CREATORBENCH: BENCHMARKING AGENTS ON MULTIMODAL CREATIVE PRODUCTION FROM BRIEF TO DELIVERABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Content studios turn structured briefs into finished creative deliverables—posters, advertisements, short dramas, interactive videos—through multi-step pipelines under hard specification constraints; a single failure in asset usage, text rendering, consistency, or technical compliance can reject the entire delivery. No existing benchmark takes a real business brief as input, covers both image and video, and evaluates the finished deliverable against that brief. CreatorBench fills this gap with 120 tasks (60 image, 60 video) abstracted from commercial studio workflows and organized into 14 task families. Each task provides a structured brief, source assets, brand facts, delivery specifications, and multi-step constraints; the agent autonomously plans, produces, and delivers final files. Human experts decompose each task into a weighted checklist of verifiable requirements. Scoring centres on per-case requirement fulfilment: a primary and an audit VLM cross-score each checklist item blindly. Deterministic corrections—objective-metric verification, defect caps, and delivery-specification gates—yield the requirement-satisfaction score R; a joint technical-quality and visual-appeal modifier produces the final score. The resulting automated ranking aligns with human Good/Same/Bad preference judgments. We evaluate five systems under two toolsets with different generation backends. The strongest system (Codex) reaches 59.67/100 under Toolset A; no system exceeds 60 and the Baseline scores 17.78/100. These results suggest that reliably turning a business brief into an acceptable deliverable still lies beyond current agent capabilities.

1 INTRODUCTION

Visual content production requires more than generating an appealing image or video. Given a business brief and source materials, a content-creation agent must determine what to produce, coordinate generation and editing tools, and assemble outputs that satisfy the requested delivery requirements. These requirements may concern exact text, visual hierarchy, subject identity, temporal continuity, or consistency across multiple files. This raises a central question: can an agent autonomously turn a production brief into a complete visual delivery that meets both functional and perceptual requirements?

The difficulty lies in satisfying these requirements together. A promotional campaign may require several layouts that preserve the same product identity and message across different aspect ratios. Each layout must contain the correct information, while making the intended selling point visually clear. Video production adds constraints across time: an edit must modify the requested regions while preserving others, and a multi-scene narrative must maintain character and story continuity. Such tasks can require agents to interpret specifications, select assets, coordinate dependent operations, inspect intermediate outputs, and revise failures. A successful tool call, or even an attractive individual output, provides only partial evidence that the overall task has been completed. Figure 1 makes this distinction concrete: Task 55 turns two supplied brand assets into a coordinated three-format advertising suite whose product identity and selling-point copy must survive across portrait, landscape, and square layouts, whereas Task 71 requires a shared-root interactive narrative whose two endings remain continuous in character identity, prop placement, and lighting across clips and shots.

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

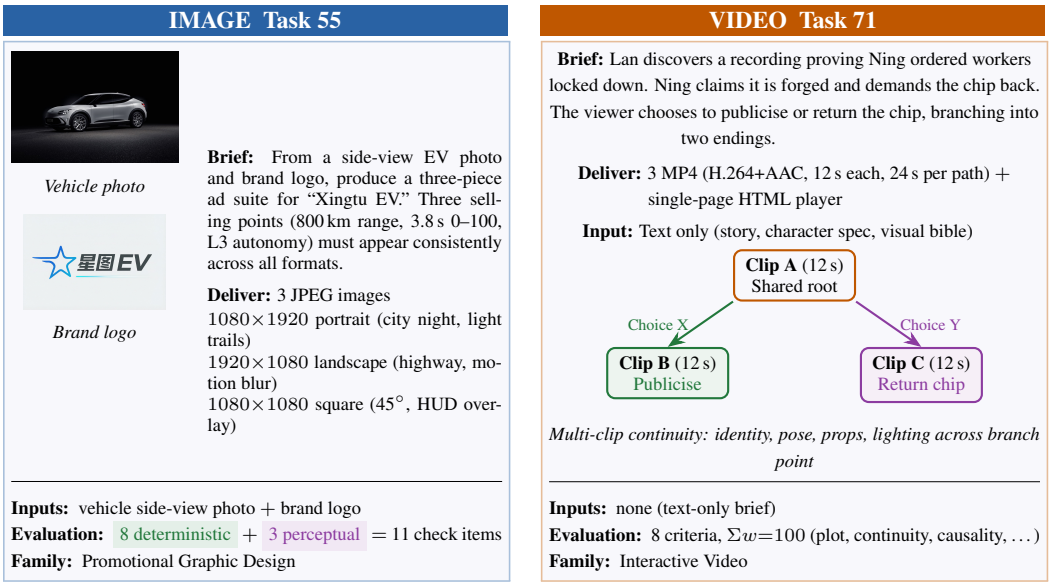

Figure 1: Two representative CreatorBench tasks. **Left**—Task 55: two brand assets yield a three-format ad suite; product identity and copy must remain consistent across layouts (11 check items). **Right**—Task 71: a text-only brief produces three branching clips plus an HTML player; evaluation enforces cross-clip identity, causal coherence, and narrative continuity (8 weighted criteria).

Existing benchmarks provide important foundations but leave this capability insufficiently characterized. General-purpose agent benchmarks evaluate execution through observable task outcomes, including environment states and software tests (Liu et al., 2023a; Xie et al., 2024; Jimenez et al., 2023). Generation and editing benchmarks assess dimensions such as instruction alignment, perceptual quality, and temporal consistency (Zhang et al., 2023; Huang et al., 2023b; Li et al., 2024). Recent design and media-production benchmarks bring evaluation closer to creative workflows within particular environments or artifact types (Sun et al., 2025; Deganutti et al., 2026; Hu et al., 2026). None of them, though, takes the delivery as its unit of judgment. What a studio hands over is a set of outputs that must satisfy the brief jointly—three campaign layouts are one delivery bound by shared constraints, not three unrelated generation requests—and quality has to be weighed at that same level, since a layout can look good on its own and still break a constraint that binds the whole set.

We introduce CreatorBench, 120 tasks drawn from commercial visual-production workflows—60 image and 60 video, across 14 families spanning graphic design, precision image editing, product visualization, narrative video, and compositional video editing. Every task ships with a structured brief giving its business context, content requirements, and delivery constraints, plus the relevant source assets. Deliverables run from single images and clips to coordinated output sets; interactive video tasks also require a player implementing the specified branching logic. Agents reach generation, editing, and visual-understanding tools through one interface and may write code for composition, conversion, or verification. We leave execution strategy open, so different production routes can be compared against the same requested outcome.

Evaluation begins after the agent submits its deliverable files. The evaluation unit is the task together with its complete delivery, including relationships among outputs where required by the brief. The protocol uses programmatic checks for verifiable properties, perceptual proxy metrics for applicable visual attributes, and task-specific vision-language-model rubrics for qualities such as information hierarchy, editing correctness, and narrative coherence. The three sources cover different failure modes: a poster can contain all required text and still communicate it poorly, while a visually compelling edit can violate a preservation constraint. Image and video tracks share the same scoring pipeline and are reported separately, with modality-specific objective metrics and explicit handling of metric applicability and missing or unscorable evidence.

Our experiments pair four agent systems and a direct-generation baseline with two alternative tool suites, so that agents can be compared under a shared tool configuration and the effect of the generation backend read off separately. The baseline anchors what agent-mediated execution is worth; resource consumption and wall-clock time are logged alongside delivery quality. In practice this lets us say where agents succeed, which delivery requirements stay hard, and how much of the outcome depends on the tools they were given.

CreatorBench contributes a task suite grounded in real production work: 120 image and video tasks across 14 families, each pinned down by a structured brief, source assets, and explicit delivery requirements. Around it we build an evaluation protocol that judges the delivery rather than the process—functional compliance and perceptual quality assessed per task, for single artifacts and coordinated multi-file outputs alike, with no execution path prescribed. We then use both as an experimental framework for content-creation agents: comparisons across agent systems and generation backends, a direct-generation baseline as reference point, and quality reported next to resource expenditure, so that any claim of superiority carries its cost.

2 Related Work

Agent and long-horizon task benchmarks. AgentBench (Liu et al., 2023a) probes decision-making across varied environments; OSWorld (Xie et al., 2024) places multimodal agents inside a real operating system and judges completion by the resulting state; SWE-bench (Jimenez et al., 2023) scores against the test suite of a resolved GitHub issue; TaskBench (Shen et al., 2023) studies structural correctness of task decomposition. LongHorizon-Harness (Ma et al., 2026) recasts long-horizon completion as persistent state management through a manage-execute-audit loop, scoring the verified delivery state rather than intermediate steps, close in spirit to the post-hoc evaluation we adopt. All of them treat the outcome as the unit of judgment (Zhou et al., 2023; Mialon et al., 2023; Ma et al., 2024), but for visual content an outcome is two things at once: a functional specification and a perceptual artifact. A poster can be dimensionally correct and textually complete, and still fail to communicate its selling point because its information hierarchy is broken. We keep the outcome-first stance and push it where creative work requires—completion is read off the delivered end state, never off conformance to a fixed plan (Qin et al., 2023; Yao et al., 2023), and the outcome has to clear both the business constraints in the brief and the visual expression of the finished files.

Multimodal generation evaluation. GenAI-Bench (Li et al., 2024) covers compositional text-to-visual requirements; VBench (Huang et al., 2023b) provides multi-dimensional video-generation quality suites; VIEScore (Ku et al., 2023) explores VLMs as evaluators of conditional synthesis. FingER (Chen et al., 2025) pushes evaluation granularity to entity-level VLM question answering, decomposing each generated video into fine-grained factual queries rather than holistic ratings. NarrLV (Feng et al., 2026) works along an independent dimension—narrative in long video—and designs Temporal Narrative Atoms grounded in film theory to assess story coherence across minutes-long sequences. On the quality-measurement side, Zhu et al. (2026a) decompose fine-grained video artifacts along appearance, motion, and temporal axes, and Peak-End-Net (Li et al., 2026) applies the psychological peak-end rule to predict aesthetic quality from critical and closing frames. These lines of work supply a rich vocabulary for perceptual quality (Heusel et al., 2017; Hessel et al., 2021; Huang et al., 2023a; Unterthiner et al., 2018), but their evaluation unit is a single generation, not a commercial delivery: the same promotion re-composed at vertical, horizontal, and square sizes is not three independent requests but one task whose constraints must hold in all frames. CreatorBench moves evaluation to the complete delivery.

Design and media production. CutVerse (Hu et al., 2026) evaluates GUI-based video editing agents within a single post-production tool, scoring operation-level correctness rather than end-to-end creative outcome. Graphic-Design-Bench (Deganutti et al., 2026) tests graphic-design generation on layered templates but incorporates no delivery specifications, no coordination across multiple tools, and no video track. GraphicBench (Ki et al., 2025) benchmarks planning for graphic design with language agents. On the production-tool front, DreamX-Creator (Zhu et al., 2026b) demonstrates native audio-video joint generation at 2K resolution within a single 7B-parameter model, and Omni-Effects (Mao et al., 2026) unifies prompt-guided visual-effects synthesis with spatial controllability—both a reminder that the generation tools behind a benchmark like ours are them-

Table 1: Task-suite composition. CreatorBench spans 14 families across image and video tracks (60 tasks each, 120 total). Full family definitions appear in Appendix A.2.

