DRAMACHAIN BENCH: AN END-TO-END BENCHMARK FOR SHORT-DRAMA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Commercial short-drama production spans multiple stages. Most benchmarks, however, assess video generation in isolation, leaving unclear whether defects originate in the evaluated stage or propagate from earlier ones. In addition, most use manually constructed inputs that do not reflect real production conditions. We present DramaChain Bench, the first short-drama benchmark evaluating pipeline-native outputs from storyboard design to multi-episode delivery. It combines an evaluation schema with three systems for generation, human annotation and automated judging. DramaChain Dimensions defines 63 leaf dimensions across six granularities and five axes: input fidelity, internal consistency, generation plausibility, visual quality and cinematic expressiveness. From supplied scripts, DramaChain Agent generates storyboards, keyframes, shot videos and finished dramas, with its workflow and output quality benchmarked against commercial platforms. DramaChain Labeling System provides three independent assessments for each of 5,569 reference items, with spatial and temporal defect localisation where applicable. DramaChain Agentic Judge combines a master–sub-agent architecture, 29 specialist modules and iterative tool use, achieving a mean model-level PLCC of 0.918 against human scores. Using this framework, we evaluate 37 generation models: 20 text LLMs, 8 image models and 9 video models. A series of judge models trained on these annotations with supervised fine-tuning and GRPO further achieve higher agreement with human ratings than the prompted agentic judge.

1 INTRODUCTION

Short drama has become one of the most important deployment scenarios for generative video. Global micro-drama revenue hit USD 11 billion in 2025 and is expected to reach USD 14 billion in 2026 (Omdia, 2025; 2026). Commercial platforms have settled on a largely shared production paradigm: they take a story premise from the user and execute a staged pipeline of script, storyboard, keyframe images, shot videos and finished short drama. Academic work has converged on the same paradigm, most typically in agent frameworks that assign each stage to a specialised role (Huang et al., 2026b; Wang et al., 2026a; Shi et al., 2025). Industry and academia are therefore building the same staged chain: each stage's input is the previous stage's output, so an upstream error caps every stage after it.

While the production pipeline has matured, evaluation methodologies have not kept pace. Early-generation benchmarks focused on individual video clips, assessing model capabilities via model-level human preference ratings (Huang et al., 2024; Zheng et al., 2025). Subsequent work expanded evaluation to cinematic language and multi-shot scenarios (Wang et al., 2026b; Shi et al., 2026a; He et al., 2026; Zhuang et al., 2026), and more recent benchmarks have structured their criteria around the end-to-end production workflow (Yang et al., 2026; Chen et al., 2026). DramaBoard evaluates storyboards, images and videos (Zhou et al., 2026a), while VideoWeaver judges execution traces alongside the final video (Wei et al., 2026). Three gaps remain for full-chain short-drama evaluation. (1) **Pipeline stages are evaluated in isolation**: most benchmarks do not jointly assess native intermediate outputs and assembled multi-episode dramas. (2) **Test inputs are pre-authored**: most benchmarks substitute curated inputs for actual upstream outputs, making it difficult to distinguish defects introduced locally from those inherited upstream. (3) **Scores do not localise the defect**: most evaluations report a score per dimension without the region, moment or reason for failure, leaving no localised evidence to verify score accuracy.

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.

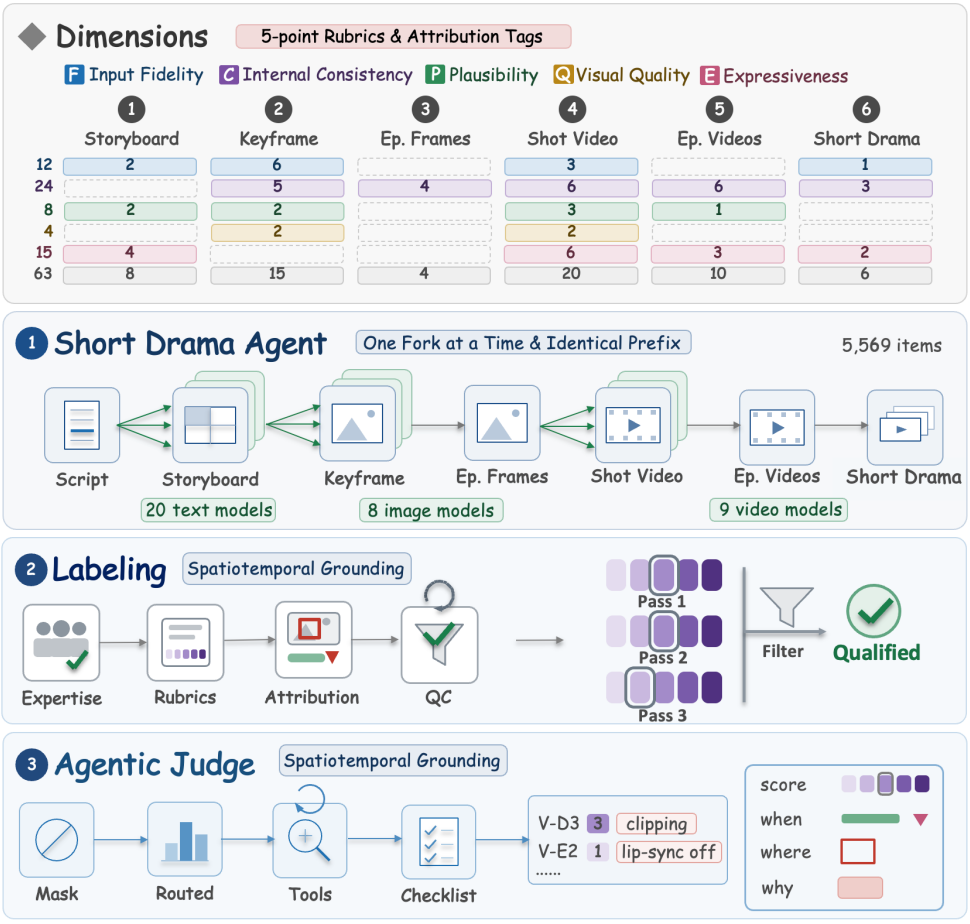


Figure 1: Overview of DramaChain Bench. The item count denotes the complete three-annotator reference set.

We present DramaChain Bench, a short-drama benchmark that evaluates pipeline-native outputs from storyboard design to multi-episode delivery. A shared evaluation schema, DramaChain Dimensions, organises 63 leaf dimensions under five axes: **F** input fidelity, **C** internal consistency, **P** generation plausibility, **Q** visual quality and **E** cinematic expressiveness. These axes are instantiated at the applicable granularities over a reference set of 5,569 triply annotated items, allowing both individual artefacts and their assembled sequences to be examined in context.

Three systems implement this schema (Fig. 1). DramaChain Agent produces evaluation items with their upstream context, enabling stage-wise comparisons on actual pipeline outputs rather than curated substitutes. DramaChain Labeling System connects human scores to dimension-specific defect tags and modality-appropriate evidence. Each deduction includes an explanation and, where applicable, a spatial region or temporal interval. DramaChain Agentic Judge applies the same rubrics through per-dimension agents and targeted evidence collection. Readers can use this evidence to check whether the cited defect exists, which rubric it violates and whether its severity supports the assigned score. The shared protocol supports comparisons between the evidence behind human annotations and automated judgements. Together they connect what was requested, what was generated and why it failed; Sec. 3 specifies the construction and evaluation protocols.

Our contributions are threefold. **(1)** We present the first benchmark that evaluates an end-to-end short-drama pipeline on pipeline-native outputs, linking defects to their upstream context and measuring how upstream weaknesses affect downstream delivery. **(2)** We diagnose 37 models stage by stage, 20 text LLMs, 8 image models and 9 video models, yielding a per-dimension profile of each model's strengths and weaknesses rather than a global ranking. **(3)** We build DramaChain Agentic Judge, an agentic scoring framework validated against three-annotator labels at a mean PLCC of 0.918, and train a judge model for lightweight, fast evaluation (Sec. 4.4).

2 Related Work

Storytelling video generation. Work on storytelling video splits in two. The first line generates the narrative inside a single model (Meng et al., 2025; 2026), and most of that literature is about memory: keyframe stores, entity banks, shot-gated slots, adaptive forgetting and the reallocation of context (Zhang et al., 2025; Vandersanden et al., 2026; Huang et al., 2026a; Wu et al., 2026; Zhou et al., 2026b; Long et al., 2026; Liu et al., 2026a). That so much of the effort goes into memory is itself evidence of the problem: long-range consistency degrades with distance. The control these models offer stays coarse as well, since a prompt, a camera path or a sketch (Wu et al., 2025a; Zhang et al., 2026d; Xu et al., 2026a) cannot re-issue a wrong requirement to the stage responsible for it. The second line decomposes the problem into staged roles that hand artefacts to one another, from animation systems (Li et al., 2024; Shi et al., 2025) to agent frameworks (Liang et al., 2025; Wang et al., 2026a; Huang et al., 2026b; Song et al., 2026a;b) and short-drama systems (Zhou et al., 2026a; Shi et al., 2026b), all of which locate the difficulty in the hand-offs. Commercial platforms run chains of this second kind, which makes per-stage evaluation both possible and necessary.

Benchmarks for video generation. The first generation scored individual clips, and VBench set the template the field still works in: decompose quality into capability dimensions and implement each with a generalist VLM or a specialist detector (Huang et al., 2024; Zheng et al., 2025). That template was then refined with MLLM-based judging and with trained judge models (Han et al., 2025; Yang et al., 2025; Wu et al., 2025b; Liang et al., 2026; Bansal et al., 2025). The second generation extended the target to cinematic language and multi-shot settings, adding cross-shot consistency, multi-reference audio-video, narrative richness and minute-scale generation (Wang et al., 2026b; Shi et al., 2026a; He et al., 2026; Zhuang et al., 2026; Li et al., 2026a; Zhang et al., 2026c; Feng et al., 2026; Zheng et al., 2026; Xu et al., 2026b; Zhang et al., 2026b; Liu et al., 2026b; Wang et al., 2025; Liu et al., 2025). The third brings the production process into the evaluation itself, defining criteria per stage, inspecting a run's checkpoints or reading an agent's execution trace (Yang et al., 2026; Chen et al., 2026; Wei et al., 2026). Yet most benchmarks in Tab. 1 score videos only; DramaBoard also scores storyboards and images. DramaChain Bench extends this coverage to assembled multi-episode dramas, jointly evaluating the native chain with 63 leaf dimensions, 5,569 triply annotated instances and spatiotemporal evidence for verifying individual deductions.

Table 1: Comparison with related benchmarks. *Full-chain* jointly evaluates native intermediate outputs and assembled multi-episode short dramas. Outputs: S storyboards, I images, V videos, D finished dramas (episode/series-level). Counts are reported or estimated ($\approx$, $<$); n/r: insufficient information; $\times$: not provided by the published protocol. †: evaluation suite in a generation-system paper. Our benchmark counts triply annotated instances.

Benchmark	Full-Chain Short-Drama Evaluation	Outputs Scored	Leaf Dimensions	Human-Annotated Items	Deductions Localised in Space & Time
VBench (Huang et al., 2024)	×	V	16	n/r	×
VBench-2.0 (Zheng et al., 2025)	×	V	18	n/r	×
FilmBench (Wang et al., 2026b)	×	V	38	2,878	×
MSVBench (Shi et al., 2026a)	×	V	20	$\approx$300	×
EntityBench (He et al., 2026)	×	V	51	200	×
EvalVerse (Yang et al., 2026)	×	V	45	n/r	×
DirectorBench (Chen et al., 2026)	×	V	40	$<$100	×
Short-Drama-Bench† (Shi et al., 2026b)	×	V	15	$\approx$1.8k	×
DramaBoard† (Zhou et al., 2026a)	×	S, I, V	9	$\approx$120	×
VideoWeaver† (Wei et al., 2026)	×	V	13	200	×
DramaChain Bench (ours)	✓	**S, I, V, D**	**63**	**5,569**	✓

3 DramaChain Bench

DramaChain Bench uses DramaChain Dimensions as a shared schema: DramaChain Agent generates pipeline-native artefacts, DramaChain Labeling System supplies human references, and DramaChain Agentic Judge automates evaluation.

3.1 Evaluation Schema: DramaChain Dimensions

Production coverage. The schema covers commercial short-drama production end to end at six evaluation granularities: storyboard design, single keyframes, episode keyframe sets, shot videos, episode videos and complete dramas (Fig. 2).

Axes and rubrics. **F input fidelity** tests adherence to script or upstream requirements; **C internal consistency** tests identity, scene and style within and across artefacts. **P generation plausibility** covers storyboard executability and appropriate elaboration, anatomical and physical plausibility, motion smoothness and causal-temporal coherence. **Q visual quality** covers clarity, artefacts, material and motion texture, and genre-appropriate aesthetics. **E cinematic expressiveness** covers framing, camera work, performance, audio and watchability. Each second-level (leaf) dimension has a five-level rubric and attribution tags, both distilled from industry experts' experience into unified, operational standards.

3.2 Data Construction: DramaChain Agent

Corpus and paradigms. Built on ViMax (Huang et al., 2026b), DramaChain Agent generates 20 dramas of 3 consecutive episodes (60 episodes), spanning genre, period, live-action/animation and Chinese/English dialogue (Sec. A). The evaluated video paradigms use first/last frames, sliced grids or character/element/scene references, covering mainstream short-drama production settings.

Controlled comparisons. Within a shared input paradigm and duration budget, stage-level comparisons vary only the tested model, reusing upstream artefacts and prompts. To trace downstream error propagation, we instead replace one upstream model while fixing the downstream model (Sec. 4.3).

Evaluation in production settings. By following commercial short-drama workflows, DramaChain Agent enables DramaChain Bench to evaluate models in realistic production settings. Expert review rated its single-pass output in the same band as three commercial platforms on one script (Fig. 7); Fig. 8 shows a three-episode output.

3.3 Human Annotation: DramaChain Labeling System

Protocol and records. Our 543-annotator pool includes raters with directing or screenwriting backgrounds for storyboards and experienced image/video annotators for other stages. The protocol targets three independent 1–5 assessments per item across applicable leaf dimensions. Scores below 5 require dimension-specific tags and explanations, with spatial or temporal localisation where applicable (Sec. B.2). Our reference set comprises 5,569 items, each with all three assessments. The full annotation collection yielded 17,488 multi-dimension records and 255,925 attribution markers; these collection-wide totals are not restricted to the complete reference set.

Quality control. Manual review is complemented by four record-level checks (speed, evidence, completeness and review outcome) and five rater-level checks: (i) score concentration, (ii) limited score diversity, (iii) repeated uniform scoring across leaf dimensions within items, (iv) low correlation with leave-one-out consensus, and (v) excessive leniency or severity (Sec. C).

Consensus and human baseline. Human consensus is the mean of available annotator scores at both leaf and composite levels. To contextualise automated agreement, the leave-one-out human baseline HB computes PLCC between each annotator's score and the mean of the remaining annotators, pooling all pairs. DramaChain Agentic Judge is compared with the annotator mean on items with both human and automated scores (Tab. 16).

3.4 Automated Evaluation: DramaChain Agentic Judge

Task allocation. The master agent selects applicable leaf dimensions from each item's granularity, paradigm and references, builds checklists, and assigns one large-model sub-agent per dimension for parallel evaluation.

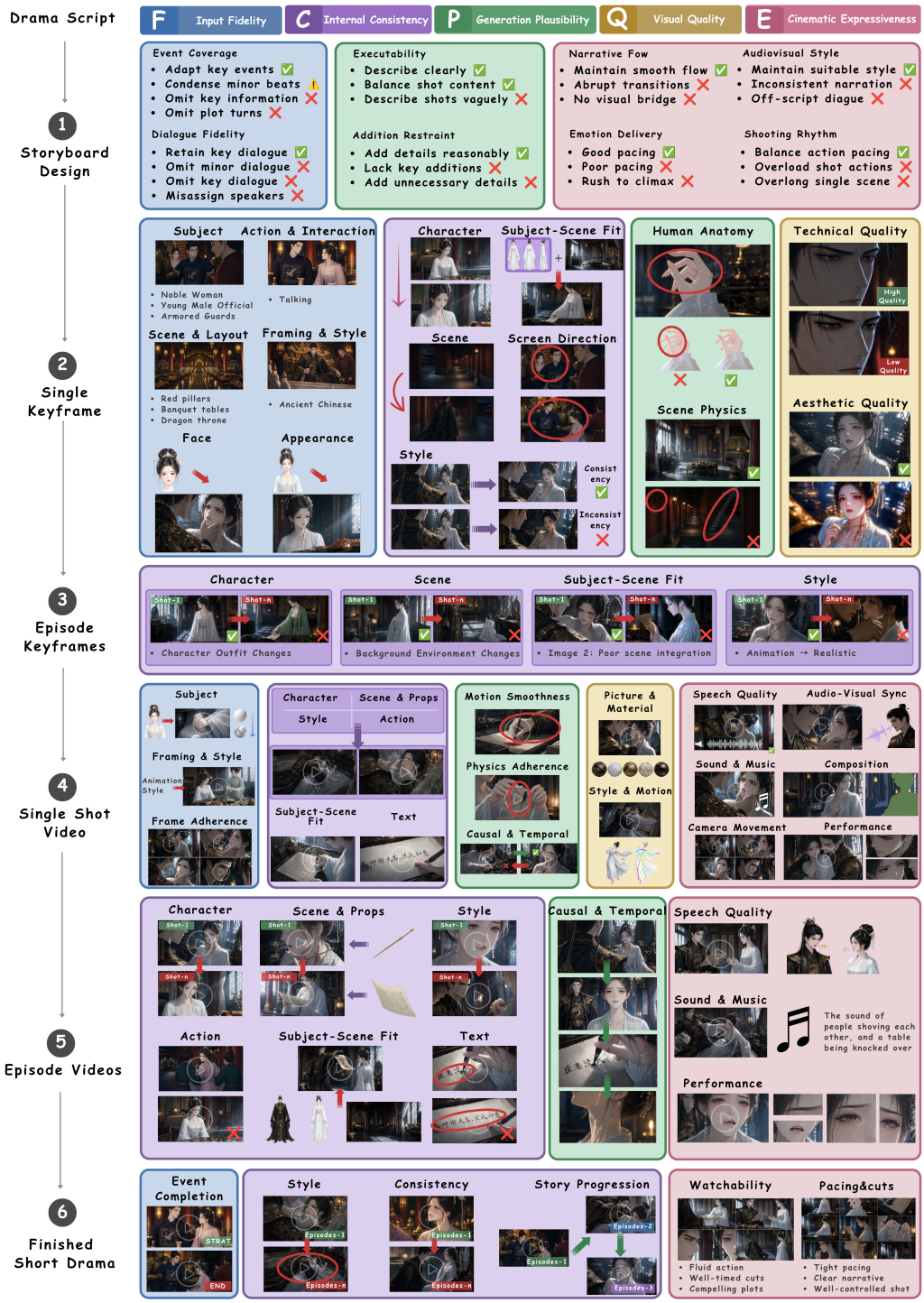


Figure 2: DramaChain Dimensions: expert-derived criteria across six production granularities and five evaluation axes.

