pyiqa.create_metric("musiq", device="cuda")` with the checkpoint `musiq_koniq_ckpt-e95806b9.pth`. The recorded environment comprised IQA-PyTorch 0.1.16, PyTorch 2.8.0+cu128, OpenCV 5.0.0, and an NVIDIA RTX A6000. The script set `torch.manual_seed(0)` before model initialization and used `torch.inference_mode()`. This seed concerns scoring execution, not the video generations. Repeat inference and cross-device numerical stability were not tested.

3.5 Aggregation and analysis

For each video, the overall score is the arithmetic mean of its nine frame scores. Each interval score is the mean of its three sampled frames. Frame-level sample standard deviations use a denominator of eight. These standard deviations describe dispersion across the sampled frames, including differences between intervals; they do not quantify uncertainty across repeated generations.

Differences are calculated as the 1080p generation mean minus the corresponding lower-setting generation mean. Values are computed before rounding. No hypothesis tests, confidence intervals, or claims of statistical significance are supplied because each tool-setting condition has only one generation. Treating nine frames as nine independent trials would create pseudoreplication.

Actual cut times and shot boundaries were not independently annotated for this draft. Results are consequently labeled by fixed intervals rather than asserted to be verified shot-size measurements. Author-reported framing deviations are documented separately and were not used to alter sampling times or exclude outputs. No objective spatial-frequency or optical-resolution test was performed.

4 Results

4.1 Native output comparison

Table 1 gives the native results. Kling had the highest overall mean at both tested settings. At the shared 1080p setting, Kling exceeded Wan by 0.313 points and Seedance by 8.746 points. Wan exceeded Seedance by 8.433 points. These are differences between the observed clips, not estimates of average model differences.

Table 1. Native output frame quality scores

Tool and setting	Dimensions	Mean	Frame SD	Frame range
Kling 720p	1280 × 720	76.551	0.410	76.061–77.207
Kling 1080p	1920 × 1080	73.564	1.259	71.747–74.640
Wan 480p	854 × 480	74.309	0.786	73.331–75.727

Tool and setting	Dimensions	Mean	Frame SD	Frame range
Wan 1080p	1920 × 1080	73.251	1.267	71.254–74.533
Seedance 480p	854 × 480	70.615	0.731	69.642–71.676
Seedance 1080p	1920 × 1080	64.818	5.386	60.147–72.112

Nine sampled frames per video. Frame SD is descriptive within-video dispersion, not uncertainty across generations.

The observed lower-setting ordering was also Kling, Wan, Seedance. However, Kling's 720p output contains more pixels than the competing 480p outputs, so this is a comparison of lowest available settings, not an equal-resolution comparison. The 1080p comparison controls decoded dimensions, but not content, encoding, or hidden service settings.

4.2 Differences between resolution settings

The 1080p-minus-lower differences were −2.987 points for Kling, −1.057 for Wan, and −5.797 for Seedance. Thus, higher nominal native resolution did not produce a higher observed mean in any tool's pair of clips. Because the pairs are separate generations with uncontrolled seeds and differing compositions, the experiment cannot attribute these differences specifically to resolution.

Higher-dimensional outputs can carry more samples without receiving higher scores from a learned quality estimator. The present data do not identify whether the differences reflect generation variability, framing, texture statistics, reference interpretation, encoding, the estimator's scale response, or a combination of these factors.

4.3 Fixed interval results

Table 2 reports scores for the three fixed intervals. The direction of change from the first to final interval was not consistent across outputs. Seedance 1080p increased from 62.027 to 71.854, while Kling 1080p decreased from 74.472 to 71.893. Wan 1080p's final interval exceeded Kling's final interval despite Kling's higher whole-clip mean. Whole-clip order therefore should not be substituted for interval-specific comparisons.

Table 2. Mean scores by fixed time interval

Tool and setting	0–2 s	2–4 s	4–6 s
Kling 720p	76.980	76.074	76.600
Kling 1080p	74.472	74.327	71.893
Wan 480p	75.145	74.089	73.693
Wan 1080p	74.273	71.605	73.875
Seedance 480p	71.231	70.889	69.725
Seedance 1080p	62.027	60.573	71.854

Each cell averages three frames. Requested framing was close-up, medium, then long; actual shot boundaries were not independently annotated.