Track	Task Family	*n*	Representative Challenges
Image	Promotional Graphic Design	11	POI metadata + album → multi-format ads
	Precision Image Editing	20	remove / replace / inpaint / super-resolve
	3D & Storyboard Composition	15	multi-view 3D scenes; storyboard panels
	Travel & Footprint Poster	6	activity log + album → retrospective poster
	Multi-format Asset Adaptation	2	spec → consistent multi-size graphics
	E-commerce Product Visualization	6	product photo → detail page / mockup
Video	Narrative Video Production	15	script → multi-episode short drama
	Compositional Video Editing	8	overlapping edits with dependency chains
	Video Generation	7	first frame → physically plausible motion
	Food & Scenic Video	5	scene photo → atmospheric commercial clip
	Commercial Cinematic Production	10	product/brand → VFX promotional film
	3D & Spatial Video	5	first frame → 360° orbit / multi-axis / FPV
	Interactive Video	5	branching narrative → clips + HTML player
	Multi-source Video Remix	5	multi-source + audio → beat-synced mashup
	Total	**120**	

selves moving quickly. CreatorBench differs where it matters for delivery: it spans image and video inside one evaluation protocol, with separate scoring tracks per modality; it treats commercial delivery constraints—resolution, duration, codec, file count—as hard gates rather than soft preferences; and it grades autonomous multi-step execution of real business briefs on the finished deliverable, through a harness that sits outside the agent.

3 Benchmark Tasks and Data

Task sources and abstraction. The 120 tasks in CreatorBench are abstracted from commercial content-production workflows. Each task originates from a real studio request that was de-identified during collection: client and brand names, internal contacts, and commercial terms were removed or substituted. What remains is the *structural complexity*—the brief, reference materials, brand facts, delivery specifications, and multi-step constraints—without the commercial identity: production-level difficulty, no proprietary content.

Composition. Tasks are organized into 14 families across two modality tracks—6 image families and 8 video families, 60 tasks each (Table 1; full definitions in Appendix A.3). These 14 families are mapped from 19 fine-grained scene or track categories in the underlying evaluation data; the full mapping appears in Appendix A.2. The distribution is non-uniform by design: Precision Image Editing accounts for 20 tasks and Narrative Video Production for 15, while Multi-format Asset Adaptation holds 2. This imbalance reflects the empirical frequency of source workflows rather than a judgment of relative importance. Tasks 001–100 form the core set; twenty additional tasks strengthen sparsely covered families, giving 120 tasks across 14 families with a verified total of $60 + 60 = 120$.

Each task is specified through a *structured brief* carrying business metadata (purpose, audience, channel), delivery constraints (resolution, aspect ratio, duration, codec), content relations (what must appear, what must not change), and references to source assets. An external agent decides how to execute the task; the benchmark prescribes no execution plan and evaluation attends only to whether the final files satisfy the brief.

What makes the tasks hard. Three categories of difficulty distinguish CreatorBench from single-generation benchmarks:

Long-context integration. Each brief composes multiple information sources: business metadata, promotional selling points, reference photo albums, brand guidelines, and technical delivery requirements. A footprint poster (Task 8) requires filtering 6 photographs to 3, generating an abstract route map, and integrating annual travel statistics—all from a single structured brief.

Multi-stage planning and dependency. Tasks require agents to sequence asset selection, content generation, editing, layout, and export with dependency ordering. A compositional video editing

task (Task 76) specifies three modifications (add mist, replace sky, remove trees) with explicit dependency constraints and must-preserve regions.

Cross-modal and temporal consistency. Video tasks demand coherence across time and modality. Interactive video (Task 71) requires three MP4 clips with a functional HTML player implementing branch logic; short drama production (Task 91) demands character-identity consistency across shots with synchronized audio; 3D spatial video requires multi-axis geometric coherence.

4 Demand-Oriented Evaluation

4.1 Checklist Construction

A team of five domain experts with backgrounds in commercial content production jointly reviewed every task. For each case the team decomposed the structured brief into atomic, verifiable *check items*—specific requirements such as "headline dominates supporting copy," "robot vacuum visible in first 3 seconds," or "character face consistent across all shots." Each item received a pre-registered weight derived from its failure severity—fatal, significant, or minor, carrying unit ratios 4:2:1—and normalized within the case so that weights sum to exactly 100. Consequently, items whose failure would cause irreparable damage to the deliverable (e.g., missing key assets or violating mandatory specifications) carry proportionally higher weight than cosmetic shortcomings with low repair cost. The checklist was frozen before any system executed, so the evaluation specification is pre-registered rather than adjusted once results were seen. This was a *joint review process*: the team worked collaboratively on each case, not as independent annotators, and no inter-annotator agreement was measured. Appendix A.4 describes the review workflow and freeze timeline.

All 100 points within a case checklist are allocated to requirement items; technical quality (Q) and visual appeal (A) are assessed by a separate quality track and enter the final score as a multiplicative modifier (Step 2 below). On the image track, atomic check items span multiple `check_type` categories covering general and scenario-specific dimensions (Appendix A.7). On the video track, each frozen criterion migrates losslessly to one check item (Appendix A.7).

Concrete example: image checklist (Task 55). Table 2(a) illustrates how the five-stage pipeline of Case 55 (§ 3) maps to atomic check items. Each ability stage contributes one to three items; most are deterministic—mechanically decidable without model judgment. The remainder are perceptual, requiring VLM assessment of qualities such as compositing realism and brand fidelity. This deterministic/perceptual split is characteristic of image tasks: rule-based guards catch specification failures early, narrowing the scope of subjective judgment to genuinely perceptual questions.

Concrete example: video checklist (Task 71). Table 2(b) shows the video-side counterpart. The eight criteria migrate one-to-one from the frozen `evaluation_criteria.json`, and all 100 points sit on requirement items. Four criteria (plot completeness, continuity, character consistency, ending validity) each carry $\geq$10% weight, reflecting the domain emphasis on narrative coherence and character identity; video quality receives only 10%, a deliberate choice because visual fidelity alone cannot rescue a causally incoherent story.

4.2 Cross-Family Scoring and Deterministic Correction

Scoring begins only after all deliverables are written to disk, governed by the frozen scoring rulebook (Appendix A). Three forces determine each item's final score: the blind judgments of two VLMs drawn from different model families, given only the task brief; objective metrics that supply deterministic verification or conflict evidence; and hard delivery-specification guards that cap or zero a case when mandatory technical requirements are violated. The pipeline realises them in four stages:

Stage 1: independent observations. Two VLMs from different model families (§ 4.3) score every frozen check item independently on a 0–5 rubric (Liu et al., 2023b; Kim et al., 2024); neither sees the other's response, and neither is told which system produced the files. In parallel, the harness computes the applicable objective metrics (file existence, dimension checks, OCR verification, hash-

Table 2: **Checklist construction examples.** (a) Image Case 55: check items across five ability stages. (b) Video Case 71: criteria from `evaluation_criteria.json`. Weights sum to 100 in both cases.

(a) Image — Task 55

Check item	w
Input asset membership verified	12
Composited artefact exists & readable	
Perspective / lighting fusion natural	25
No matting residual at edge	
Brand-element artefact exists	
Logo, slogan, model name rendered	20
Chinese selling-point copy passes OCR	
$\geq$2 adapted versions exist	25
Pixel dimensions match spec	
Final deliverable exists & readable	18
JPEG format, 3-channel	
Σw	100

(b) Video — Task 71

Check item	w
Plot completeness: full narrative arc per branch (setup, conflict, resolution), not shot splicing	20
Continuity: B/C first beat inherits A end-frame state—positions, props, costume, lighting, door status	15
Character consistency: two roles only; face, hair, costume, voice stable across A/B/C with no drift	15
Choice causality: viewer choice maps to exclusive outcomes with visible causal divergence	10
Branch divergence: end-states reversed—chip slotted vs. snapped, door open vs. sealed, beam retargets	10
Dialogue & sound: one speaker per beat, no overlap; audio matched to lip/action; synced subtitles	10
Ending validity: one positive, one negative ending, both caused by the choice; no reversal rescue	10
Video quality: motion natural, no flicker or clipping; editing rhythm serves narrative	10
Σw	100

based reuse detection). Each judge must supply clause-level evidence grounded in the media, along with any structured defects it identifies.

Stage 2: evidence gating and conservative synthesis. Scores of 4 or 5 that lack supporting evidence are capped at 3—no judge can award high marks without pointing to specific verifiable content. The two judgments are then combined conservatively: an item that neither judge lifts to 4 or above keeps the first seat's score, the requirement already being recorded as unmet, while an item that does reach the high band takes $\min(s_{\text{primary}}, s_{\text{audit}})$, the subscripts labelling seats rather than rank. A favorable reading must therefore be reproduced across the family boundary before it can count, and neither family can raise a score on its own authority. Where a fatal or significant defect is raised by only one judge, it goes to the other for cross-verification before it is allowed to move the score.

Stage 3: deterministic correction. Objective rules are then applied. These rules can only reduce or maintain item scores—never inflate them. Deterministic measurements (file format, pixel dimensions, text OCR, hash comparison) may set, zero, or cap an item score directly. Perceptual proxies such as CLIP cosine (Radford et al., 2021), LPIPS distance (Zhang et al., 2018), and aesthetic scores (Schuhmann et al., 2022) only flag conflicts for the record and are never blended into the score. Confirmed defects impose severity-weighted caps on the affected item. Finally, four case-level gates (no delivery, unreadable delivery, exact input reuse, overall mismatch) may zero an entire case, and delivery-specification failures impose case-level ceilings—for example, an image dimension mismatch caps the case at $\leq$20/100 (Appendix A.10).

Stage 4: aggregation. Aggregation proceeds in two steps once all item-level corrections are applied.

Step 1—Requirement score. Each check item's post-correction score $s_i \in [0, 5]$ is combined with its pre-registered weight w_i ($\sum_i w_i = 100$) to yield the requirement score

$$R \;=\; \textstyle\sum_i w_i \cdot s_i \,/\, 5, \qquad R \in [0, 100]. \tag{1}$$

R measures how thoroughly the deliverable satisfies every business and specification constraint in the brief.

Step 2—Quality-adjusted case score. Requirement compliance is necessary but not sufficient: in commercial production a deliverable that meets every specification yet exhibits poor rendering quality or uninspired composition still fails acceptance. We therefore introduce an independent VLM

quality assessment that produces a technical-quality score $Q \in [0, 5]$ and a visual-appeal score $A \in [0, 5]$. This quality track operates separately from the requirement checklist and is optional in the evaluation framework; all scores reported in this paper have it enabled. The case score is

$$\text{case_score} \;=\; R \;\times\; \big(0.60 + 0.20\,Q/5 + 0.20\,A/5\big). \tag{2}$$

The quality factor ranges over $[0.60, 1.00]$. The 0.60 floor ensures requirement satisfaction dominates—even when Q and A are both zero, 60% of R is preserved. Technical quality and visual appeal each modulate up to 20%, enough to separate deliverables with similar compliance but different craft. The multiplicative form guarantees $R{=}0 \Rightarrow$ case_score$=0$: an aesthetically polished submission that ignores the brief receives no credit.

