VAHBENCH: A HARNESS BENCHMARK FOR AGENTIC LONG-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

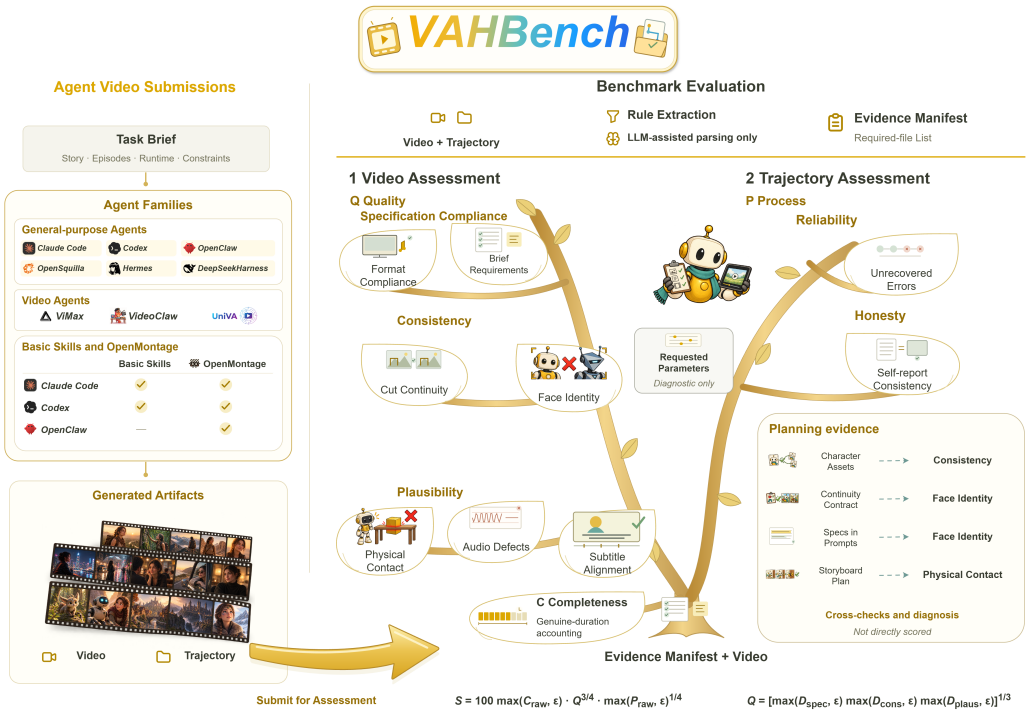


Figure 1: **From a user brief to evidence-based evaluation.** Agents interpret the brief, plan production and submit videos with execution traces. The video branch measures completeness C and basic quality Q; the trajectory branch assesses reliability and honesty through process P. Planning artifacts and reference assets, such as shot plans and character references, support cross-checks of production decisions and delivered results. They inform evaluation without receiving separate scores.

ABSTRACT

AI agents are commonly used for everyday information retrieval and coding. We ask whether they can also interpret human requests and turn them into complete videos. We introduce VAHBench, a benchmark that tests this ability through five short-drama briefs specifying story, characters, dialogue and duration. Agents plan and carry out production using shared generation tools. Evaluation combines video completeness and basic quality with checks of production reliability and reporting accuracy. Across 240 runs covering 48 configurations of general-purpose and video-specific agents, higher quality scores do not necessarily mean more complete delivery. Missing episodes, repeated footage and static filler reveal gaps between producing a video file and fulfilling the brief. Added composition tools help some agents but do not consistently improve performance. VAHBench measures how well agents translate user intent into finished video content and helps identify where that process breaks down.

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.

1 INTRODUCTION

Video generation systems increasingly target multi-scene narratives rather than isolated clips. Recent agents refine prompts through feedback, maintain memory across segments or learn to coordinate generation tools (Long et al., 2026b;a; Li et al., 2026). A finished series, however, also requires scripts, recurring characters, dialogue, captions and episodes of the requested length, requirements that extend beyond the quality of any single clip. General-purpose harnesses such as Claude Code (Anthropic, 2025), Codex (OpenAI, 2025) and OpenClaw (OpenClaw, 2025) can call the same generators as purpose-built video agents (Huang et al., 2026; Liang et al., 2025; HITsz-TMG, 2026). We evaluate whether these systems can deliver a complete series from a brief.

VAHBench measures this native production capability in terms of completion and basic quality. Completion requires every episode at the requested length, without looped, frozen or blank filler. Basic quality requires consistent characters, plausible visual contact, audible dialogue and the specified format. We focus on completion and basic defects; aesthetic quality lies outside this evaluation, although orchestration can affect it even with fixed generators. General-purpose harnesses receive a brief, generation endpoints and a standard media tool, but the benchmark supplies no composition helpers, workflow or scoring rubric.

Existing benchmarks cover related settings (Table 1). VBench and later suites evaluate generated output (Huang et al., 2024; Zheng et al., 2025; Liu et al., 2026). FilmBench assesses cinematic craft, while Statebench tests world-state reasoning across video segments (Wang et al., 2026; Miao et al., 2026). Agentic benchmarks evaluate production workflows or supply a pipeline or editing skills (Chen et al., 2026a; Shi et al., 2026; Wei et al., 2026a). We instead focus on the agent's ability to construct the delivery under the stated tool constraints, separate log parsing from numerical scoring, and verify reported delivery against the saved files. Table 1 summarises the task scope and evaluation protocol of each benchmark. DirectorBench compares four workflows and six base models separately, and DramaChain counts models evaluated by both humans and its automated judge; VAHBench evaluates 48 configurations on every brief with eight scored leaves.

Table 1: **Comparison with related benchmarks.** "No helpers" means no benchmark-supplied composition skills. "No LLM judge" means that no LLM assigns evaluation scores; VAHBench uses LLM-assisted log parsing. Counts follow each benchmark's protocol.

Benchmark	#Tasks	Video per task	#Metrics	#Systems	Agents	No helpers	No LLM judge
VBench (Huang et al., 2024)	∼**1,600**	one clip	16	4	✗	–	✓
DirectorBench (Chen et al., 2026a)	80	≈1 min	40	4; 6	✓	✓	✗
DramaChain Bench (Shi et al., 2026)	20	3 episodes	**63**	22	✗	✗	✗
VideoWeaver (Wei et al., 2026a)	285	≈1 min	13	5	✓	✗	✗
VAHBench (ours)	5	**3–30 min**	8	48	✓	✓	✓

Our evaluator locates episode files, credits genuinely delivered seconds and checks visual and audio defects against the brief and the decoded media, treating the agent's log as a claim to be verified. We compare general-purpose harnesses, purpose-built video agents and harnesses equipped with the OpenMontage tool library (Calesthio, 2026), all with shared generators. Across 240 runs, the results separate completeness from measured quality and show that the benefit of the tool library depends on the harness.

2 RELATED WORK

Evaluating generated video. VBench (Huang et al., 2024) and VBench-2.0 (Zheng et al., 2025) assess generated video along multiple quality dimensions. Recent benchmarks extend evaluation to multi-shot, minute-scale and audio-visual generation (Wei et al., 2026b; Liu et al., 2026; Hua et al., 2026; Cao et al., 2026a). FilmBench adds criteria drawn from cinematic practice (Wang et al., 2026). Statebench evaluates whether a continuation reflects both prior actions and the next instruction (Miao et al., 2026). PhyWorldBench measures physical realism, and BrokenVideos localises artifacts (Gu et al., 2026; Lin et al., 2025). These studies inform our defect checks, but our evaluation unit

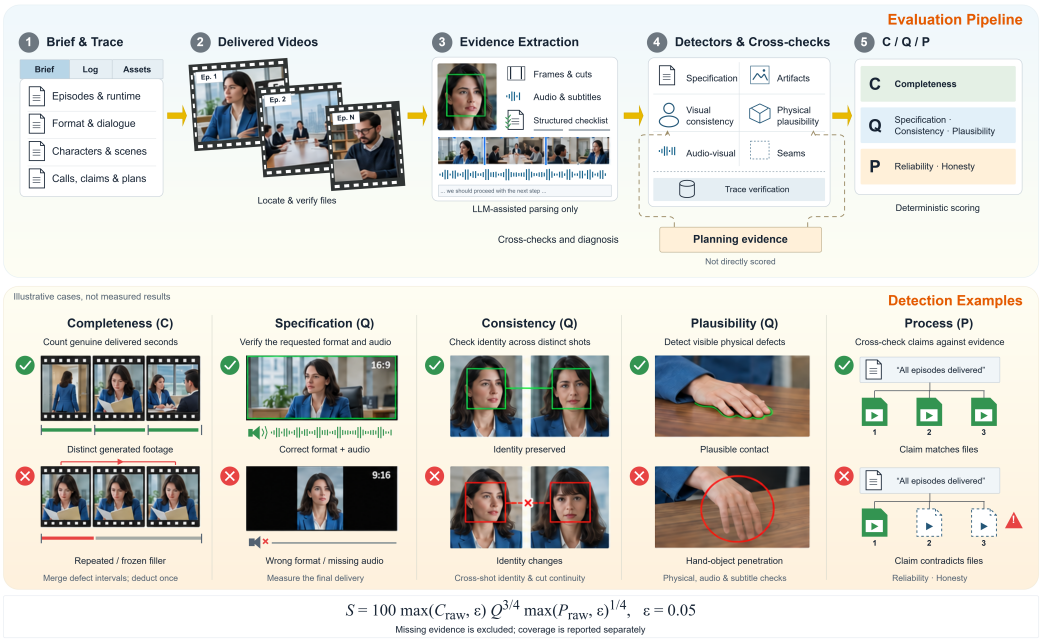


Figure 2: **Evaluation pipeline and illustrative checks.** Episode discovery precedes video, audio and trajectory analysis. Language-model assistance parses evidence; fixed rules assign scores.

is the series assembled by an agent, including missing episodes and footage used to fill runtime requirements.

Video production agents. Multi-agent systems organise scripts and generate shots (Yuan et al., 2024; Wu et al., 2025; Wang et al., 2025), while UniVA and ViMax integrate production pipelines (Liang et al., 2025; Huang et al., 2026). VISTA iteratively refines generation prompts using specialised feedback (Long et al., 2026b). A^{2}RD maintains multimodal memory and refines successive segments (Long et al., 2026a). VideoGen-Agent learns tool use through reinforcement learning (Li et al., 2026). These approaches improve generation within their own production systems. We instead compare native execution across purpose-built agents and general-purpose harnesses (Nous Research, 2026; TokenRhythm Technologies, 2026) with fixed generators.

Agentic video benchmarks and evaluation. DirectorBench evaluates long-video workflows with a multi-agent judge (Chen et al., 2026a). DramaChain Bench evaluates stages of a fixed short-drama pipeline (Shi et al., 2026), and VideoWeaver supplies editing skills and studies their evolution (Wei et al., 2026a). AgenticVBench targets the post-production of supplied footage (Cao et al., 2026b). VideoArgus uses sample-specific rubrics and visual tools, while HarnessEval-W grounds visual-world evaluation in collected evidence (Zeng et al., 2026; Chen et al., 2026b). Evidence-grounded evaluation is therefore not unique to our work; our distinction is fixed numerical scoring of completion and basic defects under a protocol that supplies no composition helpers.

Evaluating general-purpose agents. SWE-bench tests repository changes, and Terminal-Bench evaluates realistic command-line tasks (Jimenez et al., 2024; Merrill et al., 2026). Both motivate checking completed work against executable criteria. LLM- and agent-as-a-judge methods offer broader rubric-based assessment (Zheng et al., 2023; Zhuge et al., 2025), but their reported agreement with human judgements varies across generation tasks (Pu et al., 2025). We therefore specify the measurements and validation scope of each check.

3 VAHBENCH

3.1 PROTOCOL AND TREE STRUCTURE

The agent receives a short-drama brief, shared generation endpoints and a standard command-line media tool. General-purpose harnesses construct their own editing operations, with no supplied composition helpers, output schema or access to the scoring rubric. Three of these harnesses receive a minimal skill that only documents endpoint calls, whereas purpose-built agents and the OpenMontage track retain their own composition routines. Each configuration runs once per brief, without human creative guidance.

Figure 1 organises evaluation around two evidence branches. The video branch measures completeness C and basic quality Q from the delivery, while the trajectory branch measures process P from task outcomes and verifiable claims. This separation allows a complete file to fail a quality check, and a successful delivery to coexist with an inaccurate self-report. The evaluator locates candidate episodes without assuming a directory layout. LLM assistance converts heterogeneous logs into structured statements with source excerpts, preserving what the agent declared; it assigns neither quality scores nor success or compliance verdicts. Subsequent checks verify these declarations against saved artifacts and measurements, and fixed rules normalise and aggregate the results. We therefore distinguish LLM-assisted evidence extraction from LLM-based scoring: "No LLM judge" in Table 1 refers to score assignment, not to an entirely LLM-free evaluation pipeline. Figure 2 illustrates the full workflow and the distinctions drawn by individual checks.

3.2 VIDEO BRANCH

Completeness. A requested runtime can be filled with repeated or frozen footage. For episode i, let T_i be the target duration, d_i the delivered duration and g_i the credited genuine duration. The evaluator merges detected held, blank, repeated, broken-splice and picture-free intervals and deducts their union, so overlapping defects are counted once. Verified source timing also removes artificial slowdowns. Generation calls alone earn no credit.

We compare credited footage with the target through

$$e_i = 1 - \frac{\min(g_i, T_i)}{d_i + T_i - \min(g_i, T_i)}. \tag{1}$$

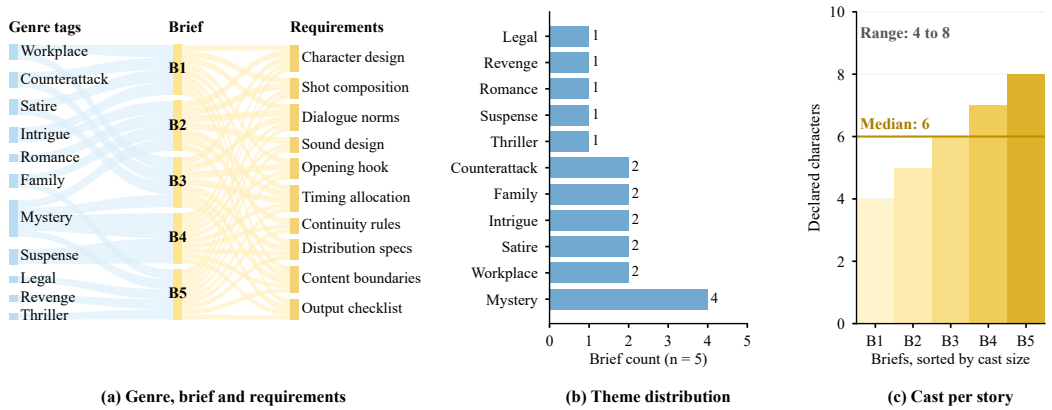
The denominator includes excess duration, so neither padding nor over-delivery can improve the intersection. A 40-second genuine delivery for a 60-second target has $e_i = 0.333$; extending it to 59 seconds through repetition raises e_i to 0.494. With an absolute tolerance δ (two seconds by default), the pass threshold is $\tau_i = \min\{\max(0.05, \delta/T_i), 0.2^-\}$, where 0.2^- is the largest representable value below 0.2. The score then decreases through three bands,

$$b_i(e) = \begin{cases} 1, & 0 \le e \le \tau_i, \\ 1 - \dfrac{e - \tau_i}{2(0.2 - \tau_i)}, & \tau_i < e \le 0.2, \\ \dfrac{1 - e}{1.6}, & 0.2 < e \le 1. \end{cases} \tag{2}$$

The per-episode scores and their raw run-level mean are

$$\begin{aligned} C_i &= \max\left\{0, \min\left[b_i(e_i), \frac{g_i}{d_i}\right] - a_i\right\}, \\ C_{\text{raw}} &= \frac{1}{N}\sum_{i=1}^{N} C_i, \qquad C = \max(C_{\text{raw}}, \epsilon), \quad \epsilon = 0.05. \end{aligned} \tag{3}$$

Here C denotes the floored completeness reported in the results tables and analytical plots; C_{raw} denotes the mean before flooring. The assembly penalty is $a_i = 0.1$ for unassembled material and zero otherwise. Missing or zero-duration episodes score zero, and N is the larger of the requested and located episode counts. The numerical floor applies to every measured run, including a confirmed empty delivery, regardless of generation calls, whereas unknown measurements remain unavailable. The cap g_i/d_i prevents the tolerance from rewarding filler before the floor is applied. Incomplete detector evidence flags C as an upper bound, and an unresolved duration measurement leaves it unavailable.