These interval results do not establish that detail necessarily deteriorates as the camera moves farther from a subject. Requested shot sizes were not independently verified, and interval scores combine content and quality differences. In particular, changes in background coverage can alter a frame-level score without representing a change in the cowboy's resolved facial detail.

4.4 Framing and informal observation

The author reported that Wan's 1080p middle shot changed to a left-angle view before returning to a frontal direction, and that Seedance's 1080p middle shot concentrated on the lower torso. The medium-shot prompt requested framing from the hat to the waist while also including hands resting on the thighs. Those boundaries introduce an ambiguity relevant to the latter deviation.

The common-direction instruction was intended to constrain angular changes.

These deviations were retained as delivered behavior. They were not scored as continuity or compliance failures in the present analysis. Retention supports a practical service-output comparison but reduces control over composition. The author’s overall viewing impression ranked Kling, Wan, and Seedance in the same order as the native-output means. No blinded viewing, standardized display protocol, formal ratings, or independent observers were used, so that agreement is anecdotal and is not treated as perceptual validation.

4.5 Expenditure and feasibility

Generation expenditure totaled 915 credits. At the author’s replacement rate of US0.0075percredit, this equals US6.8625, rounded to US\$6.86. Table 3 gives the per-condition costs. This conversion reflects the author’s discounted credit purchase, not a universal list price.

Table 3. Native generation expenditure

Tool	Lower credits / USD	1080p credits / USD	Total USD
Kling 4.0	72 / \$0.5400	106 / \$0.7950	\$1.3350
Wan 3.0 Prime	41 / \$0.3075	168 / \$1.2600	\$1.5675
Seedance 2.5	120 / \$0.9000	408 / \$3.0600	\$3.9600
Total	233 / \$1.7475	682 / \$5.1150	\$6.8625

Cloud credit consumption for the wider session was US0.35, based on the balance declining from US10.00 to US9.65. That session included setup, upload, model initialization, and scoring of both the six native files and six derivative upscales. Recorded scoring-loop elapsed time was 83.356 seconds across all 12 files; it is not a native-only runtime and excludes initial model loading. Total wider-experiment expenditure was therefore US7.2125, rounded to US\$7.21. Upscale numerical results are outside the present manuscript. Existing subscriptions, labor, local equipment, and internet access are not included in the marginal expenditure.

5 Discussion

5.1 Interpretation for production decisions

The principal empirical finding is narrow: for this reference, prompt, set of generated files, and scoring configuration, native whole-clip means ranked Kling above Wan above Seedance at both tested settings. The small 1080p gap between Kling and Wan should not be described as practically meaningful without replication and perceptual evidence. The larger separation from Seedance also remains specific to these outputs.

The study illustrates why output dimensions and learned quality scores should be reported as different properties. Choosing 1080p did not guarantee a higher score in the obtained files. This is relevant to a director deciding whether additional credits improved the measured outcome in a particular production attempt. It does not justify recommending lower resolution generally, or asserting that higher resolution reduces quality.

The contribution is a documented workflow that a practitioner can repeat cheaply, accompanied by explicit separation of frame-quality prediction from directorial compliance. It is not a new quality metric, an exhaustive benchmark, or evidence that one service dominates another across subjects and tasks.

5.2 Limitations

The strongest limitation is absence of generation replication. Frame sampling cannot recover the missing variance across independent generations. Only a single scene and reference were tested, with a mostly stationary character and little requested motion. Findings cannot be extended to action, camera movement, multiple characters, or other environments.

Resolution is confounded with stochastic generation differences and potentially with provider processing. The reference itself contains fewer pixels than a 1080p output, and no high-resolution reference or synthetic resolution chart served as a detail ground truth. Native codecs and frame rates varied. Hidden backend versions, random seeds, reference-slot records, and generation-order randomization were not captured completely.

Three frames per interval are sparse samples. Brief defects, flicker, and temporal instability may be missed. Unverified cut timing and known misframing weaken claims about distance and shot size. The metric is trained using photographic quality judgments and was not validated here on the specific synthetic domain. It may favor some texture or scale characteristics independently of human preference. No second metric or controlled perceptual evaluation was used. Numerical repeatability of the scoring implementation was not separately assessed.