The version-level total averages the image and video track means with equal weight; each track mean is computed over a fixed denominator of 60 cases (Appendix A.6).

4.3 Cross-Family Judge Pair and Conservative Synthesis

Image and video cases share one blind-review architecture, and the two judges sit on opposite sides of a model-family boundary: `qwen3.8-omni-flash` (Qwen) and `gemini-3.7-flash` (Google) score the identical frozen checklist without access to each other's responses, and emit per-item scores, clause-level evidence, and structured defects in the same schema. We chose that pairing deliberately. Self-preference is a documented failure mode of LLM judges (Panickssery et al., 2024), and one system under test is built on a Qwen backbone; a judge drawn from that family alone would have been free to favor it. Because the two judges share no weights and no training corpus, an inflated reading from one family has to survive an independent reading of the same media by the other before it reaches the score.

Synthesis between them is conservative rather than averaged, and defect traffic runs in both directions: a fatal or significant defect raised by one judge is verified by the other before it takes effect, whichever side raised it. The combined score then enters the deterministic correction sequence—objective rules, defect caps, case gates, and delivery-specification ceilings—described in § 4.2. Judge failures are explicit: if the first seat returns nothing valid the case is unavailable, whereas an invalid response from the second is recorded and flagged as a single-judge result. Appendix A.12 retains the implementation labels *primary* and *audit* for the two seats and documents media transport and auditable failure semantics; those labels record call order, not rank.

4.4 Protocol Validation against Human Judgment

We validate the automated ranking with a blind pairwise preference study. Five domain experts—the same team that constructed the checklists (§ 4.1), though they did not participate in scoring-rule tuning during this phase—independently compared deliverables for 20 cases (10 image, 10 video) sampled by stratified family selection under Toolset B. Each evaluator judged all $\binom{5}{2} = 10$ system pairs per case, labelling each pair Good (left preferred), Same, or Bad (right preferred), yielding $5 \times 20 \times 10 = 1{,}000$ valid preference triplets.

Inter-evaluator agreement reaches Fleiss' $\kappa = 0.74$ (substantial). Aggregating GSB counts into per-system win rates produces a ranking identical to the automated scores in Table 3: Kendall's $\tau = 1.0$ and Spearman's $\rho = 1.0$. The correspondence is sharpest at the extremes—Codex is preferred over the Baseline in more than 95% of pairwise judgments—while the mid-range systems (Qoder and Workbuddy) show a narrower split near 52:48, consistent with their close automated scores (44.10 vs. 42.51). The complete agreement between human preference order and automated ranking supports the validity of the scoring pipeline for system-level comparisons.

5 Experiments and Results

5.1 Systems and Setup

We pair two independent toolsets with five systems, producing a 2×5 evaluation matrix. **Toolset A** pairs Qwen-Image-3.0-Pro (text-to-image, image editing) with Wan3.0-Video (video generation/editing); **Toolset B** pairs GPT-Image-2.0 (text-to-image, image editing) with Seedance 2.0

Table 3: Overall scores (0–100) for five systems under both toolsets. Image and video track averages are combined with equal weight; only cases with valid scores are included in each track mean. R, Q, and A denote the requirement-fulfillment, quality, and aesthetic sub-scores, averaged over all included cases of both tracks with the variable denominators described in § 4.2.

(a) Toolset A *(Qwen-Image-3.0-Pro + Wan3.0-Video)*

System	Final	Image	Video	R	Q	A
Codex	59.67	64.75	54.59	69.05	3.43	3.09
DeepSeek-Harness	56.47	64.06	48.88	64.95	3.57	3.10
Workbuddy	46.59	44.79	48.39	52.77	3.61	3.25
Qoder	41.93	52.55	31.31	48.21	3.72	3.15
Baseline	14.96	19.65	10.27	17.58	3.60	3.10

(b) Toolset B *(GPT-Image-2.0 + Seedance 2.0)*

System	Final	Image	Video	R	Q	A
Codex	58.34	67.06	49.62	67.21	3.55	3.02
DeepSeek-Harness	56.21	64.45	47.97	64.51	3.51	2.93
Workbuddy	42.48	45.14	39.83	48.63	3.67	3.12
Qoder	43.51	56.90	30.11	50.38	3.65	3.08
Baseline	17.78	19.90	15.66	20.31	3.66	2.97

(video generation/editing) (Rombach et al., 2022; Betker et al., 2023). Both toolsets provide four capability categories: text-to-image generation, image editing, VLM-based visual understanding, and video generation/editing with optional audio. All tools are Python scripts; invocations are logged to JSONL telemetry files recording call signature, latency, and token usage.

The five systems are **Codex** (GPT6-Astra), **Qoder** (Qwen3.8-max), **Workbuddy** (Hy4-preview), **DeepSeek-Harness** (DeepSeek-V4.1-Flash), and the **Baseline** (No-Agent loop), where the brief enters a fixed generation pipeline for single-turn synthesis with `loop=1` enforced by harness design—no agent-driven planning, re-prompting, or adaptive tool chaining is performed (Yao et al., 2023). All systems receive the identical deliver package, share the same Python-script tool interface, and operate under frozen running conditions (Table 26, Appendix A.15). The full 120 cases are executed within each of the 10 (system $\times$ toolset) runs; every task is executed once in serial (one case at a time); artifacts are scored offline. No human raters participate in scoring.

5.2 Main Results

The component breakdown in Table 3 reveals that requirement satisfaction (R) is the dominant source of variance: across the ten system–toolset cells it spans 17.58 to 69.05, while quality (Q: 3.43–3.72) and aesthetics (A: 2.93–3.25) cluster within narrow bands on the 1–5 scale. Even the agent-free Baseline matches the four agent systems on Q and A, so under the multiplicative scoring rule there is no route by which strong generation quality could rescue a low R. Codex attains the highest R under both toolsets (69.05/67.21) with DeepSeek-Harness second (64.95/64.51), and the R ordering reproduces the final ranking in both toolsets. The pattern indicates that brief comprehension and iterative requirement fulfillment, not visual craftsmanship, separate the current generation of creative agents.

6 Quantitative and Pattern Analysis

Toolset effect. The toolset effect is mixed (Table 3): Codex ($\Delta{=}-1.33$), DeepSeek-Harness (-0.26), and Workbuddy (-4.11) score higher under Toolset A, while Qoder ($+1.58$) and the Baseline ($+2.82$) favor B. The signed Δ(B$-$A) spans a 6.93-point range centered near zero; neither toolset is uniformly superior, indicating that toolset–agent compatibility is system-specific.

Agent gain. Agent gains over the Baseline range from $+26.97$ to $+44.71$ (Toolset A) and $+24.70$ to $+40.56$ (Toolset B); since the five systems differ in their reasoning backbones, the gap reflects the aggregate contribution of planning, iteration, tool selection, and error correction.

Image–video gap. Under Toolset B, all five systems score higher on image tasks: Qoder shows the widest split (56.90 vs. 30.11, gap 26.79), followed by Codex (+17.44), DeepSeek-Harness (+16.48), Workbuddy (+5.31), and the Baseline (+4.24). Video tasks compound temporal consistency, audio–visual alignment, and—on interactive tasks—branch coherence, letting a single failure cascade under the preregistered rubric.

Score composition. Discrimination across systems comes mainly from requirement satisfaction (R): under Toolset B, mean R spans 20.31 to 67.21 (Table 3), while the quality factor $0.60 + 0.20\,Q/5 + 0.20\,A/5$ rounds to 0.86 or 0.87 for every system and Q and A each span less than 0.2 points, indicating that the bottleneck is fulfilling the brief rather than aesthetic craftsmanship.

7 Limitations and Conclusion

Limitations. Our most consequential weaknesses concern how the ground truth was produced. Five domain experts decomposed all 120 briefs into weighted check items, but they did so as one review team rather than as independent annotators: we report no inter-annotator agreement and obtained no third-party verification, so the checklists document consolidated expert judgment rather than statistically validated ground truth. The calibration study inherits that closure, tuning the rule parameters and reporting agreement against author-assigned bands on the same 15 cases (Appendix A.11); it evidences internal consistency, not generalization. A held-out validation set is left to future work.

Judge bias is constrained by design but not yet measured. The two judges come from different model families, and no item reaches the high band unless both families' independent readings support it, which closes the most direct route by which a Qwen-family judge could favor Qoder—the one system under test built on a Qwen backbone (Panickssery et al., 2024). What is missing is a quantification of the residual: we have not run the judge-swap ablation in which the two seats exchange roles and every system is re-scored under both orderings. Until that is done, the cross-family guarantee should be read as an architectural argument rather than a measured quantity.

The remaining limitations concern data rather than method. Several systems carry internal run identifiers instead of documented version numbers, and the Codex backbone in particular could not be confirmed from the logs. The tight clustering of Q and A scores may partly reflect limited discriminative power of these dimensions under the current rubric rather than genuine system equivalence; because the same VLM judge pair scores all three components, correlated biases across R, Q, and A cannot be ruled out. Family sizes range from $n{=}2$ to $n{=}20$, which leaves the smallest families disproportionately influential in any family-level breakdown.

Conclusion. What the results say most plainly is that this task is not solved. Codex, the strongest system, reaches 59.67 under Toolset A and nothing clears 60/100—even though all four agent systems beat the direct-generation baseline by +26.97 to +44.71 points, so agency helps a great deal without carrying a commercial delivery to acceptance. Our design does not separate planning from tool selection or iterative correction, so that gain can only be read as an aggregate. Two further patterns shape the gap. Toolset preference is system-specific, with Δ(B$-$A) spanning -4.11 to $+2.82$ and neither toolset uniformly superior, which means the backend cannot be chosen independently of the agent. Video remains the harder medium: all five systems score higher on images under Toolset B, by up to 26.79 points in Qoder's case, because temporal consistency and branching coherence let a single failure cascade. The component breakdown confirms that requirement satisfaction, not generation quality, drives the ranking: Q and A vary by less than a third of a point across all ten system–toolset configurations while R spans a 51-point range. Brief-level fulfillment, not generation quality, looks like the binding constraint here, and delivery-level evaluation of this kind will have to become standard before capability claims in creative production can be trusted.

AI use statement

This work studies AI systems, and we distinguish that role from AI assistance in preparing the paper. The five systems under test comprise four AI agents and one direct-generation baseline; the scoring protocol uses two cross-family judges, `qwen3.8-omni-flash` and `gemini-3.7-flash`,

combined conservatively for both modalities—these are research instruments we evaluate, not writing aids. No AI system contributed to the design of the benchmark, to the definition of check items or their pre-registered weights, or to the interpretation of results.

For manuscript preparation we used LLM assistants for literature triage and bibliography formatting, code assistance during evaluation harness development, and language editing of prose the authors had already drafted. All AI-assisted text and code was reviewed, tested, and edited by the authors. We take full responsibility for the final content.