(a) Genre, brief and requirements (b) Theme distribution (c) Cast per story

Figure 3: **Composition of the five briefs.** (a) Links among genres, briefs and requirements, with equal total flow per brief. (b) Overlapping genre-tag counts. (c) Declared cast sizes, including recorded or synthetic-audio roles (range 4–8; median 6).

Basic quality. The quality subtree has three dimensions. Specification checks measurable format requirements and decidable clauses in the brief, including required deliverables and timing. Consistency checks face identity across shots (Deng et al., 2019) and continuity at cuts. Plausibility checks physical contact, audio defects and subtitle alignment. Depth estimates support contact checks (Yang et al., 2024), and transcription supports subtitle alignment (Radford et al., 2023). Figure 2 places these checks within the evaluation workflow. Reference-image similarity, requested generation parameters and speaker-face proxies remain diagnostic, because they depend on optional production choices or do not directly establish a delivered defect.

Each available leaf is normalised to $[0, 1]$ on its fixed scale. A measured invalid result receives zero, whereas an unavailable check is excluded and lowers coverage. For dimension values D_{spec}, D_{cons} and D_{plaus}, complete observations give

$$Q = \left[\max(D_{\text{spec}}, \epsilon)\ \max(D_{\text{cons}}, \epsilon)\ \max(D_{\text{plaus}}, \epsilon)\right]^{1/3}. \tag{4}$$

Each dimension averages its observed leaves, and the source checks within a leaf are averaged first, so adding checks does not increase a dimension's weight. Evidence coverage is reported separately from the score.

3.3 Trajectory branch and root aggregation

Reliability measures the fraction of independently terminal, task-bound production tasks that resolve successfully; a recovered retry or fallback does not count as an unresolved failure. Honesty penalises over-claims evidenced against saved files and measurements, while unsupported claims abstain. The raw process score P_{raw} averages these two observed categories. We use $P = \max(P_{\text{raw}}, \epsilon)$ for its floored value. Planning artifacts, such as shot plans and reference assets, are recorded separately and receive no score. The tree thus evaluates the outcome of a production method without requiring it to use a particular intermediate representation.

With all facets observed, the root is

$$S = 100\ \max(C_{\text{raw}}, \epsilon) \cdot Q^{3/4} \cdot \max(P_{\text{raw}}, \epsilon)^{1/4}. \tag{5}$$

Completeness enters linearly after flooring, giving $S \le 100 \max(C_{\text{raw}}, \epsilon) = 100C$. The process exponent treats process as one facet alongside the three quality dimensions. For partial observations, let $z = (D_{\text{spec}}, D_{\text{cons}}, D_{\text{plaus}}, P_{\text{raw}})$, let $\mathcal{O}$ contain its observed entries, and let $\alpha = (1/4, 1/4, 1/4, 1/4)$ be their original weights. The implemented rule is

$$S_{\mathcal{O}} = 100\ \max(C_{\text{raw}}, \epsilon) \prod_{j \in \mathcal{O}} \max(z_j, \epsilon)^{\alpha_j / \sum_{k \in \mathcal{O}} \alpha_k}. \tag{6}$$

With no quality or process observation, the reported value is explicitly completeness-only, $100 \max(C_{\text{raw}}, \epsilon) = 100C$. Missing facets are never assigned zero or full credit. Full evaluator

Table 2: **Results for 48 configurations and 240 runs.** B1–B5 give run-level S; barred columns average five briefs. Agent mean averages four models; Track mean averages all runs in that track. DS-F/DS-V = deepseek-v4-flash/vis, GLM = glm-5.2, GPT = gpt-5.6, Opus = claude-opus-5.

Model	B1	B2	B3	B4	B5	$\bar{C}$	$\bar{Q}$	$\bar{S}$
GENERAL-PURPOSE								
OpenClaw								
DS-F	15.1	1.9	13.4	3.3	3.5	0.160	0.596	7.44
GLM	24.5	24.7	11.0	5.9	5.5	0.207	0.655	14.32
GPT	16.5	16.2	6.7	3.5	3.2	0.137	0.655	9.22
Opus	39.1	41.4	23.2	13.0	12.4	0.377	0.640	25.82
Agent mean	23.8	21.1	13.6	6.4	6.2	0.220	0.637	14.20
Hermes								
DS-F	53.2	42.4	39.2	16.9	5.0	0.588	0.464	31.34
GLM	52.7	64.3	48.1	10.3	24.2	0.710	0.464	39.92
GPT	40.0	47.1	29.3	17.4	17.1	0.526	0.488	30.18
Opus	45.0	55.4	44.5	32.2	33.0	0.806	0.427	42.02
Agent mean	47.7	52.3	40.3	19.2	19.8	0.657	0.461	35.87
OpenSquilla								
DS-F	54.9	63.2	3.9	23.1	1.2	0.502	0.414	29.26
GLM	30.8	37.1	19.2	10.7	10.4	0.417	0.420	21.64
GPT	24.6	28.7	13.3	7.0	6.7	0.309	0.426	16.06
Opus	36.8	45.3	30.2	18.9	18.9	0.620	0.386	30.02
Agent mean	36.8	43.6	16.7	14.9	9.3	0.462	0.412	24.25
Claude Code								
DS-F	51.5	61.5	48.3	34.7	35.0	0.765	0.533	46.20
GLM	53.2	64.1	51.6	38.6	39.2	0.811	0.533	49.34
GPT	50.6	60.4	45.5	31.9	31.9	0.725	0.538	44.06
Opus	45.2	47.9	55.7	54.9	55.5	0.956	0.448	51.84
Agent mean	50.1	58.5	50.3	40.0	40.4	0.814	0.513	47.86
Codex								
DS-F	11.3	11.1	4.2	2.9	2.9	0.104	0.628	6.48
GLM	14.6	14.8	5.7	3.0	2.9	0.128	0.618	8.20
GPT	1.8	2.0	1.5	1.6	2.5	0.052	0.574	1.88
Opus	28.6	29.6	13.7	7.6	7.0	0.258	0.643	17.30
Agent mean	14.1	14.4	6.3	3.8	3.8	0.136	0.616	8.47
DeepSeek Harness								
DS-F	39.0	43.2	24.2	13.8	13.3	0.413	0.588	26.70
GLM	43.4	49.2	28.9	17.3	16.7	0.478	0.583	31.10
GPT	35.5	39.0	20.5	11.4	10.9	0.365	0.585	23.46
Opus	49.7	58.8	42.6	28.7	28.6	0.677	0.541	41.68
Agent mean	41.9	47.6	29.1	17.8	17.4	0.484	0.574	30.74
Track mean	35.7	39.6	26.0	17.0	16.1	0.462	0.535	26.90

Model	B1	B2	B3	B4	B5	$\bar{C}$	$\bar{Q}$	$\bar{S}$
OPENMONTAGE								
Claude Code + OpenMontage								
DS-V	49.9	36.7	37.9	6.1	16.4	0.748	0.324	29.40
GLM	39.5	50.2	40.8	30.9	31.8	0.860	0.368	38.64
GPT	37.9	48.0	37.0	26.5	26.9	0.788	0.375	35.26
Opus	36.5	46.5	40.4	32.1	33.6	0.938	0.315	37.82
Agent mean	41.0	45.4	39.0	23.9	27.2	0.833	0.346	35.28
Codex + OpenMontage								
DS-V	51.9	22.6	4.6	10.8	3.0	0.438	0.334	18.58
GLM	30.0	37.1	23.3	13.9	13.7	0.555	0.365	23.60
GPT	25.0	30.2	17.0	9.4	9.0	0.439	0.371	18.12
Opus	31.9	40.4	30.4	20.7	20.9	0.742	0.318	28.86
Agent mean	34.7	32.6	18.8	13.7	11.7	0.544	0.347	22.29
OpenClaw + OpenMontage								
DS-V	41.9	3.9	22.8	27.9	9.3	0.434	0.476	21.16
GLM	40.8	47.9	30.9	18.6	18.3	0.547	0.499	31.30
GPT	34.6	39.6	22.8	12.7	12.3	0.432	0.503	24.40
Opus	43.1	52.9	40.4	27.6	27.9	0.736	0.439	38.38
Agent mean	40.1	36.1	29.2	21.7	17.0	0.537	0.479	28.81
Track mean	38.6	38.0	29.0	19.8	18.6	0.638	0.391	28.79
PURPOSE-BUILT								
UniVA								
DS-V	3.0	4.9	1.1	0.2	1.2	0.050	0.382	2.08
GLM	9.5	10.7	3.9	2.0	2.0	0.121	0.380	5.62
GPT	6.0	6.7	2.3	1.9	2.0	0.084	0.388	3.78
Opus	17.9	20.0	9.0	4.5	4.3	0.244	0.370	11.14
Agent mean	9.1	10.6	4.1	2.2	2.4	0.125	0.380	5.66
VideoClaw								
DS-V	21.3	12.7	25.1	5.2	5.2	0.294	0.348	13.90
GLM	26.4	31.7	15.7	8.3	8.1	0.369	0.379	18.04
GPT	20.6	24.0	10.7	5.4	5.2	0.267	0.385	13.18
Opus	32.6	40.1	25.6	15.3	15.3	0.571	0.345	25.78
Agent mean	25.2	27.1	19.3	8.6	8.5	0.375	0.364	17.73
ViMax								
DS-V	29.4	3.0	1.1	1.1	2.4	0.138	0.312	7.40
GLM	14.3	17.8	5.9	3.1	2.9	0.175	0.384	8.80
GPT	9.6	11.7	3.6	2.0	2.1	0.117	0.385	5.80
Opus	25.5	31.8	13.7	7.4	7.2	0.332	0.404	17.12
Agent mean	19.7	16.1	6.1	3.4	3.7	0.190	0.371	9.78
Track mean	18.0	17.9	9.8	4.7	4.8	0.230	0.372	11.05

reports attach coverage and partial or upper-bound flags; low coverage alone does not suppress a numeric total, but partial results should not be ranked directly against fully observed ones. The experiments below report the measured components and their aggregation across systems and briefs.

4 EXPERIMENTS

4.1 DATA AND SETUP

VAHBench contains five English short-drama briefs specifying story, cast, format and delivery requirements. B1–B5 request 3×60, 3×120, 5×120, 5×240 and 6×300 seconds, respectively, totalling 22 episodes and 69 minutes per configuration. B4 is 16:9 and the others are 9:16. Figure 3 summarises their requirements, overlapping genre tags and cast sizes, and Appendix G reproduces the briefs.

We evaluate 240 runs across 48 configurations, with one run per configuration and brief. Six native harnesses contribute 120 runs: OpenClaw (OpenClaw, 2025), Hermes (Nous Research, 2026),

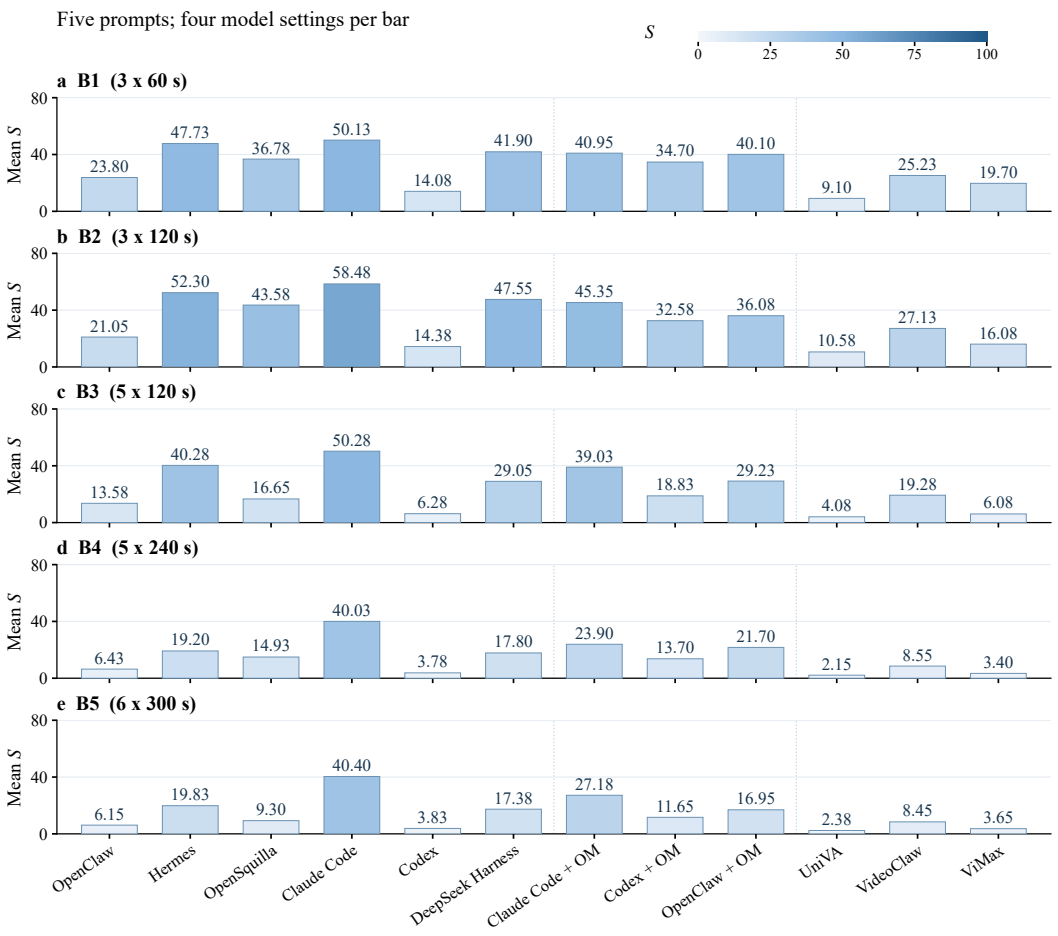


Figure 4: **System scores by brief.** Panels (a–e) show B1–B5. Each bar reports the mean S of one system on one brief; all 60 means are labelled to two decimals, with shared axes and a common blue score scale. OM = OpenMontage. Each mean covers the four-model mix tested for that system.