5.3 Replication priorities and provider reporting

A stronger follow-up would use multiple independently generated clips per tool-setting condition, several scene types, preserved generation metadata, and predetermined sampling rules. Composition could be standardized with verified framing references, and actual cut boundaries annotated before shot-level aggregation. Misframed outputs should remain visible in a separately reported compliance analysis rather than disappearing from the dataset.

Resolution effects could be investigated with a common-size scoring arm and controlled resampling of identical frames, while retaining native-size scoring as a distinct analysis. Additional image-quality estimators would test metric dependence. A separate blinded viewing study could assess perceptual agreement, and temporal metrics would be required for broader video claims. These are proposed extensions, not completed components of the present pilot.

Providers could make independent comparisons more useful by disclosing service and backend versions, generation settings, prompts, reference inputs, multiple outputs per condition, scoring implementations, and measured variation. This paper recommends such reporting; it does not establish that all providers currently omit it.

6 Conclusion

This six-generation pilot demonstrates a low-cost procedure for documenting native-output image-quality scores and generation expenditure. The observed MUSIQ ordering was Kling, Wan, then Seedance, at both the lower available settings and 1080p. Each 1080p clip had a lower mean than its tool's lower-setting clip. These findings describe the collected outputs and do not establish causal resolution effects, general tool superiority, or direct changes in resolved detail. Replicated generation and validated shot-level comparisons are needed before broader conclusions are warranted.

Declarations

Funding. The author paid the recorded generation and cloud-rental costs personally. No manufacturer-funded testing is reported in the experimental record.

Competing interests. The author develops and promotes DGP Methodology and operates jseders.net, which presents related work. This interest should be considered when interpreting the production-oriented discussion. This declaration is subject to final author confirmation before submission.

Author contribution. Jeffrey Seders conceived the production test, operated the generation services, supplied the experimental materials, and made the reported viewing observations. The author must review and approve the final manuscript and is responsible for its claims.

AI assistance. OpenAI ChatGPT assisted with test planning, scoring-code preparation, numerical aggregation, literature retrieval, and manuscript drafting. The scene reference was AI-generated, and all evaluated videos were generated using the named services. AI tools are not listed as authors. The final disclosure should be adapted to the selected journal's policy.

Ethics. No recruited participants or formal human-rating study were used. Informal author observations are identified as such and are not presented as human-subject validation.

Data and code availability. The author retains the reference image, complete prompt, six native files, frame-level scoring CSV, original scoring script, software metadata, and file hashes. The manuscript includes the prompt and identification hashes. A stable public repository record for the complete materials has not yet been assigned; a permanent deposit link must be supplied before submission. No claim of current public availability is made here.

Publication status. Draft prepared for journal submission. Not submitted, peer reviewed, or accepted as of 27 October 2026.

References

- 1] Ke J, Wang Q, Wang Y, Milanfar P, Yang F. MUSIQ: Multi-Scale Image Quality Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:5148–5157. doi:10.1109/ICCV48922.2021.00510. Author manuscript:[<https://arxiv.org/abs/2108.05997> (https://arxiv.org/abs/2108.05997?utm_source=chatgpt.com).
- 2] Chen C, Mo J. IQA-PyTorch: PyTorch toolbox for image quality assessment. Software repository, 2022–present. Version used: 0.1.16.[<https://github.com/chaofengc/IQA-PyTorch> (https://github.com/chaofengc/IQA-PyTorch?utm_source=chatgpt.com). Accessed 27 October 2026.
- 3] Hosu V, Lin H, Sziranyi T, Saupe D. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing. 2020;29:4041–4056. doi:10.1109/TIP.2020.2967829. Author manuscript:[<https://arxiv.org/abs/1910.06180> (https://arxiv.org/abs/1910.06180?utm_source=chatgpt.com).
- 4] Huang Z, He Y, Yu J, et al. VBench: Comprehensive Benchmark Suite for Video Generative Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024:21807–21818. Author manuscript:[<https://arxiv.org/abs/2311.17982> (https://arxiv.org/abs/2311.17982?utm_source=chatgpt.com).

Appendix A Complete generation prompt

Generate one 6-second photorealistic video containing THREE DISTINCT SHOTS separated by TWO INSTANTANEOUS HARD CUTS.

