VIBRICS: UNIFIED EVALUATION OF VIDEO GENERATION PIPELINES WITH REQUEST-SPECIFIC RUBRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation pipelines rewrite open-ended user requests into model-ready prompts before rendering videos. Existing evaluators focus on final outputs, providing little evidence about rewriting decisions and often missing fine-grained, request-specific requirements. We introduce **Vibrics**, a unified evaluation framework that freezes a request-specific rubric derived from the original request and reference media before observing any candidate. Its criteria evaluate the rewritten prompt, any available process trace, and the delivered video under one shared standard. Video criterion decisions yield Fulfillment (request satisfaction), Craft (visual quality), and calibrated Utility, while the full criterion record provides evidence-grounded diagnoses of unmet requirements across the pipeline. To benchmark complete pipelines at scale, we build **Vibrics-Bench** from 240 human-reviewed, open-ended production requests that reflect what users would type or paste, span eight input types and professional scenarios, and require substantive rewriting. Crossing five rewriting methods with five generators yields 25 pipelines per request and 6,000 videos, all evaluated under one frozen rubric per request. Vibrics agrees with expert consensus on 85.3% of comparable criterion decisions, while its Fulfillment, Craft, and Utility rankings reach 81%/81%/78% of human–human agreement. Experts also find that 89.6% of 1,000 diagnostic reports identify the delivered video's most important problem. On the 40-request annotated subset, Craft varies primarily with the generator, whereas Fulfillment varies with both the generator and the rewriting method. Together, Vibrics provides comparable scores and diagnoses for complete pipelines, while Vibrics-Bench provides a shared benchmark for evaluating rewriters, generators, and their combinations.

1 INTRODUCTION

Users describe the video they want as an open-ended request, not as a model-ready prompt. A video-creation system therefore first rewrites the request into a prompt, and a generator then renders the video from that prompt. The result depends on both stages. What counts as a good result varies across requests and domains, so each request has its own requirements. The rewriting stage expands a request into the content, action, and shot descriptions a generator needs. Platforms such as HeyGen and InVideo increasingly automate this step, letting users start from a short request (HeyGen, 2026; InVideo, 2026). A requirement can go unmet at either stage, and it fails in a different way at each. As Figure 1 shows, requested on-screen text such as "evaporation" can be dropped from the rewritten prompt, or kept but rendered unreadably in the video. Evaluating such a pipeline therefore requires judging both the delivered video and how each requirement is handled at every stage.

Existing work has advanced video evaluation from broad quality and alignment measures (Huang et al., 2024; Liu et al., 2024; He et al., 2024) to task-specific rubrics (Zeng et al., 2026), rubric-guided preference judgments (Jia et al., 2026), and workflow diagnosis (Chen et al., 2026). For rewrite-then-generate pipelines, two gaps remain. First, generic benchmarks score fixed dimensions that miss fine-grained, request-specific requirements. Second, for each requirement, existing evaluators do not answer two linked questions: whether the rewrite preserved it, and whether the video delivered it. They therefore cannot show where a requirement was lost. Table 1 compares these capabilities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

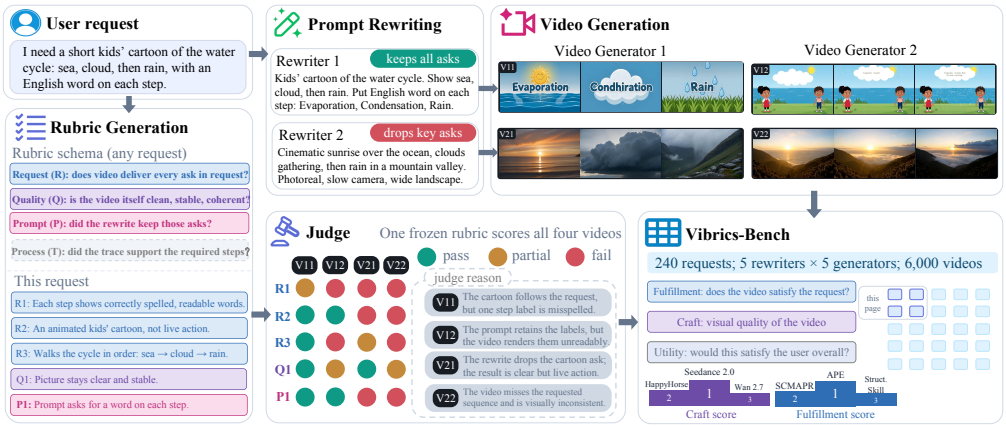

Figure 1: From one request to a shared full-pipeline standard. Vibrics freezes one rubric before evaluating four pipelines. Rewriter 2 omits the requested cartoon format and step labels from the prompt for Videos V21 and V22; Rewriter 1 retains them, but Video V12 renders the labels unreadably. Video-stage decisions yield Fulfillment, Craft, and calibrated Utility; prompt and process records add diagnostic evidence. Vibrics-Bench scales to 240 requests and 25 pipelines per request.

Table 1: Representative evaluation designs. Vibrics combines coverage of the complete pipeline, localized evidence, and request-specific criteria. ✓: supported; ×: outside scope.

Capability	VBench	VideoScore	VideoArgus	RAVEN-Eval	**Vibrics**
Video quality	✓	✓	✓	✓	✓
Input fidelity	×	✓	✓	✓	✓
Task-specific rubric	×	×	✓	✓	✓
Evidence report	×	×	✓	×	✓
Prompt & process	×	×	×	×	✓
Compares rewrites	×	×	×	×	✓

Accordingly, we introduce **Vibrics**, a unified framework that evaluates the complete pipeline against the original request. For each request, Vibrics builds a request-specific rubric from the request and its reference media. The rubric decomposes the request into independently judgeable criteria, each assigned to the rewritten prompt, an available process trace, or the delivered video. Separate rewriting trajectories, disjoint from the evaluated candidates, help sharpen these criteria while keeping each one grounded in the original request. Vibrics freezes the rubric before observing any candidate, so every pipeline is judged by one common standard even when they rewrite the request differently. A Judge Agent then executes each criterion: it locates the relevant evidence in the prompt, trace, or video, and routes it to a matched evaluator. Video criterion decisions yield *Fulfillment* (request satisfaction), *Craft* (visual quality), and calibrated *Utility* (overall). Prompt and process decisions are kept for diagnosis. Because the decisions are linked by requirement, Vibrics can show whether the rewrite preserved a requirement and whether the video delivered it, locating where each unmet requirement was lost. Figure 1 illustrates this flow.

To validate Vibrics and compare pipelines, we build **Vibrics-Bench** and collect independent expert judgments on a randomly sampled 40-request subset (1,000 videos). On this subset, calibrated Utility reaches the highest agreement with expert overall rankings among the tested automatic evaluators (Table 3), and experts judge 89.6% of the diagnostic reports to identify the video's most important problem. In a separate audit, each expert finds over 97% of criteria supported by the inputs, and Vibrics matches expert consensus on 85.3% of comparable decisions.

Applying Vibrics across all 240 requests, Seedance leads the generators and APE leads the rewriters in aggregate Utility, confirming that both stages shape the outcome. On the 40-request subset, Craft varies primarily with the generator, whereas Fulfillment varies with both the generator and the rewriting method. The retained criterion records further localize each unmet requirement to the rewritten prompt, an available process trace, or the delivered video. These patterns show why evaluation should cover the complete pipeline, separate request satisfaction from visual quality, and keep criterion-level evidence alongside aggregate scores.

Contributions. (1) We introduce **Vibrics**, a unified evaluation framework that freezes a request-specific rubric before observing any candidate and evaluates the complete rewrite-then-generate pipeline. It produces Fulfillment, Craft, calibrated Utility, and evidence-grounded diagnoses from the same criterion decisions. (2) We build **Vibrics-Bench**, a crossed benchmark of 240 human-

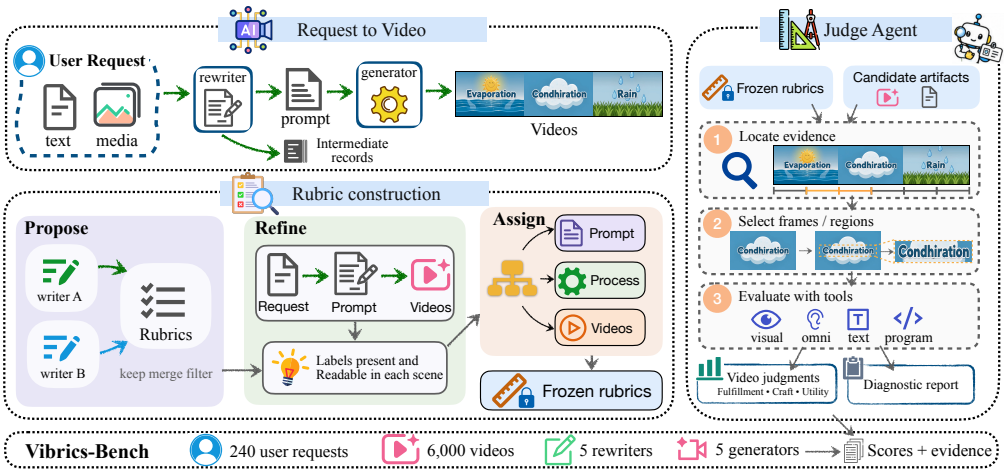

Figure 2: Vibrics workflow. **Request to Video:** a rewriter turns a user request (text and media) into a prompt that a generator renders into video, retaining intermediate records. **Rubric construction:** two writers propose criteria that are merged, refined on disjoint trajectories that never score a candidate, and assigned to the prompt, process, or video before the rubric is frozen. **Judge Agent:** for each candidate it locates evidence, selects frames or regions, and routes them to a visual, omni-modal, text, or program tool, yielding Fulfillment, Craft, and calibrated Utility plus a diagnostic report; here the video mis-renders the on-screen label *Condensation*, which the Judge localizes.

reviewed production requests spanning eight input types and professional scenarios. Each request has one frozen rubric; crossing five rewriters with five generators yields 25 pipelines and 6,000 videos. (3) We validate Vibrics with independent expert assessments. At the criterion level, decisions agree with expert consensus on 85.3% of comparable cases. At the score level, Vibrics attains the highest agreement with human overall rankings among all tested evaluators. At the report level, diagnoses identify the delivered video's most important problem in 89.6% of cases.

2 Vibrics: Unified Evaluation of Video Generation Pipelines

Vibrics evaluates a video generation pipeline end to end against the original request, from prompt rewriting and any intermediate agent operations to the delivered video. Prior evaluators judge only the delivered video against its rewritten prompt and cannot show where a requirement was lost. Vibrics instead judges every stage under one shared request-specific rubric, deriving Fulfillment, Craft, calibrated Utility, and a diagnostic report from the same frozen-rubric decisions. Fulfillment measures whether the delivered video satisfies the request by aggregating request-derived video criteria, and Craft measures its visual quality by aggregating applicable visual-quality criteria. Utility combines Fulfillment and Craft and is calibrated to the human overall scale. The report links prompt, process, and video decisions for each requirement and attributes unmet requirements to their stage of loss; Section 2.3 gives the exact aggregation.

Let x denote a request, possibly with reference media. A prompt rewriting method π produces a rewritten prompt $p = \pi(x)$, optionally exposing an intermediate process trace, and a generator g produces the video $v_{\pi,g} = g(p)$. Vibrics then proceeds in three steps. It first constructs a request-specific rubric from x and freezes it before observing any scored candidate (Section 2.1). The Judge Agent next executes the frozen rubric one criterion at a time, locating evidence at each criterion's stage and routing it to a matched evaluator (Section 2.2). Finally, video criterion decisions are aggregated into Fulfillment, Craft, and calibrated Utility, while prompt and process decisions are retained as diagnostic records (Section 2.3). Figure 2 illustrates this workflow.

2.1 Constructing a request-specific rubric

For each reviewed request x, Vibrics constructs an executable request-specific rubric before observing any scored candidate. The input consists of the request and any attached reference media. A *request requirement* is an obligation imposed by the request, whether explicitly stated, directly im-

plied, or grounded in the reference media. A *criterion* is an atomic and independently judgeable question that tests a request requirement or an applicable visual-quality property. The complete rubric records the requirements, stage-specific criteria, reference bindings, programmatic checks, challenge criteria, and the scoring or diagnostic role of each criterion. Rubric construction proceeds in three stages: *Rubric Proposal* drafts and merges candidate criteria, *Rubric Refinement* sharpens and completes them, and *Rubric Assignment* fixes each criterion's stage and role (Appendix A.5).

Rubric Proposal. Vibrics first extracts usable content from the reference media and combines it with the request. Two VLM rubric writers then independently produce initial drafts. Each draft decomposes the request into atomic requirements, identifies supported facts and derivations, and proposes concrete criteria for judging them. A separate fusion model compares the two drafts criterion by criterion. It retains well-supported items, merges overlapping items, removes unsupported proposals, and resolves differences between the two drafts. This propose-and-merge procedure reduces dependence on the interpretation of a single rubric writer.