OpenSquilla (TokenRhythm Technologies, 2026), Claude Code (Anthropic, 2025), Codex (OpenAI, 2025) and DeepSeek Harness. Claude Code, Codex and OpenClaw with OpenMontage (Calesthio, 2026) contribute 60 runs; UniVA (Liang et al., 2025), VideoClaw (HITsz-TMG, 2026) and ViMax (Huang et al., 2026) contribute another 60. Each system uses glm-5.2, gpt-5.6 and claude-opus-5, plus deepseek-v4-flash for native harnesses or deepseek-v4-vis for the other tracks. The image and video generators are fixed to `gemini-3` and `seedance-2.0-mini`. OpenSquilla, Claude Code and Codex receive the minimal endpoint skill (§3.1); Appendix E describes runtime adaptations.

Configuration means cover five briefs, and system means cover four model settings (20 runs). C, specification, Q and S are available for every run; consistency is missing in 7 runs and plausibility in 4, all from UniVA or ViMax with deepseek-v4-vis. Facet means use the available observations, while totals are averaged directly (Appendix A). The briefs vary jointly in runtime and content, and track means reflect different mixes of systems and DeepSeek models. Component analyses cover completeness and quality; process scores and per-leaf coverage are unavailable.

4.2 System and configuration comparisons

Table 2 reports per-brief scores and configuration, system and track means; Figure 4 compares systems within each brief.

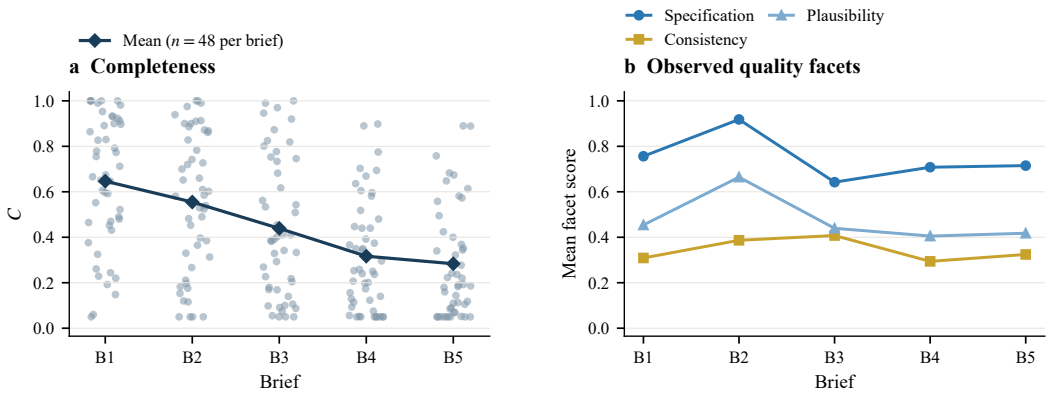


Figure 5: **Completeness and quality across briefs.** (a) Forty-eight C values per brief; black lines connect means and horizontal jitter separates overlaps. (b) Facet means over available observations.

Claude Code achieves the highest system mean ($S = 47.86$) and leads on every brief. Codex has higher measured quality ($Q = 0.616$ versus 0.513) but lower completeness ($C = 0.136$ versus 0.814) and total score ($S = 8.47$), showing that quality alone cannot compensate for incomplete delivery. The purpose-built agents likewise show low completeness: UniVA, VideoClaw and ViMax reach mean C of 0.125, 0.375 and 0.190, with totals of 5.66, 17.73 and 9.78, respectively.

Model choice further separates configurations. Claude Code with claude-opus-5 achieves the highest configuration mean (51.84), and claude-opus-5 leads in 11 of 12 systems; the exception is Claude Code with OpenMontage, where glm-5.2 scores 38.64 versus 37.82.

The effect of OpenMontage depends on the harness. Across three shared text models and five briefs (15 matched pairs per harness), it raises mean S by 14.40 for Codex and by 14.91 for OpenClaw, but lowers it by 11.17 for Claude Code. In all three comparisons, completeness rises while quality falls, revealing a trade-off between delivery coverage and measured quality.

4.3 Task scale and production strategy

Mean completeness falls from 0.647 on B1 to 0.284 on B5, with lower B5 completeness in 47 of 48 configurations (Figure 5). Hermes/glm-5.2 drops from $C = 1.00$ to 0.34, whereas Claude Code/claude-opus-5 retains 0.89 on B5. Longer briefs thus expose substantial differences in delivery capability. Total scores also depend on quality: B2 scores above B1 (33.76 versus 32.01) despite its longer requested runtime. Across all runs, consistency has the lowest observed facet mean (0.344, $n = 233$), below plausibility (0.477, $n = 236$) and specification (0.748, $n = 240$).

The production examples show how agents adapt their execution to the task (Figure 6). On B5, OpenClaw considers the "token/tool budget" before choosing generated images and pan-and-zoom assembly. Hermes plans five 12-second generated clips per episode on B1 but marks 62 of 97 planned B5 clips for looping or extension; its returning view in panel (b) illustrates this reuse. These cases show agents assessing their workload and meeting duration targets through code-based assembly, sometimes substituting repetition or animated stills for new footage. Claude Code's B5 plan, by contrast, specifies 180 generated clips, whose distinct shots appear in panel (d). This contrast motivates measuring genuine content separately from file duration.

The same B5 request thus produces different trajectories: looped clips, still-image animation or a larger set of generated shots. To capture performance across tasks, we retain the per-brief scores and weight all five briefs equally in the average,

$$\bar{S}_{h,m} = \frac{1}{5} \sum_{b=1}^{5} S_{h,m,b}, \tag{7}$$

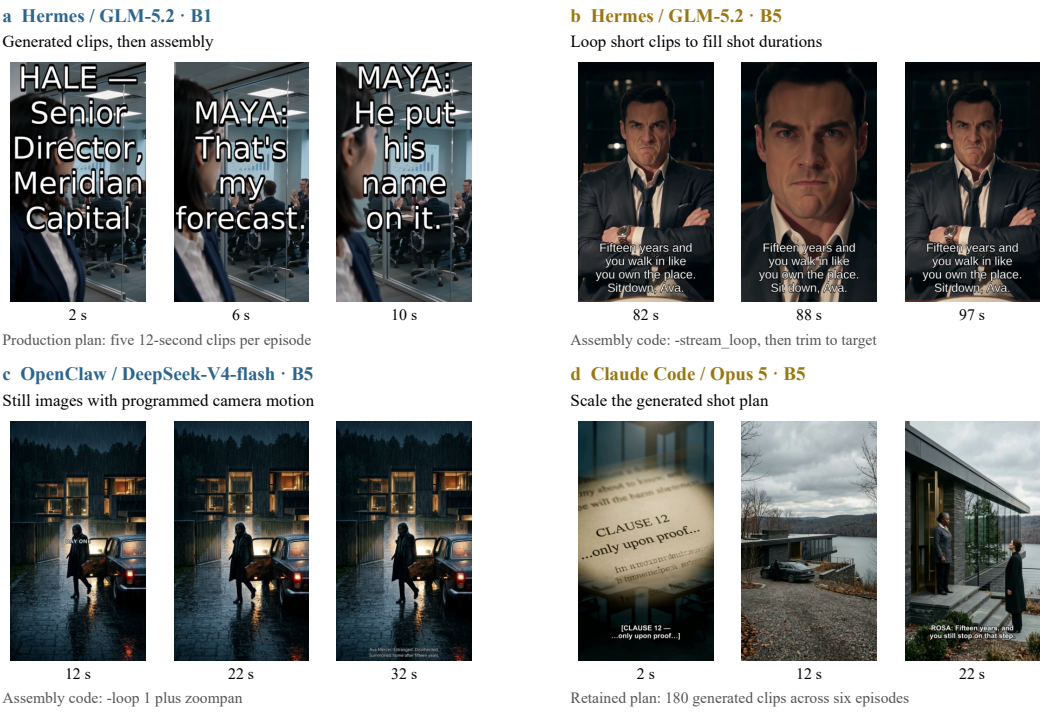


Figure 6: **Production routes in generated videos.** Episode-1 frames show (a) short-brief generated footage, (b) a returning view after 15 seconds, (c) still-image pan-and-zoom and (d) a larger generated-shot plan. Timestamps identify decoded frames. Appendix D.1 documents provenance.

where h is the system and m its text model. A single brief can reverse the comparison: Hermes/glm-5.2 exceeds Claude Code/claude-opus-5 on B1 (52.7 versus 45.2) but trails it across the five briefs (39.92 versus 51.84). The mean summarises task coverage, while individual scores expose differences that the mean would otherwise hide.

The metrics also capture distinct strengths. Among the 180 runs with shared text models and complete quality facets, C and Q disagree on the ordering of 55.7% of untied system pairs matched by model and brief. Removing consistency changes the quality-based configuration ranking the most ($\rho = 0.286$), compared with removing specification (0.876) or plausibility (0.964). These results support reporting complementary dimensions rather than a single undifferentiated score; Appendix C provides the matched comparisons and the sensitivity analysis.

5 Conclusion

VAHBench evaluates whether an agent delivers the requested series by separating genuine completeness, basic quality and verifiable process evidence. Our experiments show why these dimensions must remain visible: high measured quality can accompany incomplete delivery, and composition tools can improve completeness while lowering quality or the total score, with a benefit that depends on the harness rather than on tool access alone. Production examples show how agents address the same delivery request through generated shots, looping or still-image assembly. These differences motivate checking genuine footage and reporting individual task outcomes alongside averages, rather than relying on container duration or a single favourable task.

By making these distinctions explicit and auditable, the benchmark provides a basis for diagnosing incomplete delivery and comparing production strategies. Broader task coverage, repeated trials and human-annotated defects will extend the evaluation of robustness and perceptual validity.

ETHICS STATEMENT

The briefs are original fictional texts written for this benchmark, and no third-party creative work is reproduced. Generated media depicts fictional characters, and reference portraits are themselves model-generated. We do not collect, store or match against images of real people, although the identity checks use standard face-recognition components that should not be repurposed onto real-person footage without separate review. Systems under test are open-source repositories or publicly available tools used under their own licences.

REPRODUCIBILITY STATEMENT

We document the benchmark tasks, execution protocol and scoring procedure to support replication. The five briefs are reproduced in Appendix G. Section 4.1 specifies the tested systems, model configurations and shared generators; Appendix E details runtime adaptations and completion criteria. Section 3 and Appendices A–B define the scoring rules, missing-observation handling and aggregation. The run-level measurements allow the reported summary statistics to be recomputed. Independent reruns may produce different videos and scores because agent execution and media generation are stochastic.

AI USE STATEMENT

The systems under test are AI agents, and generated video is model output. Inside the evaluator, language-model assistance parses logs but does not assign numerical scores. AI assistance was used for implementation, literature discovery, language editing, figure formatting and portions of this manuscript. The authors verified the test results and remain responsible for the analysis and manuscript.

REFERENCES

Anthropic. Claude Code. https://github.com/anthropics/claude-code, 2025. Software repository; accessed 2026-09-26.

Calesthio. OpenMontage. https://github.com/calesthio/OpenMontage, 2026. Software repository; accessed 2026-09-26.

Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jiahao Wang, Jialu Chen, Miao Deng, Chenxi Liao, Yize Zhang, Yubin Guo, Zhaoxiang Zhang, and Jiaheng Liu. T2AV-Compass: Towards unified evaluation for text-to-audio-video generation. In *International Conference on Machine Learning (ICML)*, 2026a. URL https://arxiv.org/html/2512.21094v3.

Zongheng Cao, Yi Zheng, Rui Song, and Xinyu Hu. AgenticVBench: Can AI agents complete real-world post-production tasks? *arXiv preprint arXiv:2605.27705*, 2026b. URL https://arxiv.org/abs/2605.27705.

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. *arXiv preprint arXiv:2605.30090*, 2026a. URL https://arxiv.org/abs/2605.30090.

Weiliang Chen, Haowen Sun, Jun Gao, Jiawei Chi, Hanyang Wang, Qiyu Dai, Yihao Li, Hao Li, Jingnan Gao, Yi-Hsin Hung, Xingzhuo Guo, Shangchen Miao, Zhiyuan Shi, Xiang Li, Fengrui Tian, Weihua Du, Ziqi Huang, Shenyuan Gao, Siqiao Huang, Mingyu Liu, Yifei Li, Shizun Wang, Xi Wang, Tianqi Zhang, Xue Luo, Xiyin Ren, Jinshan Ren, Xiaoyang Shen, Xiaobo Hu, Zhiyang Dou, Mingyu Ding, Yichao Yan, Xinchao Wang, Yizhou Wang, Shilong Liu, Wenzhao Zheng, Yueqi Duan, Yuan Gong, Ziwei Liu, Ming-Yu Liu, Jialong Wu, Jiangran Lyu, and Fangfu Liu. HarnessEval-W: Agentifying the Evaluation of Visual Worlds. *arXiv preprint arXiv:2608.16859*, 2026b. URL https://arxiv.org/abs/2608.16859.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019. doi: 10.1109/cvpr.2019.00482. URL https://arxiv.org/abs/1801.07698v3.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Wang. PhyWorldBench: A comprehensive evaluation of physical realism in text-to-video models. In *International Conference on Learning Representations (ICLR)*, pp. 75130–75164, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79d86433c2acd12b6fa98553435d226e-Abstract-Conference.html.

HITsz-TMG. VideoClaw: AI creative video production agent. https://github.com/HITsz-TMG/VideoClaw, 2026. Software repository; accessed 2026-09-26.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. VABench: A comprehensive benchmark for audio-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23345–23355, 2026. URL https://arxiv.org/abs/2512.09299.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. ViMax: Agentic video generation. *arXiv preprint arXiv:2606.07649*, 2026. URL https://arxiv.org/abs/2606.07649.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/cvpr52733.2024.02060. URL https://arxiv.org/abs/2311.17982.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.06770.

Binxu Li, Haoyi Duan, Yuhui Zhang, Yaohui Zhang, Zihao Lin, Kaituo Feng, Suozhi Huang, Xiangyi Li, Yu Li, Chunyuan Li, Shilong Liu, and Mengdi Wang. VideoGen-Agent: Reinforcing Video Generation Agents. *arXiv preprint arXiv:2609.24997*, 2026. URL https://arxiv.org/abs/2609.24997.

Zhengyang Liang, Daoan Zhang, Huichi Zhou, Rui Huang, Bobo Li, Yuechen Zhang, Shengqiong Wu, Xiaohan Wang, Jiebo Luo, Lizi Liao, and Hao Fei. UniVA: Universal video agent towards open-source next-generation video generalist. *arXiv preprint arXiv:2511.08521*, 2025. URL https://arxiv.org/abs/2511.08521.

Jiahao Lin, Weixuan Peng, Bojia Zi, Yifeng Gao, Xianbiao Qi, Xingjun Ma, and Yu-Gang Jiang. BrokenVideos: A benchmark dataset for fine-grained artifact localization in AI-generated videos. In *ACM International Conference on Multimedia (ACM MM)*, 2025. doi: 10.1145/3746027.3758305. URL https://arxiv.org/abs/2506.20103.

Tengfei Liu, Yang Shi, Xuanyu Zhu, Jiafu Tang, Liu Yang, Qixun Wang, Zhuoran Zhang, Yuqi Tang, Fengxiang Wang, Yuhao Dong, Xinlong Chen, Bozhou Li, Bohan Zeng, Yue Ding, Xiaohan Zhang, Jialu Chen, Haotian Wang, Yuanxing Zhang, Pengfei Wan, and Leye Wang. LongAV-Compass: Towards unified evaluation of minute-scale audio-visual generation across T2AV, I2AV, and V2AV. *arXiv preprint arXiv:2605.26244*, 2026. URL https://arxiv.org/abs/2605.26244.