ETHICS STATEMENT

This work involves no human-subjects experiments and no crowdsourced annotation. The main evaluation is fully automated; the only human participation is (i) the five-person domain-expert team that jointly constructed check-item checklists for all 120 tasks before system execution, and (ii) the pre-registered calibration study (Appendix A.11) in which two of the authors assigned expected score bands for 15 cases. No rater-identifying data are reported.

Task briefs were distilled from commercial workflows and de-identified: client and brand names, internal contacts, and commercial terms were removed or substituted. Residual risk remains—source assets may contain faces, trademarks, and geographic identifiers, and licensing and privacy review is not yet complete; the public release will be bounded accordingly, with no personally identifiable information released beyond what appears in those assets pending review. Tasks target benign commercial creative production; no violent, sexual, hateful, or politically deceptive content is requested or rewarded.

Evaluation is uniform by construction: every system receives the identical deliver package, tool interface, and frozen rubric; the eval package is inaccessible to agents; scoring is strictly post-hoc. Anticipated risks include misuse for deceptive advertising or synthetic media at scale, judge bias including self-preference (Panickssery et al., 2024), and influence of rankings on commercial competition. We mitigate these through pre-registered check items, deterministic rule-based guards, cross-family judge pairing, documented calibration, per-case auditability, and reporting both toolsets rather than a single leaderboard number.

REPRODUCIBILITY STATEMENT

The evaluation protocol—scoring formula, check-item construction rules, capping thresholds, gate conditions, and judge configuration—is fully specified in § 4 and Appendix A. Check items and weights were frozen before any system ran. Both modalities use the same blind cross-family pair, `qwen3.8-omni-flash` and `gemini-3.7-flash`, followed by conservative high-score synthesis and deterministic correction. These assignments reflect the production scoring system.

Deterministic measurements—ffprobe specification checks, hash-based reuse detection, SSIM, LPIPS, CLIP similarity—are exactly reproducible given the pinned dependency matrix and weight-file hashes verified by pre-flight (Appendix A.14). VLM scoring is stochastic: each unit is called once with one retry on failure; variance is controlled by the pre-registered rubric and quantified by the calibration study. Sidecar records permit every capping decision to be recomputed offline without re-invoking a judge.

System backbone versions are identified by run identifiers rather than persistent version stamps (see Limitations); full bit-level reproducibility of agent outputs is therefore not claimed. What is claimed reproducible is the scoring of any given set of deliverables. Upon acceptance we will release the benchmark suite—task definitions, assets, the deliver/eval/archive packages, scoring scripts, calibration data, and cost logs, bounded by the pending licensing and privacy review.

REFERENCES

Mohammad Asim et al. Met3r: Measuring multi-view consistency in generated images. *arXiv preprint arXiv:2501.06336*, 2025. URL `https://arxiv.org/abs/2501.06336`.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, and Linjie Li. Improving image generation with better captions. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Rui Chen, Lei Sun, Jing Tang, Geng Li, and Xiangxiang Chu. Finger: Content aware fine-grained evaluation with reasoning for ai-generated videos. *arXiv preprint arXiv:2504.10358*, 2025. URL https://arxiv.org/abs/2504.10358. ACM MM 2025.

A. Deganutti et al. Graphic-design-bench: A comprehensive benchmark for evaluating ai on graphic design tasks. *arXiv preprint arXiv:2604.04192*, 2026. URL https://arxiv.org/abs/2604.04192.

Xiaokun Feng, Haiyu Yu, Meiqi Wu, Shuai Hu, Jiachen Chen, Chenrui Zhu, Jiahong Wu, Xiangxiang Chu, and Kaiqi Huang. Narrlv: Towards a comprehensive narrative-centric evaluation for long video generation. *arXiv preprint arXiv:2507.11245*, 2026. URL https://arxiv.org/abs/2507.11245. ICLR 2026.

Jack Hessel et al. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. URL https://arxiv.org/abs/2104.08718. EMNLP 2021.

Martin Heusel et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017. URL https://arxiv.org/abs/1706.08500. NIPS 2017.

H. Hu et al. Cutverse: A compositional gui agents benchmark for media post-production editing. *arXiv preprint arXiv:2605.19484*, 2026. URL https://arxiv.org/abs/2605.19484.

Kaiyi Huang et al. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023a. URL https://arxiv.org/abs/2307.06350.

Ziqi Huang et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023b. URL https://arxiv.org/abs/2311.17982.

Carlos E. Jimenez et al. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023. URL https://arxiv.org/abs/2310.06770. ICLR 2024.

Junjie Ke et al. Musiq: Multi-scale image quality transformer. *arXiv preprint arXiv:2108.05997*, 2021. URL https://arxiv.org/abs/2108.05997. ICCV 2021.

Dayeon Ki et al. Graphicbench: A planning benchmark for graphic design with language agents. *arXiv preprint arXiv:2504.11571*, 2025. URL https://arxiv.org/abs/2504.11571.

Seungone Kim et al. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*, 2024. URL https://arxiv.org/abs/2310.08491. ICLR 2024.

Max Ku et al. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023. URL https://arxiv.org/abs/2312.14867. ACL 2024.

Baiqi Li et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. URL https://arxiv.org/abs/2406.13743.

Geng Li, Haiwen Li, Rui Chen, Jing Tang, Lei Sun, and Xiangxiang Chu. Peak-end-net: A peak-end rule inspired framework for generalizable video aesthetic assessment. *arXiv preprint arXiv:2607.13941*, 2026. URL https://arxiv.org/abs/2607.13941. ACM MM 2026.

Xing Liu et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023a. URL https://arxiv.org/abs/2308.03688. ICLR 2024.

Yang Liu et al. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023b. URL https://arxiv.org/abs/2303.16634.

Chang Ma et al. Agentboard: An analytical evaluation board of multi-turn llm agents. *arXiv preprint arXiv:2401.13178*, 2024. URL https://arxiv.org/abs/2401.13178. NeurIPS 2024 Oral.

Ziyu Ma, Hailang Huang, Siyuan Zou, Yong Wang, Shunyu Yang, Yiming Hu, Fei Wei, and Xiangxiang Chu. Longhorizon-harness: Advancing long-horizon agents for real-world tasks. *arXiv preprint arXiv:2608.01964*, 2026. URL https://arxiv.org/abs/2608.01964.

Fangyuan Mao et al. Omni-effects: Unified and spatially-controllable visual effects generation. *arXiv preprint arXiv:2508.07981*, 2026. URL https://arxiv.org/abs/2508.07981. AAAI 2026.

Grégoire Mialon et al. Gaia: A benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023. URL https://arxiv.org/abs/2311.12983.

Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024. URL https://arxiv.org/abs/2404.13076.

Yujia Qin et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023. URL https://arxiv.org/abs/2307.16789.

Alec Radford et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach et al. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022. URL https://arxiv.org/abs/2112.10752. CVPR 2022.

Christoph Schuhmann et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. URL https://arxiv.org/abs/2210.08402. NeurIPS 2022 Datasets and Benchmarks Track.

Yongliang Shen et al. Taskbench: Benchmarking large language models for task automation. *arXiv preprint arXiv:2311.18760*, 2023. URL https://arxiv.org/abs/2311.18760. NeurIPS 2024.

Tao Sun et al. P2p: Automated paper-to-poster generation and fine-grained benchmark. *arXiv preprint arXiv:2505.17104*, 2025. URL https://arxiv.org/abs/2505.17104.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *arXiv preprint arXiv:2003.12039*, 2020. URL https://arxiv.org/abs/2003.12039.

Thomas Unterthiner et al. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Tianbao Xie et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024. URL https://arxiv.org/abs/2404.07972.

Shunyu Yao et al. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023. URL https://arxiv.org/abs/2210.03629. ICLR 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*, 2018. URL https://arxiv.org/abs/1801.03924. CVPR 2018.

Zicheng Zhang et al. A perceptual quality assessment exploration for aigc images. *arXiv preprint arXiv:2303.12618*, 2023. URL https://arxiv.org/abs/2303.12618.

Shuyan Zhou et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023. URL https://arxiv.org/abs/2307.13854.

Chenrui Zhu, Jiashu Zhu, Yifu Li, Meiqi Wu, Bingze Song, Chubin Chen, Jiahong Wu, Xiangxiang Chu, and Yong Wang. Artifact-aware evaluation for high-quality video generation. *arXiv preprint arXiv:2601.20297*, 2026a. URL https://arxiv.org/abs/2601.20297. ICASSP 2026.

Jiashu Zhu, He Zheng, Rui Tian, et al. Dreamx-creator: Democratizing native audio-video generation at 2k resolution. *arXiv preprint arXiv:2608.31106*, 2026b. URL https://arxiv.org/abs/2608.31106.

A Scoring Rulebook: Full Specification

This appendix supplements § 4 in three parts. Part I documents the benchmark data: the full 120-task catalogue, the 19-to-14 family mapping, ground-truth sources, the checklist construction process, and two additional worked case examples. Part II details the scoring methodology: the formula and a numeric example, atomic check-item taxonomies for both tracks, specification tolerances, the objective-metric inventory with thresholds, correction and capping rules, and the calibration protocol. Part III covers system configuration: judge roles, multi-video assessment semantics, the dependency pre-flight matrix, frozen experimental parameters, and cost-field definitions.

A.1 Task Catalogue

Table 4 lists all 120 tasks in CreatorBench with their modality, task family, a brief English summary, and required deliverables.

Table 4: Complete task catalogue for all 120 CreatorBench tasks.