Rubric Refinement. Vibrics next uses construction trajectories to refine and complete the fused draft. Each trajectory runs a complete rewrite-then-generate pipeline—prompt rewriting, any intermediate agent operations, and video generation—with rewriters and generators entirely separate from the scored benchmark matrix. Its delivered video exposes concrete ways a requirement may be expanded, preserved, or omitted across the pipeline and lets Vibrics form challenge criteria that can only be posed against a generated video. By connecting abstract user obligations to observable pipeline behavior, trajectories expose ambiguous boundaries, identify missing distinctions between requirements, and formulate independently judgeable stage-specific criteria. They can propose new criteria or refine existing ones, but cannot determine what the user requested: every retained criterion must remain independently supported by the original request or its reference media. Vibrics then removes semantically duplicate requirements, verifies facts derived from the reference media, and completes the executable set with applicable visual-quality criteria, deterministic checks for measurable constraints, and request-specific challenge criteria for plausible but incorrect outputs. These trajectories are never used as evidence for any scored candidate.

Rubric Assignment. Vibrics then assigns each criterion to the rewritten prompt, an available process trace, or the delivered video, and fixes its role in Fulfillment, Craft, or diagnosis. Request-specific video criteria contribute to Fulfillment, visual-quality criteria contribute to Craft, and prompt, process, and generic-defect criteria remain diagnostic. Vibrics finally serializes and freezes the complete rubric before observing any scored prompt, trace, or video. Every candidate pipeline for the request is then evaluated under this same standard.

2.2 Judge Agent: evidence-grounded criterion evaluation

After a candidate pipeline produces its rewritten prompt, any available process trace, and delivered video, the Judge Agent executes the frozen rubric one criterion at a time. This separates criterion construction from candidate evaluation: Section 2.1 fixes what must be judged, whereas the Judge Agent determines whether a particular candidate satisfies each fixed criterion. Scoring a whole video with a single holistic model is unreliable: such judges are often miscalibrated and sensitive to how a question is posed (Liu & Zhang, 2025; Chen et al., 2024), and one uniform pass easily overlooks evidence that is brief, localized, or fine-grained, such as a few frames of motion, small on-screen text, or a specific logo. It would also have to infer the evaluation target, gather evidence, weigh competing failures, and calibrate an overall score at once. Vibrics instead supplies an explicit target and role for every criterion and evaluates each with an agent that actively locates and re-examines the evidence a decision needs (Zhang et al., 2024). The Judge Agent cannot add, remove, merge, reweight, or reassign criteria, and no evaluator directly predicts an aggregate Fulfillment, Craft, or Utility score. For each criterion, the Judge Agent locates the required evidence at its assigned stage, routes it to a matched evaluator, and records a decision (Appendix A.6).

Judge Grounding. Each criterion specifies the stage at which it must be evaluated. At the prompt stage, the Judge Agent compares the rewritten prompt with the request to determine whether the relevant requirement is retained. At the process stage, it inspects any exposed intermediate trace, including retrieval, evidence-use, and planning records, to determine whether the requirement is

supported before generation. At the video stage, it inspects the delivered video together with the request and linked reference media. For each criterion, the Judge Agent first determines the evidence needed for a decision and then locates that evidence in the corresponding stage. It may select request facts, prompt spans, trace records, audio segments, video frames, temporal windows, or visual regions. Visual evidence is located through uniform or time-based sampling, scene segmentation, or content search and can be re-examined at higher resolution.

Judge Decision. After selecting the evidence, the Judge Agent routes it to an evaluator matched to the criterion. Programs evaluate measurable constraints, text models inspect prompts and process traces, multimodal models assess visible content and temporal events, and omni-modal models evaluate criteria that jointly depend on audio and video. Request-specific challenge criteria distinguish a satisfied requirement from a plausible but incorrect realization. Difficult or conflicting cases can be checked by a stronger evaluator before the final label is recorded. Request-specific criteria use $y_{i,s} \in \{1, 0, \varnothing\}$, where $\varnothing$ denotes an inapplicable criterion; graded visual-quality criteria receive a normalized score in $[0, 1]$. Each criterion decision retains the associated requirement, selected evidence, evaluator output, and final label. Video criterion decisions are passed to Section 2.3, while prompt and process decisions remain diagnostic records showing where a requirement lost support.

2.3 Scoring videos and generating diagnostics

Vibrics reports three complementary video scores. Fulfillment measures whether the delivered video satisfies the user's request. Craft measures the visual quality of the delivered video. Calibrated Utility combines both into a single overall score. Reporting all three prevents a polished video from masking an unmet request, or a faithful video from masking poor visual quality.

Fulfillment. Let $\mathcal{L}_T$ be the set of applicable, non-duplicate video criteria derived from request requirements that contribute to Fulfillment, and let s_ℓ be the decision score for criterion ℓ. Fulfillment is the mean of these scores:

$$F(v_{\pi,g} \mid x) = \frac{1}{|\mathcal{L}_T|} \sum_{\ell \in \mathcal{L}_T} s_\ell. \tag{1}$$

Inapplicable criteria and duplicates are excluded from the denominator. General diagnostic defects, prompt and process decisions, and criteria outside the request do not enter Fulfillment.

Craft. Craft evaluates the visual-quality aspects selected for the benchmark, including text rendering, material boundaries, composition, color, physical plausibility, motion, identity consistency, and residual artifacts. These aspects are organized into dimensions. Let $\mathcal{F}$ be the set of non-empty dimensions, D_f the available aspects in dimension f, and $s_d \in [0, 1]$ the normalized score of aspect d. Craft averages aspects within each dimension, then averages non-empty dimensions:

$$C(v_{\pi,g}) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \overline{\{s_d : d \in D_f\}}. \tag{2}$$

This dimension-level aggregation prevents a dimension with many measured aspects from receiving more implicit weight than one with fewer aspects.

Utility. Vibrics first forms an uncalibrated overall score:

$$U_0(v_{\pi,g} \mid x) = 0.6\, F(v_{\pi,g} \mid x) + 0.4\, C(v_{\pi,g}). \tag{3}$$

This fixed aggregation gives request satisfaction the larger weight while retaining visual quality as a necessary component. It is transparent, shared across all candidates, and computed solely from frozen-rubric decisions, so the rubric alone can rank candidates without relying on human overall scores. Because human overall satisfaction may vary nonlinearly with this fixed combination, equal changes in U_0 need not have equal perceptual meaning.

Vibrics therefore rescales U_0 to the human overall scale with a monotonic calibration that cannot reorder videos. Formally, the reported Utility is $U_{\text{iso}}(v_{\pi,g} \mid x) = g_{\text{iso}}(U_0(v_{\pi,g} \mid x))$, where $g_{\text{iso}} : [0, 1] \to [0, 10]$ is a non-decreasing PAV isotonic map (Barlow et al., 1972; Zadrozny & Elkan, 2002) fitted on a disjoint wave of independent human overall ratings and frozen before evaluation.

This calibration adjusts only the aggregate scale: it may pool fine-grained differences unsupported by human judgments into ties, but it does not learn criterion weights, reassign roles, or alter F or C. The main text reports uncalibrated F and C alongside calibrated U_{iso}; Appendix C.3 details the calibration protocol with auxiliary F_{iso} and C_{iso} analyses.

Diagnostics. Beyond the three scores, Vibrics compiles criterion decisions into a natural-language diagnostic report. For each requirement, it links prompt, process, and video decisions, attributing unmet requirements to the stage of loss. Grounded in retained evidence and evaluator output, the report names the delivered video's most important problem.

These outputs, the three scores and the diagnostic report, form the measurement interface that Vibrics-Bench applies to compare pipelines in Section 3.

3 VIBRICS-BENCH: A CROSSED BENCHMARK OF VIDEO PIPELINES

Vibrics-Bench instantiates Vibrics on 240 human-reviewed, open-ended production requests, together with their reference media, frozen rubrics, and candidate outputs. Unlike benchmarks with model-ready prompts and fixed output dimensions (Huang et al., 2024; Liu et al., 2024), it evaluates complete pipelines from open-ended requests under a request-specific rubric. It supports two tasks: validating automatic evaluation against independent expert judgments and comparing pipelines under one shared protocol. Experts annotate a subset; automated comparisons use all 240 requests. Crossing five rewriters with five video generators yields 25 pipelines per request and 6,000 videos, all evaluated under the same frozen rubric.

3.1 REVIEWED REQUESTS WITH MEASURABLE OBLIGATIONS

A benchmark that begins with a model-ready prompt cannot evaluate rewriting because that step has already occurred. Vibrics-Bench instead begins with what a user would type or paste, including attached files and hard constraints rather than a caption engineered for generation. Reviewers verify that each request can be executed and each attachment is accessible. They then make duration, aspect ratio, and resolution explicit. Requests are frozen before rubric construction and video generation. The 240 requests span eight input types, from text-only requests to requests with documents, reference media, or source videos. They cover professional scenarios including advertising, short-form drama, education, and healthcare. These inputs provide evidence for different requirements: an image can establish a required identity, while a document can supply relevant facts. In total, 148 requests include 172 reference items, counting documents and links as well as audiovisual materials. Every request states duration, aspect ratio, and 720p resolution, providing programmatically verifiable constraints. Appendix B reports the exact input-type counts and complete composition.

Vibrics then constructs one rubric per request, as defined in §2.1. Across all 240 rubrics, the rubrics contain 88 criteria on average, spanning prompt, process, video, challenge, and programmatic checks. Appendix A specifies the rubric format and presents a complete translated example.

3.2 SHARED-RUBRIC EVALUATION PROTOCOL

The five rewriters cover distinct mechanisms: Direct forwards the request unchanged; Structured Skill applies a fixed instruction template; SCMAPR uses multi-agent refinement (Yang et al., 2026); VPO is a learned optimizer (Cheng et al., 2025); and APE uses structured video planning with conditional material retrieval and review (Appendix B.2; broader rewriting literature in Appendix H). Wan 2.7 (Wan Team & Alibaba Group, 2025), HappyHorse, LTX 2.5, MiniMax-H3, and Seedance (Gao et al., 2025b) render every resulting prompt. The full 5×5 matrix is executed for every request. Holding the generator fixed compares rewriters through the videos their prompts produce; holding the rewriter fixed compares generators under the same prompt. Because the frozen rubric is shared across all cells, neither comparison changes the definition of success.

Applying Vibrics to each run yields F, C, U_0, and calibrated U_{iso}. Each run retains the rewritten prompt, any available process trace, the video, and criterion decisions. The scores compare systems; the records localize an unmet requirement to the prompt, an available trace, or the video. Appendix G describes the released requests, rubrics, artifacts, and Judge records.

3.3 HUMAN ANNOTATIONS AND CALIBRATION

Two blinded experts independently rate every candidate video in a randomly sampled 40-request subset (1,000 videos) for request fulfillment, visual quality, and overall satisfaction, and judge whether each diagnostic report identifies the video's most important problem. The annotations validate automatic evaluators; the same requests remain in full-benchmark comparisons but are excluded from calibration. Utility's isotonic map is fitted on 20 disjoint requests and frozen before scoring this subset. An audit of 1,032 request-specific criterion instances tests rubric grounding and Judge decisions. Appendix C.2 details the audit; Appendix C.3, the annotation protocol and calibration.

Table 2: Data scopes for comparison, evaluator validation, and calibration.

Data	Requests	Videos
Full benchmark	240	6,000
Annotated subset	40	1,000
Calibration	20	400

4 EXPERIMENTS

We evaluate Vibrics-Bench as a measurement benchmark rather than only as a collection of generated videos. The experiments ask three linked questions: whether the rubric, Judge decisions, and reported scores agree with independent experts; how complete pipelines compare under a shared rubric; and what fulfillment and quality patterns become visible in the retained evidence. These analyses remain separate because they test different claims.

4.1 EXPERIMENTAL SETUP

The comparison unit is one user request evaluated under its frozen rubric, over the data scopes in Table 2. We report uncalibrated F and C alongside calibrated U_{iso}. We rank systems by mean win rate, the share of pairwise comparisons a system wins, and by Bradley–Terry scores fit to those comparisons, and measure agreement with human rankings by per-annotator-averaged Kendall τ-b, a rank correlation that handles ties. Three separate analyses cover different parts of the measurement claim: a rubric-and-criterion audit, a subset human-ranking study, and benchmark-scale automated evaluation. On the common overall-ranking target, we compare U_{iso} with seven external evaluators (Table 3)—VideoArgus, a Qwen-VL-Max holistic judge, CLIPScore, VQAScore, VideoScore, and VBench Semantic and Quality—each of which sees only the prompt and video, not the request-specific rubric. Appendix C.1 gives the annotation protocol, calibration split, and evaluator inputs.

4.2 MEASUREMENT VALIDITY

We validate the measurement chain in order: whether the rubric criteria are warranted, whether the Judge executes them accurately, and whether the resulting scores preserve human rankings.