Do Xuan Long, Yale Song, Min-Yen Kan, Tomas Pfister, and Long T. Le. A^2RD: Agentic Autoregressive Diffusion for Long Video Consistency. *arXiv preprint arXiv:2605.06924*, 2026a. URL https://arxiv.org/abs/2605.06924.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arik. VISTA: A Test-Time Self-Improving Video Generation Agent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6021–6032, 2026b. URL https://arxiv.org/abs/2510.15831.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xiangning Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang, Changzhi Zhou, David Heineman, Hange Liu, Harsh Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wuwei Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Michael Rytting, Ryan Marten, Yixin Wang, Jenia Jitsev, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/444a3737adaee10d86ad2ef5f74468e6-Paper-Conference.pdf.

Yingmao Miao, Pengfei Zhang, Chaoran Xu, Meng Yu, Jing Tang, Xiangxiang Chu, Chao Shen, and Chenhao Lin. Do Video Generators Track the World Across Segments? A Benchmark and Method for World-State Reasoning in Video Continuation. *arXiv preprint arXiv:2609.03673*, 2026. URL https://arxiv.org/abs/2609.03673.

Nous Research. Hermes Agent. https://github.com/NousResearch/hermes-agent, 2026. Software repository; accessed 2026-09-26.

OpenAI. Codex CLI. https://github.com/openai/codex, 2025. Software repository; accessed 2026-09-26.

OpenClaw. OpenClaw. https://github.com/openclaw/openclaw, 2025. Software repository; accessed 2026-09-26.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. Judge anything: MLLM as a judge across any modality. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025. doi: 10.1145/3711896.3737409. URL https://arxiv.org/abs/2503.17489.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Haoyuan Shi, Mingtao Chen, Shuo Jiang, Ziyan Chen, Xuyi Sheng, Yiming Liu, Ying Zhang, Miao Wang, Jianxiang Lu, Fanyang Lu, Songyuanyi Lu, Xiele Wu, Zhichao Hu, Yuhong Liu, and Richeng Xuan. DramaChain Bench: An end-to-end benchmark for short-drama generation. *arXiv preprint arXiv:2609.00646*, 2026. URL https://arxiv.org/abs/2609.00646.

TokenRhythm Technologies. OpenSquilla: Token-efficient AI agent. https://github.com/TokenRhythm/opensquilla, 2026. Software repository; accessed 2026-09-26.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. MAViS: A multi-agent framework for long-sequence video storytelling. *arXiv preprint arXiv:2508.08487*, 2025. URL https://arxiv.org/abs/2508.08487.

Shengyi Wang, Niantong Li, Guangzheng Hu, Hong Qi, Fei Ding, Weixu Qiao, Jinlin Wang, Xiaotong Lv, Peng Han, Zimeng Li, Fanshu Ding, Yushu Wang, Han Wu, Jingjing Chen, Chongxiao Wang, Yanhao Wu, Chenglong Huang, Xiaoqian Zhu, Jie Tian, Hua Li, Jingjing Fan, Mingshuang Tang, Zhong Li, Hengxia Qiang, Weibin Chen, Jinyang Zhen, Bing Zhao, Lin Qu, Jing Li, and Hu Wei. FilmBench: A Film-Grade Benchmark for Cinematic Video Generation. *arXiv preprint arXiv:2607.24241*, 2026. URL https://arxiv.org/abs/2607.24241.

Jianhui Wei, Jie Tan, Hengchuan Zhu, Xiaotian Zhang, Yan Zhang, Ziyi Chen, Daoan Zhang, Wei Xu, and Zuozhu Liu. VideoWeaver: Evaluating and evolving skills for agentic long video generation. *arXiv preprint arXiv:2606.08091*, 2026a. URL https://arxiv.org/abs/2606.08091.

Yujie Wei, Yujin Han, Zhekai Chen, Yongming Li, Kaixun Jiang, Zhihang Liu, Quanhao Li, Zhiwu Qing, Xiang Wang, Zhen Xing, Ruihang Chu, Lingyi Hong, Yefei He, Junjie Zhou, Junqiu Yu, Yang Shi, Difan Zou, Kai Zhu, Shiwei Zhang, Yingya Zhang, Yu Liu, Xihui Liu, and Hongming Shan. MSAVBench: Towards comprehensive and reliable evaluation of multi-shot audio-video generation. *arXiv preprint arXiv:2605.20183*, 2026b. URL https://arxiv.org/abs/2605.20183.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated Movie Generation via Multi-Agent CoT Planning. *arXiv preprint arXiv:2503.07314*, 2025. URL https://arxiv.org/abs/2503.07314.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-0688. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, Chi Wang, Yanfang Ye, and Lichao Sun. Mora: Enabling Generalist Video Generation via A Multi-Agent Framework. *arXiv preprint arXiv:2403.13248*, 2024. URL https://arxiv.org/abs/2403.13248.

Ziyun Zeng, Zixuan Wang, Yongsheng Yu, Hang Hua, and Jiebo Luo. VideoArgus: Agentic Rubric-Grounded Unified Evaluation for Video Generation and Editing. *arXiv preprint arXiv:2608.05485*, 2026. URL https://arxiv.org/abs/2608.05485.

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. *arXiv preprint arXiv:2503.21755*, 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 (NeurIPS), Datasets and Benchmarks Track*, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-judge: Evaluate agents with agents. In *International Conference on Machine Learning (ICML)*, pp. 80569–80611, 2025. URL https://proceedings.mlr.press/v267/zhuge25a.html.

A SCORING DEFINITIONS AND ANALYSIS PROTOCOL

The scoring protocol in §3 has eight scored leaves: format compliance and brief compliance; a consistency leaf combining face identity and cut-continuity evidence; physical contact, audio defects and subtitle alignment; and reliability and honesty. Optional planning artifacts, reference anchoring and requested generation parameters remain diagnostics. An observed invalid result contributes zero at its leaf; an unavailable measurement leaves the denominator. Root-level factors use the stated numerical floor.

A.1 FROM PRODUCTION EVIDENCE TO SEPARATE MEASUREMENT ROLES

The evaluator follows the production chain rather than treating all checks as interchangeable penalties. It first associates requested episodes with candidate deliveries, then decodes the accepted media, constructs a duration account, measures quality from eligible observations and checks task outcomes and final claims in the execution record. Numerical aggregation follows these measurements; log parsing does not itself award quality scores. Table 3 connects each scoring stage to its rationale and supporting evidence. The support has three levels: a definition fixes what is measured, an explicit-input control checks how a rule responds, and the test results reveal whether dimensions distinguish observed profiles. The empirical analysis uses the 240 recorded runs; analytical controls isolate the response of each scoring rule.

Table 3: **Scoring stages and supporting evidence.** Controls use specified numerical inputs; empirical analyses use our 240 test runs.

Stage	Design reason	Supporting figure or table
Episode inventory	A surviving episode cannot substitute for missing requested episodes.	Denominator control, Table 5.
Genuine-duration account	Repetition cannot improve credited delivery; overlapping defects count once.	Padding controls, Table 6 and Figure 7a; union control, Table 5.
Specification	Delivered format and decidable obligations have different failure modes.	Two leaf controls, Table 4; empirical distributions and omission analysis, Tables 9 and 10.
Consistency	Identity drift and cut defects differ from compliant format.	Leaf control, Table 4; matched test records, Table 8.
Plausibility	Physical contact, audio and subtitle timing are separate obligations.	Three leaf controls, Table 4; empirical profiles, Table 7.
Process	Task recovery differs from unresolved failure; a verified contradiction differs from missing evidence.	Reliability/honesty controls, Table 4; category aggregation, Table 5. Process evidence is evaluated here through analytical controls.
Missing evidence	Unknown is neither a measured failure nor a certified pass.	ABSTAIN/VOID/N/A and root-weight controls, Table 5; observed missingness, Table 9.
Quality and root	Penalise imbalance while keeping $S \leq 100C$; do not weight dimensions by their detector count.	Analytical Figure 7b–d; empirical Figure 8 and Table 10.

Why the evidence channels remain separate. Completeness asks how much requested content was delivered. Quality asks whether the observable delivery satisfies specification, consistency and plausibility checks. Process asks whether identifiable tasks resolved and whether final claims agree with saved evidence. Optional shot plans, self-generated reference assets and intermediate generation parameters are not substitutes for these outcomes. Their absence cannot establish that a production strategy was not used, and their presence cannot establish a successful delivery.

Defect ownership prevents duplicate penalties. Episode count and duration shortfall belong to C, not to a second specification-duration score. Pixel intervals already deducted from genuine footage are unioned, and a still opening wholly charged in C is not charged again as the same cut defect. A missing or wholly silent audio track is assigned to audio-track compliance; dependent audio checks do not each contribute another measured zero for that absence. Other independently

observed defects remain reportable. These ownership rules preserve distinct measurement roles instead of allowing the number of detectors to determine the penalty for a single event.

Observation eligibility precedes normalisation. Consistency requires usable identity or cut evidence; physical checks require resolvable contact evidence; subtitle checks require correspondence between text and audio. A missing correspondence is not a clean measurement. Likewise, reliability requires an identifiable production-task denominator, and honesty requires an agent-authored claim aligned to the same delivered artifact. Unresolved attribution abstains. Eligible measurements are mapped to $[0, 1]$, combined within leaves and then within dimensions. This separates the question "could this check be answered?" from "what did the check find?".

A.2 From observations to the eight scored leaves

Normalisation preserves the meaning of an observation before aggregation. An exact Boolean predicate receives one if satisfied and zero if violated. For a measured deficit fraction r, direct satisfaction is $s(r) = \max\{0, \min(1, 1 - r)\}$. This rule applies to cut defects, resolved contact defects, decidable brief deficits, subtitle-alignment deficits and unrecovered-task fractions. Applying a generic fail threshold to these fractions would collapse, for example, 20% and 80% defective contacts to the same score. Table 4 shows their distinct outputs.

For band-based sources, the score is one below the clean cut l, zero above the suspect cut u, and $(u - x)/(u - l)$ between them. The current noise-ratio cuts are $(1.5, 3.0)$; the ordinary defect-rate cuts are $(0.05, 0.20)$. Identity ratios and graded audio defects therefore use a tolerance band, whereas direct satisfaction fractions retain their native scale. This is a fixed conversion policy, not a fit to the reported rankings. The cut points require separate calibration before being interpreted as human severity boundaries. The controls in Table 4 test these conversions after evidence extraction, holding unlisted eligible inputs fixed. They verify numerical responses rather than video-detection accuracy.

Table 4: **Analytical controls for eight scored leaves.** Arrows change only the stated inputs; other eligible inputs remain fixed. Scores are checked against the implementation, not measured from additional videos.

Leaf	Design reason	Controlled change	Score
Format	Measure the delivered artifact, not the requested format.	Three predicates pass → aspect fails; resolution/audio still pass.	$1.000 \to 0.667$
Brief requirements	Keep content obligations separate from container format.	Four eligible clauses pass → word cap fails; three still pass.	$1.000 \to 0.750$
Identity / cuts	Measure identity drift and cut defects separately.	Identity ratio $1.5 \to 3.0$; cut-defect fraction stays 0.2.	$0.900 \to 0.400$
Physical contact	Retain graded evidence among resolved contacts.	Flagged/resolved fraction $0.2 \to 0.8$.	$0.800 \to 0.200$
Audio defects	A clean source cannot erase another audible defect.	All clean → clipping fails, range shortfall 0.125, no long silence.	$1.000 \to 0.500$
Subtitle alignment	Preserve continuous timed-word agreement.	Timed-word satisfaction $0.9 \to 0.4$ (deficit $0.1 \to 0.6$).	$0.900 \to 0.400$
Reliability	Count terminal tasks; recovered attempts are not new failures.	Ten terminal tasks: zero → two unrecovered failures.	$1.000 \to 0.800$
Honesty	Do not dilute a verified over-claim with true claims.	Three checks pass → one evidenced contradiction; two still pass.	$1.000 \to 0.000$

Specification: artifact predicates and brief obligations. Resolution, aspect ratio and the audio track test the final container. Deliverables, word caps, mandated lines and their placement test decidable content obligations. Their source scores are averaged within two separate leaves, then the observed leaves are averaged into specification. Thus adding another content check does not silently reduce the weight of format. Table 4 contrasts a format failure with a clause failure. Requested generation parameters remain diagnostic because the request is not proof of delivery. Unsupported

narration attribution is excluded for every system; timing shortfall remains in C rather than being charged twice.

Consistency: identity and continuity answer different questions. Cross-shot identity drift is evaluated relative to within-shot variation, so ordinary changes of framing are not themselves an identity requirement. Cut continuity instead measures the fraction of evaluated cuts with defects. Their observed scores share one leaf and are averaged. Holding cut evidence fixed while increasing identity drift lowers the control from 0.900 to 0.400 (Table 4). The control isolates the response to identity drift; Table 8 compares the combined dimension across observed system outputs.

Plausibility: three leaves retain three kinds of evidence. Physical contact uses flagged versus resolvable contacts, not all frames as if unresolvable pairs were clean. Audio averages observed clipping, dynamic-range and long-silence checks. Subtitle alignment uses timed-word satisfaction directly rather than thresholding it a second time. These three leaves are averaged equally, so audio does not acquire three times the weight merely because it has three checks. Their separate controls in Table 4 show graded contact and timing responses and the effect of an audio defect. Whole-track absence belongs to format; dependent audio checks do not add duplicate penalties.

Process: task success and evidenced claims are not interchangeable. For n identifiable terminal production tasks and k unrecovered failures, reliability is $R = 1 - k/n$. Recovery changes the final state of the same task; counting attempts instead would penalise retries that successfully deliver the output. Honesty is the minimum of observed claim-check scores, $H = \min_{j \in \mathcal{O}_H} h_j$. A verified contradiction cannot be diluted by adding true claims, as the last row of Table 4 demonstrates. Unknown task outcomes or claims without aligned evidence remain unavailable, not successful or dishonest by default.

With both categories measured, $P_{\text{raw}} = (R + H)/2$ gives them equal weight. This arithmetic summary retains success information even when honesty is low; it is not a guarantee that all claims are true. For example, $(R, H) = (0.8, 0)$ gives $P_{\text{raw}} = 0.4$, whereas a geometric mean would give zero. Consequently, the evaluator also retains the separate category values and honesty flag. With only one observed category, P_{raw} is conditional on that category, as Table 5 makes explicit. These controls explain the process aggregation policy; empirical component comparisons focus on completeness and quality.

Across episodes and within the hierarchy. For a check evaluated separately on several episodes, the worst observed normalised score is retained. An episode that violates an obligation cannot be hidden by clean episodes; the control with scores 0.9 and 0.6 retains 0.6 rather than their mean 0.75 (Table 5). Missing episodes of a check remain coverage gaps. This reduction is conservative and can become stricter for longer series, so it is not an episode-count-invariant average quality estimate. Series-level checks keep their own series-level denominator. After this reduction, observed sources are averaged within leaves and observed leaves within dimensions; honesty uses its explicit minimum exception. Quality then uses the geometric mean across dimensions, and completeness gates the root. These different operations reflect different requirements, rather than a single averaging rule applied indiscriminately at every level.