Evidence collection. We combine routed small-model perception with on-demand inspection by large-model sub-agents. Depending on the dimension, specialised small models provide character similarity, optical flow, speech transcripts or audio-quality measurements, supplemented by deterministic signals such as sharpness. These small models supply evidence, not final rubric scores (Sec. D.3). Large-model sub-agents can then call tools to magnify regions, extract timestamped frames and inspect audio segments; returned evidence supports iterative inspection before scoring (Sec. B.1).

Scoring. Each sub-agent judges the original media and collected evidence against the rubric, returning `score`, `confidence`, `tags` and `reason` (Fig. 10). Deductions follow the human protocol (Sec. 3.3). Prompts and deployment settings appear in Secs. D.1 and 4.1.

4 EXPERIMENT

4.1 SETUP

Coverage. We automatically evaluate 37 generation models (20 text, 8 image and 9 video), of which 22 (9 text, 7 image and 6 video) also receive human annotation. Scores range from 1 to 5, with 3.0 marking usability. Incomplete runs are scored on their observed subsets; coverage and budget exceptions are documented in Sec. D.5.

Judging models. Storyboard scores average `gpt-5.6-sol-xhigh` and a `claude-fable-5` variant to reduce single-family dependence. Keyframes use `doubao-seed-2.1-pro`; shot videos use `gemini-3.1-pro`.

4.2 MAIN RESULTS

Table 2 reports automated scores across six granularities and five axes, alongside model-level agreement with human assessments.

The strongest chain at annotation time. Among models covered by human annotation, the strongest evaluated chain combines `gpt-5.5-xhigh` for storyboards, `gpt-image-2` for keyframes and `seedance-2.0` for video, reaching an automated finished-drama score of $3.30/5$. In the expanded set, `gpt-6-astra-max` and `gpt-image-2.5` lead storyboards and keyframes, while `seedance-2.5` has the highest automated finished-drama score ($3.70/5$).

Performance declines towards video generation. Mean scores in the human-annotated model set fall from 3.78 for storyboard design to 3.62 for single keyframes, 3.30 for episode keyframes, 3.09 for shot video and 2.93 for episode video, then recover to 3.03 for the complete drama. Stage-specific criteria reveal a shift from shot scheduling to visual specification adherence and temporal coherence: video must sustain motion and audio-visual synchrony, not merely preserve static appearance (Sec. E).

Similar totals conceal different model strengths. `kling-3.0-omni` and `pixverse-c1` differ by 0.01 overall on shot video, yet `kling-3.0-omni` is 0.16 higher on **P** and 0.22 lower on **E**. Gains and losses across axes offset in the composite, as in a close keyframe pair (Sec. E.2); each model has its own strengths and weaknesses.

Automated rankings track human assessments. DramaChain Agentic Judge achieves mean model-level PLCC 0.918 across six granularities and exact complete-drama rankings (SRCC $= 1.000$; Tab. 16). Item-level agreement reaches or exceeds the leave-one-out human baseline (HB) for text and images ($2.35\times$ for storyboards), but only 0.79–$0.85\times$ for video. Enumerable checks allow direct matching to requirements; video synchrony and smoothness also require fine temporal perception, which an explicit checklist alone cannot supply (Sec. F.2.5).

4.3 CRITICAL INSIGHTS

Beyond aggregate rankings, we examine cross-shot defects, upstream error propagation and differences in models' ability to follow production requirements.

Table 2: Main results. 5-point scores over six stages and five axes; *All* averages all applicable leaf dimensions, weighting axes by leaf count. Red: below the 3.0 usable line; **bold**: top annotated model per column; dash: axis undefined there. Stage headers give model-level PLCC/SRCC against the three-annotator human board. †: scored automatically only; ‡: partial corpus (Sec. D.5).

Model	F Input fidelity	C Internal consistency	P Generation plausibility	Q Visual quality	E Cinematic expressiveness	All
① Storyboard design text → text · PLCC 0.936 · SRCC 0.800						
gpt-5.5-xhigh	**4.84**	–	**3.89**	–	4.10	**4.23**
claude-opus-4.8-max	4.51	–	3.80	–	**4.19**	4.17
hy3	4.36	–	3.65	–	3.78	3.89
doubao-seed-2.1-pro	4.26	–	3.65	–	3.82	3.89
glm-5.2	4.01	–	3.66	–	3.74	3.79
gemini-3.1-pro	4.08	–	3.43	–	3.67	3.71
kimi-k2.6	3.66	–	3.42	–	3.74	3.64
qwen3.7-max	3.24	–	3.48	–	3.59	3.48
mimo-v2.5-pro	2.68	–	3.53	–	3.41	3.26
gpt-6-astra-max†	4.90	–	4.25	–	4.28	4.43
gpt-5.6-sol-max†	4.95	–	4.02	–	4.31	4.40
glm-5.3-max†	4.78	–	3.88	–	4.20	4.26
claude-fable-5-max†	4.55	–	3.99	–	4.20	4.24
claude-opus-5-max†	4.45	–	3.85	–	4.30	4.22
kimi-k3†	4.67	–	3.87	–	4.17	4.22
qwen3.8-max†	4.55	–	3.88	–	3.94	4.08
hy4-preview†	4.69	–	3.56	–	3.91	4.02
grok-4.5†	3.96	–	3.85	–	3.92	3.92
gemini-3.8-flash-high†	4.18	–	3.42	–	3.79	3.80
gemini-3.6-flash-high†	4.04	–	3.65	–	3.71	3.78
② Single keyframe text + refs → pixels · PLCC 0.959 · SRCC 0.750						
gpt-image-2	**3.65**	**4.32**	**4.47**	**4.13**	–	**3.99**
gpt-image-1.5	3.25	3.96	4.46	3.90	–	3.72
nano-banana-2	3.33	3.92	4.25	3.71	–	3.68
seedream-5.0-pro	3.56	3.92	3.74	3.52	–	3.64
nano-banana-pro	3.17	3.89	4.34	3.79	–	3.62
wan2.7-image-pro	2.95	3.63	4.03	3.46	–	3.38
seedream-5.0-lite‡	2.96	3.40	3.89	3.40	–	3.32
gpt-image-2.5†	3.83	4.46	4.64	4.23	–	4.12
③ Episode keyframes a whole episode of frames · PLCC 0.973 · SRCC 0.964						
gpt-image-2	–	**3.82**	–	–	–	**3.82**
gpt-image-1.5	–	3.60	–	–	–	3.60
nano-banana-2	–	3.47	–	–	–	3.47
seedream-5.0-pro	–	3.28	–	–	–	3.28
nano-banana-pro	–	3.04	–	–	–	3.04
wan2.7-image-pro	–	3.00	–	–	–	3.00
seedream-5.0-lite‡	–	2.91	–	–	–	2.91
gpt-image-2.5†	–	4.00	–	–	–	4.00
④ Shot video pixels → time + audio · PLCC 0.755 · SRCC 0.829						
seedance-2.0‡	3.44	**3.27**	3.06	**3.77**	3.05	**3.24**
happyhorse-1.1	3.34	3.21	3.04	3.43	3.08	3.18
kling-3.0-omni	3.22	3.23	**3.08**	3.28	2.88	3.11
pixverse-c1	3.07	3.09	2.92	3.42	**3.10**	3.10
wan2.7‡	**3.45**	2.96	2.87	3.22	2.92	3.04
veo-3.1‡	3.17	2.77	2.73	3.32	2.76	2.88
seedance-2.5†‡	3.24	3.40	2.97	3.71	2.85	3.17
minimax-h3†	3.43	3.20	2.83	3.46	3.11	3.17
wan-3.0†	3.51	3.26	2.74	3.42	2.97	3.15
⑤ Episode video shots in sequence · PLCC 0.935 · SRCC 0.943						
seedance-2.0‡	–	**2.96**	3.41	–	3.32	**3.11**
happyhorse-1.1	–	2.84	**3.64**	–	3.37	3.08
kling-3.0-omni	–	2.79	3.31	–	**3.48**	3.06
wan2.7‡	–	2.57	3.24	–	3.35	2.87
veo-3.1‡	–	2.47	3.27	–	3.18	2.76
pixverse-c1	–	2.42	3.08	–	3.16	2.71
seedance-2.5†‡	–	3.16	3.11	–	3.50	3.25
wan-3.0†	–	3.12	3.70	–	3.42	3.25
minimax-h3†	–	2.87	3.17	–	3.27	2.99
⑥ Short drama the finished short drama · PLCC 0.947 · SRCC 1.000						
seedance-2.0‡	**4.29**	3.38	–	–	2.68	**3.30**
happyhorse-1.1	3.82	3.38	–	–	**2.69**	3.22
kling-3.0-omni	3.37	**3.41**	–	–	2.54	3.11
wan2.7‡	4.12	2.80	–	–	2.59	2.95
pixverse-c1	3.05	2.88	–	–	2.58	2.81
veo-3.1‡	2.67	3.07	–	–	2.44	2.79
seedance-2.5†‡	4.40	3.93	–	–	3.00	3.70
wan-3.0†	4.23	3.56	–	–	2.74	3.47
minimax-h3†	3.77	3.33	–	–	2.51	3.20

Isolated scores conceal cross-shot defects. Individually usable assets do not guarantee a usable episode or finished drama. For example, the same keyframes score 3.33–4.00 individually but 2.33 as an episode; its shots score 3.23–4.00 individually but 1.67 as a continuous segment. Human scores likewise fall by nearly one band (Sec. E.4). Event completion, cross-episode progression, pacing and watchability must be assessed on the assembled work, not inferred from individual scores.

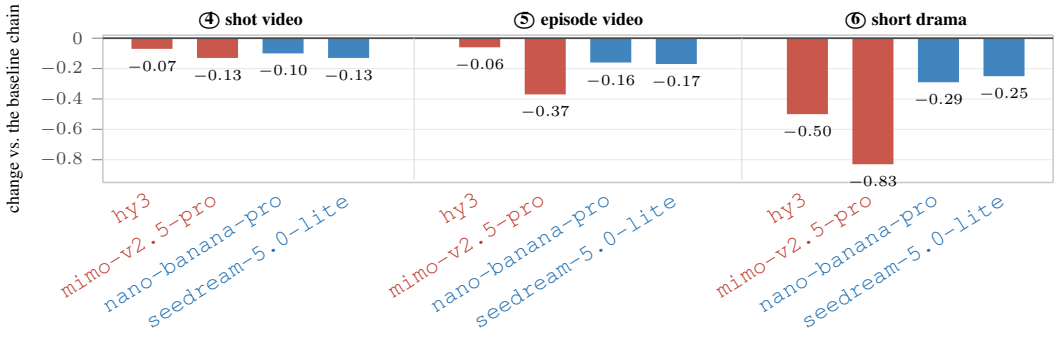


Figure 3: Upstream defects accumulate rather than staying local. One upstream stage is replaced by a weaker model, either the storyboard model (red) or the keyframe model (blue), and everything else is held unchanged; design in Sec. G.5.

Upstream weaknesses reduce downstream quality. With the video model fixed, substituting a weaker upstream model reduces scores by 0.07–0.13 on individual clips, 0.06–0.37 on continuous segments and 0.25–0.83 on multi-episode delivery (Fig. 3; design in Sec. G.5). For example, over-segmenting storyboard actions can make the video model reproduce frequent cuts rather than continuous movement, propagating upstream pacing defects.

Models differ most in fidelity and consistency. Averaged over applicable stages, top-to-bottom model gaps are 1.22 on **F** input fidelity and 0.70 on **C** internal consistency, compared with 0.64, 0.53 and 0.42 on **Q**, **P** and **E**; human scores likewise show the largest gap on **F**. These gaps suggest uneven control over prescribed content and continuity: generating plausible content does not ensure that a model preserves required details or keeps identities, scenes and styles consistent across shots.

4.4 Training a judge model

Training data. Training examples pair benchmark artefacts with mean human scores and associated attribution text. High-disagreement labels are down-weighted or excluded. Low-scoring items and instruction-mismatched negatives broaden score coverage, while general video samples supplement video-judge training. The test subsets in Sec. 4.5 are held out from both supervised fine-tuning (SFT) and GRPO.

Training algorithm. We fine-tune `Qwen3-Omni-30B-A3B` for video and `Qwen3.8-27B` for keyframes and storyboards, keeping modality encoders frozen. SFT learns structured rationales and rubric scores through token, score-regression and instruction-following objectives. Scores are continuous expectations over five band tokens, $\hat{s}_d = \sum_b v_b p_b^{(d)}$, where $v_b \in \{1, \dots, 5\}$ and $p_b^{(d)}$ is the renormalised probability of band b at dimension d. GRPO then optimises the reward

$$R = R_{\text{score}} + 0.2\, R_{\text{fmt}} + 0.3\, R_{\text{attr}}, \tag{1}$$

$$R_{\text{score}} = \sum_{\ell \in \{1,2\}} \omega_\ell \, \text{mean}_{d \in \mathcal{D}_\ell} \left[1 - \tfrac{1}{4} |\hat{s}_d - s_d^*| \right]_+, \tag{2}$$

where s_d^* is the human target and $(\omega_1, \omega_2) = (0.6, 0.4)$ weight axis-level dimensions $\mathcal{D}_1$ and their leaves $\mathcal{D}_2$. R_{fmt} checks the output schema; R_{attr} compares predicted localisations and tags with available human markers. The reward thus combines score agreement, format compliance and defect attribution. Section D.4 summarises modality-specific supervision and inference timing.

4.5 Ablation study

Setup. We compare four prompted architectures: the full system, without routed small-model evidence, without callable tools, and without either (a bare LMM). Measurement-only scoring and a trained judge serve as separate baselines. The subsets contain 210 keyframes from seven generation models and 180 shot videos from six.

Perception and tool use provide complementary gains. Table 3 compares automated and human means over 14 model–paradigm cells using PLCC/SRCC, with shared video dimensions and arm-specific keyframe coverage. With each stage's prompted judge fixed (Sec. 4.1), removing routed evidence or tools lowers keyframe PLCC from 0.937 to 0.903 or 0.891, respectively; video PLCC falls from 0.800 to 0.701 or 0.656. The full system also exceeds bare LMM and measurement-only baselines on keyframes (0.880, 0.677) and video (0.445, 0.632). This pattern is consistent with complementary roles: small models supply perceptual evidence, tools enable targeted inspection, and the LMM turns their evidence into rubric-based scores.

Training yields the strongest reported agreement. The trained judge achieves the highest PLCC/SRCC in both blocks: $0.955/0.912$ on keyframes and $0.851/0.837$ on video. Built on open-weight models, it receives only the generation inputs, the artefact and the rubric, without routed measurements or tools, yet exceeds the full stack by 0.018 and 0.051 PLCC on keyframes and video. The gains in both correlation measures are consistent with learning rubric-specific score calibration and defect cues from annotated examples, rather than relying on instructions alone.

Table 3: Scoring-architecture ablations. Bold/underline: best/second-best.

Scoring architecture	PLCC	SRCC
② **Single keyframe** 210 items · 7 generation models		
The trained scorer (Sec. 4.4)	**0.955**	**0.912**
DramaChain Agentic Judge, the full stack	<u>0.937</u>	<u>0.886</u>
without the routed measurements	0.903	0.798
without the model-callable tools	0.891	0.833
without either, a bare LMM judge	0.880	0.802
The routed measurements alone, with no LMM judge	0.677	0.771
④ **Shot video** 180 items · 6 generation models		
The trained scorer (Sec. 4.4)	**0.851**	**0.837**
DramaChain Agentic Judge, the full stack: routed measurements + model-callable tools	<u>0.800</u>	<u>0.811</u>
without the routed measurements	0.701	0.675
without the model-callable tools	0.656	0.688
without either, a bare LMM judge	0.445	0.248
The routed measurements alone, with no LMM judge	0.632	0.688

5 Conclusion

To our knowledge, DramaChain Bench is the first benchmark to evaluate short-drama generation end to end on pipeline-native artefacts, from storyboard design to multi-episode delivery. It shifts the evaluation target from isolated output quality to the fidelity and coherence of an interconnected production process. Its expert-derived criteria, evidence-linked annotations and automated judges connect capability measurement with failure diagnosis and evaluator training. Together, these contributions establish a common experimental foundation for improving short-drama pipelines as systems, rather than optimising individual generators in isolation.

6 Limitations

Each model-level correlation rests on six to nine annotated models, and several leaf dimensions have low human reproducibility (Sec. F.2.4), which weakens them as alignment targets. The taxonomy is not exhaustive: post-production, commercial outcomes, scriptless formats and judging-model comparisons are out of scope, each with its reason in Tab. 9.

ETHICS STATEMENT

All benchmark artefacts are synthetic: the twenty dramas are project-original scripts and every character sheet, keyframe and clip is a model output conditioned on them, with no real-person likeness or voice used as a reference and no commercial drama assets reused. The three commercial platforms in the pipeline comparison (Sec. 3.2) were queried through public interfaces with our own scripts, and only their returned outputs are reported. The corpus inherits the tropes the form relies on, including conflict, coercion and revenge; every item was screened against commercial release policies, and the few episodes left without an evaluable output are marked as missing rather than imputed. No broader harm audit was performed, so the benchmark does not certify any artefact as fit for release, and it is not a safety or content-policy evaluator. Annotation was performed by professional contracted staff from third-party vendors rather than volunteers or anonymous crowd-workers, recruited on the background each stage calls for (Sec. 3.3); median per-item working time was 23 minutes and compensation followed vendor terms, with hourly rates unavailable because we did not set them.

REPRODUCIBILITY STATEMENT

The methods and appendices state the reporting conventions. Section B gives every leaf dimension with its axis, the question it asks and the evidence routed to it, and Sec. D gives the scoring prompt, the tool chain and the conventions the numbers rest on: where prompts are identical across models, which models ran which slice of the corpus, and how absolute accuracy is computed. The reported judging configurations are in Sec. 4.1. Section A lists the twenty dramas the items were generated from, Sec. F reports the slice-level breakdowns behind the aggregate agreement figures, and Sec. 3.3 describes the annotation protocol and quality control, so the reference answers can be reconstructed independently. We will release all benchmark data, the judge models and the DramaChain Agent generation pipeline.

AI USE STATEMENT

Generative AI has two separate roles in this work. It is first of all the object of study: the scored items are outputs of the generation models listed in Tab. 2, and the automated scorer is itself built on large multimodal models (Sec. 3.4), whose configurations, evidence and failure modes are reported throughout the paper and in Sec. D. Second, large language models were used to polish the writing: the authors wrote the manuscript and used the models for language editing of their own text. All AI-assisted text was reviewed by the authors, who take responsibility for the final content of this work, including its text, claims and artefacts.

REFERENCES

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025. doi: 10.48550/arXiv.2503.06800. URL https://arxiv.org/abs/2503.06800.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, et al. UniPercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv preprint arXiv:2512.21675*, 2025a. doi: 10.48550/arXiv.2512.21675. URL https://arxiv.org/abs/2512.21675. ICML 2026 Spotlight.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. ArtiMuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025b. doi: 10.48550/arXiv.2507.14533. URL https://arxiv.org/abs/2507.14533.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment.