The rubric criteria are warranted and the Judge executes them accurately. Without seeing machine decisions, two experts judge whether each audited criterion is supported by its request or reference media. They independently retain 97.4% and 97.6% of the proposed criteria, supporting that the rubrics reflect user-facing inputs rather than post hoc system behavior. On the same audited scope, Vibrics matches expert consensus on 85.3% of comparable criterion decisions.

Vibrics scores align with human rankings, and calibrated Utility leads the tested evaluators. On the subset, rank agreement reaches 81%/81%/78% of human–human agreement for Fulfillment, Craft, and Utility. Calibration adjusts only the aggregate scale; the rubric, the $0.6/0.4$ combination, and reported F and C remain fixed. U_{iso} attains the highest point estimate among the automatic methods evaluated (Table 3). All scores remain positively aligned with each expert; Appendix Table 15 reports per-expert values and confidence intervals. The baselines share the human overall-ranking target but not the same inputs, and do not use Vibrics's rubric or measure F or C.

Diagnostic reports identify the most important video problem. Vibrics reports the affected requirement, available prompt/process/video decisions, and supporting evidence. Experts judge 89.6% of the 1,000 diagnostic reports in the subset to identify the delivered video's most important problem. Appendix F traces a failed Direct + LTX 2.5 automotive case and a complementary APE + MiniMax-H3 success from request to criterion evidence.

Table 3: Agreement with human overall rankings on the 40-request, 1,000-video subset. Values are request-level per-annotator-averaged Kendall τ-b; relative agreement divides each value by the human–human reference. Vibrics's isotonic map is fitted on a disjoint calibration subset.

Evaluator	Evaluation input	Score	Mean τ-b	Human agreement
Human–human reference	–	Overall rating	0.552	100%
Vibrics	Request-specific rubric + video	Calibrated U_{iso}	**0.433**	**78%**
VideoArgus (Zeng et al., 2026)	Request-specific rubric + video	Overall score	0.372	67%
Qwen-VL-Max holistic judge	Prompt + video	0–10 holistic score	0.326	59%
CLIPScore	Prompt + video	Frame text–video similarity	−0.121	−22%
VQAScore (Lin et al., 2024)	Prompt + video	Yes–No VQA score	−0.054	−10%
VideoScore v1.1 (He et al., 2024)	Prompt + video	Official overall score	−0.115	−21%
VBench Semantic (Huang et al., 2024)	Prompt + video	Overall consistency	−0.049	−9%
VBench Quality (Huang et al., 2024)	Video only	Quality overall	0.064	12%

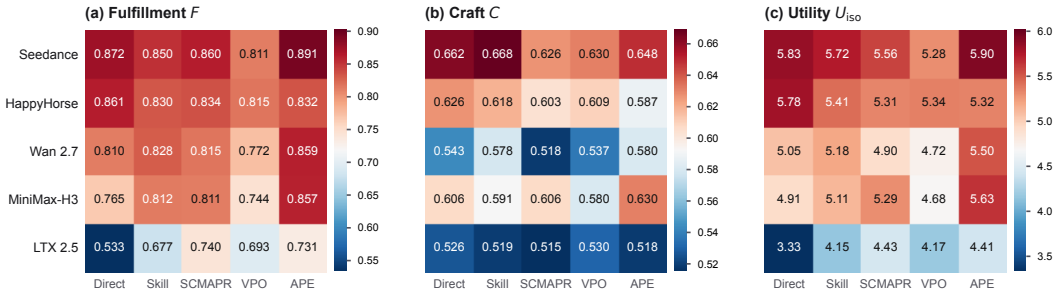

Figure 3: Held-out reference matrix. Each cell reports mean $F/C/U_{\text{iso}}$ for one rewriter (column) and generator (row), all evaluated with the same rubric per request. F and C lie in $[0, 1]$; U_{iso} uses the human overall-rating scale. Color scales are panel-specific and scaled to each panel's range.

4.3 Main results on Vibrics-Bench

Having validated the measurement chain, we first inspect the reference matrix on the subset and then apply the same procedure at benchmark scale. Every candidate for a request is evaluated under the same frozen rubric, so the matrices compare delivered outcomes under one definition of success rather than treating each rewritten prompt as a different task. Figure 3 shows the human-validated five-rewriter by five-generator reference matrix; Appendix Tables 17 and 19 report the subset and benchmark-scale values, respectively.

APE leads the evaluated rewriters. The fixed APE condition attains the highest observed rank on the subset under both Vibrics Utility and expert overall ratings. At benchmark scale, it also leads Utility (Bradley–Terry 1.67) and has the highest Fulfillment and Craft win rates (65.3% and 63.1%). The subset agreement supports using the same automated protocol at benchmark scale (Table 4).

Generator rankings agree at the top and bottom. On the subset, machine and human rankings both place Seedance first, HappyHorse second, and LTX 2.5 last. At benchmark scale, Seedance leads Utility (Bradley–Terry 2.67), Fulfillment, and Craft. HappyHorse ranks second on Utility and Craft, whereas Wan 2.7 ranks second on Fulfillment; LTX 2.5 remains last on every score. Appendix E provides the complete matrices and ranking statistics.

4.4 Benchmark findings and diagnostic analysis

Aggregate scores separate delivered outcomes but cannot show which request obligations are unmet or which visual properties differ. Vibrics-Bench therefore retains criterion-level decisions and evidence, so the same results can be inspected along both request fulfillment and visual quality. Craft varies primarily with the generator, whereas Fulfillment varies with both generator and rewriter: on the subset, generator-associated variation exceeds rewriter-associated variation for Craft (0.143 versus 0.006 of total cell-level variation), and Fulfillment varies with both generators (0.216) and rewriters (0.036); Figure 4 in Appendix D summarizes this split. Request-clustered contrasts sharpen the result: APE separates from all four alternatives on Fulfillment (Appendix D). This split is visible

Table 4: Utility mean win rate (%) on the 40-request subset under machine and human scores, and in the benchmark-scale automated evaluation. Higher is better; bold denotes first, underline second.

(a) Prompt rewriting methods

	40-request (machine)	40-request (human)	Benchmark (machine)
APE	**71.9**	**73.4**	**68.4**
Structured Skill	<u>54.4</u>	37.8	<u>61.9</u>
SCMAPR	53.1	<u>56.2</u>	56.8
VPO	30.0	43.8	34.7
Direct	40.6	38.8	28.2

(b) Video generators

	40-request (machine)	40-request (human)	Benchmark (machine)
Seedance	**86.8**	**83.4**	**83.0**
HappyHorse	<u>69.9</u>	<u>79.1</u>	<u>65.6</u>
MiniMax-H3	44.6	51.0	48.4
Wan 2.7	44.6	32.8	50.0
LTX 2.5	4.1	3.7	3.0

only because Vibrics evaluates the rewrite stage and scores Fulfillment separately from Craft; an output-only or single-score evaluation would hide it.

Criterion records localize unmet requirements to the prompt, process, or video stage. For each requirement, Vibrics retains its comparable prompt, process, and video decisions with selected evidence. Figure 5 in Appendix D counts a prompt-stage absence only when the frozen prompt criterion fails; unavailable records are excluded. Because Direct copies the request unchanged, its 1.0% absence rate estimates the check's false-positive floor. The same record structure distinguishes absent prompt support, absent support in an available trace, unrealized video content, and visual defects that affect Craft. Each diagnosis names the stage where support is missing or a defect appears, so an unmet requirement can be traced to where it was lost. Appendix D reports request-clustered contrasts and accounting, while Appendix F provides evidence records.

5 Related Work

Video generation evaluation. Video benchmarks measure broad quality and controllability, as in VBench (Huang et al., 2024), VBench++ (Huang et al., 2026), VBench-2.0 (Zheng et al., 2025), and EvalCrafter (Liu et al., 2024); compositional generation in T2V-CompBench (Sun et al., 2025); image-conditioned content in AIGCBench (Fan et al., 2023); and physical plausibility in VideoPhy (Bansal et al., 2025). Their dimensions reveal important capabilities but are fixed independently of a user's complete request. Vibrics-Bench instead evaluates request-specific delivery and visual quality under a rubric frozen from that request.

Requirement-based evaluation and agentic judges. Text–video metrics measure agreement with a supplied prompt or reference, including VQAScore (Lin et al., 2024), T2VScore (Wu et al., 2024), VideoScore (He et al., 2024), and VideoScore2 (He et al., 2026). Requirement-based evaluators decompose inputs into checkable claims, as in TIFA (Hu et al., 2023), DSG (Cho et al., 2024), FActScore (Min et al., 2023), and CheckEval (Lee et al., 2025). Model-as-judge systems such as MT-Bench (Zheng et al., 2023), MLLM-as-a-Judge (Chen et al., 2024), Evaluation Agent (Zhang et al., 2024), VideoArgus (Zeng et al., 2026), and RAVEN-Eval (Jia et al., 2026) execute such checks with models, tools, or retrieved evidence. Most related to our setting, DirectorBench (Chen et al., 2026) diagnoses the video generation workflow. Vibrics builds on these ideas but freezes one request-specific rubric before observing any candidate and executes it unchanged across prompt, process, and video, so the same decisions score every pipeline and locate lost requirements.

6 Conclusion

Vibrics unifies the evaluation of rewrite-then-generate video pipelines under one request-specific rubric frozen before any candidate is observed. On Vibrics-Bench (240 human-reviewed production requests, 25 pipelines each), experts validate the rubric, the Judge decisions, and the diagnostic reports, and calibrated Utility attains the highest agreement with human overall rankings among the tested evaluators. Craft varies chiefly with the generator, whereas Fulfillment varies with both generator and rewriter. By judging every stage against the original request, Vibrics makes heterogeneous pipelines comparable and pinpoints where each unmet requirement was lost.

AI USE STATEMENT

We used generative AI tools only to polish the English wording and grammar of the manuscript. We did not use generative AI to generate research ideas, to design the protocol or the benchmark, or to produce, select, or interpret any reported number. All AI-assisted text was checked by the authors against the underlying results, and we take responsibility for the final content of this work.

ETHICS STATEMENT

The human component of this study consists of two compensated expert annotators who rated generated videos on a 0–10 scale, blinded to method and generator identity (§4.1). Both annotators provided informed consent before participation. No personally identifiable information about the annotators is collected or released, and the released records contain only per-video scores. The requests in Vibrics-Bench describe products and scenes rather than identifiable individuals, and all videos are model outputs; we release no scraped or user-contributed media. The rewriting methods, generators, and judge model we evaluate are used under their respective terms of service, and the scores we report characterize those systems on this benchmark only, not their fitness for deployment. The authors declare no conflicts of interest.

REFERENCES

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.03520.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, New York, 1972.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24h.html.

Jiamin Chen, Qianben Chen, Jiawen Zhang, Yidi Wu, Yuchen Li, Xiaokun Zhang, Wangchunshu Zhou, and Chen Ma. DirectorBench: Diagnosing long-form video generation with personalized multi-agent evaluation, 2026. URL https://arxiv.org/abs/2605.30090.

Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. VPO: Aligning text-to-video generation models with prompt optimization. *arXiv:2503.20491*, 2025.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian Scene Graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ITq4ZRUT4a.

Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. AIGCBench: Comprehensive evaluation of image-to-video content generated by AI. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(4):100152, 2023. doi: 10.1016/j.tbench.2024.100152. URL https://doi.org/10.1016/j.tbench.2024.100152.

Bingjie Gao, Xinyu Gao, Xiaoxue Wu, Yujie Zhou, Yu Qiao, Li Niu, Xinyuan Chen, and Yaohui Wang. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3173–3183, 2025a. doi: 10.1109/CVPR52734.2025.00302. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gao_The_Devil_is_in_the_Prompts_Retrieval-Augmented_Prompt_Optimization_for_CVPR_2025_paper.html.

Yu Gao, ByteDance Seed, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv:2506.09113*, 2025b.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66923–66939, 2023. doi: 10.52202/075280-2923. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d346d91999074dd8d6073d4c3b13733b-Abstract-Conference.html.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.127. URL https://aclanthology.org/2024.emnlp-main.127/.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. VideoScore2: Think before you score in generative video evaluation. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=MpkVh4jH44.

HeyGen. HeyGen video agent. https://www.heygen.com/video-agent, 2026. Accessed: 2026-09-25.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hu_TIFA_Accurate_and_Interpretable_Text-to-Image_Faithfulness_Evaluation_with_Question_Answering_ICCV_2023_paper.html.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

InVideo. InVideo AI: Turn any idea into a video with a single prompt. https://invideo.io/ai/, 2026. Accessed: 2026-09-25.

Ziheng Jia, Jiaying Qian, Zicheng Zhang, Xiaorong Zhu, Lancheng Gao, and Xiongkuo Min. RAVEN-Eval: Rubric-guided automatic evaluation for AI video generation models based on LMM preference judgement, 2026. URL https://arxiv.org/abs/2608.09111.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15771–15798. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.796. URL https://aclanthology.org/2025.emnlp-main.796/.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision – ECCV 2024*, pp. 366–384. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72673-6_20. URL https://doi.org/10.1007/978-3-031-72673-6_20.