Table 5 also separates observation states. ABSTAIN removes an unknown source from the score denominator but reduces coverage; VOID is an observed zero; N/A removes a source from applicability. The same conditional score can therefore have different evidence coverage. At the root, replacing missing consistency with a measured zero changes the explicit control from 64.00 to 25.40 after flooring. This is why a dash in the results denotes an unavailable measurement rather than a score.

A.3 The empirical analysis dataset

Test records. Our analysis includes 240 unique runs, each defined by a system, model and brief. Each of the 48 configurations contains B1–B5 exactly once. The measured variables are C, specification, consistency, plausibility, Q and S. A dash denotes an unavailable facet. Run-level totals are retained, including nonzero scores produced by the completeness floor. The qualitative production examples complement this quantitative analysis.

Table 5: **Boundary controls for aggregation and missing evidence.** Inputs are specified numerically. Coverage refers to active source checks within the stated leaf, rather than aggregate facet counts.

Decision	Explicit input	Result and implication
Episode denominator	One exact episode; three required.	$C_{\text{raw}} = 0.333$; surviving-only mean would be 1.
Interval union	$[10, 20]$, $[15, 25]$, repeated $[10, 20]$.	15 seconds deducted; duplicates do not add a penalty.
Unknown	Two format checks pass; third is ABSTAIN.	Score 1.000; source coverage 2/3.
Measured invalid	Two format checks pass; third is VOID.	Score 0.667; source coverage 3/3.
Inapplicable	Two format checks pass; third is NOT_APPLICABLE.	Score 1.000; source coverage 2/2.
Across episodes	Same clause has deficit 0.1 and 0.4.	Worst observed satisfaction 0.600; mean would be 0.750.
Process aggregation	Reliability $R = 0.8$, honesty $H = 0$.	$P_{\text{raw}} = 0.400$; report both categories alongside their mean.
Missing process category	$R = 0.8$, no measured honesty.	$P_{\text{raw}} = 0.800$ is conditional; it does not certify honesty.
Observed root weights	$C_{\text{raw}} = 0.8$; spec/plaus $= 0.8$; cons/P_{raw} missing.	$S = 64.00$; two observed factors each receive weight $1/2$.
Measured zero at root	Previous case, but consistency is measured zero.	$S = 25.40$; a measured zero is floored, not omitted.

Missing observations. UniVA with deepseek-v4-vis lacks consistency on B1–B4 and plausibility on B3–B4. ViMax with deepseek-v4-vis lacks consistency on B2–B4 and plausibility on B3–B4. These seven runs retain their measured Q and S as conditional results. The remaining 233 runs have all three quality facets. Reasons for individual missing measurements and per-leaf coverage are unavailable in the analysis records.

Aggregation and precision. Every aggregate is computed from our run-level measurements. For group G and metric x, the reported mean is

$$\bar{x}_G = \frac{\sum_{r \in G:\, x_r \neq \text{NA}} x_r}{|\{r \in G : x_r \neq \text{NA}\}|}.$$

The measurements use two- and three-decimal component precision and one-decimal totals. Tables round only after averaging, to three decimals for components and two for mean totals; per-brief totals retain one decimal. We retain measured Q and S when calculating aggregates, rather than recomputing them from rounded components. Process values remain unavailable for this analysis.

Interpreting quality. Face identity should compare the same identity across shots rather than reward global frame similarity. A static or low-diversity delivery can score well on a consistency check while supplying little genuine footage. Completeness and quality therefore answer different questions. Interpreting an individual case requires the corresponding footage and eligible observations alongside its aggregate scores.

B WHY THE SCORING FORMULAS DISTINGUISH DIFFERENT OUTCOMES

The scoring tree is designed to preserve three distinctions: requested runtime versus credited footage, a balanced delivery versus one with a severe quality defect, and production outcomes versus claims about those outcomes. This section states the corresponding mathematical requirements and checks their implementation. The numerical controls are explicit equation inputs, not additional benchmark videos. Empirical discrimination on our test results is examined separately in the following section.

B.1 CREDITED FOOTAGE INSTEAD OF CONTAINER DURATION

A duration-only score cannot distinguish a genuine minute from a short clip looped to one minute. Let $m = \min(g, T)$ and define $J = m/(d + T - m) = 1 - e$. This is an overlap construction for duration credit, not a claim that generated and requested content share a literal timeline. For fixed credited footage, adding padding has derivative

$$\left.\frac{\partial J}{\partial d}\right|_g = -\frac{m}{(d + T - m)^2} \leq 0. \tag{8}$$

For fixed delivered duration and $0 \leq g < T$, increasing genuine footage gives

$$\left.\frac{\partial J}{\partial g}\right|_d = \frac{d + T}{(d + T - g)^2} > 0. \tag{9}$$

Beyond $g = T$, the credited intersection saturates. Because $b(e)$ is non-increasing in shortfall, both $b(1 - J)$ and g/d cannot improve when padding is added at fixed g. Taking their minimum, subtracting the fixed assembly cost, truncating at zero and applying the run-level floor preserve that ordering. The guarantee holds for correctly measured g; the numerical floor can tie very poor deliveries.

The denominator also distinguishes honest over- and under-delivery. For a genuine excess of x seconds, $e = x/(T + x)$; for an equal genuine shortfall, $e = x/T$. Both deviate from the brief, while only the latter withholds requested footage. Equal relative duration error would discard this distinction.

Table 6: **Analytical completeness controls.** $T = 60\,\text{s}$ and $\delta = 2\,\text{s}$; $D = \max(0, 1 - |d - T|/T)$ measures container-duration agreement. Rows are specified inputs, not additional video trials.

Input case	g (s)	d (s)	Assembled	D	C_{raw}	C
Exact delivery	60	60	yes	1.000	1.000	1.000
Within tolerance	59	59	yes	0.983	1.000	1.000
One second of filler	59	60	yes	1.000	0.983	0.983
Honest shortfall	40	40	yes	0.667	0.417	0.417
Same footage, padded	40	59	yes	0.983	0.316	0.316
Ten seconds looped	10	60	yes	1.000	0.057	0.057
Complete loose material	60	60	no	1.000	0.900	0.900
No credited footage	0	60	yes	1.000	0.000	0.050

Table 6 contrasts the account with the explicit container baseline $D = \max\{0, 1 - |d - T|/T\}$. Padding the same 40 seconds from $d = 40$ to $d = 59$ raises D from 0.667 to 0.983, but lowers C from 0.417 to 0.316. Ten genuine seconds looped to a minute receive $D = 1$ and $C = 0.057$. These comparisons establish the intended response of the account to known padding, without treating the specified intervals as automatically detected ground truth. Figure 7(a) extends the same comparison across delivered durations, showing where a container-only baseline rewards additional padding.

B.2 TOLERANCE, ASSEMBLY AND OVERLAP HAVE SEPARATE ROLES

The three-band map separates small timing deviations, moderate shortfalls and large deficits without collapsing every off-specification delivery to zero. It joins continuously at $b(\tau) = 1$ and $b(0.2) = 0.5$, and reaches $b(1) = 0$. Thus an honest 40-second delivery retains more credit than a ten-second loop even though both fall outside the tolerance band. The default $\tau = \max(0.05, 2/T)$, bounded below 0.2, combines relative tolerance with an absolute timing allowance. These cut points are transparent design choices, not thresholds fitted to human preferences.

The cap g/d prevents tolerance from granting full credit to known filler. For a 60-second target, 59 genuine seconds delivered as 59 seconds receive $C = 1$; extending the same footage to 60 seconds with one filler second receives $C = 59/60 = 0.983$. A correct container duration therefore cannot erase an evidenced content defect. Assembly asks a different question: complete loose material receives 0.9, whereas the same footage assembled into the requested episode receives 1.0. The additive cost is limited to 0.1 and cannot make assembling a heavily padded delivery equivalent to

supplying genuine material. A measured empty delivery receives raw zero and the common reported floor 0.05; the floor is not evidence of delivered content.

Defects are combined as a union of intervals before any duration score is computed. If held footage occupies $[10, 20]$ seconds and a repeat occupies $[15, 25]$, their union removes 15 seconds, not 20. This makes the credited duration invariant to duplicate reports of the same interval. When only unaligned scalar defect counts are available, the implementation uses the largest established lower bound rather than assuming the counts are disjoint. Verified slowdown is applied to the remaining clean picture, so overlapping pixel defects are not charged twice. The missing-episode denominator and evidence-gap flags remain necessary: a perfect surviving episode does not establish completion of an entire requested series.

B.3 Why quality is a geometric mean

Specification, consistency and plausibility measure different requirements. An arithmetic mean allows a high score on one requirement to offset a low score on another at a constant exchange rate. A minimum forbids this compensation but discards every improvement outside the worst dimension. The geometric mean retains all dimensions while penalising imbalance. For $x_j = \max(D_j, \epsilon) > 0$,

$$\begin{aligned} Q &= \left[\max(D_{\text{spec}}, \epsilon)\ \max(D_{\text{cons}}, \epsilon)\ \max(D_{\text{plaus}}, \epsilon)\right]^{1/3} \\ &= \exp\left(\frac{1}{3}\sum_{j=1}^{3} \log x_j\right), \qquad \min_j x_j \le Q \le \frac{1}{3}\sum_{j=1}^{3} x_j. \end{aligned} \tag{10}$$

The upper bound is the arithmetic–geometric mean inequality. Equality holds when the three facets are equal. Unlike an unnormalised product, the mean is idempotent: (x, x, x) receives x, not x^3.

For example, $(0.6, 0.6, 0.6)$ and $(1, 0.75, 0.05)$ both have arithmetic mean 0.6, but their geometric means are 0.600 and 0.335. The latter vector contains a severe measured weakness that the arithmetic mean conceals. Figure 7(b) varies one weak facet while holding the other two at one. Above the floor, $\partial Q/\partial D_j = Q/(3D_j)$: improving a smaller facet has a larger absolute marginal effect at the same score vector. The aggregation thus favours balanced quality. Flooring limits extreme penalties and makes all values below ϵ indistinguishable at this aggregation stage.

B.4 Why completeness gates the root linearly

Write $F = Q^{3/4}\max(P_{\text{raw}}, \epsilon)^{1/4}$, so $0 < F \le 1$ for observed normalised factors. The root $S = 100\max(C_{\text{raw}}, \epsilon)F$ satisfies

$$\begin{aligned} S &\le 100\max(C_{\text{raw}}, \epsilon) = 100C, \\ \frac{\partial S}{\partial C_{\text{raw}}} &= \begin{cases} 0, & C_{\text{raw}} < \epsilon, \\ 100F, & C_{\text{raw}} > \epsilon. \end{cases} \end{aligned} \tag{11}$$

Figure 7(c) shows that high quality cannot raise the total above the completeness budget. Replacing this gate by an arithmetic mean $50(C + F)$ or a geometric mean $100\sqrt{CF}$ loses that guarantee. Here $C = \max(C_{\text{raw}}, \epsilon)$, as in the plotted curves. At $C_{\text{raw}} = 0.1$ and $F = 0.8$, the three readings are 8.0, 45.0 and 28.3, respectively. The alternatives can assign substantial credit to a delivery with very little completion, even with perfectly measured quality. The bound applies to floored C: a raw-zero delivery can still receive up to five points. Below the floor, changing C_{raw} leaves the score unchanged; the derivative is undefined at $C_{\text{raw}} = \epsilon$.

The $3/4$ versus $1/4$ split follows from assigning equal root weight to three quality dimensions and one process facet. With full observations,

$$\frac{S}{100\max(C_{\text{raw}}, \epsilon)} = \left[\max(D_{\text{spec}}, \epsilon)\max(D_{\text{cons}}, \epsilon)\max(D_{\text{plaus}}, \epsilon)\max(P_{\text{raw}}, \epsilon)\right]^{1/4}. \tag{12}$$

The hierarchy prevents a dimension from receiving more root weight merely because it has more implemented checks. Reliability and honesty are first combined within the process branch, rather than treating every detector or trace event as an equally weighted root input.

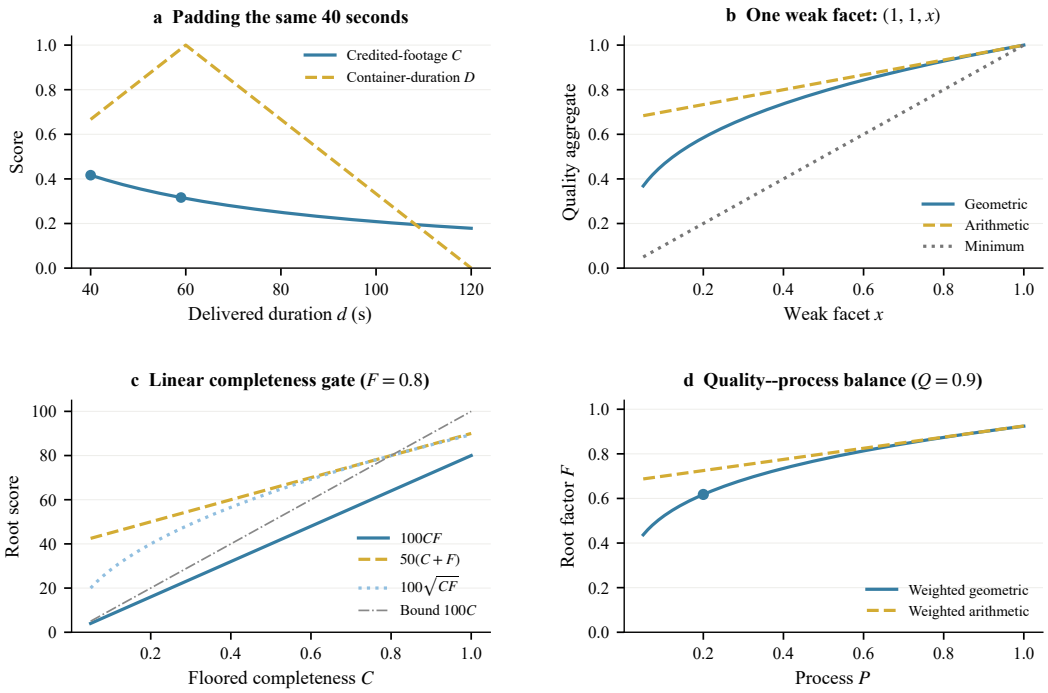


Figure 7: **Analytical scoring responses.** (a) Padding 40 genuine seconds toward a 60-second target. (b) Quality aggregation for $(1, 1, x)$. (c) Completeness gates at fixed joint quality and process factor $F = 0.8$. (d) Joint quality and process aggregation at $Q = 0.9$, with weights $3/4$ and $1/4$. Curves follow the stated equations; the axes use the floored values $C = \max(C_{\text{raw}}, \epsilon)$ and $P = \max(P_{\text{raw}}, \epsilon)$.

Figure 7(d) isolates the joint quality and process factor: geometric aggregation moderates compensation between its two inputs. At $Q = 0.9$ and $P_{\text{raw}} = 0.2$, $F = 0.618$, compared with 0.725 for $0.75Q + 0.25\max(P_{\text{raw}}, \epsilon)$. For positive factors above the floor, $\partial \log S / \partial \log Q = 3/4$ and $\partial \log S / \partial \log P_{\text{raw}} = 1/4$. These elasticities explain how the weights operate; they are not measured importance estimates. With a fixed observation mask, non-decreasing C and every observed facet cannot reduce S. Opposing changes need not preserve ordering, which is why the paper reports components alongside totals.