IEEE Transactions on Image Processing, 33:2404–2418, 2024. doi: 10.1109/TIP.2024.3378466. URL https://doi.org/10.1109/TIP.2024.3378466.

Jiamin Chen, Qianben Chen, Jiawen Zhang, Yidi Wu, Yuchen Li, Xiaokun Zhang, Wangchunshu Zhou, and Chen Ma. DirectorBench: Diagnosing Long-Form Video Generation with Personalized Multi-Agent Evaluation. *arXiv preprint arXiv:2605.30090*, 2026. doi: 10.48550/arXiv.2605.30090. URL https://arxiv.org/abs/2605.30090.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV Workshops*, 2016. doi: 10.1007/978-3-319-54427-4_19. URL https://doi.org/10.1007/978-3-319-54427-4_19.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition, 2013. ISBN 9780203771587. doi: 10.4324/9780203771587. URL https://doi.org/10.4324/9780203771587.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. doi: 10.1109/CVPR.2019.00482. URL https://doi.org/10.1109/CVPR.2019.00482.

Xiaokun Feng, Haiming Yu, Meiqi Wu, Shiyu Hu, Jintao Chen, Chen Zhu, Jiahong Wu, Xiangxiang Chu, and Kaiqi Huang. NarrLV: Towards a Comprehensive Narrative-Centric Evaluation for Long Video Generation. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/237c16c6d4352c5e0e6336143a2daf7f-Abstract-Conference.html.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, 2023. doi: 10.1109/CVPR52729.2023.01457. URL https://doi.org/10.1109/CVPR52729.2023.01457.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, et al. Video-Bench: Human-Aligned Video Generation Benchmark. *arXiv preprint arXiv:2504.04907*, 2025. doi: 10.48550/arXiv.2504.04907. URL https://arxiv.org/abs/2504.04907.

Ruozhen He, Meng Wei, Ziyan Yang, and Vicente Ordonez. EntityBench: Towards entity-consistent long-range multi-shot video generation. *arXiv preprint arXiv:2605.15199*, 2026. doi: 10.48550/arXiv.2605.15199. URL https://arxiv.org/abs/2605.15199.

Jiehui Huang, Yuechen Zhang, Bin Xia, Jiahao Wang, Xu He, Zhenchao Tang, Meng Chu, Xin Tao, Pengfei Wan, and Jiaya Jia. UnityShots: Memory-Driven Multi-Shot Audio-Video Generation with Boundary-Aware Gating. *arXiv preprint arXiv:2606.21661*, 2026a. doi: 10.48550/arXiv.2606.21661. URL https://arxiv.org/abs/2606.21661.

Lingxuan Huang, Sizhe He, Hengji Zhou, Liqiang Nie, Lianghao Xia, and Chao Huang. ViMax: Agentic Video Generation. *arXiv preprint arXiv:2606.07649*, 2026b. doi: 10.48550/arXiv.2606.07649. URL https://arxiv.org/abs/2606.07649.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026. doi: 10.48550/arXiv.2606.03748. URL https://arxiv.org/abs/2606.03748.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, 2021. doi: 10.1109/ICCV48922.2021.00510. URL https://doi.org/10.1109/ICCV48922.2021.00510.

Haitian Li, Yanghao Zhou, Heyan Huang, Liangji Chen, YiMing Cheng, Xu Liu, Dian Jin, Jiajun Xu, Jingyun Liao, Tian Lan, et al. MTAVG-Bench 2.0: Diagnosing Failure Modes of Cinematic Expressiveness in Multi-Talker Audio-Video Generation. *arXiv preprint arXiv:2605.28035*, 2026a. doi: 10.48550/arXiv.2605.28035. URL https://arxiv.org/abs/2605.28035.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-Image-Bench: From Generation to Creation in Text-to-Image Evaluation. *arXiv preprint arXiv:2605.28091*, 2026b. doi: 10.48550/arXiv.2605.28091. URL https://arxiv.org/abs/2605.28091.

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. Anim-Director: A Large Multimodal Model Powered Agent for Controllable Animation Video Generation. *arXiv preprint arXiv:2408.09787*, 2024. doi: 10.48550/arXiv.2408.09787. URL https://arxiv.org/abs/2408.09787.

Susan Liang, Chao Huang, Filippos Bellos, Yolo Yunlong Tang, Qianxiang Shen, Jing Bi, Luchuan Song, Zeliang Zhang, Jason Corso, and Chenliang Xu. Omni-Judge: Can Omni-LLMs Serve as Human-Aligned Judges for Text-Conditioned Audio-Video Generation. *arXiv preprint arXiv:2602.01623*, 2026. doi: 10.48550/arXiv.2602.01623. URL https://arxiv.org/abs/2602.01623.

Zhengyang Liang, Daoan Zhang, Huichi Zhou, Rui Huang, Bobo Li, Yuechen Zhang, Shengqiong Wu, Xiaohan Wang, Jiebo Luo, Lizi Liao, et al. UniVA: Universal Video Agent towards Open-Source Next-Generation Video Generalist. *arXiv preprint arXiv:2511.08521*, 2025. doi: 10.48550/arXiv.2511.08521. URL https://arxiv.org/abs/2511.08521.

Akide Liu, Jinbo Xing, Chaojie Mao, Ye Li, Zeyu Zhang, Yefei He, Weijie Wang, Zihan Wang, Yu Liu, Gholamreza Haffari, et al. ReCA: Multi-Shot Long Video Extrapolation via Recursive Context Allocation. *arXiv preprint arXiv:2605.26525*, 2026a. doi: 10.48550/arXiv.2605.26525. URL https://arxiv.org/abs/2605.26525.

Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Yangguang Li, Weichao Chen, et al. ShotBench: Expert-Level Cinematic Understanding in Vision-Language Models. *arXiv preprint arXiv:2506.21356*, 2025. doi: 10.48550/arXiv.2506.21356. URL https://arxiv.org/abs/2506.21356.

Tengfei Liu, Yang Shi, Xuanyu Zhu, Jiafu Tang, Liu Yang, Qixun Wang, Zhuoran Zhang, Yuqi Tang, Fengxiang Wang, Yuhao Dong, et al. LongAV-Compass: Towards Unified Evaluation of Minute-Scale Audio-Visual Generation Across T2AV, I2AV, and V2AV. *arXiv preprint arXiv:2605.26244*, 2026b. doi: 10.48550/arXiv.2605.26244. URL https://arxiv.org/abs/2605.26244.

Do Xuan Long, Yale Song, Min-Yen Kan, Tomas Pfister, and Long T. Le. A^2RD: Agentic autoregressive diffusion for long video consistency. *arXiv preprint arXiv:2605.06924*, 2026. doi: 10.48550/arXiv.2605.06924. URL https://arxiv.org/abs/2605.06924.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, et al. HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives. *arXiv preprint arXiv:2510.20822*, 2025. doi: 10.48550/arXiv.2510.20822. URL https://arxiv.org/abs/2510.20822.

Yihao Meng, Zichen Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Yue Yu, Hanlin Wang, Haobo Li, Jiapeng Zhu, Yanhong Zeng, et al. CausalCine: Real-Time Autoregressive Generation for Multi-Shot Video Narratives. *arXiv preprint arXiv:2605.12496*, 2026. doi: 10.48550/arXiv.2605.12496. URL https://arxiv.org/abs/2605.12496.

Meta AI. Segment Anything 2.1. SAM 2 release notes, 2024. URL https://github.com/facebookresearch/sam2/blob/main/RELEASE_NOTES.md.

Omdia. Microdramas to Generate $11 Billion in Global Revenues in 2025 Says Omdia. MIPCOM press release, October 2025. URL https://omdia.tech.informa.com/pr/2025/oct/microdramas-to-generate-11-billion-dollars-in-global-revenues-in-2025-says-omdia. Reports that China accounts for 83 percent of 2025 global micro-drama revenue.

Omdia. Microdramas Overtake Streamers on Mobile Engagement, Says Omdia. MIP London press release, February 2026. URL https://omdia.tech.informa.com/pr/2026/feb/microdramas-overtake-streamers-on-mobile-engagement-says-omdia. Reports USD 11 billion in global micro-drama revenue in 2025, forecasts USD 14 billion for 2026, and forecasts USD 3 billion outside China.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. doi: 10.48550/arXiv.2212.04356. URL https://arxiv.org/abs/2212.04356.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. doi: 10.48550/arXiv.2408.00714. URL https://arxiv.org/abs/2408.00714.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP*, 2021. doi: 10.1109/ICASSP39728.2021.9414878. URL https://doi.org/10.1109/ICASSP39728.2021.9414878.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. AniMaker: Multi-Agent Animated Storytelling with MCTS-Driven Clip Generation. *arXiv preprint arXiv:2506.10540*, 2025. doi: 10.48550/arXiv.2506.10540. URL https://arxiv.org/abs/2506.10540.

Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. MSVBench: Towards human-level evaluation of multi-shot video generation. *arXiv preprint arXiv:2602.23969*, 2026a. doi: 10.48550/arXiv.2602.23969. URL https://arxiv.org/abs/2602.23969.

Yufei Shi, Weilong Yan, Naixuan Huang, Yucheng Chen, Chenyu Zhang, Tao He, Si Yong Yeo, and Ming Li. One Sentence, One Drama: Personalized Short-Form Drama Generation via Multi-Agent Systems. *arXiv preprint arXiv:2605.22144*, 2026b. doi: 10.48550/arXiv.2605.22144. URL https://arxiv.org/abs/2605.22144.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. doi: 10.48550/arXiv.2508.10104. URL https://arxiv.org/abs/2508.10104.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. doi: 10.48550/arXiv.2404.01292. URL https://arxiv.org/abs/2404.01292.

Yale Song, Yiwen Song, Nick Losier, Nathan Hodson, Ye Jin, Rhyard Zhu, Yan Xu, Daniel Vlasic, Carina Claassen, Jasmine Leon, et al. Co-Director: Agentic Generative Video Storytelling. *arXiv preprint arXiv:2604.24842*, 2026a. doi: 10.48550/arXiv.2604.24842. URL https://arxiv.org/abs/2604.24842.

Yiren Song, Huilin Zhong, Kevin Qinghong Lin, Haofan Wang, and Mike Zheng Shou. Soap2Soap: Long Cinematic Video Remaking via Multi-Agent Collaboration. *arXiv preprint arXiv:2605.17423*, 2026b. doi: 10.48550/arXiv.2605.17423. URL https://arxiv.org/abs/2605.17423.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 2018. doi: 10.1109/TIP.2018.2831899. URL https://doi.org/10.1109/TIP.2018.2831899.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. doi: 10.1007/978-3-030-58536-5_24. URL https://doi.org/10.1007/978-3-030-58536-5_24.

Jente Vandersanden, Matheus Gadelha, Chun-Hao P. Huang, Hyeonho Jeong, and Yulia Gryaditskaya. EM-Vid: Training-Free Entity-Centric Memory for Efficient and Consistent Multi-Shot Video Generation. *arXiv preprint arXiv:2605.23610*, 2026. doi: 10.48550/arXiv.2605.23610. URL https://arxiv.org/abs/2605.23610.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *ICASSP*, 2018. doi: 10.1109/ICASSP.2018.8462665. URL https://doi.org/10.1109/ICASSP.2018.8462665.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. MAViS: A Multi-Agent Framework for Long-Sequence Video Storytelling. In *EACL*, pp. 2273–2295, 2026a. doi: 10.18653/v1/2026.eacl-long.101. URL https://aclanthology.org/2026.eacl-long.101/.

Shengyi Wang, Niantong Li, Guangzheng Hu, Hong Qi, Fei Ding, Weixu Qiao, Jinlin Wang, Xiaotong Lv, Peng Han, Zimeng Li, et al. FilmBench: A film-grade benchmark for cinematic video generation. *arXiv preprint arXiv:2607.24241*, 2026b. doi: 10.48550/arXiv.2607.24241. URL https://arxiv.org/abs/2607.24241.

Xinran Wang, Songyu Xu, Xiangxuan Shan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Yanhua Huang, Kongming Liang, and Zhanyu Ma. CineTechBench: A Benchmark for Cinematographic Technique Understanding and Generation. *arXiv preprint arXiv:2505.15145*, 2025. doi: 10.48550/arXiv.2505.15145. URL https://arxiv.org/abs/2505.15145.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Jianhui Wei, Jie Tan, Hengchuan Zhu, Xiaotian Zhang, Yan Zhang, Ziyi Chen, Daoan Zhang, Wei Xu, and Zuozhu Liu. VideoWeaver: Evaluating and Evolving Skills for Agentic Long Video Generation. *arXiv preprint arXiv:2606.08091*, 2026. doi: 10.48550/arXiv.2606.08091. URL https://arxiv.org/abs/2606.08091.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. doi: 10.1109/ICCV51070.2023.01843. URL https://doi.org/10.1109/ICCV51070.2023.01843.

Mingqiang Wu, Weilun Feng, Zhefeng Zhang, Haotong Qin, Yuqi Li, Guoxin Fan, Xiaokun Liu, Zhulin An, Libo Huang, Yongjun Xu, et al. Echo-Forcing: A Scene Memory Framework for Interactive Long Video Generation. *arXiv preprint arXiv:2605.16003*, 2026. doi: 10.48550/arXiv.2605.16003. URL https://arxiv.org/abs/2605.16003.

Xiaoxue Wu, Xinyuan Chen, Yaohui Wang, and Yu Qiao. ShotDirector: Directorially Controllable Multi-Shot Video Generation with Cinematographic Transitions. *arXiv preprint arXiv:2512.10286*, 2025a. doi: 10.48550/arXiv.2512.10286. URL https://arxiv.org/abs/2512.10286.

Xiele Wu, Zicheng Zhang, Mingtao Chen, Yixian Liu, Yiming Liu, Shushi Wang, Zhichao Hu, Yuhong Liu, Guangtao Zhai, and Xiaohong Liu. Q-Save: Towards Scoring and Attribution for Generated Video Evaluation. *arXiv preprint arXiv:2511.18825*, 2025b. doi: 10.48550/arXiv.2511.18825. URL https://arxiv.org/abs/2511.18825.

Chuanzhi Xu, Huiqi Liang, Bang Shi, Huiming Zhang, Yifan Xiao, Guangcheng Lin, Haodong Chen, Qiang Qu, Zhicheng Lu, and Weidong Cai. DrawVideo: Generating Long Video from Storyboard Keyframe Sketches. *arXiv preprint arXiv:2605.23508*, 2026a. doi: 10.48550/arXiv.2605.23508. URL https://arxiv.org/abs/2605.23508.

Songyu Xu, Xin Wang, Qiang Chen, Xinran Wang, Muxi Diao, Yuxuan Zhang, Kongming Liang, Rui Lin, and Zhanyu Ma. VGIF-Score: Interpretable and Diagnostic Evaluation of Spatio-Temporal Instruction Following in Video Generation. *arXiv preprint arXiv:2607.13527*, 2026b. doi: 10.48550/arXiv.2607.13527. URL https://arxiv.org/abs/2607.13527.

Songlin Yang, Haobin Zhong, Ruilin Zhang, Xiaotong Zhao, Shuai Li, Kai Zheng, Xuyi Yang, Zhe Wang, Zhenchen Tang, Yang Li, et al. EvalVerse: Pipeline-Aware and Expert-Calibrated Benchmarking for Professional Cinematic Video Generation. *arXiv preprint arXiv:2605.23271*, 2026. doi: 10.48550/arXiv.2605.23271. URL https://arxiv.org/abs/2605.23271.

Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, FeiLin Han, Wei Zhai, Wei Liu, Yang Cao, and Zheng-Jun Zha. VideoGen-Eval: Agent-based System for Video Generation Evaluation. *arXiv preprint arXiv:2503.23452*, 2025. doi: 10.48550/arXiv.2503.23452. URL https://arxiv.org/abs/2503.23452.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. OpenS2V-Nexus: A Detailed Benchmark and Million-Scale Dataset for Subject-to-Video Generation. *arXiv preprint arXiv:2505.20292*, 2025. doi: 10.48550/arXiv.2505.20292. URL https://arxiv.org/abs/2505.20292.

Haojie Zhang, Di Wu, Bingyan Liu, Linjie Zhong, Yuancheng Wei, Xingsong Ye, Nanqing Liu, and Yaling Liang. MuSS: A large-scale dataset and cinematic narrative benchmark for multi-shot subject-to-video generation. *arXiv preprint arXiv:2604.23789*, 2026a. doi: 10.48550/arXiv.2604.23789. URL https://arxiv.org/abs/2604.23789.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. StoryMem: Multi-shot Long Video Storytelling with Memory. *arXiv preprint arXiv:2512.19539*, 2025. doi: 10.48550/arXiv.2512.19539. URL https://arxiv.org/abs/2512.19539.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, et al. MBench: A Comprehensive Benchmark on Memory Capability for Video World Models. *arXiv preprint arXiv:2606.00793*, 2026b. doi: 10.48550/arXiv.2606.00793. URL https://arxiv.org/abs/2606.00793.

Xiaohan Zhang, Yuqing Wen, Junlin Chen, Yuqi Tang, Yiting He, Lizhuo Shao, Weiming Zhu, Tengfei Liu, Yang Shi, Jialu Chen, et al. MultiRef-Compass: Towards Comprehensive Evaluation of Multi-Reference-to-Audio-Video Generation. *arXiv preprint arXiv:2607.14189*, 2026c. doi: 10.48550/arXiv.2607.14189. URL https://arxiv.org/abs/2607.14189.

Zhida Zhang, Jie Ma, Zhan Peng, Haoxue Wu, Yang Han, Jun Liang, Jie Cao, and Jing Li. SmartDirector: Keyframe-Conditioned Cinematic Video Generation with Narrative Pacing Control. *arXiv preprint arXiv:2605.27891*, 2026d. doi: 10.48550/arXiv.2605.27891. URL https://arxiv.org/abs/2605.27891.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. doi: 10.48550/arXiv.2503.21755. URL https://arxiv.org/abs/2503.21755.

Xiangqing Zheng, Chengyue Wu, Kehai Chen, and Min Zhang. LoCoT2V-Bench: Benchmarking Long-Form and Complex Text-to-Video Generation. In *ICML*, 2026. URL https://icml.cc/virtual/2026/poster/65294.

Hengji Zhou, Sijie Liu, Jianrun Chen, Xingchen Zou, Lianghao Xia, and Liqiang Nie. DramaDirector: Geometry-Guided Short Drama Generation. *arXiv preprint arXiv:2606.24107*, 2026a. doi: 10.48550/arXiv.2606.24107. URL https://arxiv.org/abs/2606.24107.

Jinsong Zhou, Yihua Du, Xinli Xu, Luozhou Wang, Zijie Zhuang, Yehang Zhang, Shuaibo Li, Xiaojun Hu, Bolan Su, and Ying-cong Chen. VideoMemory: Toward Consistent Video Generation via Memory Integration. *arXiv preprint arXiv:2601.03655*, 2026b. doi: 10.48550/arXiv.2601.03655. URL https://arxiv.org/abs/2601.03655.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. ViStoryBench: Comprehensive benchmark suite for story visualization. In *CVPR*, pp. 9455–9467, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhuang_ViStoryBench_Comprehensive_Benchmark_Suite_for_Story_Visualization_CVPR_2026_paper.html.