ID	Mod.	Family / Track	Brief	Deliverables
001	Image	Ad Creative	Design a 3:4 feed ad for a Sichuan hotpot restaurant with promo offer	1 image 1080×1440
002	Image	Ad Creative	Create multi-size ad suite (3 formats) for a coffee shop promotion	3 images (3 sizes)
003	Image	Ad Creative	Restore low-quality bakery photos, then create a new-product launch ad	1 image 1080×1080
004	Image	Ad Creative	Remove expired badge from tea shop ad; add new promo text	1 image 1080×1080
005	Image	Ad Creative	Correct perspective distortion in car wash storefront; create opening ad	1 image 1080×1080
006	Image	Ad Creative	Denoise and super-resolve noodle photo; create breakfast promo ad	1 image 1080×1080
007	Image	Ad Creative	Design a Mid-Autumn festival banquet booking ad for a Chinese restaurant	1 image 1080×1440
008	Image	Footprint Report	Create annual footprint poster from 12 cities, 86 POIs, and 6 photos	1 image 1080×1920
009	Image	Footprint Report	Design a hiking season report poster with 6-summit elevation visualization	1 image 1080×1920
010	Image	Footprint Report	Create coastal road-trip route poster across 4 seaside cities	1 image 1080×1920
011	Image	Footprint Report	Compose skyline collection poster from 7-city silhouette photographs	1 image 1920×1080
012	Image	Footprint Report	Design cross-province self-driving trip report with route map and photos	1 image 1080×1920
013	Image	Footprint Report	Create first-year city exploration milestone poster for a newcomer	1 image 1080×1440
014	Image	User Growth	Design senior-mode feature poster for a map app (large font, warm tone)	1 image 1080×1440
015	Image	User Growth	Create campus-edition banner and splash screen for a map app launch	2 images (2 sizes)

continued on next page

Table 4 – *continued from previous page*

ID	Mod.	Family / Track	Brief	Deliverables
016	Image	Precise Editing	Remove 2 photobombers from travel photo; 2x super-resolution; keep subject	1 enhanced photo
017	Image	Precise Editing	Erase expired watermark from promo poster; add new sale copy in same style	1 image (edited)
018	Image	Precise Editing	Restore a 1990s family photo: denoise, remove scratches, colorize, 2x SR	1 restored photo
019	Image	Precise Editing	Update all prices from ¥35 to ¥42 on a restaurant menu image	1 image (edited)
020	Image	Precise Editing	Add 4-level text hierarchy to product ad blank area	1 image (edited)
021	Image	Precise Editing	Correct perspective of tilted storefront; re-render signboard with new name	1 image (edited)
022	Image	Precise Editing	Replace handbag product background with pure white; preserve shadow	1 image (white bg)
023	Image	Precise Editing	Swap middle person's clothing in group photo using reference garment image	1 image (edited)
024	Image	Precise Editing	Clean hand-drawn marks from floor plan; redraw 3 furniture icons	1 image (edited)
025	Image	Precise Editing	Denoise and enhance a dark night-scene photo; preserve neon colors	1 image (enhanced)
026	Image	Precise Editing	Composite 2 friends into a beach scene with consistent perspective/lighting	1 composite image
027	Image	Precise Editing	Relocate a red official seal to correct position on a scanned document	1 image (edited)
028	Image	Precise Editing	Change status-bar time from 09:41 to 14:30 on an app screenshot	1 image (edited)
029	Image	Precise Editing	Re-typeset all text in an event poster with new brand font; keep content	1 image (re-typeset)
030	Image	Precise Editing	Replace overcast sky with sunset; adjust ground lighting to match	1 image (sky swap)
031	Image	Precise Editing	4x super-resolve sneaker image; change sole color; convert to tennis shoe style	1 image (4× SR)
032	Image	Precise Editing	Denoise video keyframe; cinematic color grade; correct to frontal angle	1 promo still
033	Image	Precise Editing	Correct handwritten-note perspective; extract line 9 as calligraphy artwork	2 images
034	Image	Precise Editing	Remove power lines, trash can, and road sign from storefront street photo	1 image (edited)
035	Image	Precise Editing	Restore yellowed menu photo (denoise, 2x SR); OCR-verify all text accuracy	1 image + OCR report
036	Image	3D & Storyboard	Generate 3-view 3D rendering of 2D mascot; match original design exactly	3-view 3D render
037	Image	3D & Storyboard	Create 3D metallic logo (brushed metal) in multi-angle brand showcase	3D logo renders
038	Image	3D & Storyboard	Render tea-canister 3D packaging mockup from flat design for e-commerce	1 3D mockup
039	Image	3D & Storyboard	Generate 6-panel live-action storyboard for rainy-night drama reunion scene	6-panel storyboard
040	Image	3D & Storyboard	Create 16-panel manga-style storyboard for a superpower-awakening prologue	16-panel storyboard
041	Image	3D & Storyboard	Generate 6-pose expression sheet from IP character for sticker use	6-pose sheet
042	Image	3D & Storyboard	Convert landmark photo to tilt-shift 3D miniature model rendering	1 miniature render
043	Image	3D & Storyboard	Transform roast-duck photo into LEGO-style 3D brick scene render	1 LEGO-style render
044	Image	3D & Storyboard	Generate 4-pose character turnaround from actor costume-test photo	4-pose sheet

continued on next page

Table 4 – *continued from previous page*

ID	Mod.	Family / Track	Brief	Deliverables
045	Image	3D & Storyboard	Render brand name as 3D brushed-silver metallic text on dark background	1 3D text render
046	Image	3D & Storyboard	Convert live-action drama keyframe to anime cel-shading style	1 anime-style image
047	Image	3D & Storyboard	Create isometric top-down 3D doll-house rendering of restaurant interior	1 isometric render
048	Image	3D & Storyboard	Generate exploded-view 3D render of mechanical watch components	1 exploded-view render
049	Image	3D & Storyboard	Create 4-frame progressive storyboard: city to door-push cinematic sequence	4-frame sequence
050	Image	3D & Storyboard	Composite 3D mascot into real store-front scene with matching evening light	1 composite image
101	Image	E-commerce	Create 5-image e-commerce detail-page suite for a cosmetics foundation	5 e-com detail images
102	Image	E-commerce	Design 618-sale multi-format hero visuals for a home furniture brand	3 promo visuals
103	Image	E-commerce	Adapt running-shoe product shot to 3 platform-specific e-commerce formats	3 platform images
104	Image	E-commerce	Produce 4-piece livestream material kit (backdrop, overlay, popup, price tag)	4 livestream assets
105	Image	Ad Design	Create 3-format EV car ad suite: portrait, landscape, and square layouts	3 ad images (3 formats)
106	Image	Ad Design	Composite 2 furniture cutouts into a Nordic living-room with AI soft furnishing	2 scene composites
107	Image	Ad Design	Generate 4 A/B test variants of an existing skincare ad	4 A/B variants
108	Image	Viral Replication	Replicate 2 viral Xiaohongshu cover styles with new sunscreen product	2 Xiaohongshu covers
109	Image	Viral Replication	Transfer competitor's visual style to own brand under strict VI constraints	2 brand-transfer ads
110	Image	Premium Commercial	Design 3-piece product-launch suite for a smartwatch: backdrop, invite, teaser	3 launch visuals
051	Video	Food & Scenery	Produce 8s vertical commercial promo for a Sichuan fish dish brand	8s MP4 3:4
052	Video	Food & Scenery	Produce 6s square-format grilled-skewer commercial for delivery platforms	6s MP4 1:1
053	Video	Food & Scenery	Create 12s horizontal hotel banquet-hall showcase for wedding marketing	12s MP4 4:3
054	Video	Food & Scenery	Create 10s horizontal mountain-resort brand atmosphere video for OTA	10s MP4 4:3
055	Video	Food & Scenery	Produce 15s vertical ancient-town festival teaser for social platforms	15s MP4 3:4
056	Video	3D & Spatial	360-degree orbit then kung-fu action of warrior figure from first frame	8s MP4 1:1
057	Video	3D & Spatial	Dual-axis 360-degree orbit (horizontal then vertical) of a drum kit	8s MP4 1:1
058	Video	3D & Spatial	360-degree orbit of tank; then tank rams open the garage door	8s MP4 16:9
059	Video	3D & Spatial	Full 360-degree scene orbit of an auditorium interior	8s MP4 16:9
060	Video	3D & Spatial	First-person retreating camera with 45-degree left-right pan of bicycle scene	8s MP4 16:9
061	Video	Interactive Video	Interactive branching video: 3-segment A→B/C rain-night story with player	3×12s MP4 + player
062	Video	Interactive Video	Interactive branching video: 3-segment sci-fi lounge story with player	3×12s MP4 + player
063	Video	Interactive Video	Interactive branching video: 3-segment clockwork-city story with player	3×12s MP4 + player

continued on next page

Table 4 – *continued from previous page*

ID	Mod.	Family / Track	Brief	Deliverables
064	Video	Interactive Video	Interactive branching video: 3-segment orbital-greenhouse story with player	3×12s MP4 + player
065	Video	Interactive Video	Interactive branching video: 3-segment birthday-party story with player	3×12s MP4 + player
066	Video	Video Editing	Add morning mist to river scene in source video; preserve original audio	5s MP4 (edited)
067	Video	Video Editing	Recolor runway surface in 4K source video; maintain no-audio track	5s MP4 4K (edited)
068	Video	Video Editing	Remove tourists from castle scene in 4K source video; preserve audio	5s MP4 4K (edited)
069	Video	Video Editing	Replace background through mesh/grid in source video; preserve audio	5s MP4 (edited)
070	Video	Video Editing	Convert to B&W film look, keep sweater color, replace window to rain bokeh	5s MP4 9:16 (edited)
071	Video	Video Editing	Replace sky with sunset; propagate lighting; remove clutter in pastoral video	5s MP4 16:9 (edited)
072	Video	Video Editing	Recolor black alarm clock to gold; replace table to walnut; keep dial motion	5s MP4 9:16 (edited)
073	Video	Video Editing	Replace container, add ring, remove logos, change countertop in coffee video	5s MP4 4K (edited)
074	Video	Video Generation	Simulate coin precession and settling physics from first-frame setup (9s)	9s MP4 16:9
075	Video	Video Generation	Simulate lotus-leaf water-droplet merge, roll-off, and ripple physics (9s)	9s MP4 16:9
076	Video	Video Generation	Simulate sand-pendulum drawing nested shrinking ellipses on disk (12s)	12s MP4 16:9
077	Video	Video Generation	Simulate rotating prism with continuous spectrum sweep on screen (10s)	10s MP4 16:9
078	Video	Video Generation	Simulate maple samara autorotation and landing in slow motion (9s)	9s MP4 16:9
079	Video	Video Generation	Simulate slinky spring descending 4 stair steps after wedge release (12s)	12s MP4 16:9
080	Video	Video Generation	Simulate cork ball buoyancy release, surfacing, and damped oscillation (10s)	10s MP4 16:9
081	Video	Short Drama	Short drama: delivery courier and teashop owner share a quiet evening tea	24s MP4 (1 ep.)
082	Video	Short Drama	Short drama: intern asks security guard to verify late arrival at ground floor	22s MP4 (1 ep.)
083	Video	Short Drama	Short drama: child and snail deliver a dewdrop to dry moss in grandma's yard	27s MP4 (1 ep.)
084	Video	Short Drama	Short drama: estranged couple meet at late-night laundromat; washing a curtain	75s MP4 (1 ep.)
085	Video	Short Drama	Short drama: engineers investigate false flood-gate alarm in underground station	80s MP4 (1 ep.)
086	Video	Short Drama	Short drama: ancient courier and ledger officer verify emergency dispatch token	70s MP4 (1 ep.)
087	Video	Short Drama	Short drama: two grad students discover a shared unnamed audio recording	60s MP4 (1 ep.)
088	Video	Short Drama	Short drama: space-station crew defrost solar-sail rail to survive proton storm	85s MP4 (1 ep.)
089	Video	Short Drama	Short drama (3 eps): repairing an umbrella reveals a decade-old rescue story	180s MP4 (3 eps.)
090	Video	Short Drama	Short drama: health inspector solves cold-storage sabotage at fish market	65s MP4 (1 ep.)
091	Video	Short Drama	Short drama: single-episode narrative (brief in input_context.json only)	75s MP4 (1 ep.)