Ming Liu and Wensheng Zhang. Is your video language model a reliable judge? *CoRR*, abs/2503.05977, 2025. doi: 10.48550/arXiv.2503.05977. URL https://arxiv.org/abs/2503.05977.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024. doi: 10.1109/CVPR52733.2024.02090. URL https://doi.org/10.1109/CVPR52733.2024.02090.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arık. VISTA: A test-time self-improving video generation agent, 2025. URL https://arxiv.org/abs/2510.15831.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization, 2024. URL https://arxiv.org/abs/2403.17804.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Yiwen Song, Tomas Pfister, and Yale Song. VQQA: An agentic approach for video evaluation and quality improvement, 2026. URL https://arxiv.org/abs/2603.12310.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025. doi: 10.1109/CVPR52734.2025.00787. URL https://doi.org/10.1109/CVPR52734.2025.00787.

Wan Team and Alibaba Group. Wan: Open and advanced large-scale video generative models. *arXiv:2503.20314*, 2025.

Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, and Mike Zheng Shou. Towards a better metric for text-to-video generation. *CoRR*, abs/2401.07781, 2024. doi: 10.48550/arXiv.2401.07781. URL https://arxiv.org/abs/2401.07781.

Chengyi Yang, Pengzhen Li, Jiayin Qi, Aimin Zhou, Ji Wu, and Ji Liu. SCMAPR: Self-correcting multi-agent prompt refinement for complex-scenario text-to-video generation, 2026. URL https://arxiv.org/abs/2604.05489.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

Ziyun Zeng, Zixuan Wang, Yongsheng Yu, Hang Hua, and Jiebo Luo. VideoArgus: Agentic rubric-grounded unified evaluation for video generation and editing, 2026. URL https://arxiv.org/abs/2608.05485.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. *arXiv:2412.09645*, 2024.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL https://arxiv.org/abs/2503.21755.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Appendix

A. Vibrics Method Details 15
A.1 What a Frozen Rubric Contains 15
A.2 Scoring Rules and Edge Cases 15
A.3 An Illustrative Automotive Rubric Instance 16
A.4 Stage-Separated Execution 17
A.5 Rubric-Construction Prompts 17
A.6 Judge-Execution Prompts 18
B. Vibrics-Bench Construction and Execution 18
B.1 What Was Fixed in Execution 19
B.2 APE Architecture and Fixed Configuration 19
B.3 Versioning and Evaluation Manifests 19
C. Human Annotation and Measurement Evidence 21
C.1 Annotation Protocol 21
C.2 Rubric Warrant and Criterion-Level Execution 21
C.3 Human Ranking Study and Isotonic Calibration 22
C.4 Automatic Evaluator Results 22
D. Crossed-Matrix Statistics 24
E. Large-Scale Ranking and Cross-Request Aggregation 25
F. Complete Evidence Records for Two Archived Runs 27
F.1 C01: Agent-Rewritten Sleeve-Dance Request 27
F.2 C02: Direct Automotive-Ad Request 29
G. Evidence Boundaries and Reproducibility 33
H. Extended Related Work: Prompt Rewriting for Visual Generation 33

A VIBRICS METHOD DETAILS

Vibrics serializes each rubric before observing a scored prompt, trace, or video. Beyond request requirements, the rubric records narrative structure, reference links, stage roles, evidence-localization instructions, programmatic checks, and request-tied hard negatives. Construction-only trajectories may clarify how a supported requirement can be judged, but cannot add requirements or supply evidence from a scored matrix cell. Generic defects remain diagnostic and never enter Fulfillment, Craft, or Utility.

For visual criteria, localization may use uniform or time-based sampling, scene segmentation, or content search, followed by higher-resolution verification. Programs evaluate measurable constraints, text models inspect prompts and traces, multimodal models inspect visible content, and omni-modal models handle audiovisual evidence. Request-tied hard negatives distinguish fulfillment from a plausible but incorrect rendering; difficult or conflicting visual evidence can be rechecked before the final criterion decision.

A.1 WHAT A FROZEN RUBRIC CONTAINS

Table 5 separates the full executable rubric from its request requirements. Requirements define what must hold; the remaining fields specify how each obligation is observed, localized, and adjudicated.

Table 5: Components of a frozen Vibrics rubric. The final column indicates whether each component can contribute directly to video-side scores.

Component	What is frozen before candidates are observed	Role in $F/C/U$
Reviewed request specification	Input track, domain, hard duration/aspect/resolution constraints, and minimum narrative structure.	Applicable video-side criteria may enter F.
Narrative and shot structure	Ordered beats, shot functions, and dependencies between required events.	Supports criterion construction and temporal evidence; not a separate score.
Request requirements	Stated requirements, supported derivations, material facts or fidelity, literal on-screen text, and verbatim requirements.	Applicable video-side criteria enter F.
Reference and evidence links	Attached assets, linked requirement identifiers, required forms, and evidence or localization instructions.	Conditions how a requirement is assessed; no independent score.
Criterion evaluation layer	Criterion-specific questions for identity/reference, event/effect, cinematic, execution, and technical criteria.	Produces the decisions used by Fulfillment or Craft aggregation.
Objective metrics	Programmatic checks for hard constraints and selected visual properties.	May supply a frozen video-side decision or Craft aspect.
Request-tied challenges	Explicit hard-negative probes tied to a request requirement and, when applicable, a time window.	Tied content probes may enter F; unbound diagnostics do not.
Universal defect set	Request-independent failures such as garbling, residue, or collapse.	Diagnostic only; excluded from $F/C/U$.

A.2 SCORING RULES AND EDGE CASES

The serialized records fully determine every score. Fulfillment averages applicable, non-duplicate request criteria; Craft averages available aspects within each dimension and then across non-empty dimensions. Vibrics computes $U_0 = 0.6F + 0.4C$ and reports $U_{\text{iso}} = g_{\text{iso}}(U_0)$. Prompt/process records, universal diagnostics, duplicates, and inapplicable criteria are excluded. Appendix C.3 details calibration.

Table 6: Score components and exclusion rules. Every reported score is computed from video-side decisions only.

Component	Included evidence	Exclusion or aggregation rule
Fulfillment F	Applicable video criteria for subject, event, supporting content, trademark, literal text, specification, and request-tied probes.	Equal mean over applicable criteria; duplicates and not-applicable criteria are excluded.
Craft C	Available video-side visual aspects, including text rendering, material edges, composition, color, motion, identity, and residual defects.	Mean within each dimension, then equal mean across non-empty dimensions; divide the 0–10 scale by 10.
Utility U_{iso}	Machine F and C; disjoint human overall ratings are used only to fit the frozen calibrator.	First compute $U_0 = 0.6F + 0.4C$, then apply the frozen isotonic calibrator. No third criterion set is used.
Diagnostic records	Prompt coverage, process trace, universal defects, and report text.	Do not enter F, C, or U; retained to locate evidence for a failure.

A.3 An illustrative automotive rubric instance

We use an archived automotive request to show the rubric's three layers: user obligations, executable criteria, and stage-specific records. The case includes hard constraints, narrative and literal-render requirements, localized challenge probes, and programmatic checks.

The request specifies a 15-second, 16:9 BYD electric-vehicle commercial with a cyberpunk city, drone tracking, HUD range and acceleration information, tunnel light trails, and a frozen logo ending. The rubric converts this specification into four narrative beats, four planned shots, nine stated requirements, two licensed derivations, and two literal-text requirements. Table 7 lists the stated requirements; the release contains the full serialization.

Table 7: All nine stated request requirements in the illustrative automotive rubric. Each row is fixed from the reviewed request, not inferred from a candidate prompt or video.

ID	Frozen requirement	Primary evidence and execution route
A1	The delivered video lasts 15 seconds.	Container metadata; programmatic check.
A2	The delivered video uses a landscape 16:9 aspect ratio.	Container metadata; programmatic check.
A3	The featured subject is a BYD electric vehicle, not an unspecified car.	Video evidence; visual content check.
A4	The setting is a high-tech, futuristic cyberpunk city.	Video evidence; visual style check.
A5	A drone-view camera tracks the moving vehicle.	Temporally localized video evidence; camera motion check.
A6	A HUD displays both driving range and 0–100 km/h acceleration time.	Localized frames and rendered text; video and text check.
A7	The vehicle enters a tunnel and the lighting produces flowing light trails.	Temporally localized video evidence; event check.
A8	The ending freezes on the BYD brand logo.	End-of-video evidence and literal form check.
A9	The commercial foregrounds range and acceleration performance.	Evidence aggregated from the request-specific performance elements above.

The archived rubric further binds each licensed derivation, required HUD label, and hard-negative probe to a named requirement and evidence strategy. Its execution layer contains 32 VQA records (19 identity/reference, 2 event/effect, 2 cinematic, 4 execution, and 5 technical), 16 objective metrics, five request-tied challenges, and 17 universal diagnostics. Duration and aspect ratio appear in VQA records but are routed to programmatic checks. The released JSON therefore captures both *what* the request demands and *how* each frozen criterion is executed.

A.4 Stage-separated execution does not change the standard

A requirement can be inspected at several stages, but only its video-side decision can affect a score. In this case, the rewritten prompt omits the 16:9 constraint, so prompt-side A2 fails, while the delivered video is programmatically verified as 16:9. The prompt record measures retention; the video record measures delivery. This example illustrates the evidence trail, not human validation of stage attribution.

Process checks may inspect the request, trace, retrieved evidence, and prompt; prompt checks inspect the request and rewritten prompt; video checks inspect the delivered video and criterion-linked references. Prompt and process decisions remain diagnostic, video decisions may enter F or C, and no stage record can alter the frozen requirement.

A.5 Rubric-construction prompts

The templates below construct a frozen rubric following the process in Section 2.1: two independent drafts are proposed from the request and its reference media, merged into one, refined on disjoint construction trajectories, and extended with request-tied challenge criteria. Placeholders in `<angle brackets>` denote inputs. The templates are lightly edited for length and aligned to this paper's terminology; the released code contains the exact strings.

Criterion proposal.

> **Role.** You construct a request-specific rubric for one video-creation request. **Task.** From the request and the independently extracted reference media, list the criteria the request implies. A criterion is an atomic, independently judgeable question tied to a stated or directly derivable requirement. **Rules.** (1) Ground every criterion in the request text or reference media; do not add facts the request never asks to appear in the video. (2) Tag each criterion with its requirement type (named entity, verifiable fact, visual or audio style, camera, action or plot, reference-material use, intent, or hard constraint) and the stage that can evidence it (prompt, process, or video). (3) Record a minimum narrative structure—ordered beats and a shot-count floor—only when the request implies distinct shots. (4) Assign each criterion an importance, justified by the request. **Inputs.** `<request>`, `<reference-media summary>`. **Output.** A list of criteria, each with type, stage, evidence instruction, and importance.

Draft fusion.

> **Role.** You merge two independently proposed rubric drafts, A and B, into one. **Task.** Produce a single rubric and re-derive importance rather than averaging the drafts. **Rules.** (1) Keep a criterion only if it is derivable from the request or reference media; drop anything both drafts invented but the request never asks to appear. (2) Merge duplicates that state the same requirement, keeping the most complete evidence instruction and combining alternative references into one; keep genuinely distinct facts as separate criteria and never join them with "or". (3) Take the union of hard constraints, forbidden patterns, and reference assets, resolving conflicts toward the draft that better follows the construction rules. (4) Re-derive each criterion's importance from its role in this request, with a one-line reason. **Inputs.** `<request>`, `<draft A>`, `<draft B>`. **Output.** The merged rubric.

Refinement on disjoint trajectories.

> **Role.** You validate a merged rubric against three separate construction trajectories that are never scored as candidates. **Task.** Make the rubric fair and general—not a replica of these three runs—using a three-state rule. **Rules.** (1) *Consensus*: if all three trajectories do

> something and the rubric already judges it, leave it unchanged; do not copy shared behavior into a new criterion. (2) *Deficiency*: if at least one trajectory omits a requirement the request or media supports, strengthen the corresponding criterion (raise importance or add an evidence point), because it is easy to miss. (3) *Divergence*: if the trajectories differ on a free implementation choice not required by the request, such as camera movement, set no standard. Never add, remove, or rewrite request-derived structure from trajectory behavior. **Inputs.** `<merged rubric>`, `<three construction trajectories>`. **Output.** The refined rubric.

Challenge-criterion synthesis.