Appendix Roadmap

Section A lists the drama scripts; Sec. B details the dimensions, implementations and attribution tags. Section C presents the generation and annotation workflows, and Sec. D describes automated evaluation, evidence tools, metric validation and judge training. Section E reports stage-specific results; Sec. F analyses content-dependent performance and the reliability of automated scoring; Sec. G presents case studies.

A Drama Manifest

The script corpus comprises 20 dramas of 10 episodes each, spanning audience channels, periods/themes and genres. It includes 10 live-action and 10 animated dramas, with 16 in Chinese dialogue and 4 in English. The video subset uses the first three episodes of each drama, totalling 60 episodes. Table 4 lists the manifest; style and language analyses are reported in Secs. F.1.1 and F.1.2.

Each drama has a shared cast list and episode outlines of three to five hundred characters. The scripts follow Hongguo's shooting-script format, specifying scene numbers, locations and times, cast members, actions, camera notes and dialogue with emotional cues.

B Dimension System: Leaf-level Dimensions and Implementation

DramaChain Dimensions comprises 63 leaf dimensions across five evaluation axes and six production granularities. Figure 4 places the axes along rows and the granularities in production order along columns, with the applicable dimensions in each cell.

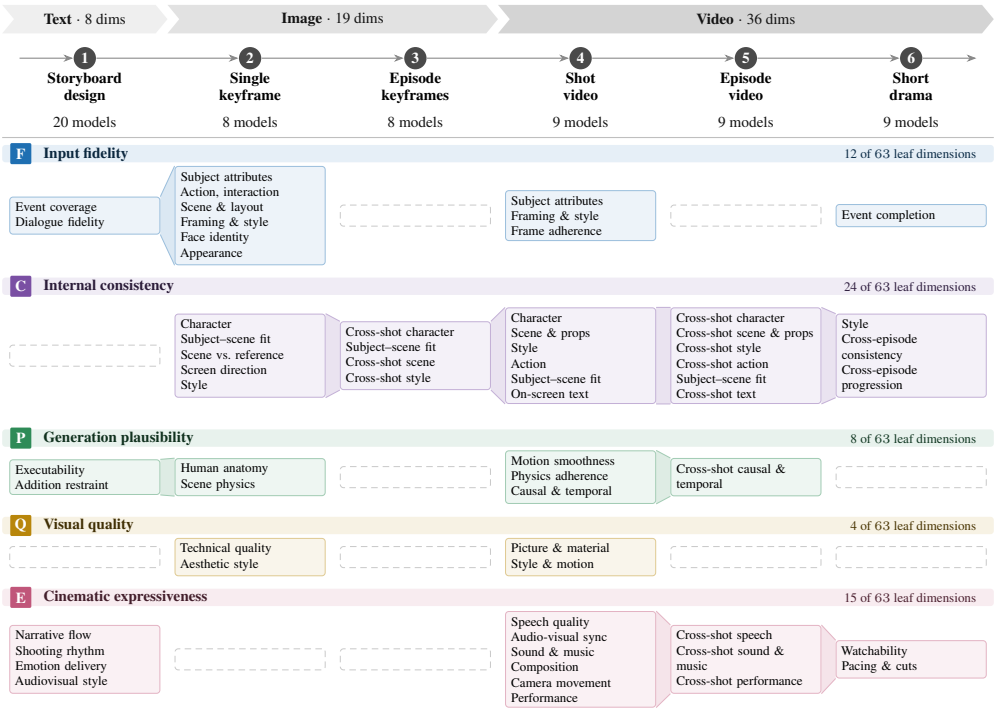


Figure 4: The overview of DramaChain Dimensions. Five evaluation axes are defined once and instantiated at the six production granularities, resolving into 63 leaf dimensions.

Each dimension has a five-level rubric with explicit decision criteria and a dimension-specific attribution-tag vocabulary, used by both annotators and the automated scorer. The vocabularies appear in Sec. B.2; the full rubric texts are omitted for length. Per-dimension scores and agreement are reported in the stage-specific tables of Sec. E.

Table 4: Drama manifest. English titles are followed by romanised originals for Chinese-language dramas. The four English-dialogue dramas use English for dialogue and on-screen text.

Channel	Period / theme	Genre	Lang.	Style	Ep.
Male-oriented					
	Contemporary urban	War god returns — *Blade Reclaimed* (Gui Ren)	zh	animated	3
	Contemporary urban	Live-in son-in-law rises — *Bladeless Son-in-Law* (Zhuixu Wufeng)	zh	live	3
	Contemporary urban	Sudden-wealth flex — *The Squander Decree* (Sancai Ling)	zh	live	3
	Historical costume	Court intrigue — *Chronicle of Buried Injustice* (Chenyuan Lu)	zh	live	3
	Historical costume	Time-slip revenge — *A Ming Dynasty Hotpot Tale* (Mingchao Huoguo Ji)	zh	animated	3
	Fantasy premise	Post-apocalyptic survival — *Permafrost Bastion* (Dongtu Baolei)	zh	animated	3
	Fantasy premise	Rebirth revenge — *No Second Crossing* (Chonglai Budu)	zh	live	3
All audiences					
	Fantasy premise	Isekai settlement — *Homesteading the Cretaceous* (Bai'e Kaihuang)	zh	animated	3
	Modern period	Campus youth — *The Year of Cicadas* (Nanian Chanming)	zh	live	3
	Modern period	Rural farming — *Granaries Full of Wheat* (Maidun Mancang)	zh	animated	3
Female-oriented					
	Urban romance	Marriage first, love later — *It Had to Be You* (Pianpian Nuan Ni)	zh	live	3
	Urban romance	Swapped heiresses — *I Don't Want It* (Bu Xihan)	zh	animated	3
	Urban romance	Dynastic tragic romance — *Moon Understudy* (Ti Yue)	zh	animated	3
	Historical romance	Costume time-slip — *The Lie-Flat Consort* (Tangying Niangniang)	zh	live	3
	Historical romance	Forced possession — *I Insist on Taking You* (Pianyao Qiang Qing)	zh	animated	3
	Female lead, no CP	Coming of age — *Over the Mountain* (Fan Shan)	zh	live	3
Overseas, male-oriented					
	Contemporary urban	Revenge rise — *The Janitor's Billion-Dollar Revenge*	en	animated	3
	Contemporary urban	Hidden billionaire — *The Bellboy Owns This Hotel*	en	live	3
Overseas, female-oriented					
	Contemporary urban	Vampire romance — *The Vampire Lord's Forbidden Blood*	en	animated	3
	Contemporary urban	Contract marriage — *The Billionaire's Two-Year Wife*	en	live	3
Total					**60**

B.1 IMPLEMENTATION OF EVERY LEAF DIMENSION

Each dimension specifies its rubric, tag vocabulary and a *method* describing the required checks. The router uses this method to select evidence modules without dimension-specific code. **Routed evidence** is collected before scoring by zero to three of 29 specialist modules, which provide small-model perception and deterministic measurements. **Model-callable tools** support iterative inspection through 15 operations: six for visual inspection (cropping, region annotation, counting and enhancement), three for temporal inspection (metadata, frame extraction and shot detection), and six for audio and text inspection (transcription, audio quality and events, audio extraction, subtitles and OCR; Tab. 7).

The implementation table identifies the tool set available to each dimension. For images, `zoom` provides cropping and `detail` adds boxes; `count`/`scene` add counting, `fine` adds sharpening and super-resolution, and `lowlight` adds contrast enhancement. Video sets start with metadata and frame extraction: `v_frame` adds cropping and boxes; `v_count`, `v_text` and `v_cut` add counting, OCR and shot detection, respectively. `v_sync`, `v_audio` and `v_speech` support task-specific audio inspection. Large-model sub-agents may use these sets over multiple rounds. Pixel-altering operations are excluded from image-quality assessment, and `cut_detect` supplies shot-boundary evidence with the base context rather than through per-dimension routing.

Table 5 lists the checks and evidence routes under the default configuration.

Table 5: Implementation and evidence routes for all 63 dimensions under the default configuration. **Implementation**: check method; **routed evidence**: modules available at the given granularity; **model-callable**: tool set (tool count). "—": none. Model-callable operations are listed in Tab. 7; module outputs are filtered as described in Sec. D.3.

Code	Dimension	Axis	Implementation	Routed evidence	Model-callable
① Storyboard design					
S-A1	Event coverage	F	master agent extracts beats → sub-agent aligns per shot for coverage	`event_align`	—
S-A2	Dialogue fidelity	F	dialogue text alignment (edit distance + embedding similarity)	`dialogue_align`	—
S-B1	Description executability	P	VLM: phrasing clarity / parsable content / spatial-physical check	—	—
S-B2	Restraint in additions	P	LLM per shot: surplus detail + transition elements + key gaps; overstep or consistency harm	`supplement_audit`	—
S-C1	Narrative flow	E	LLM: cross-scene hard cuts + sequence breaks + shot order + OS misplacement; seven-form banding	`narrative_flow_audit`	—
S-C2	Shooting rhythm	E	action-verb count and duration spread per shot + VLM feasibility	`shot_stats`	—
S-C3	Emotional expression	E	LLM per beat: eight audiovisual channels + director calibration	`emotion_beats`	—
S-C4	Audiovisual style treatment	E	shot-language classes + genre lexicon + director calibration	—	—
② Single keyframe					
I-A1	Technical quality	Q	zoom to check three defect classes; sharpness dominant	—	`zoom(1)`

Table 5 (continued)

Code	Dimension	Axis	Implementation	Routed evidence	Model-callable
I-A2	Aesthetic style	Q	CSD (Somepalli et al., 2024) style purity + genre lexicon; colour harmony, cheap AI look	style_csd	—
I-B1	Subject attributes	F	YOLO26 (Jocher et al., 2026) detection + SAM 2 (Ravi et al., 2024; Meta AI, 2024) segmentation + zoom, 0/1/2 per facet	sam_count	count(3)
I-B2	Action and interaction	F	per-facet scoring	—	detail(2)
I-B3	Scene and spatial layout	F	scene-type check + programmatic bearing match	scene_consistency	count(3)
I-B4	Shot framing and visual style	F	shot-attribute classes + composition, lighting + CSD style	style_csd framing_check	—
I-C1	Face identity	F	ArcFace (Deng et al., 2019) for live action; anime-ID model or DINOv3 feature similarity + CIDS (Siméoni et al., 2025; Zhuang et al., 2026), selected by this audit	identity_sim	fine(4)
I-C2	Appearance and costume	F	zoom to compare against the character sheet; accessories need zoom	—	fine(4)
I-D1	Character consistency across panels	C	pairwise CIDS / DINOv3 feature similarity, aggregated	identity_sim	fine(4)
I-D2	Subject–scene integration	C	CPR (Zhuang et al., 2026) / NaturalScore (Yuan et al., 2025) + light and perspective blend; adversarial I-D1	copypaste_var	detail(2)
I-D3	Scene and layout vs. reference	C	cross-image scene-bank embedding + spatial-structure recheck	scene_consistency	scene(3)
I-D4	Screen-direction axis	C	per-shot subject facing and motion emitted + rule-based axis check	axis_check	detail(2)
I-D5	Style consistency	C	cross-image CSD style embedding + grid and style-clash checks	style_csd	detail(2)
I-E1	Human anatomy plausibility	P	body and hand keypoints + zoom for deformity + pose naturalness	body_structure	count(3)
I-E2	Environmental physics	P	per-item detection: gravity / collision / perspective / lighting	—	detail(2)
③ Episode keyframes					
MI-D1	Cross-shot character consistency	C	VLM compares character identity across keyframes	—	fine(4)
MI-D2	Subject–scene integration	C	alert on excess cross-shot framing similarity; adversarial to MI-D1	framing_check copypaste_var	detail(2)
MI-D3	Cross-shot scene consistency	C	cross-shot scene-bank embedding + zoom on set dressing per item	scene_consistency	lowlight(4)

Table 5 (continued)

Code	Dimension	Axis	Implementation	Routed evidence	Model-callable
MI-D5	Cross-shot style consistency	C	cross-shot CSD style embedding + style-break zoom; tone by scene	style_csd	detail(2)
④ Shot video					
V-A1	Image quality and material	Q	DOVER (Wu et al., 2023) / MUSIQ (Ke et al., 2021) perceptual quality + Laplacian sharpness + material fit to style	quality_metric	v_frame(4)
V-A2	Aesthetic style and motion texture	Q	aesthetic regression + style vs. genre; optical-flow motion texture	quality_metric aesthetic_metric style_csd	—
V-B1	Subject attribute accuracy	F	frame sampling + SAM + detection counts; satisfaction rate	sam_count	v_count(5)
V-B2	Shot framing and visual style	F	shot-attribute classes + composition and lighting + cut count	framing_check	v_frame(4)
V-B3	Keyframe adherence	F	SSIM (Wang et al., 2004) / DINO vs. given frame + subject-box match	frame_follow	v_frame(4)
V-C1	Character appearance consistency	C	per-frame consistency, flicker, gap-decay, costume; fidelity-gated	identity_sim	v_frame(4)
V-C2	Scene and prop consistency	C	scene-consistency anchor + prop embedding or cross-shot match	scene_consistency	v_frame(4)
V-C3	Style consistency	C	CSD style embedding per frame and cross-shot + style-break check	style_csd	v_frame(4)
V-C4	Action consistency	C	action recognition and cross-shot pose match + action continuity	body_structure	v_frame(4)
V-C5	Subject–scene integration	C	ACP-Var (Zhang et al., 2026a) / CPR + light and perspective blend; adversarial V-C1	copypaste_var	v_frame(4)
V-C6	On-screen text stability	C	per-frame OCR of text regions + adjacent-frame edit distance	ocr_track	v_text(5)
V-D1	Motion smoothness	P	inter-frame RAFT optical flow (Teed & Deng, 2020), less jitter	quality_metric	v_frame(4)
V-D3	Physical plausibility	P	VideoPhy-AutoEval (Bansal et al., 2025) physical commonsense, then plausibility on the footage	videophy	v_frame(4)
V-D4	Causal and temporal plausibility	P	sub-agent per time window → master agent: causal order, completion	—	v_frame(4)

Table 5 (continued)

Code	Dimension	Axis	Implementation	Routed evidence	Model-callable
V-E1	Speech quality	E	DNSMOS (Reddy et al., 2021) quality + Whisper (Radford et al., 2022) ASR WER / CER on lines + GE2E voiceprint (Wan et al., 2018) timbre	dialogue_align asr_wer	v_speech(7)
V-E2	Audio-visual sync	E	SyncNet (Chung & Zisserman, 2016) lip-sync + ImageBind (Girdhar et al., 2023) audio-visual alignment	syncnet imagebind_av	v_sync(5)
V-E3	Sound effects and music	E	fit of music and sound effects to scene emotion and on-screen action	—	v_audio(4)
V-F1	Composition	E	shot size / depth / blocking + subject-box head-crop check	framing_check	—
V-F2	Camera movement plausibility	E	camera-move intent, position emotion, momentum + cut density	—	—
V-F3	Character performance	E	acting and expression naturalness	—	—
⑤ Episode video					
MV-C1	Cross-shot character consistency	C	VLM compares character identity and costume across shots	—	v_frame(4)
MV-C2	Cross-shot scene and props	C	cross-shot scene-bank embedding + prop-state embedding or match	scene_consistency	v_frame(4)
MV-C3	Cross-shot style consistency	C	cross-shot CSD style embedding + style-break check	style_csd	v_frame(4)
MV-C4	Cross-shot action continuity	C	cross-shot action and pose match + facing/position, camera, axis check	axis_check body_structure	v_frame(4)
MV-C5	Subject–scene integration	C	alert on excess cross-shot framing similarity; adversarial to MV-C1	framing_check copypaste_var	v_frame(4)
MV-C6	Cross-shot text stability	C	OCR per shot's text region; same element compared across shots	ocr_track	v_text(5)
MV-D4	Cross-shot causal/temporal plausibility	P	sub-agent in shot order → master agent judges causality + closure	—	v_frame(4)
MV-E1	Cross-shot speech quality	E	cross-shot voiceprint + audio-visual alignment + ASR continuity	imagebind_av asr_wer	v_speech(7)
MV-E3	Cross-shot sound and music	E	cross-shot coherence and style consistency of music and sound effects	—	v_audio(4)
MV-F3	Cross-shot performance continuity	E	per-segment expression, emotion strength + transitions	—	—

Table 5 (continued)

Code	Dimension	Axis	Implementation	Routed evidence	Model-callable
⑥ Short drama					
O-A1	Event completion	F	key-event list from the script → per-event completion on the cut	—	v_frame(4)
O-B1	Style	C	CSD whole-episode style embedding; tone grouped by scene	style_csd	v_frame(4)
O-B2	Cross-episode consistency	C	VLM checks character and scene consistency across episodes, with scene-bank evidence	scene_consistency	v_frame(4)
O-B3	Cross-episode progression	C	against script beats: motivated progression vs. arbitrary change	—	—
O-C1	Watchability of the cut	E	calibrated whole-episode score	—	—
O-C2	Pacing and transitions	E	human-cut distribution gap + transition naturalness; cut rhythm	—	v_cut(5)

B.2 THE ATTRIBUTION-TAG VOCABULARY

Human and automated scores below 5 require at least one dimension-specific tag and a textual explanation. Table 6 lists the shared vocabularies for all 63 leaf dimensions. The full annotation collection contains 255,925 human attributions (Sec. 3.3).

Table 6: Dimension-specific attribution tags across six production granularities, shared by human and automated scoring. Each dimension also offers *other*, omitted here.