continued on next page

Table 4 – *continued from previous page*

ID	Mod.	Family / Track	Brief	Deliverables
092	Video	Short Drama	Short drama: single-episode complex narrative (brief in input_context.json)	120s MP4 (1 ep.)
093	Video	Short Drama	Short drama: 3-episode long-form narrative (brief in input_context.json)	270s MP4 (3 eps.)
094	Video	Short Drama	Short drama (2 eps): observatory meteorite theft investigation in snowstorm	240s MP4 (2 eps.)
095	Video	Short Drama	Short drama: theater rigging crisis solved with improvised prism-light moon	150s MP4 (1 ep.)
096	Video	Video Remix	Video remix: 4-clip fog-ridge slow-train travel short with voiceover and BGM	MP4 9:16 30fps
097	Video	Video Remix	Video remix: 4-clip seaside cycling lifestyle ad with voiceover and BGM	MP4 9:16 30fps
098	Video	Video Remix	Video remix: 4-clip coffee-cup lifestyle ad with voiceover and BGM	MP4 9:16 30fps
099	Video	Video Remix	Video remix: 4-clip bakery morning routine ad with voiceover and BGM	MP4 9:16 30fps
100	Video	Video Remix	Video remix: 4-clip lakeside camping lifestyle short with voiceover and BGM	MP4 9:16 30fps
111	Video	E-com Promo	Create 15s vertical cosmetics product showcase video (rose-gold theme)	15s MP4 9:16
112	Video	E-com Promo	Produce 30s vertical unboxing video for flagship wireless earbuds	30s MP4 9:16
113	Video	E-com Promo	Create 20s vertical smart robot-vacuum feature demo video	20s MP4 9:16
114	Video	Ad Production	Produce 15s vertical brand TVC for a Chinese fast-food chain	15s MP4 9:16
115	Video	Ad Production	Create 10s vertical feed ad for budget BB cream (¥49.9 BOGO)	10s MP4 9:16
116	Video	Ad Production	Produce 30s vertical franchise recruitment video for kids' coding brand	30s MP4 9:16
117	Video	Viral Replication	Create 15s beat-synced photo montage video from 5 input images	15s MP4 9:16
118	Video	Viral Replication	Produce 22s vertical food-vlog style restaurant exploration video (no assets)	22s MP4 9:16
119	Video	Ad Production	Create 15s vertical beauty serum seeding ad with 3 selling points	15s MP4 9:16
120	Video	Ad Production	Produce 20s vertical Double-11 shopping festival promotional video	20s MP4 9:16

A.2 Fine-Grained Category to Family Mapping

The 14 paper-level task families are mapped from 19 fine-grained scene (image) or track (video) categories defined in the evaluation data. The mapping preserves all cases and sums to 120.

Key merges: on the image side, Advertisement + Ad Design + High-end Commercial $\rightarrow$ Promotional Graphic Design ($7 + 3 + 1 = 11$); E-commerce Marketing + Viral Remake $\rightarrow$ E-commerce Product Visualization ($4 + 2 = 6$). On the video side, E-commerce Promo + Ad Production + Viral Remake $\rightarrow$ Commercial Cinematic Production ($3 + 5 + 2 = 10$).

A.3 Ground-Truth Sources

All scoring is anchored to frozen per-case artefacts; no evaluator discretion or generic aesthetic preference enters the pipeline.

Image track. These four content classes are atomised into requirement items: the business brief supplies requirement verbatim text, key review points supply hard-constraint verbatim text, the

Table 5: Mapping from 19 fine-grained categories to 14 paper-level families.

Modality	**Fine-grained category**	**Paper family**	n
Image	Advertisement	Promotional Graphic Design	7
Image	Ad Design	Promotional Graphic Design	3
Image	High-end Commercial	Promotional Graphic Design	1
Image	Precision Editing	Precision Image Editing	20
Image	3D & Storyboard	3D & Storyboard Composition	15
Image	Footprint	Travel & Footprint Poster	6
Image	User Growth	Multi-format Asset Adaptation	2
Image	E-commerce Marketing	E-commerce Product Visualization	4
Image	Viral Remake (image)	E-commerce Product Visualization	2
Video	Short Drama	Narrative Video Production	15
Video	Video Editing	Compositional Video Editing	8
Video	Video Generation	Video Generation	7
Video	Food & Scenic	Food & Scenic Video	5
Video	E-commerce Promo	Commercial Cinematic Production	3
Video	Ad Production	Commercial Cinematic Production	5
Video	Viral Remake (video)	Commercial Cinematic Production	2
Video	3D & Spatial	3D & Spatial Video	5
Video	Interactive Video	Interactive Video	5
Video	Video Remix	Multi-source Video Remix	5
Total (19 $\rightarrow$ 14 families)			120

Table 6: Ground-truth source classes for the image track.

Source class	**Ground truth contributed**
Case metric mapping	Task objective, verbatim business brief, and key review checkpoints for each case.
Step checklist	Step-by-step acceptance checklist with per-item `check_type` and `params`.
Structured delivery specification	Delivery contract: dimensions, file count, format, and channel requirements.

checklist supplies step-by-step acceptance actions, and the structured specification supplies the delivery contract.

Table 7: Ground-truth source classes for the video track.

Source class	Ground truth contributed
Evaluation criteria	Criterion text, weights, and verbatim `pass_condition` acceptance details.
Business metric configuration	Objective metric backends, weight paths, and sampling parameters.
Track declarations	Track gate rules (allowlist / blocklist / schema quoted verbatim per track).

Video track. Video criteria are never rewritten: each criterion is migrated losslessly into one check item; weights are redistributed proportionally within the 100-point requirement budget while preserving the relative proportions declared in the frozen criterion file.

A.4 Checklist Construction Process

Team composition. Five domain experts with backgrounds in commercial content production, visual design, and video post-production participated in checklist construction. Their identities are not disclosed to preserve double-blind review.

Workflow. The team worked collaboratively (not independently) through four stages:

1. **Brief decomposition**: For each of the 120 tasks, the team jointly reviewed the structured brief, source assets, and delivery specifications, extracting each requirement as an atomic, verifiable check item with a natural-language pass condition.
2. **Weight assignment**: Weights reflect each item's commercial importance relative to the brief's intent and the cost of remediation; technical quality and visual appeal are assessed separately by the independent quality track and enter through the multiplicative modifier of Eq. (2).
3. **Objective metric binding**: Where a pass condition corresponds to a deterministic measurement (resolution, duration, codec, file count), a deterministic rule was attached with registered tolerances (Appendix A.8).
4. **Freeze**: The complete checklist—items, weights, pass conditions, and metric bindings—was frozen as a single artefact; no modifications were made after the freeze date.

Limitations of this process. Because the team worked collaboratively rather than as independent annotators, no inter-annotator agreement (IAA) statistic was measured. The checklist reflects joint expert judgment, not statistically validated ground truth. Independent third-party annotation and IAA measurement are left to future work.

A.5 Additional Case Examples

This section provides two supplementary case examples—one image, one video—illustrating the diversity of task structures within CreatorBench.

Image Case 008: Annual Footprint Report Poster.

Video Case 081: Short Drama — Misunderstanding and Reconciliation at a Herbal Tea Shop.

A.6 Scoring Formula and Worked Example

Per-item scoring. Each check item is independently scored from 0–5 by the primary and audit VLMs against the same frozen rubric. After evidence gating, conservative dual-model synthesis, objective correction (§ A.10), and defect and delivery-specification capping yield `final_item_score`.

Field	Content
Family	Data-driven Visual Narrative
Brief	User's 2025 annual footprint report: visited 12 cities, 86 POIs, 23 400 km total. 6 representative photos provided; select the 3 most representative, combine with an abstract route map, and produce a year-in-review poster.
Inputs	6 photos (008_1–008_6.jpg): Chengdu Chunxi Rd, Chongqing Hongyadong, Dali Erhai Lake, Shanghai Bund, Xi'an City Wall, Beijing Forbidden City
Deliver	1 image, 1080×1920 annual footprint poster
Objective metrics	aesthetic_score, clip_score, style_consistency, identity_consistency, sr_gain, size_compliance, compliance_pass (7 metrics)
Subjective dims	semantic_accuracy, sel_rationality, layout_hierarchy, communication, aesthetic_appeal, compliance_judge (6 dimensions)
Ratio	objective 55% / subjective 45%

Table 8: Image Case 008 task specification. The agent must perform photo filtering (6→3), abstract map generation, and data-visualization layout—a multi-stage pipeline combining selection, generation, and composition.

Table 9: Video Case 081 task specification.

Field	Content
Family	Narrative Video Production
Brief	Produce a 24-second, 3-shot short drama set in a herbal tea shop: a misunderstanding arises, the visitor reveals their purpose, and the scene resolves over shared tea. Character consistency (face, costume, voice) and single-location continuity are mandatory.
Deliver	1 video per episode (E01), 24 s, 3 shots, H.264 MP4

	Check Item	w	Pass Condition (abridged)
1	Plot & causal completeness	20	Clearly establish misunderstanding, reveal purpose, resolve over tea within 24 s
2	Character consistency	20	Face, costume, voice, mannerisms stable throughout
3	Scene & prop continuity	20	Single tea-shop space; window axis, post-rain lighting, background objects unchanged
4	Shot continuity	20	3 shots: action, gaze, sugar-packet and cup state continuous; no axis crossing
5	Dialogue & performance	15	Speaker–lip-sync match; restrained warmth in pauses; audio mandatory; subtitles if dialogue used
6	Duration & delivery compliance	5	Exactly 24 s, 3 shots, no timecode gaps or overlaps
Σw		100	

Table 10: Video Case 081 check-item decomposition. All items are perceptual (VLM-judged) except item 6 which includes deterministic duration/timecode verification. The narrative structure (three-act, single-location) tests temporal coherence and character identity preservation—capabilities orthogonal to single-frame generation quality.

Case-level aggregation. As defined in the main text (Eqs. (1)–(2)):

$$R = \sum_i w_i \cdot s_i \,/\, 5, \qquad R \in [0, 100]$$

$$\text{case_score} = R \;\times\; \big(0.60 + 0.20\, Q/5 + 0.20\, A/5\big)$$

where w_i is the weight of requirement check item i (all items within a case sum to 100), s_i is its final item score on the 0–5 scale after dual-model synthesis and deterministic correction, $Q \in [0, 5]$ is the technical-quality score, and $A \in [0, 5]$ is the visual-appeal score, both produced by the independent quality track.

Constraints.

- All check-item weights within a case sum to exactly 100; all items are requirement items.
- Technical quality (Q) and visual appeal (A) are scored independently by VLM on a 0–5 scale, outside the requirement checklist.
- The quality factor $(0.60 + 0.20\, Q/5 + 0.20\, A/5)$ ranges over $[0.60, 1.00]$; at least 60% of R is always preserved.
- If $R = 0$, the case score is 0 regardless of quality or appeal—the multiplicative form enforces requirement-first scoring.
- If the quality track is unavailable, the case score falls back to R alone (factor = 1.0). All results reported in this paper use the full formula with quality enabled.
- **Missing items are not renormalised.** If any requirement item cannot be scored, the case receives `null`—never a higher score due to fewer evaluated items.