> **Role.** You write request-tied challenge criteria (request-tied hard negatives) that probe plausible but incorrect realizations. **Task.** Turn near-miss failure modes for this request into yes/no criteria phrased so that the questioned state is the wrong one. **Polarity rule.** Read each criterion twice: a video that fully meets the request must answer "no," and a video that failed in the intended way must answer "yes." Discard any criterion a fully correct video would answer "yes." **Rules.** (1) Describe the defect state concretely—an object, its wrong appearance, and where or when to look—so it can be located in frames. (2) Tie each challenge criterion to a specific request requirement. (3) Do not state how to score; the verdict maps "yes" to an unmet requirement. **Inputs.** `<request>`, `<frozen criteria>`, `<candidate observations>`. **Output.** A list of challenge criteria, each with a defect-state description.

A.6 JUDGE-EXECUTION PROMPTS

The Judge Agent executes the frozen rubric one criterion at a time (Section 2.2): it locates evidence at the criterion's stage, selects frames or regions, routes the evidence to a matched evaluator, and records a verdict; a final step aggregates the video-stage decisions into scores and a diagnostic report.

Per-criterion evaluation.

> **Role.** You evaluate one frozen criterion against the located evidence for one candidate. **Routing.** The criterion is answered by the evaluator matched to its stage and modality: a program for measurable constraints, a text model for the prompt or process trace, a multimodal model for visible content, or an omni-modal model for audiovisual content. **Verdict.** Return *satisfied*, *unmet*, or *not-assessable*, with a short evidence note. Judge only from the provided evidence; do not infer or use outside knowledge. Mark *satisfied* only when the evidence directly shows the questioned content; mark *unmet* when a full scan shows it absent or contradicted; mark *not-assessable* when the target is present but too small, blurred, or occluded to confirm. **Existence vs. quality.** If the criterion asks whether required content appears, its absence is *unmet*; if it asks whether an object's quality is adequate and that object never appears, it is *not-assessable*. **Inputs.** `<criterion>`, `<located evidence: frames, regions, prompt, or trace>`. **Output.** A verdict with an evidence note.

Aggregation and diagnosis.

> **Role.** You aggregate the per-criterion verdicts for one candidate into scores and a diagnostic report. **Rules.** (1) Fulfillment aggregates the request-specific video criteria and Craft aggregates the visual-quality video criteria; prompt- and process-stage decisions are diagnostic evidence and never enter the scores. (2) Report uncalibrated Fulfillment and Craft; calibrated Utility is produced by the frozen isotonic map, not by this prompt. (3) For each unmet requirement, thread its prompt, process, and video decisions, attribute the loss to the stage where it occurred, and name the single most important problem in the delivered video. (4) Ground every statement in the retained criterion decisions and their evidence. **Inputs.** `<criterion verdicts with evidence>`. **Output.** Uncalibrated Fulfillment and Craft, and a natural-language diagnostic report.

B VIBRICS-BENCH CONSTRUCTION AND EXECUTION

Table 8 gives the released request-bank composition. The eight input types comprise 92 text-only requests, 45 with documents, 34 with reference images or videos, 26 with source videos to edit, 18

with mixed inputs, 11 with audio, 10 with links, and 4 with first or last frames. Scenario and action entries count requests, while media entries count attached objects because one request may contain several. Durations are 15, 10, and 5 seconds for 113, 106, and 21 requests; aspect ratios are 16:9, 9:16, and 1:1 for 171, 47, and 22 requests; all specify 720p. Attached media additionally grounds identity, factual, and fidelity obligations unavailable to text-only requests.

Table 8: Full composition of the 240 human-reviewed requests. Scenario and action counts are requests; media counts are attached items.

Scenario category	Requests
Advertising and marketing	27
Short-form drama	25
Motion graphics and UI	21
Popular science	21
Dance and music video	21
Education and training	18
Video editing	15
Camera movement	15
Corporate and business	14
Travel and lifestyle	13
Entertainment and culture	12
Public service	12
News	12
Healthcare	10
Frame-conditioned	4

Attached media	Items
Document or other file	53
Reference image	46
Reference video	31
Reference audio	21
Link	15
First frame	4
Last frame	2
Requested action	**Requests**
Direct generation	182
Edit or composite	38
Effect enhancement	20

B.1 What was fixed in execution

Table 9 records the evaluated systems and Judge channels. Rubric construction occurs once per request; criterion execution begins only after freezing. Execution channels cannot add requirements, change decision rules, or observe aggregate scores, and routing is fixed across matrix cells.

B.2 APE architecture and fixed configuration

APE is a tool-using prompt rewriter driven by Qwen3.7-Max. Given an original request and optional reference media, it first normalizes hard generation constraints and routes the request to either a text-only or reference-grounded workflow. Text-only requests proceed directly to structured planning. Requests that require additional grounding trigger parsing of supplied documents or links and, when needed, textual and visual retrieval. A material-review stage filters the collected items for relevance and binds retained evidence to the request.

APE next constructs a structured video plan that specifies the temporal organization, shot composition, camera behavior, audio, and reference-media bindings. A staged prompt compiler converts this plan into one generator-ready prompt and checks its structural consistency and hard constraints. The resulting record contains the original and enhanced prompts, generation parameters, selected reference bindings, structured plan, and available tool trace; Vibrics uses the trace only as diagnostic process evidence and never as a video score.

The architecture, prompt templates, tool set, and execution configuration were frozen before any machine or human evaluation results were inspected and were not changed in response to Vibrics feedback. For each request, APE produced one final enhanced prompt that was reused unchanged across all five generators. The scored APE campaign was stored separately from the construction-only trajectories used during rubric development, and no scored APE prompt, trace, or video was used to construct or revise its rubric. Reported APE results therefore characterize this single evaluated configuration.

B.3 Versioning and evaluation manifests

Each result is bound to versions of the request bank, rubric, candidate artifacts, and Judge configuration. Its manifest hashes the request, prompt, video, rubric, VQA records, sampling settings, code,

Table 9: Systems and channels used in the reported evaluation. Programmatic checks use container metadata and standard video measures rather than an LLM. Construction, execution, and diagnostic localization are distinct operations.

Component	Implementation or channel	Fixed responsibility
Direct	request forwarded unchanged	Reference rewriting column; estimates the prompt-side false-positive floor.
Structured Skill	instruction-based rewrite, Qwen3.7-Max	Fixed instruction baseline.
SCMAPR	multi-agent refinement, Qwen-Plus	Compared rewriting method.
VPO	released optimizer, VPO-5B	Compared rewriting method.
APE	tool-using rewriter, Qwen3.7-Max	Fixed rewriting condition with structured planning and conditional evidence acquisition (Appendix B.2).
Wan 2.7 / HappyHorse / LTX 2.5 / MiniMax-H3 / Seedance	hosted and local generator channels	The five generator columns of every request.
Text	Qwen3.7-Max	Process and rewritten-prompt criteria.
Video	Qwen-VL-Max; Qwen3.8-Max for hard visual cases	Video-side visual content and defect checks.
Omni-modal	Qwen3.5-Omni-Flash	Requirements that must be heard rather than inferred from frames.
Programmatic	Container metadata and video measures	Duration, aspect ratio, resolution, and designated objective checks.
Localization and transcription	Qwen3-VL-Plus; Qwen3-VL-Flash	Locate frame evidence and transcribe literal on-screen text; provides inputs to a verdict.
Rubric construction	Qwen3.7-Max and GLM-5.2 propose-and-merge	Compile a request-specific rubric once, before candidates are observed.

and prompt bundle. A result is reusable only when all input-relevant fields match, making rerun decisions explicit after either candidate or evaluator changes.

Table 10: Manifest fields required to reproduce a reported evaluation.

Field group	Frozen record
Request and rubric	Request-bank version, request and reference hashes, rubric hash, VQA hash.
Candidate artifact	Rewriting method, generator, prompt hash, video hash, and available trace hash.
Evaluator	Judge-channel versions, prompt-bundle source hash, sampling settings, and aggregation version.
Reuse decision	Input-relevant manifest equality; code-only changes are recorded separately for audit.

C Human annotation and measurement evidence

C.1 Annotation protocol

For each request, the blinded interface randomizes candidate videos and hides rewriter and generator identities. Experts independently assign 0–10 ratings for request fulfillment and visual quality and separately provide an overall rating; the overall rating is elicited directly rather than derived from the first two ratings. Optional fields collect failure tags, notes, and best/worst selections. Condition labels are decoded only after collection.

The 40-request subset comprises 40 requests evaluated by five rewriters and five generators, yielding 1,000 delivered videos available for annotation.

The statistical unit is a within-request ranking. We compute Kendall τ-b between machine scores and each expert separately, average the two correlations, and then average across requests; human–human agreement uses the same candidates. A Craft request is omitted only when an expert gives all candidates the same score and rank correlation is undefined. Both experts gave informed consent and were compensated.

Table 11: Human annotation fields and their role in the evidence chain.

Field	Role
Fulfillment rating	Independent human ranking target for machine F.
Craft rating	Independent human ranking target for machine C.
Overall rating	Independent human ranking target for machine U; not derived from the other fields.
Best/worst selections and notes	Qualitative context for case studies; not used to fit reported scores.
Failure tags	Coarse human observations; not a gold standard for report prose or stage attribution.

Diagnostic-report check. Across all 1,000 diagnostic reports from the 40-request subset, 89.6% were judged to identify the most important problem in the delivered video. This check supports critical-problem identification on the evaluated subset; it does not establish that a report exhaustively lists every failure or identifies the component that caused a failure.

C.2 Rubric warrant and criterion-level execution

A separate blinded audit covers 1,032 request-specific criterion instances. Experts first assess whether each fixed criterion is warranted and then answer it without seeing Vibrics's decision. They independently retain 97.4% and 97.6%, and both mark only three instances as unnecessary. Among 967 in-scope instances independently labeled by both experts, 843 receive determinate labels from both and 687 reach expert consensus. Of these, 640 have comparable Vibrics outputs; Vibrics agrees with the consensus on 546 instances (85.3%), with macro-$F_1 = 0.761$.

Table 12: Criterion-level agreement with the two-expert consensus. Comparable criterion instances exclude not-applicable decisions.

Audit scope	Comparable	Agreement	Macro-F_1
All comparable criterion instances	640	85.3%	0.761

Separately, among 792 criterion instances with determinate labels from both experts and comparable Vibrics outputs, Vibrics agrees with the two experts on 76.1% and 80.9% of instances, compared with 80.8% expert–expert agreement. The audit therefore supports the complete Judge pipeline on this scope, not any component's causal contribution.

C.3 Human ranking study and isotonic calibration

Two blinded experts rate every video in the 40-request subset on Fulfillment and Craft and independently provide an overall rating. Because candidates are grouped by request, the primary statistic is within-request Kendall τ-b averaged across requests. Table 13 reports human–human agreement; Table 15 gives the five-generator reliability analysis.

Table 13: Human–human reference agreement on the subset. Kendall τ-b is computed within each request and then averaged.

Axis	Requests	Human–human τ-b
Fulfillment	40	0.528
Craft	40	0.287
Utility	40	0.552

Cross-wave calibration on three axes. Human ratings never modify the rubric, the $0.6/0.4$ aggregation in Eq. 3, or reported F and C. On 20 Wave 2 requests (400 videos), we fit PAV maps from F, C, and U_0 to their corresponding independent human ratings. Subset requests are used only for evaluation. Predictions clip outside the observed calibration range; each map is non-decreasing, so it can merge unsupported differences into ties but cannot reverse input order. As a post-hoc sensitivity check on the subset, the mean request-level Kendall τ-b with human overall ratings is 0.406, 0.426, and 0.426 under F/C weights of $0.5/0.5$, $0.6/0.4$, and $0.7/0.3$, respectively. Thus, the observed alignment is stable around the predefined weighting.

Table 14 reports subset alignment under the same request-level statistic. F_{iso} and C_{iso} are auxiliary analyses; benchmark reporting remains uncalibrated F and C with calibrated U_{iso}.

Table 14: Cross-wave isotonic calibration on the subset. PAV maps are fitted on a disjoint calibration subset (400 videos) and evaluated on the 40-request subset. Values are mean request-level per-annotator-averaged Kendall τ-b.

Axis	Input	Human target	Requests	Raw	Isotonic
Fulfillment	F	Fulfillment	40	0.427	0.430
Craft	C	Craft	40	0.233	0.244
Utility	U_0	Overall	40	0.426	0.433

All 40 requests are retained.

Table 15 summarizes subset reliability for each score, reporting request-level Kendall τ-b, inter-annotator agreement (IAA), and their ratio on the subset. Fulfillment attains the highest ratio, and calibrated Utility recovers most of the inter-annotator agreement.

C.4 Automatic evaluator results and statistical intervals

Table 16 adds uncertainty intervals and paired differences to the main comparison. Text-conditioned baselines receive the actual candidate prompt supplied to the generator; VBench Quality receives video only. We report each evaluator's predefined official score rather than select a favorable subscore after observing human ratings.

Table 15: Subset reliability. Panel A reports request-level Kendall τ-b averaged over the two experts, human–human agreement (IAA), and their ratio on the subset. Panel B reports each expert separately, together with 95% confidence intervals from 2,000 bootstrap resamples clustered by request.