Code	Dimension	Attribution tags
① Storyboard design		
S-A1	Event coverage	key event omitted · minor beat omitted · action or voiceover omitted · events do not join up · surplus plot invented · plot left vague
S-A2	Dialogue fidelity	key line omitted · line rewritten · wrong speaker · wrong emotional register
S-B1	Description executability	ambiguous description · viewpoint unclear · blocking unclear · several actions in one shot · shot carries too little · verbose padding · blocking not physically possible · blocking too complex
S-B2	Restraint in additions	key detail not supplied · transition not supplied · surplus detail added · implausible content invented
S-C1	Narrative flow	abrupt transition · jarring join · order reversed within a shot · misplaced voiceover · mixed shot sizes in one shot · setup gives away the payoff · redundant narration · self-contradiction
S-C2	Shooting rhythm	too much action in one shot · over-fragmented cutting · key scene under-covered · transitional scene drags · no reaction shot · monotonous camera work
S-C3	Emotional expression	flat, no tension · emotion not externalised · conflict poorly staged
S-C4	Audiovisual style treatment	style mismatched to genre · inconsistent style · key scene not dramatised · wrong shot grammar
② Single keyframe		
I-A1	Technical quality	blurred · visible noise · artefacts and colour blocks · ghosting · watermark or garbled glyphs · structural distortion
I-A2	Aesthetic style	aesthetic mismatched to genre · cluttered frame · unbalanced composition · colour and lighting clash · mixed art styles · cheap AI look

Table 6 (continued)

Code	Dimension	Attribution tags
I−B1	Subject attributes	subject missing · subject feature wrong or missing · wrong subject colour · wrong subject material · wrong subject count · extra subject · wrong subject position · left/right mirrored · wrong expression
I−B2	Action and interaction	action not performed · wrong action · action mirrored · wrong interaction target · stiff pose
I−B3	Scene and spatial layout	wrong scene type · wrong bearing · wrong occlusion · wrong relative position · scene element missing
I−B4	Shot framing and visual style	wrong shot size or viewpoint · wrong composition · wrong lighting mood · wrong art style
I−C1	Face identity	features do not match · wrong face shape · wrong expression · distinguishing feature lost · identity drift · face breakdown
I−C2	Appearance and costume	wrong hairstyle · wrong hair colour · wrong garment style · wrong garment colour · accessory missing or wrong · wrong to station or period
I−D1	Character consistency across panels	face differs across panels · costume differs across panels · accessories differ across panels · identity drift · face breakdown
I−D2	Subject–scene integration	cut-out, pasted-on look · subject and scene lighting mismatch · floating, not grounded · pose or expression repeated · panels copy-pasted
I−D3	Scene and layout vs. reference	spatial structure changed · object bearing changed · light direction inconsistent · prop placement changed
I−D4	Screen-direction axis	characters swapped left/right · position drift · excessive displacement across panels · eyelines inconsistent · motion direction reversed
I−D5	Style consistency	art style differs across panels · colour grade differs within a scene · animation and live action mixed
I−E1	Human anatomy plausibility	extra or missing fingers · limbs twisted or fused · limb clipping · limb missing · facial proportions off · impossible pose · stiff, unnatural pose
I−E2	Environmental physics	object floating · clipping through geometry · perspective or scale distorted · lighting or reflection contradictory
③ Episode keyframes		
MI−D1	Cross-shot character consistency	face differs across shots · costume differs across shots · accessories differ across shots · identity drift · face swap across shots
MI−D2	Subject–scene integration	cut-out, pasted-on look · subject and scene lighting mismatch · shots look alike · composition and pose unchanged
MI−D3	Cross-shot scene consistency	layout change · furnishing change · lighting or colour shift · disordered scene cuts
MI−D5	Cross-shot style consistency	art style jumps · colour grade differs within a scene · lighting differs within a scene · inconsistent texture
④ Shot video		
V−A1	Image quality and material	blurred, low resolution · noise and artefacts · material distorted · waxy, plastic skin
V−A2	Aesthetic style and motion texture	poor composition and colour · art style mismatched to genre · cheap AI look · cheap, coarse motion
V−B1	Subject attribute accuracy	subject missing · appearance differs from the reference · wrong subject colour · wrong subject material · wrong subject count · extra subject
V−B2	Shot framing and visual style	wrong shot size or viewpoint · wrong composition · camera move does not follow the instruction · wrong lighting mood · wrong art style
V−B3	Keyframe adherence	deviates from the given keyframe · drifts from the reference during motion
V−C1	Character appearance consistency	subject flicker · slight appearance change · face breakdown or swap · garment colour change · garment style change
V−C2	Scene and prop consistency	background layout change · furnishing change · colour grade change · prop state jumps
V−C3	Style consistency	art style jumps · colour grade jumps
V−C4	Action consistency	action jumps at the join · stiff action join · action contradicts itself
V−C5	Subject–scene integration	cut-out, pasted-on look · subject and scene lighting mismatch · floating, not grounded · pose repeated and static · frames duplicated
V−C6	On-screen text stability	garbled glyphs · glyph deformation · text flicker
V−D1	Motion smoothness	stutter and dropped frames · teleporting · jitter

Table 6 (continued)

Code	Dimension	Attribution tags
V-D3	Physical plausibility	weightless floating · clipping · unconvincing collision · unnatural inertia · appears or vanishes from nowhere
V-D4	Causal and temporal plausibility	effect precedes cause · events out of order · cause and effect do not match · action not performed · action wrong or incomplete
V-E1	Speech quality	robotic, vocoded · noise and clipping · muffled · unstable loudness · line misread or words dropped · lines out of order · wrong speaker · unintelligible · timbre drift
V-E2	Audio-visual sync	lips out of sync · mouth moves, no sound · sound effect misaligned
V-E3	Sound effects and music	music missing · music does not fit the emotion · sound effect missing · sound effect does not fit the picture
V-F1	Composition	wrong shot size · subject not salient · flat, no depth · foreground and background confused
V-F2	Camera movement plausibility	almost no camera movement · aimless drifting camera · gratuitous camera moves
V-F3	Character performance	blank, monotonous expression · wrong eyeline · overacted, out of character
⑤ Episode video		
MV-C1	Cross-shot character consistency	face differs across shots · face breakdown or swap · garment colour change · garment style change · identity drift
MV-C2	Cross-shot scene and props	background layout change · furnishing change · colour grade change · prop state jumps
MV-C3	Cross-shot style consistency	art style jumps to another medium · colour grade differs within a scene · animation and live action mixed
MV-C4	Cross-shot action continuity	action discontinuity at the join · action contradicts itself · axis crossed, left/right swapped · eyelines inconsistent · motion direction reversed · abrupt camera-position change
MV-C5	Subject–scene integration	cut-out, pasted-on look · subject and scene lighting mismatch · shots look alike · camera and composition unchanged
MV-C6	Cross-shot text stability	text changes across shots · garbled glyphs · glyph deformation
MV-D4	Cross-shot causal/temporal plausibility	effect precedes cause across shots · events out of order across shots · cause and effect do not match across shots · action never completed across shots
MV-E1	Cross-shot speech quality	timbre drift across shots · unstable loudness across shots · audio quality differs across shots · dialogue broken across shots
MV-E3	Cross-shot sound and music	music drops out across shots · abrupt music change · music style jumps · sound effects differ across shots
MV-F3	Cross-shot performance continuity	emotion breaks between shots · emotional intensity jumps · performance style shifts · out of character across shots
⑥ Short drama		
O-A1	Event completion	key event missing · event not played out · main line broken · characters misidentified · invented, departs from the script
O-B1	Style	style break between episodes · colour grade differs within a scene · inconsistent art style
O-B2	Cross-episode consistency	character changes across episodes · signature feature inconsistent · scene changes across episodes · face swap across episodes
O-B3	Cross-episode progression	character does not develop · costume does not change · appearance or scene changed arbitrarily · change unmotivated
O-C1	Watchability of the cut	cut incomplete · looks cheap · too many hard defects to watch
O-C2	Pacing and transitions	key scene rushed · transitional scene drags · uneven pacing · stiff transition

C DramaChain Agent and Labeling System

DramaChain Agent generates the benchmark items, and DramaChain Labeling System collects and reviews their human assessments. Their workflows extend the overview in Sec. 3; DramaChain Agentic Judge is described separately in Sec. D.

Figure 5 shows the generation workflow, while Fig. 6 illustrates its outputs and their coverage of dramas, styles, languages and input paradigms (Sec. 3.2).

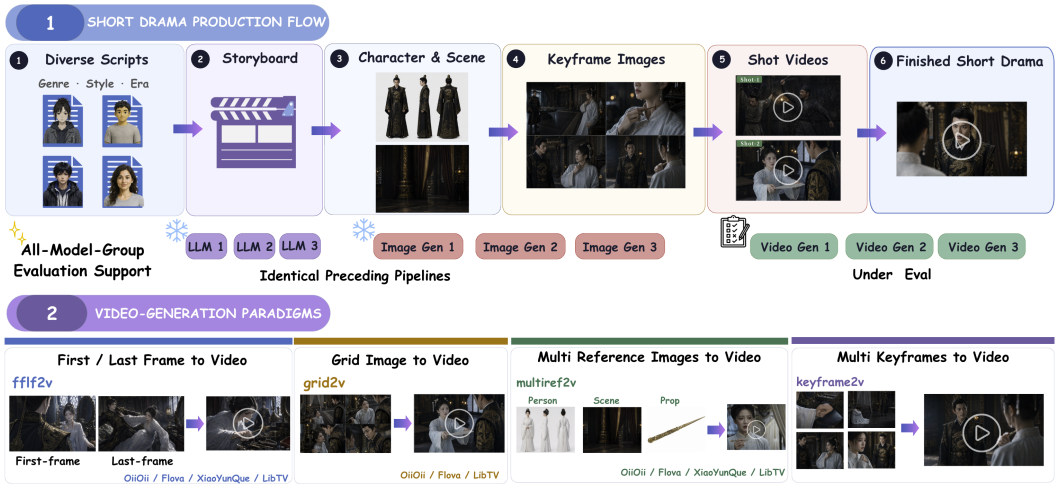


Figure 5: DramaChain Agent workflow. Branching occurs only at the evaluated stage, giving competing models identical upstream artefacts. Completed dramas are compared with outputs from commercial short-drama platforms.



live action · zh
court intrigue

live action · zh
urban romance

live action · zh
campus youth

live action · **en**
contract marriage

animated · zh
dynastic romance

animated · zh
swapped heiresses

animated · zh
rural farming

animated · zh
forced possession

Coverage	Split	Coverage	Split
Channel	male / all-audience / female	Visual style	live action 10 : animated 10
Period	contemporary, historical, modern, fantasy	Dialogue language	Chinese 16 : English 4, scene descriptions all Chinese
Genre	20 distinct sub-genres, one per drama	Input paradigm	`fflf2v`, `grid2v`, `multiref2v`

Figure 6: Example benchmark items and corpus composition. The full manifest is given in Sec. A.

Figure 7 compares outputs from DramaChain Agent and three commercial short-drama platforms using the same script of three scenes and eight shots. Each system runs fully automatically in a single pass (Sec. 3.2). This example illustrates a practical production setting, not a controlled platform comparison. Figure 8 then shows one drama's three episodes, sampled at equal intervals to illustrate the continuity required beyond individual shots.

The generated items are assessed through DramaChain Labeling System. Figure 9 shows its interface, the evidence required for each deduction, and the review process before annotations enter the qualified corpus (Sec. 3.3).

Quality-control rules. Four record-level checks detect **A1** unusually short completion times relative to the median for the annotation type and granularity; **A2** a dimension score of 2 or below without any attribution marker in the record; **A3** more than half the dimensions marked N/A; and **A4** rejection by manual review.

Five rater-level checks detect **B1** excessive concentration on a single score; **B2** too few distinct scores across submissions; **B3** repeated identical scores across leaf dimensions within items; **B4**

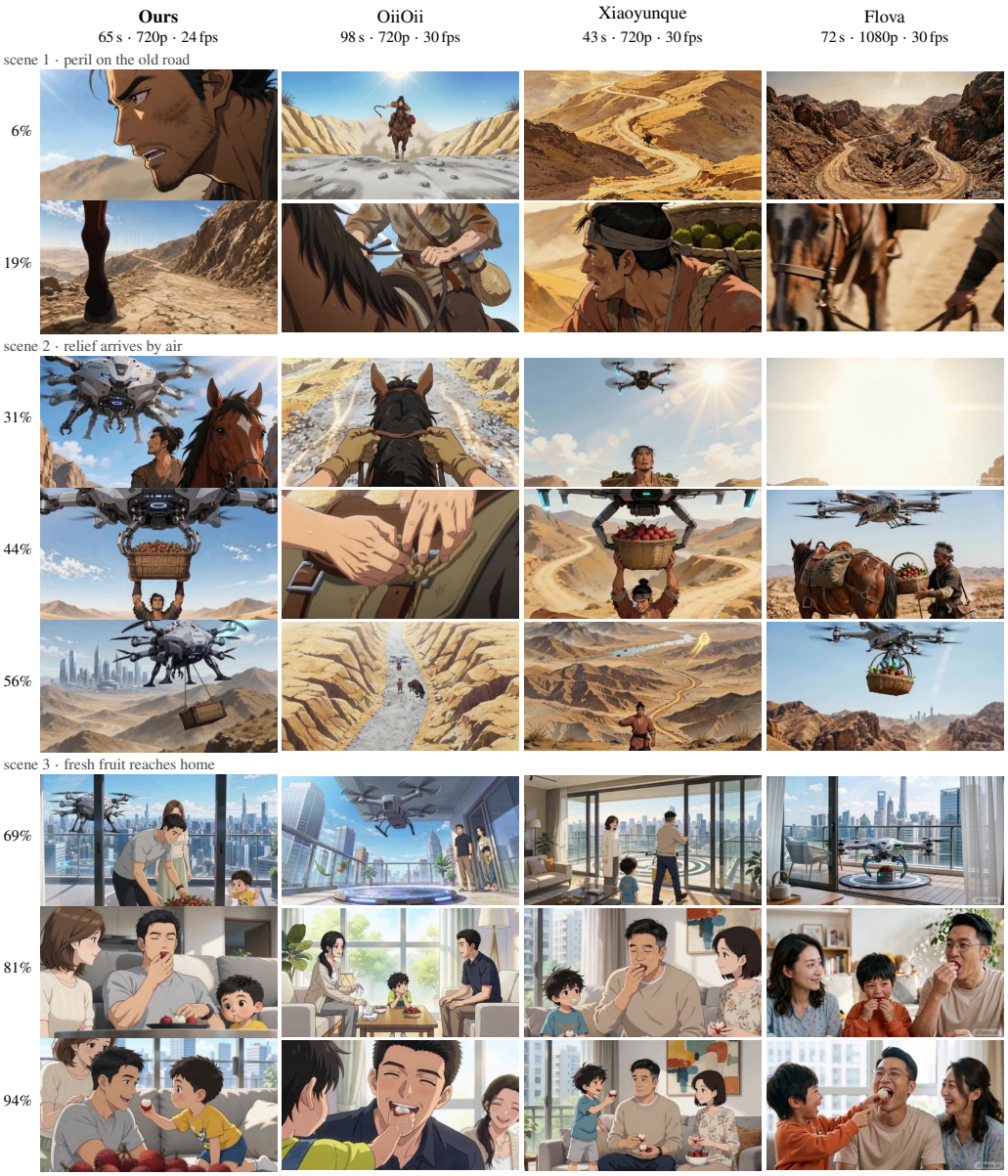


Figure 7: DramaChain Agent (**Ours**) and three commercial platforms process the same script of three scenes and eight shots, fully automatically in a single pass without regeneration or manual editing. Each system determines its own cut (43–98 s), so rows align relative progress rather than timestamps. Expert review rated all four in the same band for narrative pacing, overall consistency and cinematic expressiveness, with differences in stylistic orientation. This single-script example under default settings illustrates a practical production setting, not a controlled comparison, general platform equivalence or ranking.

EP1 the contract expires, the stand-in exits 89.8 s



EP2 “what you stand in for is me” 92.8 s



EP3 the truth about to surface, and she is gone 116.0 s



Figure 8: DramaChain Agent’s output for *Moon Understudy* (Ti Yue), an animated, Chinese-dialogue urban romance for a female audience (Tab. 4). In a three-year contract marriage, the heroine believes she is a substitute for a tycoon’s first love, who moved abroad: her room and routines are modelled after that woman. When the contract expires, she discovers that the photograph he has kept for three years depicts her. She returns the keys before learning why his first love left. Video generation uses each drama’s first three episodes; each block shows 20 evenly sampled frames in reading order.

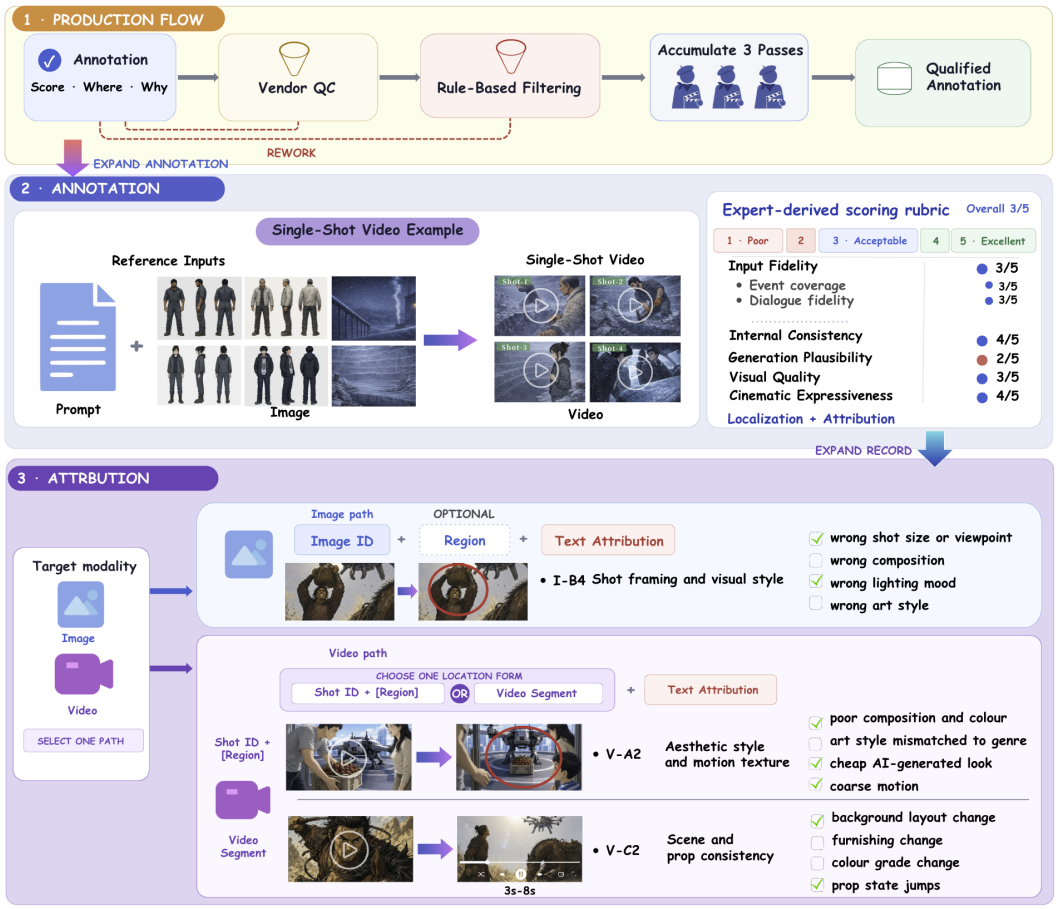


Figure 9: DramaChain Labeling System workflow. Each item is assigned three independent assessments using the expert-derived rubric. Scores below full marks require a dimension-specific tag and explanation, with spatial or temporal localisation where applicable. Annotations undergo vendor quality control and rule-based filtering before entering the qualified corpus.

low correlation with leave-one-out consensus (the other raters' scores); and **B5** extreme leniency or severity relative to that consensus.

When filtering is enabled, record-level failures exclude individual assessments, while rater-level failures exclude the rater's assessments within the relevant group. A separate item-level rule requires a configured minimum number of remaining raters; it is not one of the nine checks. Consensus means are recomputed from retained assessments.

D AGENTIC EVALUATION SYSTEM

DramaChain Agentic Judge applies the dimensions in Sec. B through a shared scoring prompt, routed evidence and model-callable tools. The following sections describe this process, metric validation, scorer training and the conventions used to report results.