Worked example. Consider a case with four requirement items (weights 30, 25, 25, 20; $\Sigma w = 100$). After dual-model scoring and deterministic correction the final item scores are: $s_1 = 4$, $s_2 = 3$, $s_3 = 5$, $s_4 = 2$ (each on a 0–5 scale).

$$R = (30 \times 4 + 25 \times 3 + 25 \times 5 + 20 \times 2) \,/\, 5 = (120 + 75 + 125 + 40) \,/\, 5 = 72.0$$

The independent quality track yields $Q = 3.5$ and $A = 4.0$:

$$\begin{aligned} \text{case_score} &= 72.0 \times \big(0.60 + 0.20 \times 3.5/5 + 0.20 \times 4.0/5\big) \\ &= 72.0 \times (0.60 + 0.14 + 0.16) = 72.0 \times 0.90 = 64.8 \end{aligned}$$

Without the quality modifier the score would be $72.0 \times 0.60 = 43.2$ (floor) or simply $R = 72.0$ (if the quality track is disabled). The multiplicative form ensures that high quality can lift the score by up to $0.40\, R$ above the $0.60\, R$ floor—reaching R itself when $Q = A = 5$—while poor quality can reduce it to $0.60\, R$ when $Q = A = 0$. Table 11 itemises the contributions to R.

	Req-1 (w=30)	Req-2 (w=25)	Req-3 (w=25)	Req-4 (w=20)
`final_item_score` s_i	4	3	5	2
$w_i \cdot s_i/5$	24.0	15.0	25.0	8.0

Table 11: Per-item contributions to R in the worked example. Weights sum to 100 and every item is a requirement item; $R = 72.0$.

A.7 Atomic Check Items

Image track. Every image case is decomposed into requirement check items drawn from the following `check_type` categories; each category names its judgment source and the quality dimension it maps to.

Table 12: Image atomic check-item categories, judgment sources, and mapped dimensions.

check_type	**Judgment source**	**Mapped dimension**
delivery_integrity	Task objective + delivered files	G1 Requirements & Delivery
brief_fulfillment	Business brief verbatim	G1 Requirements & Delivery
constraint_preserved	Key review checkpoints (per item)	G1 Requirements & Delivery
source_membership	Designated input asset requirement	S7 Asset Usage
no_residual	Designated removal target	S2 Editing & Removal
text_exact_match	Designated text, character-exact	S1 Text
size_compliance	Delivery dimensions	G4 Delivery Specification
format_compliance	Delivery format & channel count	G4 Delivery Specification (supporting)

Fixed quality dimensions. Two universal quality dimensions are assessed for every image case by the independent quality track (not as checklist items, and carrying no share of the 100-point requirement budget):

- **Image quality and naturalness** (dimension G2) $\rightarrow$ technical quality $Q \in [0, 5]$;
- **Aesthetics and overall impression** (dimension G3) $\rightarrow$ visual appeal $A \in [0, 5]$.

These two scores enter the case score only through the multiplicative modifier of Eq. (2).

Budget allocation.

- All 100 points are allocated to requirement items; every check item in the frozen per-case checklist is a requirement item.
- Item weights reflect the severity of failure (fatal / significant / minor, unit ratio $4 : 2 : 1$) jointly with the cost of remediation; within a case they are normalised to sum to exactly 100.
- Supporting items (e.g. format_compliance) receive a small share of the same 100-point budget; when a case declares no supporting item, its share flows back to the core requirement items.
- Technical quality and visual appeal are scored separately on a 0–5 scale by the independent quality track and enter the case score via the multiplicative modifier (Eq. (2)); they carry no share of the 100-point checklist budget.

Video track. Each frozen video criterion is migrated to exactly one check item. All task-specific criteria are requirement items; the full 100-point budget is allocated among them, preserving the relative proportions declared in the frozen criteria file. Universal quality assessment (creative organisation, engagement, image quality and aesthetics, video quality, final-cut quality) is handled by the independent quality track, which produces the technical-quality score Q and the visual-appeal score A on a 0–5 scale and enters the case score through the multiplicative modifier of Eq. (2).

Table 13: Video quality dimensions and their scope.

Code	**Dimension name**	**Scope**
V1	Requirement fulfilment	Core task goals
V2	Fidelity & consistency	Cross-frame subject/scene coherence
V3	Temporal & technical quality	Motion smoothness, encoding integrity
V4	Specification & compliance	Hard delivery specs (resolution, fps, codec)

Video dimension mapping. Weight distribution across the requirement items preserves the *relative* proportions declared in the frozen criteria file; the declared weights are rescaled once, at migration time, so that the per-case budget sums to exactly 100. No renormalisation occurs at scoring time (§ A.6).

A.8 Specification Auto-Extraction and Tolerances

When the frozen `pass_condition` explicitly states a delivery specification, the system automatically attaches a deterministic rule. Specifications not explicitly stated are never inferred.

Table 14: Deterministic delivery-specification tolerances.

Specification	Judgment	Tolerance	Notes
Resolution $W \times H$	exact	none	Evaluation canvases (e.g. 518×518 analysis canvas) are excluded
Frame rate N fps	between	± 0.5 fps	—
Duration N seconds	between	$\pm \max(0.35\text{ s},\ 3\%)$	—
Codec H.264	exact	none	—
Audio track	per task direction	none	Direction determined by task brief (see § A.10)
Delivery count	exact	none	Must equal declared count

A.9 Objective Metrics: Full Inventory

Metrics are organised in a three-layer evidence model:

- **Deterministic** (rule-verifiable facts): mechanically decidable, no ambiguity; can zero, cap, or veto.
- **Perceptual** (model-judged proxies): noisy continuous metrics; serve *only* as conflict evidence—never linearly blended into scores.
- **context**: human-readable auxiliary (OCR text, loudness); retained for audit but never mechanically determines scores.

All metrics carry a unified status: `ok` / `not_applicable` / `unavailable` / `error`. Unavailable-and-not-applicable metrics incur no penalty; applicable metrics returning `unavailable` or `error` cause the case to be marked `evaluation_unavailable`.

Table 15: Deterministic image metrics: judgment, action on failure, and case-level cap.

Metric	Judgment	Action on fail	Case-level cap
`size_compliance`	Delivered dimensions match spec	cap item at 2.0/5	$\leq$20/100
`count_compliance`	Delivered file count matches declared	zero item	$\leq$15/100
`format_compliance`	Encoding format & channel count	hard_item: pass=5, fail=0	—
`not_reused`	SHA-256/pHash confirms input not reused verbatim	zero item	—

Image: deterministic metrics (4).

Image: perceptual metrics (7). Thresholds are pre-registered extreme values pending human calibration, so a flag requires a gross discrepancy and conflict review fires rarely—predominantly for `identity_cosine` and `style_raw_similarity`, the two metrics defined only when an input reference exists. Backends: CLIP (Radford et al., 2021), LPIPS (Zhang et al., 2018), LAION aesthetic predictor (Schuhmann et al., 2022).

Design constraints: only raw-scale values are used. Metric names must reflect their actual implementation (CLIP proxies are not labelled ArcFace/DINO; Laplacian ratios are not labelled PSNR/SSIM). FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) are dataset-level metrics that do not apply at the single-case level.

Table 16: Perceptual image metrics with fail/pass thresholds and caps.

Metric	Backend	Direction	fail	pass	cap
clip_cosine	CLIP text-image cosine	high_good	0.15	0.75	2.5
lpips_distance	LPIPS (AlexNet)	low_good	0.6	0.1	2.5
aesthetic_raw_score	LAION aesthetic head	high_good	3.0	7.0	2.5
identity_cosine	CLIP image-image cosine	high_good	0.2	0.8	2.0
style_raw_similarity	CLIP + histogram sim.	high_good	0.2	0.8	2.5
denoise_noise_reduction_ratio	Noise sigma drop ratio	high_good	-0.2	0.6	2.5
sr_sharpness_ratio	Laplacian variance ratio	high_good	0.7	1.25	2.5

Image: context (4 metrics). ocr_text_accuracy, erase_residual, compliance_pass, watermark_residual. These provide human-readable auxiliary evidence for audit records; they never mechanically determine scores.

Table 17: Deterministic video metrics and tolerances.

Metric	Judgment	Tolerance
video_width / video_height	Matches task-stated resolution	exact
video_fps	Matches task-stated frame rate	± 0.5 fps
video_duration_seconds	Matches task-stated duration	$\pm \max(0.35\text{ s}, 3\%)$
video_codec	H.264	exact
audio_stream_present	Audio track presence matches task direction	exact
loop_detected	dHash + SSIM fixed-lag repetition	—
black_screen_detected	Sampled-frame luminance black screen	—

Video: deterministic metrics (7).

Video: perceptual metrics (9). Frame consistency via DINOv2 (Oquab et al., 2023)/CLIP (Radford et al., 2021) encoding; motion via RAFT (Teed & Deng, 2020) optical flow; quality via MUSIQ (Ke et al., 2021).

Table 18: Perceptual video metrics with thresholds, caps, and semantics.

Metric	Direction	fail	pass	cap	Semantics
dino_similarity	high_good	0.45	0.88	2.5	Subject/character inter-frame consistency
clip_frame_similarity	high_good	0.45	0.88	2.5	Background/scene/prop continuity
compensated_flicker	low_good	0.35	0.08	2.5	Motion-compensated flicker
middle_frame_interpolation_error	low_good	0.35	0.08	2.5	Mid-point frame interpolation error
dynamic_degree	high_good	0.01	0.08	2.0	Motion magnitude (too low = frozen)
musiq	high_good	20	65	2.5	No-reference image quality
fixed_aesthetic	high_good	0.15	0.75	2.5	Fixed aesthetic prediction
met3r_3d_consistency	low_good	0.7	0.15	2.5	Geometric inconsistency (Asim et al., 2025) (3D/spatial tasks only)
hard_cut_ratio	low_good	0.55	0.10	2.5	Hard-cut proportion (fast-paced editing may legitimately trigger; conflict review only)

Video: context (2 metrics).

- subtitle_ocr_samples: per-frame OCR text with confidence scores.
- audio_loudness_ebur128: integrated loudness, LRA, and true peak.

A.10 Correction and Capping Rules

The deterministic correction layer receives *only* the objective metric evidence cards mapped to each check item by declaration; each card carries the raw metric value, its status (ok / not_applicable / unavailable / error), the registered fail and pass thresholds, any triggered cap, and limitation notes. Only metrics explicitly mapped to a check item enter its deterministic correction path, preventing cross-semantic "evidence contamination".

DETERMINISTIC ACTIONS

Table 19: Deterministic correction actions.

Action	Effect
`zero`	The bound check item is set to 0/5 unconditionally.
`cap`	The bound check item is capped at the registered value (e.g., 2.0/5 or 1.2/5).
`veto`	Triggers a case-level gate; all items are zeroed (see Case-Level Gates below).

PERCEPTUAL CONFLICT LOGGING

- **Trigger conditions:** an objective metric signals *fail* while the synthesized VLM score is > 3.5, or signals *pass* while that score is < 1.5.
- A trigger records a `review_required` action and the conflicting evidence in the audit record.
- In the production `eval_v2` path, perceptual conflicts do not invoke another model call, replace the synthesized score, or impose a mechanical penalty.
- `unavailable`, unknown metrics, and non-finite values never trigger deduction.