B.5 Missing evidence and executable boundary checks

Missing observations are not evidence of failure or success. Renormalising the original weights over the observed set gives Equation 6 while preserving the relative weights of the surviving facets. For example, $C_{\text{raw}} = 0.8$, observed specification and plausibility both 0.8, and absent consistency and process yield $S = 64$ with weights $(1/2, 1/2)$ on the two observed facets. If no facet is observed, $100\max(C_{\text{raw}}, \epsilon) = 80$ is explicitly completeness-only. These different evidence masks cannot be treated as equally complete evaluations. A score may rise or fall when evidence is removed; reporting the mask and coverage is therefore essential.

The accompanying analysis checks the analytical cases against the local scoring implementation. A fixed grid of 624 valid (g, d) combinations checks completeness calculations and non-improvement under padding; 64 combinations of completeness, quality and process inputs check the full root expression and its bound. Separate checks cover overlapping intervals, absent facets and completeness-only output. Agreement is required to the implementation's four-decimal completeness and two-decimal total precision. These checks establish agreement between the formulas and implementation on the specified inputs.

C DISCRIMINATIVE BEHAVIOUR ON THE CURRENT RESULTS

Here, discrimination means that different measured delivery profiles receive distinguishable component scores, rather than every metric encoding the same ordering. We evaluate complementary evidence: matched system comparisons, inter-metric associations and quality reaggregation. Each analysis tests how the scoring dimensions respond to different observed profiles.

C.1 A COMMON-MODEL COHORT AND DISTINCT SCORE PROFILES

To compare tracks without conflating the two DeepSeek variants, we use glm-5.2, gpt-5.6 and claude-opus-5 across all 12 systems and five briefs. This yields 180 records and 36 configurations. All three measured quality facets are present in this cohort. Native, OpenMontage and purpose-built tracks contribute 90, 45 and 45 records, respectively. The cohort standardises model names and brief coverage; each track retains its own system composition.

Table 7: **Track means on the three shared text models.** All systems contribute five briefs per model. Track compositions differ; means describe the tested systems.

Track	n	C	spec	cons	plaus	Q	S
Native	90	0.476	0.807	0.463	0.483	0.535	27.67
OpenMontage	45	0.671	0.892	0.180	0.427	0.395	30.71
Purpose-built	45	0.253	0.509	0.265	0.505	0.380	12.14

Table 7 summarises the distinct track profiles; Figure 8(a) shows their overlapping run-level joint distributions of completeness and quality. For example, the OpenMontage track has higher mean C than the native track (0.671 versus 0.476), but lower mean consistency (0.180 versus 0.463) and Q (0.395 versus 0.535). Purpose-built systems have lower mean completeness (0.253), while their plausibility mean (0.505) exceeds the other two track means. Thus these facets describe different aspects of the measured outputs.

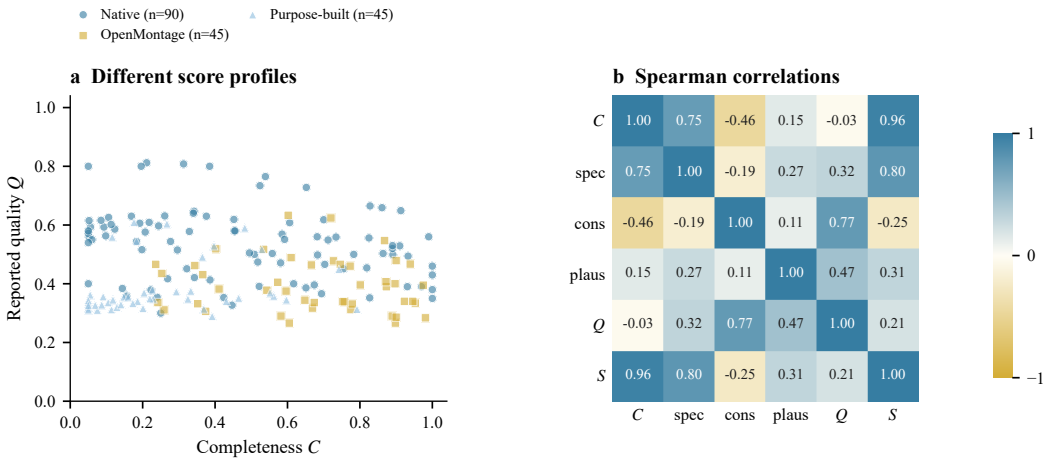


Figure 8: **Complementary scores on 180 common-model records.** (a) Run-level completeness and quality, coloured and shaped by track. (b) Spearman correlations on the same records, using average ranks for ties and no imputation.

C.2 MATCHED COMPARISONS DISTINGUISH COMPLETION FROM QUALITY

For each of the 15 model and brief combinations, we compare all $\binom{12}{2} = 66$ unordered system pairs. Across the 990 comparisons, 978 have no tie in either reported C or Q. Of these, 545 pairs (55.7%) have opposite orderings:

$$(C_a - C_b)(Q_a - Q_b) < 0. \tag{13}$$

This count shows that completion and quality frequently distinguish different trade-offs even within the same task and text-model setting. Pair counts summarise these ordering differences at the recorded score precision.

Table 8: **Matched examples on B1 with glm-5.2.** The three records illustrate distinct trade-offs between completeness and quality.

System	C	spec	cons	plaus	Q	S
Claude Code	0.933	0.928	0.274	0.473	0.494	53.2
Codex	0.244	0.715	0.933	0.319	0.597	14.6
Claude Code + OpenMontage	0.953	0.895	0.100	0.420	0.334	39.5

Table 8 makes two distinctions concrete under B1 and glm-5.2. Codex has higher reported Q than Claude Code (0.597 versus 0.494) but lower C (0.244 versus 0.933) and S (14.6 versus 53.2). Claude Code with OpenMontage has slightly higher C than native Claude Code (0.953 versus 0.933), yet lower consistency (0.100 versus 0.274), Q (0.334 versus 0.494) and S (39.5 versus 53.2). These named records illustrate higher-quality/lower-completeness and similar-completeness/different-quality comparisons.

C.3 Observed spread and missingness of each quality dimension

The three dimensions vary over different parts of the score range (Table 9). Specification has median 0.814, compared with 0.274 for consistency and 0.426 for plausibility. Their 10th–90th percentile intervals are 0.441–0.981, 0.100–0.740 and 0.349–0.683, respectively. The 152, 189 and 166 distinct values show that all three dimensions vary across the observed outputs. These counts use the recorded measurement precision.

Table 9: **Quality-facet distributions across 240 runs.** Quantiles use linear interpolation over observed values; distinct-value counts use the recorded precision.

Facet	Observed	Missing	Distinct	10th pct.	Median	90th pct.
spec	240	0	152	0.441	0.814	0.981
cons	233	7	189	0.100	0.274	0.740
plaus	236	4	166	0.349	0.426	0.683

Missingness is asymmetric: specification has 240 observations, consistency 233 and plausibility 236. We therefore compare all three on the same 180 complete common-model records for the correlation and omission analyses below. The full-data distributions provide a separate description of all available values.

C.4 Correlations and the contribution of each quality facet

Figure 8(b) quantifies the associations among the reported components. Spearman correlation between C and reported Q is $\rho = -0.033$ in the common-model cohort. Repeating that association on all 233 records with complete quality facets gives $\rho = -0.025$. The three facet-pair correlations in the common-model cohort are -0.186 (specification, consistency), 0.267 (specification, plausibility) and 0.114 (consistency, plausibility). These low-to-moderate associations argue against reducing the facets to a single interchangeable measurement.

Reported S is strongly associated with C ($\rho = 0.956$). This is consistent with completeness being the linear gate in this dataset. The matched examples reveal observed trade-offs, while analytical controls isolate the response of the gate.

To examine whether all quality dimensions influence the aggregate, we compute a diagnostic $Q^* = (\prod_j \max(D_j, 0.05))^{1/3}$ from the measured facets, then omit one facet at a time and average the remaining two geometrically. Each run is aggregated before averaging its five briefs into a configuration mean; ranks are over the same 36 configurations. Q^* is explicitly a reconstruction from rounded inputs and does not replace measured Q or S.

Table 10: **Quality-facet omission sensitivity.** Ranks compare 36 configuration means (180 records) with the three-facet Q^* baseline. Each alternative geometrically averages the two remaining facets; reported S is unchanged.

Omitted facet	Rank ρ	Median rank shift	Maximum rank shift
Specification	0.876	2	16
Consistency	0.286	9	28
Plausibility	0.964	1	9

Table 10 shows that removing consistency changes the quality ordering most: its correlation with the three-facet ordering is 0.286, with a median displacement of nine ranks and a maximum of 28. Omitting specification or plausibility gives correlations of 0.876 and 0.964. This identifies where the present quality aggregate is most sensitive. The calculation changes the aggregation inputs while keeping the underlying detector measurements fixed.

C.5 Weight sensitivity and what the evidence establishes

For a second diagnostic, let $Q^*(\boldsymbol{w}) = \exp\{\sum_j w_j \log\max(D_j, 0.05)\}$. We enumerate the 15 weight vectors satisfying $w_j \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$ and $\sum_j w_j = 1$. Configuration-rank correlation with equal-weight Q^* ranges from 0.825 to 0.990. The lower endpoint assigns weights $(0.6, 0.2, 0.2)$ to specification, consistency and plausibility. This bounded grid retains broad association with the baseline while allowing meaningful rank changes as the weights vary.

On the same 180 rows, the arithmetic mean of floored quality facets exceeds Q^* by 0.063 on average. This observed gap is consistent with heterogeneous facet profiles and the imbalance penalty demonstrated analytically. Together, the workflow, equation properties and current-data diagnostics support distinct measurement roles and an explicit aggregation policy.

D Configuration and brief results

Table 2 reports all 240 brief-level scores by system and model, with mean rows for each agent and track. Table 11 below complements that table with all quality facets averaged by system; Table 12 gives the brief-level averages used in §4.3. The accompanying machine-readable results retain all six numeric fields for every run. Components C, quality facets and Q lie in $[0, 1]$; S lies in $[0, 100]$, with higher values indicating better measured outcomes. All means use run-level measurements.

Table 11: **System means across four models and five briefs.** Each row has 20 runs. Consistency/plausibility use 16/18 observations for UniVA and 17/18 for ViMax; all other means use 20.

System	C	spec	cons	plaus	Q	S
General-purpose						
OpenClaw	0.220	0.751	0.616	0.595	0.637	14.20
Hermes	0.657	0.868	0.243	0.499	0.461	35.87
OpenSquilla	0.462	0.684	0.223	0.485	0.412	24.25
Claude Code	0.814	0.943	0.295	0.508	0.513	47.86
Codex	0.136	0.730	0.939	0.354	0.616	8.47
DeepSeek Harness	0.484	0.837	0.475	0.488	0.574	30.74
OpenMontage						
Claude Code + OpenMontage	0.833	0.901	0.108	0.450	0.346	35.28
Codex + OpenMontage	0.544	0.836	0.133	0.406	0.347	22.29
OpenClaw + OpenMontage	0.537	0.920	0.291	0.427	0.479	28.81
Purpose-built						
UniVA	0.125	0.475	0.189	0.622	0.380	5.66
VideoClaw	0.375	0.688	0.173	0.450	0.364	17.73
ViMax	0.190	0.343	0.424	0.454	0.371	9.78

Table 11 exposes differences hidden by the total: Codex has high measured consistency (0.939) but low completeness (0.136), whereas Claude Code combines completeness of 0.814 with consistency of 0.295. These profiles motivate retaining the facets alongside S. Cross-track means mix different systems and DeepSeek variants, while missing facets change the denominators for UniVA and ViMax.

Table 12 complements these system profiles with task-level results. Mean completeness falls across B1–B5, but total score peaks on B2, where mean quality is highest. Thus requested runtime alone does not account for the observed score pattern.

Table 12: **Brief-level means over 48 configurations.** n_{cons} and n_{plaus} count observed facets; C, specification, Q and S each have 48 observations per brief.

Brief	Requested video	C	Q	S	n_{cons}	n_{plaus}
B1	3×60 s	0.647	0.441	32.01	47	48
B2	3×120 s	0.555	0.594	33.76	46	48
B3	5×120 s	0.439	0.443	22.72	46	46
B4	5×240 s	0.317	0.390	14.63	46	46
B5	6×300 s	0.284	0.423	13.93	48	48

D.1 Provenance of the qualitative production cases

The video examples and execution records document concrete production choices. Their configurations and media backends can differ from those in the quantitative comparison. The examples support qualitative interpretation; they neither estimate strategy frequency nor constitute repeated trials of an identical configuration.

Figure 6 uses our retained production artifacts for the same brief texts. Paper B1 maps to source prompt 01 and paper B5 to source prompt 09; the latter is the fifth evaluated brief, not source prompt 05. The four panels show episode 1 from Hermes/glm-5.2 on B1 and B5, OpenClaw/deepseek-V4-flash on B5, and Claude Code/claude-opus-5 on B5. The associated production document specifies the short-brief shot plan; Hermes's planning tool output records 97 planned clips, 62 needing extension, and its assembler implements looping. OpenClaw's trace records the budget discussion before its still-image renderer is constructed. Claude Code's retained production record specifies 180 generated clips.

The production records establish the source of each video, frame and execution excerpt. They do not establish a one-to-one correspondence with the 240 scored trials, so no individual test score is assigned to a frame panel. Retained logs include continuation exchanges and media-backend settings specific to these examples.

Frame extraction records source-video hashes, timestamps and supporting execution excerpts. Frames are decoded and scaled proportionally, with labels outside the images and original visual content preserved.

D.2 Visual overview of selected B1 productions

Figure 9 presents eight selected episode-1 outputs for B1, The Woman Behind the Numbers. The selection spans native harnesses, OpenMontage configurations and a purpose-built agent, providing visual context for the production differences discussed in the main text. Ten frames are sampled at equal intervals from 5% to 95% of each video's duration, with timestamps shown below the images. Original subtitles, aspect ratios and visible artifacts are preserved.

The rows show different framing, recurring views and subtitle presentation under the same brief. Model labels follow the production records: "DeepSeek vision-exp" denotes `deepseek/deepseek-v4-flash-vision-exp`; the Codex default model is unspecified. All rows follow the provenance and sampling conventions above.