Figure 10 illustrates a recorded evaluation, including the selected evidence, tool calls and per-dimension judgements. The criteria at each granularity are summarised in Fig. 2.

D.1 THE SCORING PROMPT

The master agent first determines which of the 63 dimensions apply to the item's granularity, input paradigm and available references, then dispatches sub-agents for the selected dimensions. The sub-agents share a scoring template with dimension-specific criteria and evidence in fixed slots. The excerpt below shows its scoring fields after dimension selection.

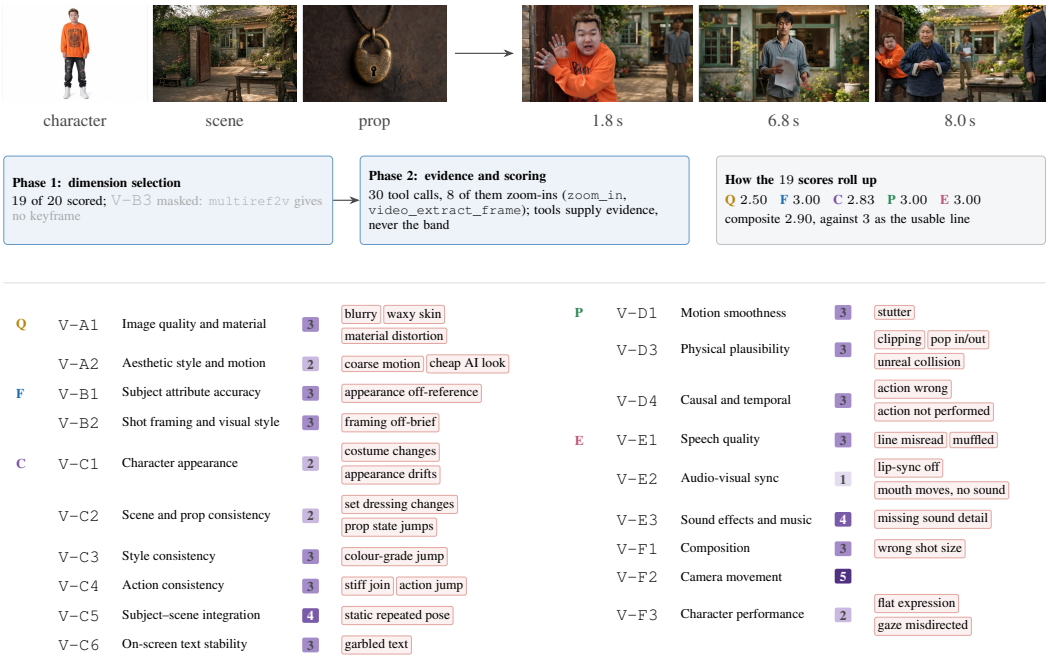


Figure 10: One real automated-evaluation record.

```
You are a short-drama generation quality sub-agent.  Score
the single dimension [{code} {name}] (axis {axis}, tier
{tier}) on the five-band scale, and nothing else.
[Five band anchors]:  {rubric[5..1]}
[Per-shot checkpoints and banding rules]:  {guide, optional
-- carries the checklist}
[Attribution tags] (at least one required when the score is
below 5):  {tags}
[Automated method hint]:  {method}
[Item] stage={...} scope={...} paradigm={...}
[Source script (ground-truth reference, not the artefact
under evaluation)]:  {script, conditional}
[Objective tool evidence]:  {evidence, one line per
measurement}
[Hard scoring requirements (apply to every dimension)]:
{constant}
Output strictly JSON only:  {"score", "confidence", "tags",
"reason"}
```

Four shared requirements constrain scoring. The first two are quoted verbatim below. The first prohibits unsupported heuristics that artificially reshape score distributions: *Never score using unevidenced, formulaic hedges such as "by the score-compression principle", "to be conservative", "full marks are not usually given" or "good overall but still docking a point", and never pull a score towards the middle or the extremes artificially. The score is determined only by real defects in the artefact and the evidence; make no distributional adjustment.* The second calibrates deductions to overall severity rather than adding a penalty for every defect: *The size of the deduction must match the defect's actual effect on the overall impression… With several blemishes, band by their combined effect on that impression rather than stacking a deduction for each one and ending up too low.* The remaining requirements link scores to evidence. Scores below 5 must identify a specific, verifiable defect and select at least one dimension-specific attribution tag. For image dimensions with a checklist, the explanation must begin with its conclusions in checklist order rather than replace them with a general summary.

Multi-round scoring. Dimensions with model-callable tools are evaluated over multiple rounds. In rounds 0 to $n-1$, the model may invoke the available tools; results are appended to the conversation,

and cropped images become media inputs for the next round. Tools are removed in the final round to request a judgement. Repeating an identical call more than twice returns an error. If the run ends without valid JSON, it receives one additional tool-free round before falling back to single-pass scoring.

D.2 THE EVIDENCE TOOL CHAIN

Table 7 lists the 15 model-callable tools by function. Specialist evidence modules and their dimension-specific routes are listed separately in Tab. 5.

Table 7: Model-callable inspection tools, grouped by function. Tools return evidence rather than final rubric scores.

Family	Tools
Vision	`zoom_in` (crop and magnify), `draw_bbox` (annotate a region), `count_marker` (numbered point counting), and three pixel-altering ones withheld from `pixel_sensitive` dimensions (`sharpen`, `super_resolution`, `enhance_contrast`). Grid panels are sliced into sub-panels before consistency is assessed.
Temporal	`extract_frame` (sample by timestamp; extracted frames can be zoomed again), `detect_scenes` (shot segmentation), `video_probe` (container metadata).
Audio/text	`asr_segment` (transcription), `dnsmos` (no-reference audio quality), `detect_audio_events`, `extract_subtitles` (burned-in subtitle track), `extract_clip_audio`, `ocr_frame`.

Cost and reliability. Median wall-clock time is 14.6 minutes per item. Concurrent execution makes elapsed time shorter than cumulative tool time, whose median is 198.7 seconds *per dimension*. Each job uses a median of 20 sub-agent calls and about 19 measurement calls. Of the eight primary evidence modules, only `ocr_track` has availability below 98.9%, at 71.1%. Tool execution accounts for nearly all runtime, so additional reasoning rounds add little overhead compared with invoking another tool.

D.3 VALIDATING CANDIDATE METRICS

We evaluated candidate metrics against human consensus before using them as scoring anchors. The audit used collected annotations at each granularity, testing applicable evidence modules against all relevant dimensions. SRCC and p were computed both within each generation paradigm and on pooled data. Each scalar output was tested separately.

Validation results. Most candidates fail the scoring-anchor threshold of $|\mathrm{SRCC}| \geq 0.3$ over at least 100 items. Eight representative quality-scoring results illustrate this limitation. For general-purpose metrics, SRCC is -0.152 for NIMA aesthetics (Talebi & Milanfar, 2018), -0.144 for Laplacian sharpness, -0.067 for TOPIQ-NR (Chen et al., 2024) and -0.057 for MUSIQ (Ke et al., 2021). For AIGC-specific models, it is 0.004 for Q-Judger (Li et al., 2026b), 0.030 for ArtiMuse (Cao et al., 2025b), and 0.190 and 0.110 for UniPercept's image-quality and aesthetic heads (Cao et al., 2025a).

Genre-dependent quality criteria. These shortfalls suggest that image quality in short dramas depends on genre and art direction, not only on individual frames. High saturation may be deliberate in a xianxia power fantasy but appear artificial in another genre. This contextual dependence may explain errors in both metric groups: general-purpose metrics are calibrated on natural photographs, while AIGC-specific models assess visual appeal without necessarily accounting for the script's requirements.

Using metrics as evidence. Table 5 lists evidence routes, not a complete inventory of outputs shown to the judge. Outputs are filtered by granularity and criterion; routing a module does not mean treating all its measurements as scoring anchors. Retained measurements support rubric-guided judgement alongside the artefact and its context, rather than replacing it.

D.4 Training with human annotations

Training follows the SFT–GRPO procedure in Sec. 4.4. Each training example is a prompt for an item, evaluation axis and output form, with score and attribution supervision from human annotations. Outputs retain DramaChain Agentic Judge's axis-organised structure, including leaf-level scores, brief rationales and a composite score. The test items in Sec. 4.5 are excluded from both training phases.

The keyframe and storyboard scorer omits the audio pathway. Attribution rewards use the human annotations available for each item; missing annotations are not treated as failures. Storyboard training omits the attribution reward.

With greedy decoding, inference takes 6–9 s without reasoning and 13–27 s with reasoning per item–axis–output-form prompt. These timings describe individual prompts, not a full evaluation across all applicable axes. They are therefore reported separately from DramaChain Agentic Judge's per-item wall-clock time in Sec. D.

D.5 Coverage gaps

Model names follow vendor naming; the suffixes `-xhigh`, `-max` and `-high` indicate reasoning-effort settings rather than model variants. Coverage is incomplete and non-random because of provider-side content moderation and generation failures. `seedream-5.0-lite` covers 10 of the 20 dramas; `veo-3.1` and `wan2.7` cover only the first/last-frame paradigm. `seedance-2.0` covers only animation, and `seedance-2.5` only animated, multi-reference generation. Storyboard coverage is nearly complete: all nine models cover the corpus, except for one to three missing units in three models. Scores use each model's available items rather than the full corpus; incomplete coverage is marked with ‡ in Tab. 2.

Duration budgets affect comparability. `minimax-h3` uses the standard 15 seconds and reuses the other models' storyboards and first/last frames, preserving shot identity. `wan-3.0` uses 30 seconds with replanned storyboards, so its units match by drama, episode, paradigm and shot index rather than by identical shot content. The same convention applies to `veo-3.1` at 10 seconds; the difference is in duration budgeting, not the sampling rule.

D.6 Absolute-score accuracy and stage bias

Table 8 reports absolute score accuracy. MAE measures error magnitude without direction, whereas signed bias shows that the automated scorer is more lenient than human consensus at five stages and stricter at one. Bias is $+0.28$ for storyboards, $+0.41$ for single keyframes, $+0.57$ for episode keyframes, -0.35 for shot video, $+0.14$ for episode video and $+0.42$ for short dramas.

Table 8: Absolute accuracy per stage. Scoring configurations are as in Sec. 4.1; the reference is the three-annotator consensus; item counts are pairs scored on both sides.

Stage	MAE points, lower is better	within ±0.5	within ±1.0	paired items
① Storyboard design	**0.416**	**67.5%**	**95.3%**	533
② Single keyframe	0.529	54.2%	87.6%	2,125
③ Episode keyframes	0.733	38.2%	75.1%	429
④ Shot video	0.436	62.5%	94.6%	1,149
⑤ Episode video	0.433	66.0%	92.4%	238
⑥ Short drama	0.425	66.7%	92.0%	201

D.7 Scope exclusions

Table 9 lists the exclusions and their rationale, expanding on Sec. 6.

Table 9: What the benchmark does not cover, and why.

	Not covered	Why, and what follows
Stages	Post-production: music and sound design, editing and colour grading, subtitling and thumbnails.	No unified automated approach to these stages exists yet.
Dimensions	All commercial outcome measures: completion, retention, conversion.	These outcomes are hard to quantify offline.
Content	Hybrid, documentary and other non-narrative formats.	The axes assume a shot script to be faithful to.
Systems	Text-to-video without an intermediate keyframe.	The chain always produces keyframes, so a model that skips them cannot enter as it stands.
Judgment	Broad comparisons across judging-model families.	The reported ablations cover two stages, not a comprehensive comparison of judging models.

E Results by Production Stage

This appendix expands the per-stage analysis in Sec. 4.2, organised by the three *modalities* across all six granularities: storyboard design covers ①, the text stage; keyframe images cover ② and ③; and video covers ④ to ⑥. Each modality-specific section examines model performance and automated evaluation, with overall scores in Tab. 2. A final subsection considers the unit of judgement.

E.1 Storyboard design

Nine text-based large language models convert episode scripts into storyboard scripts, yielding 533 evaluation items. Their automated scores range from 3.26 to 4.23 (Tab. 2), the widest gap among the three single-asset granularities. After the annotation round, eleven additional models were evaluated on the same item set.

Table 10: Storyboard design, all eight leaf dimensions: human mean (averaging the three annotators), automated mean, and per-item agreement on both sides. HB is the leave-one-out human baseline. Ordered by code.

		Mean score		Per-item PLCC		
Dimension	**Axis**	human	judge	HB	judge	ratio
`S-A1` event coverage	**F**	4.11	4.01	0.150	0.497	3.3×
`S-A2` dialogue fidelity	**F**	4.17	3.91	0.442	**0.728**	1.6×
`S-B1` executability	**P**	3.41	3.89	0.046	0.076	1.7×
`S-B2` restraint in additions	**P**	3.96	3.34	−0.023	0.110	—
`S-C1` narrative flow	**E**	4.03	4.12	0.199	0.256	1.3×
`S-C2` shooting rhythm	**E**	3.47	2.74	0.092	0.307	3.3×
`S-C3` emotional expression	**E**	3.93	4.18	0.128	0.332	2.6×
`S-C4` audiovisual style	**E**	4.11	4.09	0.041	0.265	6.5×
Composite		3.51	3.79	0.217	0.509	2.35×

Production planning is the main bottleneck. Seven of the eight leaf dimensions reach the usable line; six score between 3.89 and 4.18. These scores indicate faithful reproduction of core script elements, including plot events, character dialogue and scene presentation. The two weakest dimensions concern production logic: shooting rhythm at 2.74 and restraint in additions at 3.34 (Tab. 10). The main limitations are thus production scheduling and duration budgeting rather than language generation. Duration budgeting follows arithmetic constraints that can be checked before visual rendering, allowing the dominant defect to be identified and corrected in pre-production.

Nine of the additional models have an earlier generation from the same family in the comparison. All nine score higher than their predecessors, by +0.02 to +0.60; the newest, `gpt-6-astra-max` at 4.43, has the highest score in this stage.

Agreement ratios depend on human reproducibility. Per-item agreement is 0.509 against a human baseline of 0.217 ($2.35\times$), and every dimension exceeds its own baseline (Tab. 10). Baseline multiples reflect variation in human reproducibility, not simply automated scoring performance. Dialogue fidelity can be checked word by word against the script and has high agreement for both the automated scorer and humans (0.728 against 0.442). In contrast, audiovisual style treatment has the largest ratio, at $6.5\times$, despite automated agreement of 0.265, because the human baseline is 0.041. For restraint in additions, the baseline of -0.023 makes the ratio uninterpretable. A large multiple supports automatability only if the human criterion is reproducible; otherwise, higher automated agreement does not establish validity. These dimensions still need clearer operational criteria (Tab. 20).

Additional models are scored automatically. The eleven additional models enter Tab. 2 with per-axis scores and no agreement statistics of their own, at zero incremental human annotation cost. Their scores extend the judge validated on the annotated models and are not themselves checked against human ratings.

E.2 KEYFRAME IMAGES

Seven image models render storyboard scripts as keyframes, yielding $2{,}125$ single-panel items and 429 episode-level items. Models receive similar single-panel scores, but the human composite first falls below the usable line at episode-keyframe level (2.74; Tab. 11).

Specification adherence is the main deficit. Three of the fifteen single-panel dimensions fall below the usable line. Each concerns adherence to specified content: subject attributes against the character sheet (2.75), the scene against its reference image (2.72), and camera framing and visual style against the shot description (2.91). Across all fifteen dimensions, scores range from 2.72 to 4.59; both picture-quality dimensions, both generation-plausibility dimensions and every style-consistency dimension stay above the line (Tab. 11). This pattern points to conditioning and reference handling as the main limitations, rather than rendering capacity.

Cross-shot consistency breaks down first. At episode level, two of the four dimensions fall below the usable line: scene consistency (2.50) and character consistency (2.87). The lower scene score reflects shifts in layout, furnishings and lighting rather than changes to cast members (Tab. 11). Evaluating the same panels as a full episode exposes a different weakness: maintaining entities across frames. This relational property is not captured by judging frames individually.

Similar composites hide different strengths. The single model pair with a composite-score difference no larger than 0.03 illustrates the axis-based capability structure in Sec. 4.2. `seedream-5.0-pro` and `nano-banana-pro` differ by 0.02 in automated composite score. Yet `seedream-5.0-pro` leads by 0.39 on **F** input fidelity and trails by 0.60 on **P** generation plausibility. These axis-level gaps are an order of magnitude larger than the composite difference. The differences align with capability axes rather than isolated dimensions, so composite scores may distinguish model tiers while obscuring suitability for specific production needs.

Explicit references support automated agreement. The results do not follow a simple objective–subjective divide. Nine of the fifteen single-panel dimensions meet or exceed the human baseline, as do all four episode-level dimensions, but the four substantial shortfalls occur on the **Q** and **P** criteria (Tab. 11). Three of these four (technical quality, human anatomy and environmental physics) are among the taxonomy's most physically objective criteria. The distinction appears to be whether the item supplies a reference for the judgement. Technical quality seems readily measurable, yet a soft, low-contrast frame could be deliberate texture or a defect, and the item does not specify which was intended. Automated agreement is 0.167 against a human baseline of 0.275. By contrast, action and interaction has the strongest agreement, 0.527 against 0.401: the shot description specifies the intended action.

Table 11: The keyframe stages, every leaf dimension, ordered by code: human (three-annotator average) and automated mean scores, and per-item agreement on both sides. HB is the leave-one-out human baseline.

		Mean score		Per-item PLCC		
Dimension	**Axis**	human	judge	HB	judge	ratio
② Single keyframe						
I-A1 technical quality	Q	3.10	3.92	0.275	0.167	0.6×
I-A2 aesthetic style	Q	3.14	3.53	0.194	0.146	0.8×
I-B1 subject attributes	F	3.05	2.75	0.362	0.525	1.5×
I-B2 action and interaction	F	2.94	3.20	0.401	0.527	1.3×
I-B3 scene and spatial layout	F	3.28	3.28	0.306	0.443	1.4×
I-B4 shot framing and visual style	F	2.82	2.91	0.224	0.296	1.3×
I-C1 face identity	F	3.25	3.96	0.440	0.485	1.1×
I-C2 appearance and costume	F	3.69	3.72	0.409	0.471	1.2×
I-D1 cross-panel character	C	3.50	3.97	0.391	0.459	1.2×
I-D2 subject–scene integration	C	3.15	4.59	0.218	0.274	1.3×
I-D3 scene vs. reference	C	2.75	2.72	0.189	0.182	1.0×
I-D4 screen-direction axis	C	3.35	3.75	0.166	0.317	1.9×
I-D5 style consistency	C	4.12	4.52	0.131	0.130	1.0×
I-E1 human anatomy	P	3.21	3.87	0.439	0.311	0.7×
I-E2 environmental physics	P	3.26	4.50	0.234	0.152	0.7×
composite		3.21	3.62	0.449	0.536	1.19×
③ Episode keyframes						
MI-D1 cross-shot character	C	2.59	2.87	0.339	0.456	1.3×
MI-D2 subject–scene integration	C	2.80	4.37	0.110	0.229	2.1×
MI-D3 cross-shot scene	C	2.42	2.50	0.098	0.160	1.6×
MI-D5 cross-shot style	C	3.22	3.60	0.240	0.304	1.3×
composite		2.74	3.30	0.233	0.344	1.48×

E.3 SHOT VIDEO AND SHORT DRAMA

Six video models animate keyframes, and the resulting clips are assembled into episodes, yielding 2,186 shot-level items, 306 episode-video items and 206 short dramas. The gap between the highest- and lowest-scoring models widens from 0.36 points on single shots to 0.51 on short dramas.