DEFECT CAPS

Defects cap only the requirement item they are explicitly bound to; other items are unaffected.

Table 20: Dual-axis defect caps applied to the bound requirement item.

Defect situation	Max score for bound item	Normalised
Fatal × natively hard to fix	0.5/5	10%
Fatal × low-cost fixable	1.5/5	30%
Single significant defect	2.5/5	50%
Multiple significant + natively hard to fix	2.0/5	40%
Multiple significant (other)	2.5/5	50%
Minor defect	3.5/5	70%
Multiple minor + natively hard to fix	3.0/5	60%

Defect deduplication. Defects are merged by the composite key: *requirement* × *defect type* × *asset* × *spatiotemporal position*. For both modalities, matching reports from the primary and audit VLMs are merged into one confirmed defect. A fatal or significant defect reported by only one model is cross-verified by the other; a refuted report is discarded, while a confirmed report enters capping.

Important. "Fatal × natively hard to fix" does **not** by itself trigger a case-level veto (zero). It only caps the bound item at 0.5/5.

CASE-LEVEL GATES

Table 21: The four registered case-level gates.

Gate rule	Trigger condition	Effect
`no_final_delivery`	No final deliverable exists	All items → 0
`all_deliveries_unreadable`	Every delivered file is unreadable	All items → 0
`exact_input_reuse`	Input asset reused verbatim as output	All items → 0
`overall_case_mismatch`	Delivery fundamentally mismatches the task brief	All items → 0

Unregistered rules cannot trigger a case gate. Every gate activation must carry auditable `evidence` and `source` fields.

DELIVERY-SPECIFICATION CEILINGS

When a deterministic specification guard fires (fail), beyond per-item handling the *entire case score* is capped at a registered ceiling (0–100 scale):

Table 22: Delivery-specification case-level ceilings.

metric_id	**Trigger**	**Case cap**
count_compliance	Delivery count mismatch	≤15/100
size_compliance	Image dimensions mismatch	≤20/100
video_width / video_height	Video resolution mismatch	≤20/100
video_fps / video_duration_seconds / video_codec	Frame rate, duration, or codec mismatch	≤30/100
audio_stream_present	Audio presence contradicts task requirement	≤30/100

Implementation technique. Each check item (0–5 scale) is capped at limit/20. Because $R = \sum(w \times s)/5$ and the weights sum to 100, the requirement score is structurally guaranteed not to exceed the registered cap; since the quality factor of Eq. (2) is at most 1.0, case_score $\leq R$ and the cap therefore also bounds the case score. This holds under any recomputation.

Scope restriction. Only specifications explicitly stated in the frozen task brief *and* already attached to a check item participate. Unstated specifications are never inferred.

Mutual exclusion with veto. When a case-level gate (zero) fires, the specification cap is not applied additionally.

Audio direction rule. The direction is determined by the task brief: "retain source audio" / "must have sound" requires audio presence; "keep silent" / "no audio track" requires audio absence. A video-editing case declaring "keep no audio track" is never penalised in reverse.

Canvas exclusion. Evaluation canvas dimensions (e.g., "518×518 analysis canvas for MUSIQ…") are internal measurement parameters, not delivery specifications, and are excluded from spec judgment.

A.11 CALIBRATION PROTOCOL

Calibrator composition. 15 cases: 5 image (one per sub-track) + 10 video (one per sub-track), each of moderate difficulty. Used to observe overall score level and cross-version discrimination.

Table 23: Calibration results across two system versions against design targets.

Version	**Image mean**	**Video mean**	**Total**	**Design target**
baseline_mode2	23.87	19.86	**21.87**	∼20 (image slightly above video)
codex_mode2	37.22	37.37	**37.29**	35–40

Calibration results.

Source of discrimination. The gap is driven primarily by delivery specification case-level caps:

- baseline: image 4/5 cases trigger size_compliance cap; video 6/10 trigger resolution cap.
- codex: image 2/5 trigger (one count, one size); video 2/10 trigger (one resolution, one audio).

Positioning. The calibrator is a *ruler for tuning rules*, not a report card, and not a validation set: because the rule parameters are tuned on the same 15 cases against which the result is reported, the exercise evidences *internal consistency* of the parameterisation rather than generalization to unseen systems. The design bands are assigned by two of the authors, so they carry confirmation-bias risk; independent third-party annotation and a held-out calibration split are left to future work. After any rule adjustment, the same 15 cases should be recomputed offline (without re-invoking the VLM) to confirm the score level matches design intent before launching full-scale formal scoring.

A.12 Judge Roles and Conservative Synthesis

Unified Model Roles

Table 24: Unified VLM judge roles across modalities.

Role	Model	Trigger and behaviour
Primary	`qwen3.8-omni-flash`	Independently scores every frozen item for image and video cases, returning clause-level evidence and structured defects. A contract-invalid response after retries makes the case `evaluation_unavailable`.
Audit	`gemini-3.7-flash`	Blindly scores the same items without the primary response. After evidence gating, it lowers an item only when the primary score is at least 4; an invalid audit response yields an explicit primary-only result.
Cross-verifier	Opposite VLM	Verifies a fatal or significant defect reported by only one model: primary-reported defects go to audit, and audit-reported defects go to primary.

The role assignment is identical across modalities. For each item, evidence gating precedes the asymmetric high-min rule

$$s_{\text{synth}} = \begin{cases} \min(s_{\text{primary}}, s_{\text{audit}}), & s_{\text{primary}} \geq 4 \text{ and audit valid,} \\ s_{\text{primary}}, & \text{otherwise.} \end{cases}$$

The synthesized score then passes through objective rules (which can only reduce or maintain item scores, never inflate them), confirmed-defect caps, case gates, and delivery-specification caps. No human raters participate in this scoring path.

Audit metadata. Each case record preserves the raw primary and audit responses, evidence-gate and high-min decisions, synthesis mode and any audit-unavailable reason, cross-model verification requests and results, objective actions, confirmed defects, caps, call counts, and errors.

A.13 Multi-Video Assessment

Default interpretation. Multiple output videos within one case are treated as **component parts of a single work** (e.g., short-drama segments, interactive branches, main feature + supplementary clips). They are submitted jointly in one request following the manifest order; the prompt labels each file by name and role. They are *not* compared as alternative candidate versions, nor is the minimum score simply taken.

Joint assessment. The model judges overall completeness, cross-segment continuity, branch coverage, and per-part quality. Local defects remain bound to the specific media file and requirement item.

A.14 Dependency Preflight

A one-time dependency preflight executes before any API scoring call and verifies the full dependency matrix:

Table 25: One-time dependency preflight matrix.

Dependency	Requirement
ffmpeg / ffprobe	Must be on PATH; used for codec, duration, fps, audio stream probing.
Tesseract	`chi_sim+eng` language packs installed; used for OCR context metrics.
torch / torchvision	GPU-capable installation; backend for LPIPS, DINOv2, CLIP, RAFT, MUSIQ.
open_clip	Pre-trained CLIP weights at pinned revision hash.
lpips	AlexNet backbone weights cached locally.
pyiqa	MUSIQ and aesthetic model weights at pinned revisions.
transformers	HuggingFace model weights for DINOv2 at pinned revision hash.
Pillow	Image I/O and basic transforms.

The preflight outputs a matrix: `metric / applicable / backend / weight revision-hash / status`. If any required dependency is missing, the system **halts before any API scoring call**. In formal mode, implicit downloading or model substitution is prohibited.

A.15 Frozen Experimental Parameters

The configuration items frozen before the production run are enumerated below. Items determined by the scoring rulebook are filled; remaining items are documented as available from production run logs.

A.16 Cost Field Definitions

Table 3 in the main text reports Toolset B execution statistics. This section defines each field and documents coverage limitations.

Field definitions.

- **Coverage (Cov.)**: Number of cases for which execution data (elapsed time, token count, or tool-call count) exists in the log source.
- **Success (Succ.)**: Number of cases producing at least one deliverable file (partial deliveries count as success; empty outputs do not).
- **Avg Time**: Mean per-case wall-clock elapsed time in seconds, computed over covered cases.
- **Avg Tokens**: Mean per-case total token count. For systems with partial token reporting, this includes only reported tokens. Unreported tokens are noted, never treated as zero.
- **Avg Calls**: Mean per-case tool invocations.
- **Flag**: Data-quality annotation. Blank means standard single-source coverage.

Per-system coverage notes. Codex, DeepSeek-Harness, and Baseline each reach 120/120 case coverage (token coverage 85.4%, 81.6%, and 65.3% respectively); Baseline records 118 successes after four retry batches. Qoder aggregates 120/120 cases from seven source batches (673 attempts), where the `wan-video` batches report zero tokens—*unreported*, not zero—and elapsed time is inflated by multi-round retries. Workbuddy is flagged HETEROGENEOUS_COVERAGE: Table 3 reports the more representative `toollog` source (115/120 cases) rather than the last-batch-only `execution_log.jsonl` (60 cases), which is unsuitable for cross-system comparison.

Per-case score–cost pairing. Per-case score data for the full 120-case runs are not available as structured data on disk for most systems. Video sidecars exist only for DeepSeek-Harness (60 video cases). Consequently, Pearson or Spearman correlations between per-case quality and per-case cost

Table 26: Experimental parameters. Items determined by the scoring rulebook are filled; remaining items are documented as available from production run logs.

Configuration item	Status
System versions, backbone models, and access mode	Codex (GPT6-Astra, run ID `gpt6_dsd`), Qoder (Qwen3.8-max), Workbuddy (Hy4-preview), DeepSeek-Harness (DeepSeek-V4.1-Flash)
Generation and understanding tool versions	Toolset A: Qwen-Image-3.0-Pro + Wan3.0; Toolset B: GPT-Image-2 + SeedAnce (`DSD-2.0`)
Judge model and scoring scheme	Both modalities: primary `qwen3.8-omni-flash`, blind audit `gemini-3.7-flash`; high-min when primary ≥ 4
System sampling parameters and number of repeat runs	1 run per system×toolset; serial per-case execution
Permissions, runtime isolation, and network access scope	Agent-level Python sandbox; network access to API endpoints only; eval package inaccessible to agents
Token, call-count, cost, and time budgets	Not pre-specified; execution telemetry collected post-hoc (Table 3)
Timeout, retry, and termination rules	90 QPM per model, max 8 in-flight; upload fail $\rightarrow$ `evaluation_unavailable`
Per-case weight budget and scoring formula	Requirement 100 (full budget); quality factor $0.60 + 0.20\,Q/5 + 0.20\,A/5$; fixed denominator
Video rating scale and aggregation weights	0–5 per check item; per-case requirement budget 100
Overall definition and cross-modal weights	image $\times$ 0.5 + video $\times$ 0.5
Human references and validity-validation plan	Five-person expert checklist (§ 4.1); author calibration bands (§ A.11)
Statistical testing and missing-data analysis plan	Descriptive system-level comparisons; per-case correlation not feasible (§ A.16)

cannot be computed across the full benchmark; only system-level descriptive comparisons are reported in § 5.2.