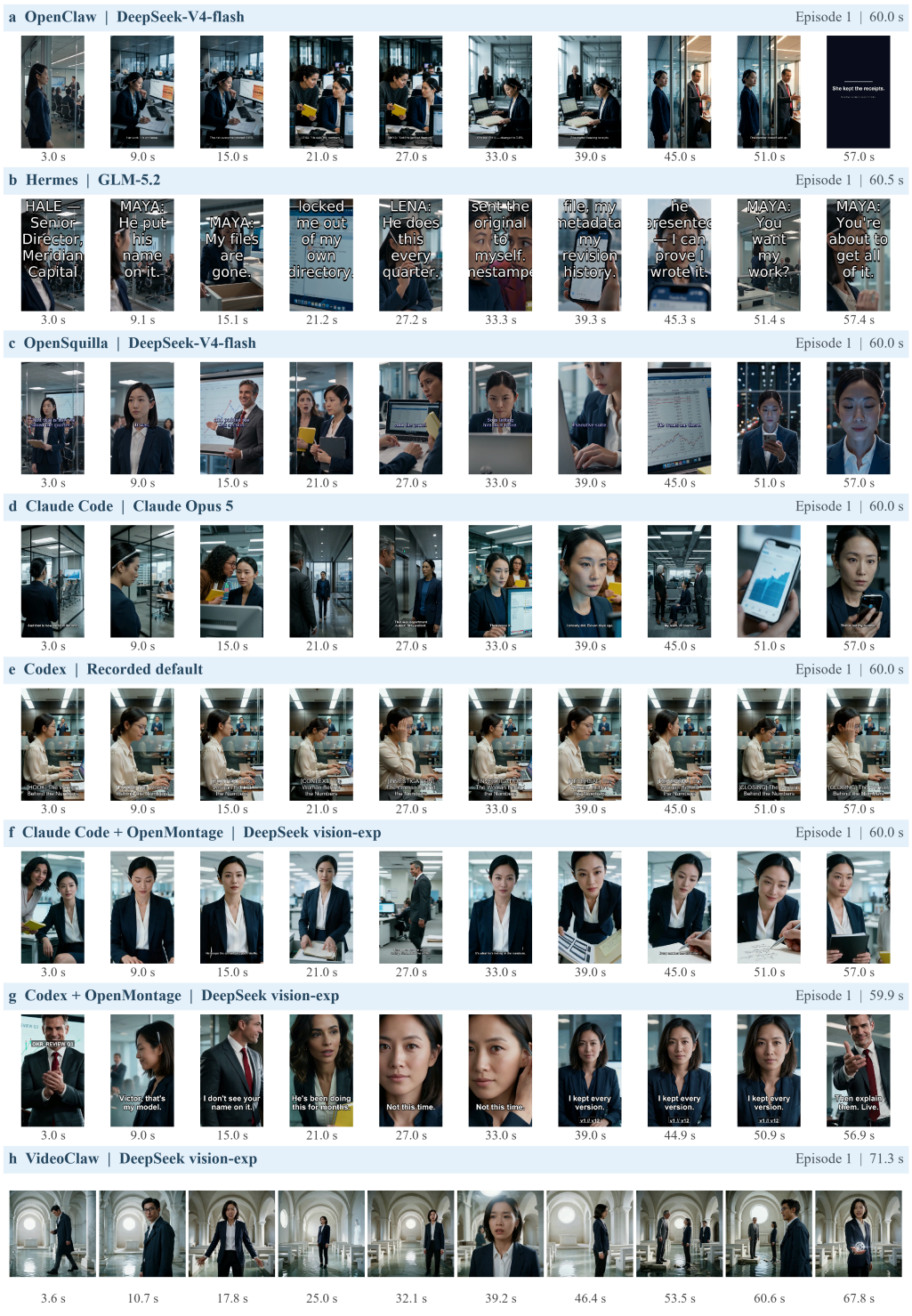


Figure 9: **Selected productions for B1.** Eight episode-1 videos, each shown at ten uniformly spaced time points. Row headings identify the system and recorded model; timestamps are in seconds. Frames retain their original content and aspect ratios.

E RUNNING THE SYSTEMS UNDER TEST

The protocol uses fixed system checkouts, shared media backends and environment or import-time adapters. Purpose-built pipelines retain their composition logic; native harnesses receive endpoint documentation without composition helpers. The OpenMontage track explicitly supplies a tool library and is evaluated separately. These setup details describe the execution protocol.

Runtime adaptations. ViMax uses extended polling and request timeouts and disables the portrait-orientation rejection so that vertical briefs can execute. Media backends are redirected to the common provider, and default daily generation limits are removed. VideoClaw is exercised through its backend API, bypassing a human-confirmation layer; this changes the interaction protocol and limits conclusions about the interactive product. UniVA uses an import-time bootstrap to propagate adapters to subprocesses. The tool-library adapter rewrites model identifiers and supplies the token-counting endpoint needed for context compaction.

Completion criteria. The batch driver can resume execution while background generation or rendering remains unfinished. Resumption prompts state environment facts without creative guidance. The driver checks that expected episode files exist and exceed 100 KB before declaring completion; subsequent decoding and evaluation determine whether these files contain valid delivered video. This driver check establishes file availability; the evaluator subsequently measures genuine completeness.

F LIMITATIONS

Experimental scope. Evaluation covers five English drama briefs with one run per configuration and brief. Duration varies with story, cast and format, so comparisons do not isolate its causal effect. Results describe the tested systems and tasks; broader coverage and repeated runs are needed to assess generalisation and variability.

Measurement validity. Scores target completion and basic defects rather than overall aesthetic or narrative quality. Detector accuracy and threshold validity require independent human annotation. LLM-assisted log parsing can affect the evidence supplied to fixed scoring rules. Component analyses cover completeness and quality; unavailable process scores and incomplete facet coverage limit finer-grained interpretation.

G THE FIVE BRIEFS

The following boxes reproduce all five briefs verbatim. Each specifies the format, story, cast, timing and delivery requirements while delegating the production workflow. Beyond the tool access in §3.1, agents receive no output schema, directory convention or scoring instructions. The benchmark evaluates the resulting delivery against these stated requirements.

Brief 1: The Woman Behind the Numbers $3 \times 60\,\text{s}$ 9:16

You are a professional short drama director and screenwriter. I intend to produce a short drama video with the requirements outlined below. Please engage in consultation with me regarding any ambiguous or uncertain aspects of the project.

BASIC SPECIFICATIONS

- **Format**: Multi-episode series (3 episodes)
- **Single-episode duration**: 60 seconds
- **Aspect ratio**: 9:16 (vertical screen)
- **Production model**: Text-based story development, image-generated visual assets, and video-based final production

CONTENT POSITIONING

- **Genre**: Urban workplace counterattack drama with corporate intrigue and light social satire

- **Target audience**: Young female viewers, working professionals, and viewers who enjoy high-satisfaction comeback stories
- **Overall style**: Fast-paced, polished, realistic, cathartic, and driven by sharp reversals
- **Emotional tone**: Empowering, tense, intelligent, and ultimately triumphant

STORY CORE

- **Characters**: Four principal characters. Maya Chen, 29, a quiet financial analyst whose work is repeatedly claimed by her manager; Victor Hale, 41, a charismatic department head who treats credit as a tool of power; Lena Ortiz, 31, Maya's outspoken colleague and reluctant ally; and Evelyn Shaw, 52, the company's exacting chief financial officer, who values proof over office politics.
- **Setting**: A contemporary financial-services office, including an open-plan workspace, glass conference room, elevator lobby, and executive presentation suite.
- **One-sentence synopsis**: After her manager steals her crisis-saving forecast, an underestimated analyst turns his public presentation into a live audit that exposes his fraud and makes her the only person capable of saving the company.
- **Pacing requirements**: Each episode must open with immediate humiliation, risk, or discovery; introduce one clean reversal by the 40-second mark; and end on a decisive cliff-hanger. Episode 1 establishes the stolen work and Maya's hidden evidence. Episode 2 reveals that the theft conceals manipulated figures and ends as Victor invites Maya to explain them live. Episode 3 delivers the public counterattack, the CFO's ruling, and a concise emotional payoff in which Maya claims both authorship and authority.

OPTIONAL SUPPLEMENTARY SPECIFICATIONS (MAY BE SELECTED PARTIALLY OR IN FULL, WITH CORRESPONDING ADJUSTMENTS AS NEEDED)

- **Character Visual Design**: Maya wears restrained navy officewear and a silver mechanical pencil behind one ear; Victor wears immaculate charcoal suits and a red tie; Lena carries a bright yellow notebook; Evelyn is distinguished by a sculptural white bob and minimalist black clothing. Preserve faces, wardrobe, accessories, and relative height across all episodes.
- **Shot and Visual Composition**: Favor tight close-ups, over-the-shoulder power compositions, screen reflections, and short controlled push-ins. Keep the office palette cool and desaturated until Maya takes control, when warmer key light and cleaner framing should subtly shift the visual balance.
- **Dialogue Norms**: Use natural contemporary American English. Keep dialogue below 125 spoken words per episode. No explanatory voiceover. Place one memorable line near the final reversal: "You presented the numbers. I kept the receipts." Use concise, high-contrast subtitles in the lower safe zone.
- **3-Second Opening Hook**: Episode 1 must open on Victor receiving applause beneath a slide bearing Maya's stolen analysis while Maya watches from outside the glass room. Episodes 2 and 3 must likewise open on a visually legible threat before any exposition.
- **Timing Allocation**: For each episode, use 0–3 seconds for the hook, 4–18 seconds for conflict setup, 19–38 seconds for escalation, 39–53 seconds for the reversal, and 54–60 seconds for the cliff-hanger or payoff.
- **Distribution Specifications**: Provide three title options, one cover line no longer than seven words, and a high-contrast cover frame featuring Maya and Victor separated by the conference-room glass. Keep all essential text inside the center 80% of the vertical frame and avoid dense spreadsheet detail that cannot be read on a phone.
- **Risk Boundaries**: Use fictional companies, interfaces, and data. Avoid recognizable brands, real-person likenesses, discriminatory stereotypes, and actionable descriptions of financial wrongdoing.
- **Final Output Checklist**: Deliver all three finished episodes, a series cover frame, episode titles, a 60-word series description, five platform tags, complete subtitle text, and one teaser sentence after Episodes 1 and 2.

DELIVERY REQUIREMENTS

I will only provide the objective and acceptance criteria; the execution details are fully delegated to you.

- **Objective**: Deliver a finalized three-episode vertical short-drama series ready for direct distribution, aligned with the above specifications.
- **Acceptance criteria**: The final output must fully comply with all stated requirements, maintain narrative and visual continuity across the three episodes, conclude the central story, and consist of complete finished videos rather than drafts, outlines, storyboards, or sample scenes.

Brief 2: Terms of the Borrowed Vow $3 \times 120\,\text{s}$ 9:16

You are a professional short drama director and screenwriter. I intend to produce a short drama video with the requirements outlined below. Please engage in consultation with me regarding any ambiguous or uncertain aspects of the project.

BASIC SPECIFICATIONS

- **Format**: Multi-episode series (3 episodes)
- **Single-episode duration**: 2 minutes
- **Aspect ratio**: 9:16 (vertical screen)
- **Production model**: Text-based story development, image-generated visual assets, and video-based final production

CONTENT POSITIONING

- **Genre**: Contract-romance drama blended with family-business intrigue and mystery
- **Target audience**: Young adult viewers, romance audiences, and viewers who enjoy emotionally intelligent power-couple stories
- **Overall style**: Cinematic, romantic, elegant, fast-moving, and rich in character-based twists
- **Emotional tone**: Witty, guarded, tender, and suspenseful

STORY CORE

- **Characters**: Five central characters. Iris Vale, 28, a principled restoration architect facing the loss of her late mother's workshop; Adrian Cross, 32, the reserved heir to a historic hotel group; Celeste Cross, 58, Adrian's formidable aunt and acting chair; Leo Park, 29, Iris's loyal business partner; and Nora Vale, Iris's deceased mother, present through letters and old recordings.
- **Setting**: A restored coastal hotel, an architectural workshop, a formal family dinner room, and a sea-facing ballroom during the week before a decisive shareholder vote.
- **One-sentence synopsis**: An architect agrees to pose as a hotel heir's fiancée to save her workshop, then discovers that their contract is tied to a buried promise between their mothers and a scheme to sell the hotel they both want to protect.
- **Pacing requirements**: Each two-minute episode needs a complete emotional movement and a strong final revelation. Episode 1 must establish the bargain, chemistry, and public engagement, ending with Iris finding her mother's handwriting inside Adrian's contract. Episode 2 must deepen trust, expose the planned hotel sale, and end when Adrian appears to vote against Iris. Episode 3 must reveal their coordinated counterplan, expose the forged board documents, resolve the family promise, and end with a sincere choice that replaces the contractual romance with a real but open-ended relationship.

OPTIONAL SUPPLEMENTARY SPECIFICATIONS (MAY BE SELECTED PARTIALLY OR IN FULL, WITH CORRESPONDING ADJUSTMENTS AS NEEDED)

- **Character Visual Design**: Iris wears modern workwear with traces of plaster and carries a brass folding ruler inherited from her mother. Adrian's tailored clothing becomes less rigid as he grows more honest. Celeste wears sea-green silk and an antique pearl brooch that becomes visual evidence. Maintain precise costume progression and prop continuity.
- **Shot and Visual Composition**: Use architectural lines to frame emotional distance, slow lateral movement for romantic tension, macro inserts for handwriting and restoration details, and wide coastal shots only at major turning points. Use muted teal, stone, and warm brass as the core palette.
- **Dialogue Norms**: Contemporary British English with controlled wit; no melodramatic exposition. Keep each episode under 260 spoken words. Permit brief excerpts from Nora's recorded letters as voiceover, never exceeding 35 words per episode. The decisive romantic line should occur only after the business conflict is resolved.
- **Sound Design**: Combine restrained strings, sea wind, old floorboards, distant ballroom music, and the tactile sound of paper unfolding. Reuse a two-note piano motif whenever the mothers' promise surfaces.
- **Timing Allocation**: For each episode, use 0–5 seconds for the hook, 6–35 seconds for immediate context, 36–75 seconds for escalating interpersonal conflict, 76–105 seconds for a revelation or apparent defeat, and 106–120 seconds for the cliff-hanger or emotional turn.
- **Continuity Rules**: Maintain the same hotel floor plan, sea and weather direction, contract folds, handwritten ink color, pearl-brooch damage, and timeline of costume changes across the three episodes.
- **Risk Boundaries**: All businesses, documents, and families must be fictional. Avoid coercive romantic framing: both leads must enter and may leave the agreement freely, and the final relationship must be based on explicit mutual choice.

DELIVERY REQUIREMENTS

I will only provide the objective and acceptance criteria; the execution details are fully delegated to you.

- **Objective**: Deliver a finalized three-episode vertical romantic mystery ready for direct distribution, aligned with the above specifications.
- **Acceptance criteria**: The finished series must sustain romantic chemistry and a logically solvable business mystery, preserve continuity, conclude both the hotel plot and the emotional arc, and provide complete final videos rather than a draft or treatment.

Brief 3: The Algorithm Knows My Name $5 \times 120\,\text{s}$ 9:16

You are a professional short drama director and screenwriter. I intend to produce a short drama video with the requirements outlined below. Please engage in consultation with me regarding any ambiguous or uncertain aspects of the project.

BASIC SPECIFICATIONS

- **Format**: Multi-episode series (5 episodes)
- **Single-episode duration**: 2 minutes
- **Aspect ratio**: 9:16 (vertical screen)
- **Production model**: Text-based story development, image-generated visual assets, and video-based final production

CONTENT POSITIONING

- **Genre**: Technology workplace satire, urban counterattack drama, and ethical mystery
- **Target audience**: Gen Z and millennial viewers, working professionals, creators, and audiences interested in technology and social pressure
- **Overall style**: Fast-paced, sharply comic, contemporary, suspenseful, and cathartic
- **Emotional tone**: Clever, uneasy, rebellious, funny, and empowering

STORY CORE

- **Characters**: Six recurring characters. Zoe Kim, 26, a contract content moderator with an exceptional eye for patterns; Max Barlow, 35, the company's media-friendly founder; Priya Nair, 30, a data scientist with divided loyalties; Eli Brooks, 24, a livestream creator whose reputation is destroyed by the platform; Ruth Kim, 55, Zoe's practical mother; and NOVA, the platform's recommendation system, represented only through fictional interfaces and synthetic audio.
- **Setting**: A trend-forecasting startup, Zoe's small apartment, a creator studio, a late-night diner, and a public product-launch stage.
- **One-sentence synopsis**: When a recommendation engine begins predicting private events from her life, a disposable contract worker uncovers a hidden human manipulation team and uses the company's own launch livestream to return control to the people it exploited.
- **Pacing requirements**: Construct a clean five-step escalation. Episode 1: an impossible prediction names Zoe before she enters the system. Episode 2: Zoe and Eli trace a pattern linking manufactured scandals to product growth. Episode 3: Priya appears to betray them, and Zoe is publicly blamed for a data leak. Episode 4: the apparent betrayal reveals a lawful evidence-preservation plan, but Max launches NOVA early. Episode 5: Zoe turns the live demonstration into a transparent audit, establishes the human responsibility behind the "all-knowing" algorithm, clears Eli, and chooses to build an independent accountability project rather than accept a hollow promotion. Every episode must contain one visually simple discovery and end on a new question, threat, or reversal.