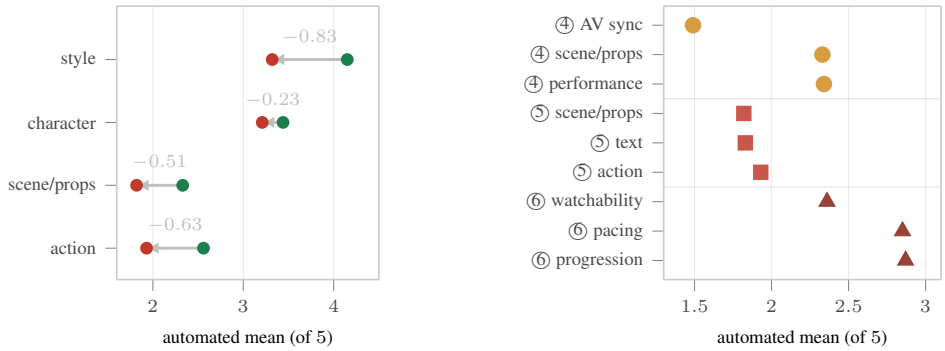
Temporal behaviour remains weakest. The twenty shot-level leaf dimensions span 1.49 to 4.15 (Tab. 12). Audio-visual sync scores lowest at 1.49, and character performance also ranks near the bottom at 2.34. The three highest scores are for style consistency (4.15), keyframe adherence (4.14) and image quality (4.11). These extremes distinguish static appearance from temporal behaviour: models perform better on the former. Still-image evaluation cannot capture the latter defects.

Larger judgement units reveal different defects. For four consistency dimensions among the ten defined at both shot and episode granularity, scores fall as the judgement unit expands: style consistency from 4.15 to 3.32, action consistency from 2.56 to 1.93, scene and prop consistency from 2.33 to 1.82, and character appearance consistency from 3.44 to 3.21 (Fig. 11). The type of deduction also changes. It can concern execution within a clip, a mismatch among clips in an episode video, or narrative content invented by an episode in a multi-episode drama. This last class cannot be assessed from an isolated shot. Larger units therefore reveal different kinds of defects, rather than simply more instances of the same kind. Clip-level results cannot establish whether a short drama meets delivery requirements.

Input constraints shape paradigm strengths. No paradigm leads across all shot-level criteria (Tab. 13). Their profiles reflect different input constraints. Grid panels combine multiple shots in one image, sharing visual information about characters and backgrounds. This paradigm has the highest scores for subject–scene integration (4.28), character appearance consistency (3.67) and style consistency (4.32). However, each shot uses a local cell rather than an independent frame, and cell adherence (3.97) trails frame-pair adherence (4.28). First/last frame supplies both shot endpoints and

Table 12: The three video granularities, every leaf dimension, ordered by code: human (three-annotator average) and automated mean scores, and per-item agreement on both sides. HB is the leave-one-out human baseline.

		Mean score		Per-item PLCC		
Dimension	**Axis**	human	judge	HB	judge	ratio
④ Shot video						
V-A1 image quality	Q	3.27	4.11	0.191	0.076	0.4×
V-A2 aesthetic style and motion	Q	3.25	2.72	0.169	0.096	0.6×
V-B1 subject attributes	F	3.54	3.43	0.314	0.264	0.8×
V-B2 shot framing and visual style	F	3.33	2.55	0.163	0.155	0.9×
V-B3 keyframe adherence	F	3.48	4.14	0.380	0.324	0.8×
V-C1 character appearance	C	3.80	3.44	0.282	0.231	0.8×
V-C2 scene and prop consistency	C	3.39	2.33	0.304	0.194	0.6×
V-C3 style consistency	C	4.09	4.15	0.245	0.149	0.6×
V-C4 action consistency	C	3.38	2.56	0.211	0.247	1.2×
V-C5 subject–scene integration	C	3.40	3.61	0.144	0.127	0.9×
V-C6 on-screen text	C	3.44	2.41	0.345	0.393	1.1×
V-D1 motion smoothness	P	3.60	3.31	0.208	0.069	0.3×
V-D3 physical plausibility	P	3.10	3.16	0.331	0.231	0.7×
V-D4 causal and temporal	P	3.35	2.49	0.158	0.136	0.9×
V-E1 speech quality	E	3.54	2.85	0.474	0.354	0.8×
V-E2 audio-visual sync	E	3.15	1.49	0.448	0.075	0.2×
V-E3 sound effects and music	E	3.38	3.58	0.252	0.061	0.2×
V-F1 composition	E	3.76	4.01	0.216	0.118	0.6×
V-F2 camera movement	E	3.49	3.66	0.147	0.070	0.5×
V-F3 character performance	E	3.07	2.34	0.183	0.122	0.7×
composite		3.44	3.09	0.380	0.301	0.79×
⑤ Episode video						
MV-C1 cross-shot character	C	2.55	3.21	0.288	0.283	1.0×
MV-C2 cross-shot scene and props	C	2.53	1.82	0.194	0.099	0.5×
MV-C3 cross-shot style	C	3.11	3.32	0.299	0.306	1.0×
MV-C4 cross-shot action	C	2.65	1.93	0.166	−0.022	−0.1×
MV-C5 subject–scene integration	C	3.19	4.11	0.123	0.027	0.2×
MV-C6 cross-shot text	C	3.10	1.83	0.227	0.333	1.5×
MV-D4 cross-shot causal/temporal	P	3.09	3.35	0.083	0.087	1.0×
MV-E1 cross-shot speech	E	2.75	3.06	0.407	0.019	0.1×
MV-E3 cross-shot sound	E	3.76	3.00	0.172	0.025	0.1×
MV-F3 cross-shot performance	E	3.17	3.94	0.122	0.017	0.1×
composite		2.79	2.93	0.240	0.203	0.85×
⑥ Short drama						
O-A1 event completion	F	3.00	3.58	0.200	0.363	1.8×
O-B1 style	C	3.19	3.67	0.461	0.302	0.7×
O-B2 cross-episode consistency	C	2.79	3.13	0.433	0.306	0.7×
O-B3 cross-episode progression	C	2.96	2.87	0.130	0.149	1.1×
O-C1 watchability	E	2.48	2.36	0.343	0.074	0.2×
O-C2 pacing and transitions	E	2.90	2.85	0.152	−0.112	−0.7×
composite		2.61	3.03	0.450	0.384	0.85×



(a) Shot level ● to episode ●: the four cross-shot consistency dimensions all fall.

(b) ④ shot video, ⑤ episode video, ⑥ short drama.

Figure 11: Different judgement units reveal different defects. (a) The same four consistency criteria score lower at episode level. (b) The three lowest automated means are selected separately at each granularity, not from a shared set of dimensions: sync, scene/props and performance for shots; cross-shot scene/props, text and action for episodes; watchability, pacing and progression for complete dramas.

leads on keyframe adherence (4.28). The intervening content is less constrained, and this paradigm has the lowest subject–scene integration (3.01) and camera movement (3.30) scores. Multi-reference has no fixed frame to reproduce, leaving more freedom for camera movement. It leads on camera movement plausibility (4.09) and sound and music (3.62), but ranks last on style consistency (4.04), 0.28 below the leader. These strengths and weaknesses are consistent with the constraints each input format imposes or relaxes, making the format a substantive deployment choice (Sec. G.4).

Table 13: Mean automated shot-video scores by input paradigm. **Bold**/red indicate the highest/lowest score in each row. *n/a*: multi-reference provides no fixed keyframe.

Dimension	first/last frame	grid panel	multi-reference
V-C5 subject–scene integration	3.01	**4.28**	3.96
V-F2 camera movement	3.30	3.82	**4.09**
V-C3 style consistency	4.12	**4.32**	4.04
V-B3 keyframe adherence	**4.28**	3.97	*n/a*
V-C1 character appearance	3.29	**3.67**	3.45
V-E3 sound and music	3.59	3.51	**3.62**

Reference availability also matters for video. Per-item agreement shows a similar reference-related pattern, distinct from mean model scores. Event completion, which can be checked against the script's event list, reaches 0.363 against a human baseline of 0.200. Watchability of the cut, which has no textual reference, reaches 0.074 against 0.343 (Tab. 12). Speech quality and audio-visual sync offer a comparison within the same modality and tool chain, with comparable human baselines (0.474 and 0.448) but automated agreement of 0.354 and 0.075, respectively. The script's dialogue provides a textual anchor for speech quality, whereas no such reference exists for sync. Audio-visual sync has both the lowest shot-level mean score and the largest relative shortfall from its human baseline (0.17×). Its mean score should therefore be interpreted alongside the scorer's limited agreement, rather than as an unqualified ranking of generation models.

E.4 The unit of judgement

Figure 12 illustrates the granularity argument in Sec. 4.3 with two items. The artefacts remain fixed while the evaluation question changes, so the resulting deductions concern relationships among the artefacts.

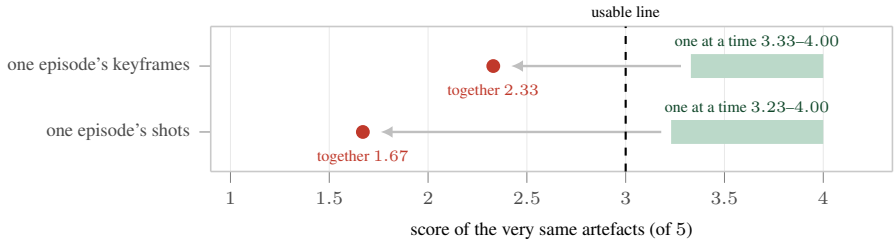


Figure 12: The same artefacts scored individually and as an episode. Changing the judgement unit exposes cross-shot defects that individual scores do not capture.

F Automated Evaluation Analysis

Automated evaluation raises two distinct questions: how model performance varies with production conditions, and how reliably the scores reflect human judgements. Section F.1 addresses the first through content-conditioned comparisons; Sec. F.2 addresses the second through agreement analyses and their implications for automation.

F.1 Performance across content conditions

The three slices compare visual style, dialogue language and model input paradigm (Tab. 14). Comparisons are restricted to models evaluated at every level of the relevant slice, keeping configuration coverage consistent. The largest differences in automated composite scores occur across input paradigms, followed by visual style and dialogue language (Tab. 15).

Table 14: Content slices and observed differences in automated scores.

Slice	Comparison	Difference in automated score
Visual style	live action versus animation	+0.08 to +0.22 in favour of live action at five of six stages
Dialogue language	English versus Chinese dialogue	at most 0.12, concentrated on **F**
Input paradigm	first/last frame, grid panel, multi-reference	up to 0.42, with the lead reversing between keyframe stages

Table 15: Automated composite-score differences across content slices and the leading input paradigm. Each comparison includes only models evaluated at every level of that slice.

Granularity	live action − animation	English − Chinese	paradigm spread	leading paradigm
① Storyboard design	+0.14	+0.07	0.09	*n/a*
② Single keyframe	−0.08	+0.02	0.37	first/last frame
③ Episode keyframes	+0.08	+0.12	0.42	grid panel
④ Shot video	+0.15	+0.04	0.14	grid panel
⑤ Episode video	+0.22	+0.12	0.19	grid panel
⑥ Short drama	+0.14	+0.08	0.29	grid panel

F.1.1 Scores by visual style

Among models evaluated in both styles, the automated composite score is higher for live action at five of the six granularities, by 0.08 to 0.22; single keyframes instead favour animation by 0.08 (Tab. 15). These composite differences mask opposing axis-level patterns. At shot video, **C** is 0.32 higher for live action and **Q** is 0.25 higher for animation, while **F** is identical. The dimensional gaps are larger: `V-C3` style consistency is over a point higher for live action, while `V-A1` image quality and material is two thirds of a point higher for animation. These results associate style with different defect profiles, rather than uniformly easier or harder items.

F.1.2 SCORES BY DIALOGUE LANGUAGE

Scene descriptions are Chinese throughout; the language contrast concerns dialogue and on-screen text, with four English-dialogue dramas and sixteen Chinese-dialogue dramas balanced by visual style. English-dialogue dramas have higher automated composite scores at every granularity, but the gap never exceeds 0.12 (Tab. 15). Almost all of the difference lies on **F**, which is 0.64 higher for English at short-drama level, while **Q** is identical to two decimals at both stages where it is defined. The most affected dimensions involve language (event completion, speech quality, dialogue fidelity), whereas cross-shot character and cross-episode consistency favour Chinese. The differences thus concentrate in language-dependent dimensions rather than visual quality.

F.1.3 SCORES BY INPUT PARADIGM

Input paradigm shows the largest average composite-score differences, with leadership changing along the chain. First/last frame exceeds grid panel by 0.37 on single keyframes but trails it by 0.42 on episode keyframes. Grid panel ranks first at all three video granularities, with narrow leads on shot video and short drama (Tab. 15). The single-keyframe comparison also depends on dimension coverage: **C** is defined only for grid panel, and restricting the comparison to the ten shared dimensions widens the first/last-frame lead to 0.61. These patterns are consistent with each paradigm's input constraints. First/last frame provides both endpoint images and leads on **F** input fidelity at the keyframe stage (3.57 against 2.73) and on frame-anchored dimensions at shot video (scene and prop consistency, shot framing and visual style). Its lower camera plausibility is consistent with the constraints those endpoints place on motion. Grid panel places an episode in one image and leads on cross-shot consistency (episode keyframes 3.53 against 3.11) and composition, but has lower audio-visual sync. At shot video, multi-reference has no fixed frame to reproduce and ranks last on **F** (2.92 against 3.39), while leading on camera plausibility, audio-visual sync and **E** cinematic expressiveness.

F.2 RELIABILITY OF AUTOMATED SCORING

The preceding comparisons describe differences in automated scores, not the reliability of those scores. We therefore assess agreement with human ratings before identifying where automation is appropriate. Table 16 reports the stage-level validation used in Sec. 4.2: model-level score correlation, per-item agreement relative to the leave-one-out human baseline, and absolute score accuracy.

Table 16: Agreement with human ratings by stage. **Model Level**: model-score correlations; **Item Level**: per-item correlation and the leave-one-out human baseline (HB); **Score Acc**: bias, MAE and accuracy within tolerance. *Paired items* have human and automated scores: all collected text and image items, 1,149 of 2,186 shot videos, 238 of 306 episode videos and 201 of 206 short dramas. Counts describe collection and pairing, not the complete reference set.

		Model Level		**Item Level**			**Score Acc**		
Stage	Paired items	PLCC	SRCC	HB	Judge	Ratio	Bias	MAE	Within ±0.5
① Storyboard design	533	0.936	0.800	0.217	0.509	2.35×	+0.28	0.416	67.5%
② Single keyframe	2,125	0.959	0.750	0.449	0.536	1.19×	+0.41	0.529	54.2%
③ Episode keyframes	429	**0.973**	0.964	0.233	0.344	1.48×	+0.57	0.733	38.2%
④ Shot video	1,149	0.755	0.829	0.380	0.301	0.79×	−0.35	0.436	62.5%
⑤ Episode video	238	0.935	0.943	0.240	0.203	0.85×	+0.14	0.433	66.0%
⑥ Short drama	201	0.947	**1.000**	0.450	0.384	0.85×	+0.42	0.425	66.7%
Mean		**0.918**	**0.881**	0.328	0.379			0.495	

Correlations use composite scores. *Judge* is per-item PLCC with the three-annotator consensus; HB is the leave-one-out human baseline, and *Ratio* is Judge / HB. Green meets or exceeds the baseline; red falls below it.

The next four analyses use the same corpus to examine agreement by stage, evaluation axis, modality and human reproducibility (Tab. 17). Agreement means per-item PLCC between DramaChain Agentic Judge and the three-annotator consensus (mean at both leaf and composite levels; Sec. 3.3), reported directly or relative to the leave-one-out human baseline. Content differences in Sec. F are expressed

in score units rather than standardised effects (Cohen, 2013), retaining comparison with the 3.0 usable line.

Table 17: Agreement slices and their main patterns.

Slice	Comparison	Agreement pattern
Pipeline stage	the six granularities	above the human baseline for text and images, below it for video
Evaluation axis	the five evaluation axes	lower along checkable → perceptual → aesthetic
Modality	text, pixels, time + audio	agreement within an axis varies by modality
Human consensus	reproducibility of human judgements	higher automated agreement where humans agree more

F.2.1 Agreement by stage

Relative to the leave-one-out human baseline, automated per-item agreement is 2.35× on storyboard design, 1.19× on single keyframes and 1.48× on episode keyframes, then 0.79×, 0.85× and 0.85× on the three video granularities (Tab. 16). Text and image stages therefore span 1.19–2.35×, compared with 0.79–0.85× for video. The 1× boundary lies between images and video, not between single- and multi-asset units: episode keyframes require cross-panel comparisons but exceed the baseline, whereas a single clip does not. This pattern points to modality rather than depth along the chain as the relevant distinction (Fig. 14).

F.2.2 Agreement by evaluation axis

Averaging per-item agreement over the 63 leaf dimensions gives **F** 0.423, **C** 0.235, **P** 0.146, **E** 0.132 and **Q** 0.121 (Tab. 18). Only **F** exceeds its axis-level human baseline, by +0.107, with 9 of its 12 dimensions meeting or exceeding their own baselines. In contrast, **Q** meets the baseline on none of its four dimensions and is 0.086 below it at axis level. This ordering is consistent with the availability of an external reference: an upstream output for **F**, a peer artefact for **C**, and no equivalent for **Q**. It also parallels the failure of small-model quality metrics as scoring anchors (Sec. D.3). These averages pool three modalities and can conceal modality-specific differences.

F.2.3 Agreement by modality

Where Tab. 18 pools modalities, Tab. 19 separates each axis by modality, as discussed in Sec. 4.2. The clearest contrast is **E**: storyboard text stands +0.175 above the human baseline, with 4 of 4 dimensions reaching it, whereas video with time and audio stands 0.190 below it, with 0 of 11 reaching it. These are different leaf criteria within the same axis, not repeated measurements of one criterion. For **F**, the margin over the baseline shrinks across the three modalities (+0.317 on text, +0.101 on pixels, +0.012 on time and audio), and **Q** is below the baseline in both modalities where it is defined. Agreement established for an axis in one modality therefore does not establish validity in another; reports should specify both.

Table 18: Per-item agreement by evaluation axis, pooled across six stages. HB denotes the leave-one-out human baseline; **judge** ≥ HB counts dimensions meeting it. The final column describes the available references.