Panel A: Aggregate reliability

Score	Mean τ	IAA	Ratio
Fulfillment F	0.427	0.528	81%
Craft C	0.233	0.287	81%
Utility U_0	0.426	0.552	77%
Utility U_{iso}	0.433	0.552	78%

Panel B: Stability across experts on the five-generator panel

Score	Expert 1 τ	95% CI	Expert 2 τ	95% CI	Mean τ	95% CI	Gap	95% CI
Fulfillment F	0.443	[0.396, 0.491]	0.411	[0.345, 0.472]	0.427	[0.374, 0.477]	0.032	[−0.015, 0.080]
Craft C	0.271	[0.209, 0.329]	0.194	[0.129, 0.253]	0.233	[0.188, 0.277]	0.077	[−0.009, 0.162]
Utility U_0	0.441	[0.400, 0.477]	0.411	[0.364, 0.451]	0.426	[0.388, 0.459]	0.030	[−0.011, 0.070]
Utility U_{iso}	0.452	[0.410, 0.497]	0.414	[0.369, 0.461]	0.433	[0.395, 0.473]	0.038	[−0.004, 0.084]

Expert gap is Expert 1 minus Expert 2. All gap intervals include zero. All 40 requests are retained.

Table 16: Auxiliary comparison of the uncalibrated U_0 baseline with external evaluators. Values are mean request-level per-annotator-averaged Kendall τ-b on the subset. The U_0 difference is U_0 minus the baseline, with request-clustered 95% bootstrap intervals.

Evaluator	Input	Mean τ-b [95% CI]	U_0 difference [95% CI]	Interpretation
Vibrics-U_0	Request-specific rubric + video	0.426 [0.389, 0.462]	–	Unfitted auxiliary baseline.
Qwen-VL-Max holistic judge	Prompt + video	0.326 [0.265, 0.386]	0.100 [0.057, 0.147]	Closest prompt baseline.
CLIPScore	Prompt + video	−0.121 [−0.160, −0.083]	0.547 [0.484, 0.608]	Frame text–video similarity.
VQAScore	Prompt + video	−0.054 [−0.099, −0.009]	0.480 [0.424, 0.538]	Yes–No VQA score.
VideoScore v1.1	Prompt + video	−0.115 [−0.165, −0.061]	0.541 [0.486, 0.593]	Official overall score.
VBench Semantic	Prompt + video	−0.049 [−0.090, −0.004]	0.475 [0.414, 0.536]	Overall consistency.
VBench Quality	Video only	0.064 [0.009, 0.116]	0.362 [0.305, 0.418]	Quality overall.

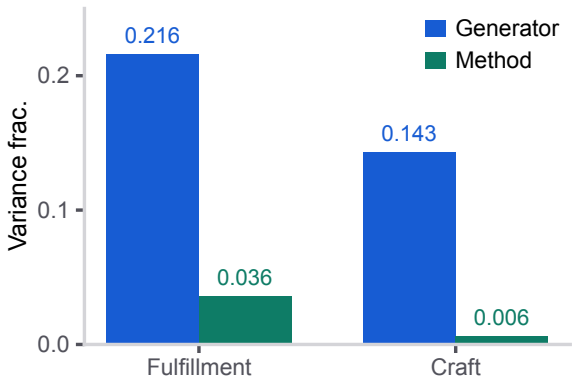

Figure 4: Held-out variance shares associated with generator and rewriter.

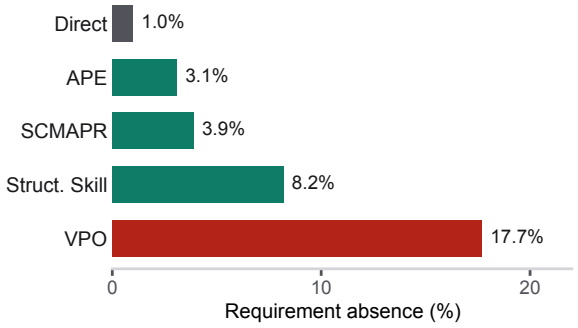

Figure 5: Prompt-stage requirement absence; Direct (1.0%) estimates the false-positive floor.

Additional VideoScore and VBench dimensions are released but omitted because they are not the designated overall scores. T2VScore-A/Q are absent because no verifiable official runnable scorer and weights were available at evaluation time.

D CROSSED-MATRIX STATISTICS

For each score, we decompose variation in the frozen 40-request five-generator matrix, with request as a blocking factor. Each reported share is a factor's sum of squares divided by total sum of squares and is descriptive, not causal. Pairwise intervals use request-clustered bootstrap samples that retain every candidate video for each sampled request.

Table 17 reports the numeric values underlying Figure 3. Among all 25 cells, Seedance × APE attains the highest U_{iso} (5.897) and LTX 2.5 × Direct the lowest (3.329). Relative to Direct, APE improves U_{iso} for every generator except HappyHorse, where Direct yields a marginally higher value.

Table 17: Numeric subset reference matrix underlying Figure 3. Cells report $F/C/U_{\text{iso}}$; F and C are uncalibrated in $[0, 1]$, while U_{iso} uses the human overall-rating scale.

Generator	Prompt rewriting method				
	Direct	Structured Skill	SCMAPR	VPO	APE
Seedance	.872 / .662 / 5.830	.850 / .668 / 5.722	.860 / .626 / 5.562	.811 / .630 / 5.282	.891 / .648 / 5.897
HappyHorse	.861 / .626 / 5.784	.830 / .618 / 5.414	.834 / .603 / 5.313	.815 / .609 / 5.337	.832 / .587 / 5.324
Wan 2.7	.810 / .543 / 5.052	.828 / .578 / 5.183	.815 / .518 / 4.897	.772 / .537 / 4.718	.859 / .580 / 5.503
MiniMax-H3	.765 / .606 / 4.908	.812 / .591 / 5.105	.811 / .606 / 5.289	.744 / .580 / 4.677	.857 / .630 / 5.630
LTX 2.5	.533 / .526 / 3.329	.677 / .519 / 4.150	.740 / .515 / 4.433	.693 / .530 / 4.168	.731 / .518 / 4.412

Generator-associated variation is 0.216 for Fulfillment and 0.143 for Craft, versus 0.036 and 0.006 for the rewriter. Pairwise intervals prevent treating every mean difference as resolved: APE separates from all alternatives on Fulfillment; Structured Skill and SCMAPR each separate from Direct but not from one another; VPO does not separate from Direct. Eight of ten rewriter contrasts for Craft include zero, and the two separated contrasts are at most 0.020. The matrix therefore supports APE's Fulfillment advantage, but not a broad Craft ranking among rewriters.

Table 18: Paired contrasts on the frozen 40-request crossed matrix. Intervals are request-clustered bootstraps; a filled marker denotes an interval excluding zero.

Pair	Δ	95% CI	Separated
Rewriting methods, Fulfillment			
APE – Direct	$+0.065$	$[+0.045, +0.086]$	•
APE – VPO	$+0.066$	$[+0.041, +0.094]$	•
APE – Structured Skill	$+0.034$	$[+0.015, +0.055]$	•
APE – SCMAPR	$+0.021$	$[+0.005, +0.037]$	•
SCMAPR – Direct	$+0.045$	$[+0.028, +0.064]$	•
SCMAPR – VPO	$+0.045$	$[+0.026, +0.066]$	•
SCMAPR – Structured Skill	$+0.013$	$[-0.005, +0.035]$	
Structured Skill – Direct	$+0.032$	$[+0.012, +0.051]$	•
Structured Skill – VPO	$+0.032$	$[+0.011, +0.056]$	•
VPO – Direct	-0.001	$[-0.022, +0.018]$	
Generators, Fulfillment			
Seedance – HappyHorse	$+0.024$	$[+0.012, +0.036]$	•
Seedance – Wan 2.7	$+0.037$	$[+0.023, +0.051]$	•
Seedance – MiniMax-H3	$+0.059$	$[+0.041, +0.077]$	•
Seedance – LTX 2.5	$+0.174$	$[+0.145, +0.200]$	•
HappyHorse – Wan 2.7	$+0.018$	$[+0.004, +0.032]$	•
HappyHorse – MiniMax-H3	$+0.037$	$[+0.019, +0.054]$	•
HappyHorse – LTX 2.5	$+0.160$	$[+0.131, +0.189]$	•
Wan 2.7 – MiniMax-H3	$+0.019$	$[-0.001, +0.037]$	
Wan 2.7 – LTX 2.5	$+0.142$	$[+0.110, +0.174]$	•
MiniMax-H3 – LTX 2.5	$+0.123$	$[+0.099, +0.148]$	•
Generators, Craft			
Seedance – HappyHorse	$+0.032$	$[+0.016, +0.049]$	•
Seedance – Wan 2.7	$+0.089$	$[+0.071, +0.109]$	•
Seedance – MiniMax-H3	$+0.039$	$[+0.019, +0.058]$	•
Seedance – LTX 2.5	$+0.120$	$[+0.100, +0.139]$	•
HappyHorse – MiniMax-H3	$+0.006$	$[-0.011, +0.025]$	
HappyHorse – Wan 2.7	$+0.059$	$[+0.044, +0.073]$	•
HappyHorse – LTX 2.5	$+0.086$	$[+0.067, +0.107]$	•
MiniMax-H3 – Wan 2.7	$+0.053$	$[+0.036, +0.070]$	•
MiniMax-H3 – LTX 2.5	$+0.080$	$[+0.063, +0.097]$	•
Wan 2.7 – LTX 2.5	$+0.028$	$[+0.009, +0.048]$	•

In summary, the pairwise contrasts support the main-text finding (§4.4): APE's Fulfillment advantage over every alternative is statistically resolved, whereas rewriter differences on Craft are negligible. Among generators, Seedance separates from all others on both scores, and LTX 2.5 is consistently last.

E Large-scale ranking and cross-request aggregation

This section expands Table 4 to the benchmark-scale automated evaluation over all five generators. Mean win rate averages, over opponents, the fraction of pairwise-available requests on which a system's per-request mean is higher. Bradley–Terry strengths are maximum-likelihood fits to the same request-level pairwise outcomes and are normalized to mean one. Mean rank is one plus the expected number of opponents ranked ahead under these pairwise outcomes.

The matrix in Table 19 reports the benchmark-scale automated results for every rewriter–generator combination, extending the human-validated reference matrix in Table 17.

Table 19: Benchmark-scale outcome matrix. Each cell reports mean $F/C/U_{\text{iso}}$ over all available automated evaluations for one rewriter and one generator.

Generator	Prompt rewriting method				
	Direct	Structured Skill	SCMAPR	VPO	APE
Seedance	.882 / .648 / 5.775	.858 / .650 / 5.646	.865 / .638 / 5.647	.806 / .639 / 5.291	.865 / .645 / 5.666
HappyHorse	.841 / .613 / 5.470	.842 / .622 / 5.493	.837 / .599 / 5.350	.776 / .617 / 5.074	.832 / .615 / 5.446
Wan 2.7	.800 / .534 / 4.870	.869 / .601 / 5.585	.843 / .557 / 5.218	.776 / .560 / 4.823	.861 / .607 / 5.585
MiniMax-H3	.752 / .598 / 4.829	.842 / .620 / 5.483	.834 / .592 / 5.319	.766 / .588 / 4.890	.827 / .636 / 5.452
LTX 2.5	.529 / .528 / 3.227	.702 / .510 / 4.211	.739 / .513 / 4.440	.693 / .525 / 4.170	.717 / .511 / 4.273

Aggregating these cells, Table 20 reports Bradley–Terry strength and mean rank for the rewriters and generators. The fixed APE condition attains the highest observed rank among the evaluated rewriters in both the subset and benchmark-scale analyses. Seedance leads the generators in both analyses, while LTX 2.5 ranks last.

Table 20: Bradley–Terry strength and mean rank on calibrated Utility. Held-out results use machine scores on the human-validated subset; benchmark-scale results use all pairwise-available automated evaluations.

	Held-out (machine)		Benchmark (machine)	
	BT strength	Mean rank	BT strength	Mean rank
Prompt rewriting methods				
APE	1.91	2.12	1.67	2.25
Structured Skill	1.02	2.83	1.32	2.51
SCMAPR	0.99	2.85	1.10	2.71
VPO	0.43	3.77	0.51	3.60
Direct	0.64	3.38	0.40	3.86
Video generators				
Seedance	3.00	1.51	2.67	1.68
HappyHorse	1.22	2.19	1.17	2.38
MiniMax-H3	0.38	3.22	0.54	3.07
Wan 2.7	0.38	3.22	0.59	3.00
LTX 2.5	0.03	4.84	0.03	4.88

Table 21 reports generator round-robin win rates on calibrated Utility. The resulting order is consistent across scopes: Seedance wins a majority against every generator, HappyHorse ranks second, and LTX 2.5 loses almost all comparisons.

Table 21: Generator round-robin Utility win rates (row over column, %). Top: benchmark-scale automated evaluation using all pairwise-available results; bottom: human-validated subset evaluation.