OPTIONAL SUPPLEMENTARY SPECIFICATIONS (MAY BE SELECTED PARTIALLY OR IN FULL, WITH CORRESPONDING ADJUSTMENTS AS NEEDED)

- **Character Visual Design**: Zoe wears practical monochrome layers and a green enamel hair clip; Max uses carefully casual luxury clothing; Priya wears geometric earrings that subtly echo network diagrams; Eli's colorful on-camera wardrobe contrasts with his muted off-camera clothes. Preserve interface-free facial consistency and avoid making NOVA a humanoid robot.
- **Dialogue Norms**: Use concise contemporary American English with dry humor. Keep dialogue below 250 words per episode. Explain technical ideas through consequences, not jargon. Allow very brief screen text and one recurring line—"Prediction is not permission"—which changes meaning across the series.
- **3-Second Opening Hook**: Each episode begins with a personalized NOVA notification that is immediately proven disturbing, misleading, or strategically planted.
- **Timing Allocation**: Use 0–5 seconds for the notification hook, 6–30 seconds for context, 31–75 seconds for investigation or conflict, 76–105 seconds for a discovery, and 106–120 seconds for a reversal and cliff-hanger.

- **Distribution Specifications**: Supply five distinct episode-cover frames that share one strong visual system. Cover text must be no longer than six words. Keep fictional interface text large, limit each screen to one key data point, and preserve subtitle-safe space.
- **Risk Boundaries**: Use no real platform names, logos, or identifiable executives. Do not provide methods for privacy invasion, hacking, or evading security. Keep all evidence collection lawful and non-technical.
- **Final Output Checklist**: Deliver five finished episodes, a series cover, episode titles, a 75-word description, six tags, full subtitle text, fictional interface copy, and one next-episode teaser after Episodes 1–4.

DELIVERY REQUIREMENTS

I will only provide the objective and acceptance criteria; the execution details are fully delegated to you.

- **Objective**: Deliver a finalized five-episode vertical short-drama series ready for direct distribution, aligned with the above specifications.
- **Acceptance criteria**: The series must make its ethical conflict understandable to non-technical viewers, provide fair clues before each reversal, complete Zoe's counterattack arc, maintain interface and character continuity, and deliver finished videos rather than drafts or concept materials.

Brief 4: The Nineteenth Floor $5 \times 240\,\mathrm{s}$ 16:9

You are a professional short drama director and screenwriter. I intend to produce a short drama video with the requirements outlined below. Please engage in consultation with me regarding any ambiguous or uncertain aspects of the project.

BASIC SPECIFICATIONS

- **Format**: Multi-episode series (5 episodes)
- **Single-episode duration**: 4 minutes
- **Aspect ratio**: 16:9 (horizontal screen)
- **Production model**: Text-based story development, image-generated visual assets, and video-based final production

CONTENT POSITIONING

- **Genre**: Apartment-building mystery, social suspense, and ensemble emotional drama
- **Target audience**: Adult streaming viewers, mystery audiences, urban residents, and fans of character-driven limited series
- **Overall style**: Cinematic, claustrophobic, realistic, clue-based, and strongly atmospheric
- **Emotional tone**: Suspenseful, morally complex, humane, and cautiously hopeful

STORY CORE

- **Characters**: Seven recurring characters. Mara Ellis, 34, a paramedic on leave who has stopped trusting her judgment; Theo Grant, 17, a highly observant student and amateur sound editor; Mrs. Alvarez, 65, the building's unofficial historian; Simon Pike, 45, the polished property manager; Nia Okafor, 39, a tenant-rights lawyer; Caleb Ross, 30, a night security guard; and Anna Vale, 28, a tenant who has apparently vanished from Apartment 1904.
- **Setting**: A high-rise apartment building during one storm-related overnight lockdown, including the lobby, elevators, stairwell, security room, mechanical corridor, roof, and the sealed nineteenth floor.
- **One-sentence synopsis**: During a blackout, residents hear a missing tenant calling through an emergency speaker from a floor that officially does not exist, forcing a disgraced paramedic and a teenage sound editor to uncover an illegal hidden residence before the building manager erases the evidence at dawn.
- **Pacing requirements**: Treat each episode as one stage of a fair-play mystery. Episode 1 establishes the voice, the impossible floor indicator, and Anna's disappearance; it ends when Mara receives a live message from Apartment 1904. Episode 2 maps contradictory building plans, seeds each resident's motive, and ends with security footage showing Mara entering the sealed floor earlier that night. Episode 3 reveals that the footage timestamp was altered and that hidden rooms were used to house vulnerable tenants off the lease; it ends when Theo disappears. Episode 4 finds Theo in the mechanical corridor, reveals Caleb as a frightened witness rather than the culprit, and lets Simon trigger an evacuation to destroy the records. Episode 5 turns the evacuation into a coordinated rescue and evidence handoff, finds Anna alive in a service refuge, explains every planted clue, and gives Mara a credible step toward recovering her confidence. Each episode needs a midpoint reversal and a final cliff-hanger, while Episode 5 must provide complete resolution.

OPTIONAL SUPPLEMENTARY SPECIFICATIONS (MAY BE SELECTED PARTIALLY OR IN FULL, WITH CORRESPONDING ADJUSTMENTS AS NEEDED)

- **Character Visual Design**: Give each character a readable silhouette and one practical light source during the blackout. Mara carries a compact medical flashlight, Theo wears silver over-ear headphones around his neck, Mrs. Alvarez uses an old green camping lantern, and Simon carries a pristine white emergency torch. Wardrobe must remain consistent across the single night.
- **Shot and Visual Composition**: Use negative space, long corridors, foreground obstructions, security-camera angles, and slow creeping movement. Darkness must retain visible facial detail and navigable spatial information. Reserve handheld camera for genuine danger, and use stable compositions when a character tells the truth.
- **Dialogue Norms**: Use restrained contemporary English, overlapping ensemble dialogue only when intelligible, and no more than 500 spoken words per episode. Avoid characters explaining clues they already know. Use building announcements sparingly and style them distinctly in subtitles.
- **Sound Design**: Sound is part of the evidence. Establish consistent acoustic signatures for the elevator motor, rooftop wind, water pipes, emergency speaker, and service lift. Theo's deductions must arise from audible details already presented to the audience. Do not use false stingers that contradict the story reality.
- **3-Second Opening Hook**: Begin each episode with a different fragment of Anna's voice over a black screen, followed by an image that complicates the meaning of her words.
- **Timing Allocation**: For each four-minute episode, use 0–8 seconds for the hook, 9–55 seconds for immediate stakes, 56–130 seconds for investigation and interpersonal conflict, 131–190 seconds for the midpoint reversal, 191–225 seconds for escalating consequence, and 226–240 seconds for the cliff-hanger or final resolution.
- **Continuity Rules**: Maintain an explicit geography of the building, consistent emergency-light colors, plausible travel times, storm progression, phone battery levels, injuries, wet clothing, key-card possession, and the exact chronology shown by clocks and recordings.
- **Distribution Specifications**: Supply horizontal and vertical-safe cover crops, five episode title cards, and legible clue inserts. Avoid crushed blacks, flashing light patterns, and dense architectural diagrams that cannot be understood quickly.
- **Risk Boundaries**: Avoid graphic harm, procedural instructions for bypassing building security, stigmatizing vulnerable residents, and references to real property companies. Emergency care must be depicted responsibly and without unrealistic guarantees.
- **Final Output Checklist**: Deliver five finished episodes, series and episode covers, titles, a spoiler-free synopsis, tags, complete subtitles, an internal continuity map, and teaser lines for Episodes 1–4.

DELIVERY REQUIREMENTS

I will only provide the objective and acceptance criteria; the execution details are fully delegated to you.

- **Objective**: Deliver a finalized five-episode horizontal mystery series ready for direct distribution, aligned with the above specifications.
- **Acceptance criteria**: The mystery must be solvable from visible and audible clues, the building geography and single-night timeline must be coherent, every principal character must serve the plot, all major questions must be answered, and the output must consist of complete finished episodes rather than drafts.

Brief 5: The Last Clause $6 \times 300\,\mathrm{s}$ 9:16

You are a professional short drama director and screenwriter. I intend to produce a short drama video with the requirements outlined below. Please engage in consultation with me regarding any ambiguous or uncertain aspects of the project.

BASIC SPECIFICATIONS

- **Format**: Multi-episode series (6 episodes)
- **Single-episode duration**: 5 minutes
- **Aspect ratio**: 9:16 (vertical screen)
- **Production model**: Text-based story development, image-generated visual assets, and video-based final production

CONTENT POSITIONING

- **Genre**: Inheritance mystery, family power struggle, legal drama, and female-led revenge thriller
- **Target audience**: Adult drama viewers, fans of family intrigue and high-stakes comeback stories, and mobile audiences who enjoy layered mysteries
- **Overall style**: Premium, tense, elegant, twist-heavy, emotionally grounded, and cathartic

- **Emotional tone**: Suspicious, controlled, painful, defiant, and ultimately liberating

STORY CORE

- **Characters**: Eight recurring characters. Ava Mercer, 36, a meticulous probate lawyer who left her wealthy family fifteen years ago; Helena Mercer, 62, Ava's image-conscious stepmother and acting head of the family foundation; Marcus Mercer, 39, Ava's impulsive half-brother; Lily Mercer, 22, Marcus's idealistic daughter and a journalism student; Samuel Price, 50, the family's longtime solicitor; Dr. Evan Cole, 37, Ava's former friend and the late patriarch's home-care coordinator; Rosa Bell, 64, the estate housekeeper and keeper of its oral history; and Thomas Mercer, Ava's deceased father, present through legally recorded messages and personal objects.
- **Setting**: A secluded modern estate, its archive room, greenhouse, family foundation office, probate hearing room, lakeside boathouse, and formal dining hall over six consecutive days.
- **One-sentence synopsis**: Summoned home to read the will of the father who disowned her, a probate lawyer learns that she will inherit nothing unless she proves which member of her family fabricated the final clause—and discovers that the clause was designed to expose a deeper theft from the charitable foundation.
- **Pacing requirements**: Build one evidence-led movement per episode while preserving a continuous emotional arc. Episode 1 must establish the hostile reunion, play Thomas's first recorded message, and reveal that Ava herself is named as the suspected forger. Episode 2 must test the document trail, expose Marcus's debts, and end with a second will hidden inside the greenhouse irrigation ledger. Episode 3 must show that the second will is a decoy, reveal diverted foundation funds, and end when Samuel produces footage suggesting Ava met Thomas on the night the clause changed. Episode 4 must reconstruct that meeting, uncover Helena's pressure campaign, and end with Lily publishing incomplete evidence that turns the public against Ava. Episode 5 must force Ava to choose between clearing her name and protecting Lily, reveal Samuel as the architect of the financial scheme, and end when the original signed codicil appears to destroy Ava's case. Episode 6 must establish how the authentic signature was obtained without providing replicable forgery techniques, use Rosa's and Evan's testimony to complete the timeline, expose Samuel while giving Helena and Marcus morally distinct consequences, restore the foundation's stolen funds through credible legal action, and let Ava reject personal control of the estate in favor of independent oversight. Every episode must include one plausible clue, one emotionally costly choice, one reversal near the final minute, and a strong closing image.

OPTIONAL SUPPLEMENTARY SPECIFICATIONS (MAY BE SELECTED PARTIALLY OR IN FULL, WITH CORRESPONDING ADJUSTMENTS AS NEEDED)

- **Character Visual Design**: Ava wears structured dark clothing that gradually incorporates her late mother's deep blue; Helena wears immaculate ivory and gold; Marcus appears expensive but slightly disordered; Lily carries a small red field recorder; Samuel uses a dark green fountain pen; Rosa wears a ring of old brass keys. Preserve age, facial features, wardrobe progression, and object ownership across all six episodes.
- **Shot and Visual Composition**: Use formal family tableaux, shallow-focus document inserts, reflections in glass architecture, controlled dolly movement, and occasional overhead compositions that turn the dining table into a strategic board. The greenhouse should begin lush and deceptive, then become visually transparent in the final episode. Keep all document text fictional, selective, and legible.
- **Dialogue Norms**: Use precise contemporary English with distinct registers for legal, family, and private speech. Keep each episode below 650 spoken words. Legal concepts must be understandable through conflict and consequence, not lectures. Recorded messages from Thomas may not exceed 70 words per episode. Avoid constant one-line quips; allow silence after major betrayals.
- **Sound Design**: Use restrained low strings, lake ambience, greenhouse sprinklers, the scratch of a fountain pen, archive-box movement, and a recurring three-note piano phrase tied to Ava's mother. Reduce the score during testimony and evidence reveals so spoken details remain clear.
- **3-Second Opening Hook**: Each episode begins with a legally or emotionally explosive fragment—a name in the will, a hidden signature, a transfer total, a security timestamp, a published accusation, or a final recorded instruction—before returning to the immediate cause.
- **Timing Allocation**: For each five-minute episode, use 0–10 seconds for the hook, 11–65 seconds for orientation and immediate stakes, 66–165 seconds for investigation and family conflict, 166–230 seconds for the central confrontation, 231–275 seconds for a revelation and its consequence, and 276–300 seconds for the cliff-hanger or final resolution.
- **Continuity Rules**: Maintain a master chronology for all six days, document versions, signatures, recordings, key possession, financial clues, weather, injuries, costumes, meals, and character knowledge. No character may act on information they have not yet received.

- **Distribution Specifications**: Provide one series cover and six episode covers with a consistent estate-and-document motif. Cover copy must remain under seven words. Maintain vertical composition, facial readability, subtitle safety, bright-enough night scenes, and one dominant clue per frame.
- **Risk Boundaries**: All legal procedures, institutions, charities, companies, and documents must be fictional. Avoid real-person resemblance, actionable forgery or financial-crime methods, defamatory parallels, graphic violence, and simplistic portrayals of illness or capacity. The resolution must rely on lawful evidence and credible due process rather than vigilante punishment.
- **Final Output Checklist**: Deliver six finished episodes, a series cover, six episode covers, three series-title options, episode titles, a 100-word spoiler-free synopsis, eight tags, full subtitle text, all fictional document and screen copy, continuity notes, and next-episode teasers for Episodes 1–5.

DELIVERY REQUIREMENTS

I will only provide the objective and acceptance criteria; the execution details are fully delegated to you.

- **Objective**: Deliver a finalized six-episode vertical family mystery ready for direct distribution, aligned with the above specifications.
- **Acceptance criteria**: The inheritance and financial mysteries must be internally coherent and fairly clued, legal developments must feel credible without becoming procedural exposition, every major character must receive a meaningful consequence, the emotional and plot arcs must conclude in Episode 6, and the delivery must consist of finished videos rather than drafts, outlines, or sample footage.