REFERENCE AUTHORITY

Use the supplied cowboy still to establish the man's identity, facial features, beard, hat, clothing, seated posture, bench, timber building, rope, fencing, vegetation, mountains, lighting, and colors.

The cowboy remains seated in the same physical position throughout. Only the camera framing changes at the two hard cuts.

SHOT 1 — CLOSE-UP — 0.00–2.00 SECONDS

Stationary, eye-level close-up showing the cowboy's entire hat, face, beard, shoulders, and upper chest. Preserve natural skin texture, individual beard hairs, felt texture, and shirt weave. He looks toward the camera with a relaxed, neutral expression.

INSTANTANEOUS HARD CUT AT 2.00 SECONDS.

SHOT 2 — MEDIUM — 2.00–4.00 SECONDS

Stationary, eye-level medium shot showing the same cowboy from the top of his hat to his waist, including his forearms and hands resting on his thighs. Preserve his posture and expression. Show the shirt pattern, seams, leather belt, and surrounding timber texture.

INSTANTANEOUS HARD CUT AT 4.00 SECONDS.

SHOT 3 — LONG — 4.00–6.00 SECONDS

Stationary, eye-level long shot showing the cowboy's entire seated figure, bench, and surrounding ranch landscape. His figure, from hat to boots, occupies approximately one-third of the image height. Include the timber building, coiled rope, gravel, fencing, trees, and distant mountain ridges.

Hold this framing through the final frame.

CAMERA AND CONTINUITY REQUIREMENTS

All three viewpoints face the cowboy along the same viewing direction at eye level. Preserve the spatial relationships among the cowboy, bench, building, props, and landscape.

No camera movement within any shot. No zoom, dolly, pan, tilt, orbit, drift, reframing, dissolve, morphing, or animated travel between viewpoints. Change shot size only through the two specified hard cuts.

PERFORMANCE AND VISUAL REQUIREMENTS

The cowboy remains nearly still, with only subtle natural breathing and blinking. His hands remain on his thighs. No dialogue or additional actions.

Maintain consistent daylight, exposure, color, clothing, anatomy, and scene geometry. Keep the cowboy and visible surroundings in focus, including the distant landscape. No intentional shallow depth of field, rack focus, motion blur, added film grain, or stylized effects.

Render fine textures naturally at each shot size. Do not add characters, objects, text, logos, or watermarks.

Complete exactly three shots and two hard cuts within six seconds.

Appendix B File identification and scoring record

The following SHA-256 values identify the retained native video files. Each value is shown on one line.

Kling 720p

1aff1cd90b0967951837e507fea8da20a3764d69dbbab8476f5187d1b9e45e72

Kling 1080p

e8304a3063ba9571b23114a71ab2d879ba50882c1ea461b09644f541a7b2790c

Wan 480p

c83e8edde9293c7b35d441a61666ef68141c700a622dc15102fbfbf0647884e6

Wan 1080p

743708bd56e7ebbb0da47cb4e4bb6b8b38955d47c20dc7e51c3e0dd230c45131

Seedance 480p

9ac13844f03e20d195e2b08b1950e099f496701f55aeb275ecc09373a53222f4

Seedance 1080p

9cc28b5b41d34d98365f1310b861dc34e2ea970cbd63cce4d92c3c309f01affa

Reference image

82a5c6baee0c41d7f06c0a00075d13b707ae05c3d310fb6e5b37e4fa787cb250

The original scoring session included 12 files. The frame-level CSV contains 108 observations, of which 54 are the six native files analyzed here. The retained native rows can be selected by excluding filenames containing “4K”. The session metadata file is MUSIQ_run.json and the frame-level file is MUSIQ_frames.csv. The scoring script was recovered from the downloaded file named Unconfirmed 706157.crdownload; its Python syntax was checked successfully. Original video bytes were preserved when filename labels were corrected.

For each native file, the logged fields are filename, segment number, requested time, selected frame index, width, height, and MUSIQ score. Overall means use all nine rows without exclusions. The original script sets frame position using OpenCV CAP_PROP_POS_FRAMES; it records the requested index rather than independently validating decoded presentation timestamps. Exact decoder builds were not frozen beyond the recorded OpenCV version.