Benchmark	Wan 2.7	HappyHorse	MiniMax-H3	LTX 2.5	Seedance
Wan 2.7	–	34.9	50.0	95.8	19.3
HappyHorse	65.1	–	70.5	98.3	28.6
MiniMax-H3	50.0	29.5	–	94.8	19.2
LTX 2.5	4.2	1.7	5.2	–	0.8
Seedance	80.7	71.4	80.8	99.2	–
Held-out	**Wan 2.7**	**HappyHorse**	**MiniMax-H3**	**LTX 2.5**	**Seedance**
Wan 2.7	–	27.0	51.4	91.9	8.1
HappyHorse	73.0	–	75.7	97.3	33.8
MiniMax-H3	48.6	24.3	–	97.3	8.1
LTX 2.5	8.1	2.7	2.7	–	2.7
Seedance	91.9	66.2	91.9	97.3	–

F COMPLETE EVIDENCE RECORDS FOR TWO ARCHIVED RUNS

This appendix traces two archived runs from request to verdict. C01 pairs APE with MiniMax-H3 on a reference-conditioned sleeve-dance request and is broadly successful; C02 pairs Direct with LTX 2.5 on an automotive advertisement and misses several request-specific obligations. Each record includes the reviewed request, frozen rubric, candidate prompt, available process trace, complete video-side Judge record, and independent expert ratings. Prompt and process evidence is diagnostic only and never affects the scores.

F.1 C01: AGENT-REWRITTEN SLEEVE-DANCE REQUEST

Reviewed request and frozen rubric. The request asks for a 15-second, 16:9 sleeve-dance video grounded in supplied imagery. It specifies the dancer's clothing and water sleeves, the studio appearance, three dance actions, guzheng music, and a progression from a medium-wide studio view to a dancer view and sleeve close-up. Vibrics freezes these obligations as follows:

- **Format (`a9--a11`).** Deliver a 15-second, landscape 16:9, 720p video.
- **Style and reference (`a1--a3`).** Create a graceful classical sleeve-dance music video grounded in the supplied dancer and studio reference image.
- **Wardrobe and setting (`a4--a5`).** Show the specified light-blue top, black trousers, soft shoes, and green sleeves in a studio with wood flooring, grey walls, and white ballet-barre supports.
- **Action and camera (`a6--a7`).** Show cloud hands, circular walking steps, and a controlled turn; progress from the studio view to the dancer and then to the sleeve trajectory.
- **Music and detail (`a8, d1--d3`).** Use guzheng accompaniment, a graceful classical visual mood, and specific descriptions of the requested dance motions.

Complete candidate prompt. C01 uses APE to rewrite the request and MiniMax-H3 to render it. The complete archived prompt is:

> Create a 15-second classical sleeve-dance music video using the supplied reference image. In a bright dance studio with a wood floor, grey wall, and white-supported ballet barre, a young dancer wears a light-blue cross-collar frog-button top, black wide-leg trousers, flesh-toned soft shoes, and turquoise-green water sleeves. Begin with a level medium-wide view of the studio. To guzheng music, she performs cloud hands, circular walking steps, and a controlled turn; the sleeves form continuous, graceful arcs. The camera slowly dollies from the studio view to a medium view of the dancer and ends in a close view tracking the sleeve trajectory. Use soft side light and maintain a calm classical mood throughout.

Complete agent process (diagnostic). The six-turn trace reads the supplied plan, enters a reference-to-video branch, selects first-frame conditioning, develops the dance description, finalizes generation instructions, and writes the generation plan. It contains no separate visual-analysis or text-search action, while the final submission parameters specify 15 seconds, 16:9, and 720p. This trace is diagnostic evidence only.

Prompt and process trace (diagnostic). The prompt retains the subject, wardrobe, studio, actions, camera progression, and guzheng accompaniment, but realizes the requested three-view progression as one continuous shot. The diagnostic record preserves both the retained content and this structural change without altering video-side scores.

Complete archived video Judge record. Table 22 lists every non-universal video check for C01. PASS always denotes a satisfied requirement or absent defect, FAIL the reverse, and N/A an unavailable check. Table 24 reports the shared universal diagnostics.

Table 22: Complete non-universal video Judge record for C01. The route records the archived evaluation path.

ID	Frozen video question or check	Verdict	Route
Request fulfillment			
vq1	Are the dance motions graceful, fluid, and extended?	PASS	Video
vq2	Does the video present a classical sleeve-dance music-video style?	PASS	Video
vq3	Does the content remain focused on classical sleeve dance?	PASS	Video
vq4	Do the dancer and studio match the supplied reference image's core visual features?	PASS	Video
vq5	Does the dancer wear the specified light-blue top, black trousers, and green sleeves?	PASS	Video
vq6	Does the studio contain wood flooring, grey walls, and white-supported ballet barres?	PASS	Video
vq7	Are the requested wardrobe and studio details visually specific rather than generic?	PASS	Video
vq8	Is the light-blue cross-collar frog-button top visible?	PASS	Video
vq9	Are black wide-leg trousers visible?	PASS	Video
vq10	Are flesh-toned soft shoes visible?	PASS	Video
vq11	Are turquoise-green water sleeves visible?	PASS	Video
vq12	Are the requested colors and garment details clearly rendered?	PASS	Video
vq13	Is the scene recognizably a dance rehearsal studio?	PASS	Video
vq14	Is the floor visibly wood?	PASS	Video
vq15	Is the rear wall grey?	PASS	Video
vq16	Are white-supported ballet barres visible behind the dancer?	PASS	Video
vq17	Does the dancer perform cloud hands?	PASS	Video
vq18	Does the dancer perform circular walking steps?	PASS	Video
vq19	Does the dancer perform the requested controlled turn?	PASS	Video
vq20	Do the water sleeves produce flowing, continuous motion?	PASS	Video
vq21	Does the video begin with a medium-wide studio view?	PASS	Video
vq22	Does the camera move toward the dancer?	PASS	Video
vq23	Does the video contain a close view of the sleeve trajectory?	PASS	Video
vq24	Archived style check unavailable in the released record.	N/A	Video
vq25	Does the delivered video last 15 seconds?	PASS	Program
vq26	Does the delivered video use landscape 16:9?	PASS	Program
vq27	Does the delivered video use 720p resolution?	PASS	Program
Evidence enrichment			
vq1	Is the sleeve-dance performance clearly situated in the studio?	PASS	Video
vq2	Are wardrobe and studio details visually specific and legible?	PASS	Video
vq3	Is unrelated content absent from the reference-conditioned scene?	PASS	Video
vq4	Is the light-blue cross-collar top visible?	PASS	Video
vq5	Are black wide-leg trousers visible?	PASS	Video
vq6	Are flesh-toned soft shoes visible?	PASS	Video
vq7	Are turquoise-green water sleeves visible?	PASS	Video
vq8	Is wood flooring visible?	PASS	Video
vq9	Is a grey wall visible?	PASS	Video
vq10	Are white-supported ballet barres visible?	PASS	Video
Cinematic and execution checks			
vq1	Does the presentation progress from the studio view to the dancer and then a sleeve close-up?	PASS	Video
vq2	Does it use a continuous dolly toward the dancer?	PASS	Video
vq3	Does the wardrobe remain visually consistent through the progression?	PASS	Video
vq1	Does the opening establish the requested studio?	PASS	Video
vq2	Is the forward camera movement executed?	PASS	Video
vq3	Does a late sleeve-detail view appear?	PASS	Video
Technical checks			
vq1	Does the dancer remain anatomically coherent?	PASS	Video
vq2	Is the video temporally stable?	PASS	Video
vq3	Does the dancer's appearance remain consistent?	PASS	Video
vq4	Is the camera motion smooth?	PASS	Video

Table 22, continued.

ID	Frozen video question or check	Verdict	Route
Request-tied challenge checks			
cq1	Do the water sleeves remain extended and continuously flowing?	PASS	Targeted
cq2	Does the cross-collar frog-button style remain consistent?	PASS	Targeted
cq3	Is the sleeve close-up free of severe defocus or ghosting?	PASS	Targeted
Prompt–video faithfulness checks			
faith_entity	Does the delivered subject preserve the candidate prompt entity?	PASS	Video
faith_composition	Does the delivered video preserve the requested composition?	PASS	Video
faith_sequence	Does the delivered video preserve the requested sequence?	PASS	Video
faith_tone	Does the delivered video preserve the requested tone?	PASS	Video
faith_motion	Does the delivered video preserve the requested motion?	PASS	Video

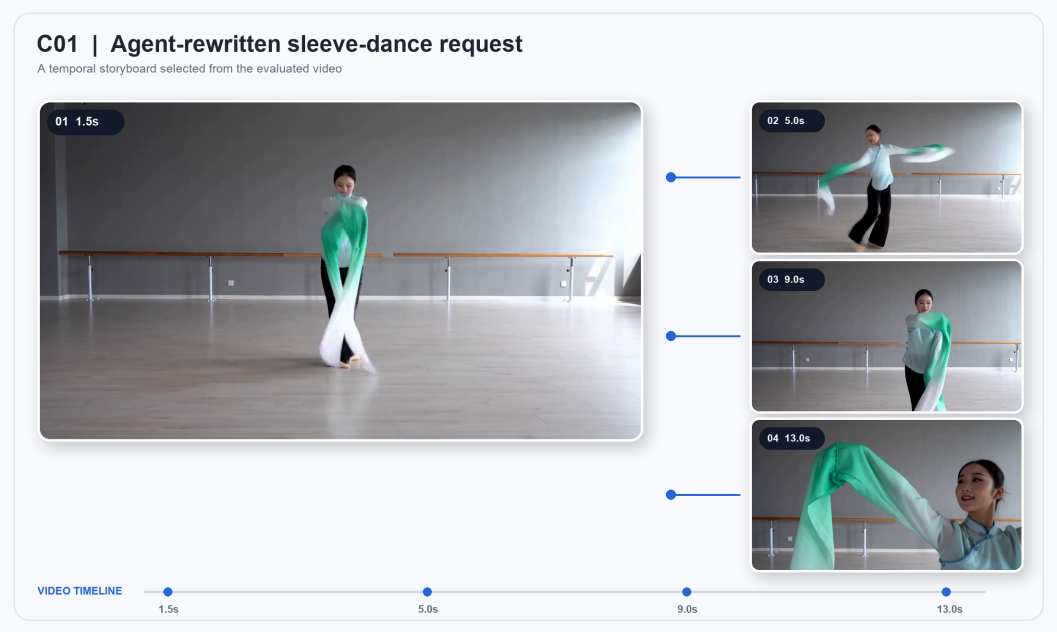

Figure 6: C01 storyboard from APE + MiniMax-H3, showing the requested studio, dancer, and sleeve-focused progression.

Diagnostic report. APE rewrites the request into a single continuous prompt, and the delivered video satisfies every request-specific obligation: the subject, wardrobe, studio, the three dance actions, the camera progression, and the format constraints all pass, so Fulfillment localizes no loss to the prompt, process, or video. The requested progression from a studio view to the dancer and then a sleeve close-up is realized as one continuous dolly rather than three distinct shots; the diagnostic record notes this as a structural choice at the rewriting stage that violates no frozen requirement. The residual issues are matters of visual quality rather than unmet requirements: the universal diagnostics flag occasional physically implausible motion, one less coherent transition, and fewer distinct shots than planned. The most important problem is therefore one of craft, not fulfillment—the requested performance and setting are delivered, while motion smoothness and shot structure leave room to improve.

F.2 C02: Direct automotive-ad request

Reviewed request and frozen rubric. The request asks for a 15-second, 16:9 BYD electric-vehicle advertisement with a cyberpunk city, drone-following view, range and acceleration HUD, tunnel light trails, and frozen logo ending. Vibrics freezes the following obligations:

- **Format (`a1--a2`).** Deliver a 15-second landscape 16:9 video.
- **Subject and world (`a3--a4`).** Depict a BYD electric vehicle in a technological, futuristic cyberpunk city.
- **Camera and selling point (`a5--a6`).** Use a drone-following view and present HUD data for range and 0–100 km/h acceleration.
- **Closing action (`a7--a8`).** Send the vehicle into a tunnel with light trails, then freeze on a BYD brand-logo ending.
- **Marketing intent (`d1--d2`).** Specify cyberpunk visual details and concrete range and acceleration values.

Complete Direct record. C02 uses Direct with LTX 2.5. Because Direct performs no rewriting, the candidate prompt is the complete reviewed request:

> Create a 15-second, landscape 16:9 advertisement for a BYD electric vehicle. Set the video in a futuristic cyberpunk city and use a drone view to follow the moving vehicle. Show a HUD that presents range and 0–100 km/h acceleration-time information. The vehicle enters a tunnel with light trails, and the video ends on a frozen BYD brand logo.

The record contains no agent trace, retrieval output, or rewritten expansion. Vibrics records this as a process-stage property, not a video failure.

Complete archived video Judge record. Table 23 lists every non-universal video check using C01's favorable verdict convention. Prompt–video faithfulness rows are diagnostic and do not override request-specific Fulfillment decisions; a PASS there does not imply that every request obligation is satisfied.

Table 23: Complete non-universal video Judge record for C02.

ID	Frozen video question or check	Verdict	Route
Request fulfillment			
`vq1`	Does the delivered video last 15 seconds?	PASS	Program
`vq2`	Does the delivered video use landscape 16:9?	PASS	Program
`vq3`	Is a BYD brand identifier visible?	**FAIL**	Video
`vq4`	Does the vehicle exhibit electric-vehicle characteristics?	**FAIL**	Video
`vq5`	Does the vehicle have recognizable BYD-specific appearance cues?	**FAIL**	Video
`vq6`	Are technological visual elements present?	PASS	Video
`vq7`	Does the setting convey a futuristic or science-fiction atmosphere?	PASS	Video
`vq8`	Is the background a cyberpunk-style city?	PASS	Video
`vq9`	Is a drone or aerial viewpoint used?	PASS	Video
`vq10`	Does the camera dynamically follow the vehicle?	PASS	Video
`vq11`	Is a HUD or comparable data-popup visual present?	**FAIL**	Video
`vq12`	Is range information visibly displayed?	**FAIL**	Video
`vq13`	Is 0–100 km/h acceleration information visibly displayed?	**FAIL**	Video
`vq14`	Does the vehicle enter a tunnel?	PASS	Video
`vq15`	Are vehicle or environmental lights rendered as trailing streaks?	PASS	Video
`vq16`	Does a BYD logo appear at the end?	**FAIL**	Video
`vq17`	Is the ending logo presented as a frozen image?	**FAIL**	Video
`vq18`	Is range performance communicated clearly?	**FAIL**	Video
`vq19`	Is acceleration performance communicated clearly?	**FAIL**	Video
Evidence enrichment			
`vq1`	Is rendered range text complete and legible?	N/A	OCR
`vq2`	Is rendered acceleration text complete and legible?	N/A	OCR
Cinematic and execution checks			
`vq1`	Does the video advance through city following, HUD display, tunnel light trails, and a logo ending?	**FAIL**	Video
`vq2`	Does the vehicle remain visually consistent across the requested sequence?	PASS	Video
`vq1`	Does the drone viewpoint follow the moving vehicle?	PASS	Video

Table 23, continued.

ID	Frozen video question or check	Verdict	Route
vq2	Does the HUD show both range and acceleration data?	**FAIL**	Video
vq3	Does tunnel entry include the required light-trail effect?	PASS	Video
vq4	Does the video end on a frozen BYD-logo frame?	**FAIL**	Video
Technical checks			
vq1	Is rendered range text structurally complete and readable?	N/A	OCR
vq2	Is rendered acceleration text structurally complete and readable?	N/A	OCR
vq3	Does the vehicle remain structurally consistent?	PASS	Video
vq4	Is the drone-following camera motion smooth?	PASS	Video
vq5	Do the city and tunnel retain coherent spatial structure?	PASS	Video
Request-tied challenge checks			
cq1	Are acceleration data, if a HUD is present, stable time values rather than percentages?	N/A	Targeted
cq2	Do the wheels remain naturally in contact with the ground?	PASS	Targeted
cq3	Does vehicle color and reflectance remain consistent across the drive?	PASS	Targeted
cq4	Does the aerial view dynamically follow rather than remain static?	PASS	Targeted
cq5	Do the license plate and brand mark remain visually consistent?	PASS	Targeted
Prompt–video faithfulness checks			
faith_entity	Does the delivered subject preserve the candidate prompt entity?	PASS	Video
faith_sequence	Does the delivered video preserve the requested sequence?	PASS	Video
faith_composition	Does the delivered video preserve the requested composition?	PASS	Video
faith_motion	Does the delivered video preserve the requested motion?	PASS	Video
faith_tone	Does the delivered video preserve the requested tone?	PASS	Video

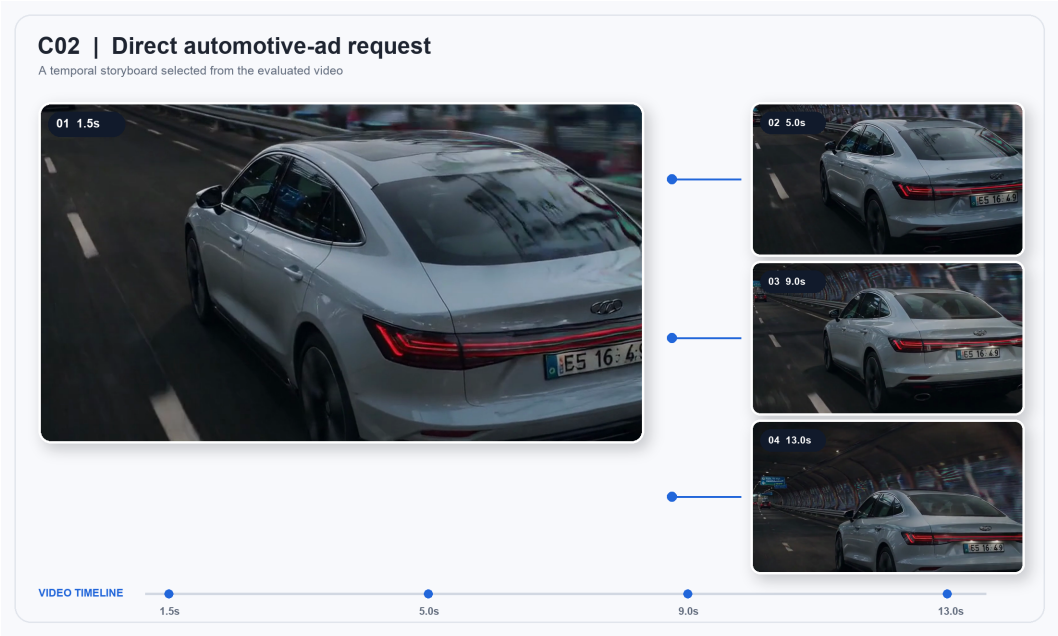

Figure 7: C02 storyboard from Direct + LTX 2.5. The vehicle-following shot is coherent, but the requested brand, electric-vehicle identity, HUD, and logo ending are absent.

Diagnostic report. Because Direct forwards the reviewed request unchanged, the candidate prompt preserves every stated obligation—the brand, the electric-vehicle identity, the drone-following view, the HUD range and acceleration data, the tunnel light trails, and the frozen-logo ending—and no process trace is available to add or remove content, so every unmet obligation localizes to the video stage. The delivered video renders a coherent cyberpunk city and a smooth

drone-following shot and enters a tunnel with light trails, fulfilling the requested world, camera, and mid-sequence action. It nonetheless fails to establish the requested subject: no BYD brand identifier is visible, the vehicle shows neither electric-vehicle nor BYD-specific appearance cues, and the closing frozen BYD logo never appears. It also omits the requested on-screen information: no HUD is rendered, so neither the range nor the 0–100 km/h acceleration value is displayed, and the challenge criterion for stable acceleration time values is consequently inapplicable. The single most important problem is that the advertisement's core payload is missing—the BYD electric-vehicle identity and its selling-point data are absent—even though the surrounding scene is visually competent. Because the prompt itself carried every requirement, the retained evidence attributes these losses to generation rather than to rewriting.

Shared universal diagnostics and independent expert ratings. Table 24 applies the same universal diagnostics to both videos; these checks do not independently contribute to Fulfillment, Craft, or Utility. Table 25 reports the independent expert ratings.

Table 24: Complete shared universal diagnostic record. Every check is phrased favorably: PASS means no named defect was found.

ID	Universal diagnostic check	C01	C02
uq_render_1	No thumbnails, previews, or unintended multi-panel layout enter the video.	PASS	PASS
uq_render_2	No excessive blank or solid background crowds out intended content.	PASS	PASS
uq_render_3	No black, white, or prolonged frozen screen occurs.	PASS	PASS
uq_render_4	No mosaic, tearing, severe compression, or color-block corruption occurs.	PASS	PASS
uq_subject_1	The core subject remains consistent across the video.	PASS	PASS
uq_subject_2	No human-body or face deformation occurs.	PASS	PASS
uq_physics_1	Motion remains physically plausible.	**FAIL**	PASS
uq_physics_2	Camera motion remains smooth and continuous.	PASS	PASS
uq_text_1	Any displayed text is readable and semantically valid.	PASS	PASS
uq_text_2	No unrelated watermark, platform mark, timestamp, or debug UI occurs.	PASS	PASS
uq_transition_1	Transitions remain coherent and non-repetitive.	**FAIL**	PASS
uq_color_1	Color, white balance, and lighting remain consistent within a scene.	PASS	PASS
uq_ai_texture	No conspicuous AI texture or malformed-object artifact occurs.	PASS	**FAIL**
uq_shot_count	The video contains the planned minimum number of shots.	**FAIL**	**FAIL**
uq_image_quality	No low-detail, blurred, or incompletely rendered region occurs.	PASS	PASS
uq_ending	The video reaches a complete ending rather than cutting off mid-content.	PASS	PASS
uq_edit_degeneracy	Editing degeneracy check.	N/A	N/A
audio_style_a8	Guzheng accompaniment check.	N/A	N/A

Table 25: Machine scores and independent expert ratings for the two complete records. F and C are uncalibrated; $U_0 = 0.6F + 0.4C$.

Case	F	C	U_0	Expert 1 Align	Craft	Overall	Expert 2 Align	Craft	Overall
C01	1.000	0.683	0.873	7	9	7	6	8	7
C02	0.581	0.652	0.609	2	2	0	3	4	1

Together, the records show what Vibrics adds beyond an overall score. C01 links a reference-conditioned rewrite to fulfilled obligations and residual physics, transition, and shot-structure defects. C02 separates a visually coherent car-following shot from failure to deliver the requested brand, vehicle type, information display, and ending. These cases illustrate evidence structure; they are not controlled system comparisons.

G EVIDENCE BOUNDARIES AND REPRODUCIBILITY

Vibrics separates evidence roles without treating them as causal stages. A prompt omission is diagnostic; Fulfillment changes only if the corresponding video criterion fails. Localized video evidence can establish an unmet obligation or reference mismatch and enter F. A rendering defect can enter F, C, or both only through its predefined criterion roles. Missing process traces never affect a score. These records identify where follow-up is needed without proving which component caused the failure.

Release scope. The release includes reviewed requests, frozen rubrics, criterion prompts and rules, machine records, anonymized human ratings, and aggregation and audit scripts. Rubric construction, routing, aggregation, and candidate outputs are versioned separately through the manifests in Appendix B.3. Exact denominators remain attached to each claim because score alignment, crossed-matrix analysis, the criterion audit, and report evaluation use distinct samples.

Anonymous code availability. An anonymized implementation and evaluation harness are available at `https://anonymous.4open.science/r/vibrics-523B/`.

Artifact inventory. Table 26 maps each claim to its released evidence. Before reuse, the request-bank version and rubric/Judge manifests must match, preventing changed prompts, evaluators, or candidate artifacts from being mixed with earlier records.

Table 26: Artifact inventory and the scope of public reproducibility.

Artifact	Reproducibility role	Claim or audit it supports
Versioned request bank and media manifest	Restores the reviewed request, attached media, and hard constraints.	All request-specific rubric and coverage claims.
Fixed rubric and VQA records	Restores each request requirement, criterion, decision rule, evidence-selection record, and channel.	Criterion-level execution and score-definition claims.
Candidate prompts, traces, and videos	Restores the artifacts evaluated under each rubric.	Matrix, retention, and case-study claims.
Judge records and evidence references	Restores video decisions, localized evidence, and diagnostic records.	Criterion-level audit, score alignment, and report traceability.
Anonymized per-video human ratings	Associates each released video identifier with its independent human ratings.	Human-score inspection and inter-annotator agreement audits.
Aggregation and audit scripts	Recomputes F/C/U and crossed-matrix statistics from released evaluator records.	Score-definition and matrix-statistics claims.

Boundaries. The criterion audit validates the complete Judge pipeline, not the causal value of each routing or localization component. Report evaluation tests whether the most important video problem is identified, not exhaustive coverage or causal attribution. Released human ratings support inspection and agreement audits but do not independently reconstruct condition-level alignment statistics. These limits define the scope of the claims.

H EXTENDED RELATED WORK: PROMPT REWRITING FOR VISUAL GENERATION

Prompt rewriting spans direct optimization such as Promptist (Hao et al., 2023), OPT2I (Mañas et al., 2024), and VPO (Cheng et al., 2025); retrieval-augmented optimization in RAPO (Gao et al., 2025a); multi-agent refinement in VISTA (Long et al., 2025) and SCMAPR (Yang et al., 2026); and video-grounded iteration in VQQA (Song et al., 2026). These methods improve generation inputs but do not define a common request-level standard for comparing videos across rewriters and generators. Vibrics-Bench supplies that standard.