Axis	dims	HB	judge	Δ	judge ≥ HB	What the automated scorer has to work with
F Input fidelity	12	0.316	**0.423**	+0.107	9/12	the upstream artefact, checkable line by line
C Internal consistency	24	0.239	0.235	−0.004	12/24	another shot, as a referent
P Generation plausibility	8	0.184	0.146	−0.038	3/8	no referent; common sense only
Q Visual quality	4	0.207	0.121	−0.086	0/4	pure perception, no checkable answer
E Cinematic expressiveness	15	0.225	0.132	−0.093	4/15	pure impression, no checkable answer

Values are means of leaf-level PLCCs: judge against the three-annotator consensus, and the leave-one-out human baseline. Δ is judge − HB. Low agreement on **Q** is consistent with the difficulty of using low-level metrics as quality-scoring anchors (Sec. D.3).

F.2.4 Agreement by human consensus

The final slice relates automated agreement to the reproducibility of human judgements. Across the 63 leaf dimensions, the human baseline correlates with automated per-item agreement at PLCC

Table 19: Per-item agreement by axis × modality. Each cell gives the human baseline and the automated scorer (PLCC), their difference, and how many leaf dimensions reach the baseline.

	text			pixels			time + audio		
Axis	HB / judge	Δ	win	HB / judge	Δ	win	HB / judge	Δ	win
F input fidelity	.296 / **.613**	+.317	2/2	.357 / .458	+.101	6/6	.264 / .277	+.012	1/4
C internal consistency	—	—	—	.209 / .279	+.070	7/9	.257 / .208	−.049	5/15
P generation plausibility	.011 / .093	+.082	2/2	.337 / .231	−.105	0/2	.195 / .131	−.064	1/4
Q visual quality	—	—	—	.235 / .157	−.078	0/2	.180 / .086	−.094	0/2
E cinematic expressiveness	.115 / .290	+.175	4/4	—	—	—	.265 / .075	−.190	0/11

0.418 (SRCC 0.416): higher human agreement tends to accompany higher automated agreement (Fig. 13). This association does not establish that either score is accurate. Where both are low, clearer operational criteria may be needed before adding annotations or tuning prompts. Human reproducibility, rather than the automated score alone, should guide decisions about retaining a dimension: unreliable human judgements offer a weak alignment target.

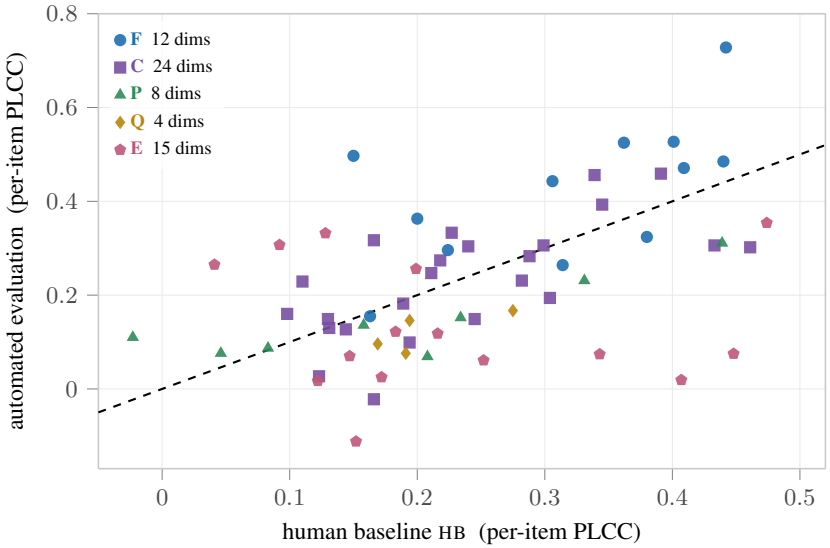


Figure 13: Automated agreement and the leave-one-out human baseline across all 63 leaf dimensions, measured by per-item PLCC. The dashed diagonal marks equal agreement; points above it exceed the human baseline. The two measures are positively correlated. Most **F** dimensions lie above the diagonal, while most **E** and all **Q** dimensions lie below it.

F.2.5 Where automated agreement exceeds the human baseline

Figure 14 summarises automated per-item agreement relative to the human baseline at each stage. Table 20 groups leaf dimensions by two properties relevant to automated agreement: whether the criterion can be expressed as a checklist, and whether judgement requires fine-grained perception (Sec. 4.2).

G Case Study

The cases examine differences hidden by aggregate scores (Sec. G.1, Sec. G.2), how defects arise and propagate (Sec. G.3–G.5), model-specific strengths and weaknesses (Sec. G.6, Sec. G.7), and a failure of automated scoring (Sec. G.8).

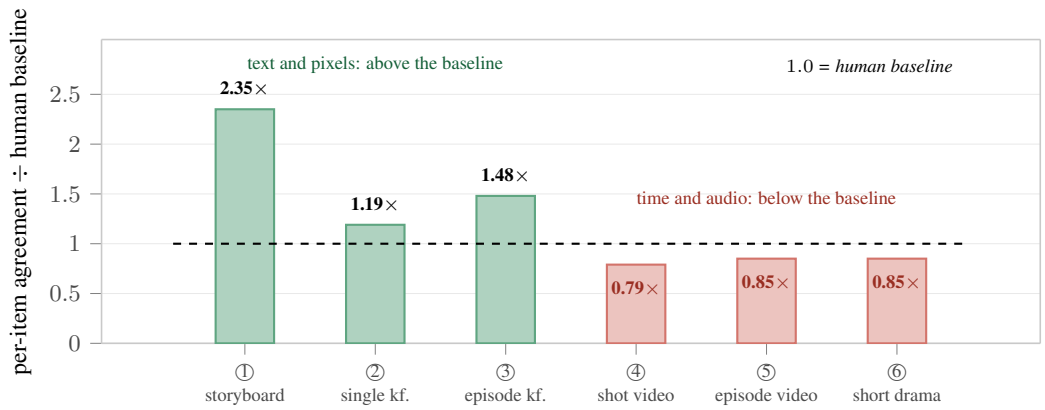


Figure 14: Automated per-item agreement as a multiple of the leave-one-out human baseline, using PLCC. Values above 1 indicate agreement exceeding the baseline.

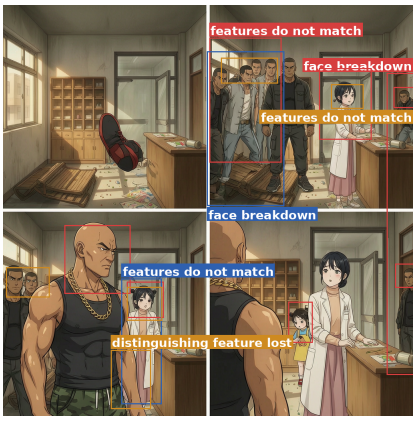
Table 20: Dimensions grouped by perceptual requirements (rows) and whether criteria can be expressed as checklists (columns). Cells report automated / human-baseline per-item PLCC, their ratio, and implications for automation.

	Criterion enumerable as check items	**Criterion not enumerable**
Perception not needed	**Above the baseline.** S−A2 dialogue fidelity 0.728/0.442 1.6× S−A1 event coverage 0.497/0.150 3.3× S−C4 audiovisual style 0.265/0.041 6.5×* *An external answer exists to compare against: the event list, the verbatim dialogue, the style instruction.*	**Both sides near zero: repair the criterion.** S−B2 restraint in additions 0.110/−0.023 — I−D5 style consistency 0.130/0.131 0.99× *The obstacle is the operability of the criterion, not annotation cost; tuning prompts here yields little.*
Perception needed	**Below the baseline: add perception.** V−E2 audio-visual sync 0.075/0.448 0.17× V−D1 motion smoothness 0.069/0.208 0.33× V−A1 image quality 0.076/0.191 0.40× I−E2 environmental physics 0.152/0.234 0.65× *All perceptual, and the human baseline is itself normal or high: the problem is real and people can see it, the scorer cannot.*	*No dimension falls here.*

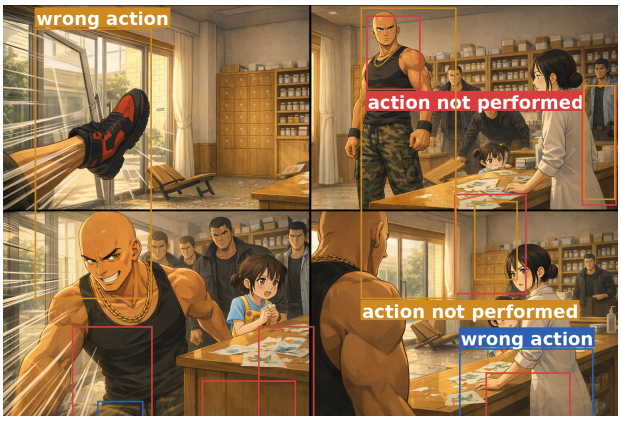
* Inflated by a near-zero human baseline (0.041); see Sec. E.1.

G.1 Similar scores, different failure modes

The four items in Fig. 15 each score 1.3–1.7 on a dimension, yet their boxed defects affect different aspects: facial identity, limbs during action, unwanted people, and hands with the wrong number of fingers. Within each pair, the models exchange strengths: the model with an identity error scores 3.0 on action, whereas the model with an action error scores 3.33 on identity. A composite could group these items together while concealing the reason for a low score. The cases illustrate the need for dimension-specific attribution and provide an item-level counterpart to Sec. E.2.

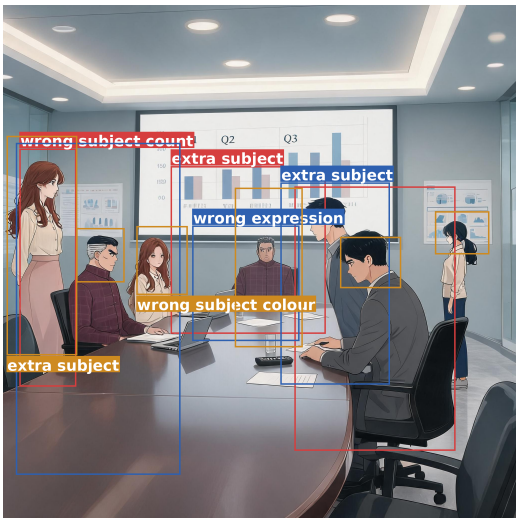


(a) `wan2.7-image-pro`: `I-C1` face identity 1.67, 15 boxes, every one on a face; action interaction scored 3.0.

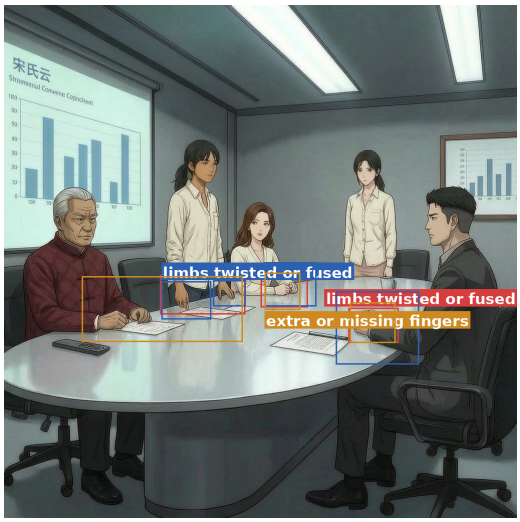


(b) `gpt-image-1.5`: `I-B2` action interaction 1.33, 13 boxes, every one on a limb; face identity scored 3.33.

Figure 15: Four items with similarly low scores (1.3–1.7) on different dimensions. Attribution boxes identify distinct defects, and each model pair fails on different criteria. Boxes are drawn independently by three annotators, distinguished by colour.



(c) `seedream-5.0-lite`: `I-B1` subject attributes 1.33, 12 boxes, all "extra subject"; human anatomy scored 3.0.



(d) `wan2.7-image-pro`: `I-E1` human anatomy 1.33, 10 boxes, all broken hands; subject attributes scored 3.33.

Figure 15: Attribution examples (continued).

G.2 Cross-shot consistency is a separate capability

Two models generate the same episode from the same character sheets (Fig. 16). The top-tier output maintains the identities of its two leads across shots; the bottom-tier output does not, leaving a 2.66-point gap on `MI-D1`. In the failing row, shot 7 is plausible in isolation but inconsistent with shot 5. Individual panels can thus look acceptable while the episode fails continuity. Single-asset evaluation measures a different property, not simply a less sensitive version of cross-shot quality.



Figure 16: Top- and bottom-tier models use the same episode and character sheets but differ by 2.66 points on MI-D1. The identity changes become apparent across shots rather than within individual panels.

G.3 Shot count and storyboard executability

Table 21 compares three models given the same episode script, with the pipeline differing at the storyboard stage. The three-shot output compresses the episode into 30 s of self-declared duration and 10 dialogue turns. All three annotators assign 1.00 to its executability (S-B1) and shooting rhythm (S-C2), marking repeated production defects: multiple actions within a shot, overly complex blocking and missing reaction shots. The six-shot output scores higher on both **F** dimensions; the twelve-shot output matches those scores and scores higher on executability (Fig. 17). These are cross-model observations, not an isolated test of shot count.

Table 21: Three models generate different shot counts from the same episode script. Scores are means of the three annotators' scores. The pipeline forks at the storyboard stage; all models receive the same script.

Dimension	Axis	mimo-v2.5-pro 3 shots	gpt-5.5-xhigh 6 shots	claude-opus-4.8-max 12 shots	Observed pattern
S-A2 dialogue fidelity	F	1.33	**5.00**	**5.00**	the 6-shot output loses no lines
S-A1 event coverage	F	2.67	**4.67**	**4.67**	higher in both longer outputs
S-B1 executability	P	1.00	3.00	**4.00**	scores increase across these outputs
S-C2 shooting rhythm	E	1.00	**3.67**	**3.67**	the 6-shot output ties the longest
Overall impression		1.67	**4.00**	**4.00**	—
Problems marked		23	18	15	fewer marks, with different defect types
Annotator disagreement		1.75	0.75	0.63	lower for the higher-scoring outputs

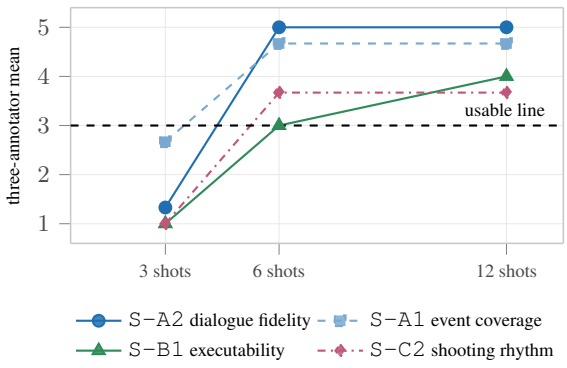
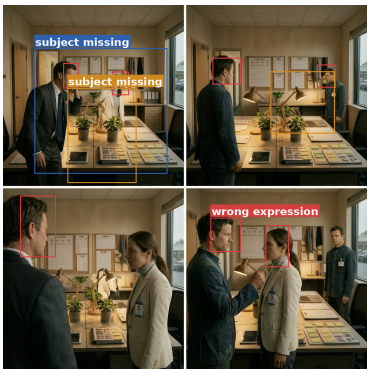
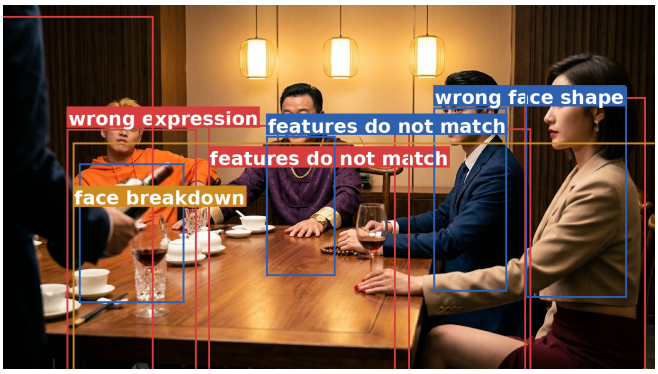


Figure 17: Three models produce different shot counts from the same 1,861-character script, branching only at storyboard design (Sec. G.3). The 6-shot output reaches the usable threshold on all four dimensions shown and achieves full dialogue fidelity. Relative to this output, the 12-shot output improves only executability among these dimensions.



(a) Grid panel omits people. `wan2.7-image-pro`, `I-B1` subject attributes 1.33: cells that should hold three people hold two.



(b) Identity drift between endpoint frames. `nano-banana-pro`, `I-C1` face identity 1.33: all three annotators independently marked the identity mismatch.



(c) The input for (b): the shot's own first frame (left) and three character sheets.

Figure 18: Examples of distinct failure modes under the two input paradigms. The numbers of below-threshold dimensions are $1/10$ for first/last frame and $9/15$ for grid panel.

G.4 Input paradigms expose different failure modes

Figure 18 compares grid-panel and first/last-frame outputs. The grid panels omit people, whereas the first/last-frame pairs change identity between endpoints. These defects reflect different constraints: a grid image must fill several cells, while separately generated endpoint images may not preserve identity. The cases illustrate paradigm-specific difficulty that visual style alone does not explain; pooling paradigms can obscure the grid-panel failures.

On video, a grid-panel `pixverse-c1` shot scored 1.67 on keyframe adherence. All three annotators marked the same defect class, "deviates from the given keyframe", localised to 0.29–11.04 s, 2.98–4.87 s and 6.56–8.15 s. The prompt describes only camera moves and dialogue, leaving appearance and blocking anchored by the grid image. This example illustrates the difficulty of maintaining adherence to individual grid cells, consistent with the lower grid-panel adherence in Tab. 13.

G.5 Upstream defects persist downstream



`gpt-image-1.5`, four consecutive shots

`MI-D3` scene **2.00** vs. `MI-D1` character **3.67**

Figure 19: Scene consistency fails before character consistency: all three annotators marked set-dressing changes across these consecutive shots, while the characters remained consistent. Across the corpus, scene consistency has the lowest human mean (2.42) among the four episode-keyframe dimensions.

The propagation experiment in Fig. 3 uses four dramas (two live action, two animated) under the first/last-frame paradigm. The video model is fixed, and the baseline chain leads at each stage (3.03–3.29 across the three granularities). Each run replaces one upstream stage and holds the rest unchanged: the storyboard stage with `hy3` or `mimo-v2.5-pro`, or the image stage with

nano-banana-pro or seedream-5.0-lite, giving the 12 configurations plotted there. Figure 19 illustrates an upstream consistency defect: set dressing changes across consecutive keyframes while character identities remain stable.

G.6 HY4-PREVIEW AND HY3: FIDELITY AND PRODUCTION EXECUTABILITY

Table 22 expands the automatic storyboard comparison in Tab. 2. hy4-preview scores higher than hy3 overall (4.02 versus 3.89), with higher **F** (4.69 versus 4.36) and **E** (3.91 versus 3.78) but lower **P** (3.56 versus 3.65). Better script reproduction need not mean better production executability. hy3 and doubao-seed-2.1-pro both score 3.89 overall, but the latter leads in emotional expression and narrative flow.

Table 22: Automatic storyboard means ($n = 60/57/60$ in column order). Samples are not fully paired; hy4-preview has automatic scores only.

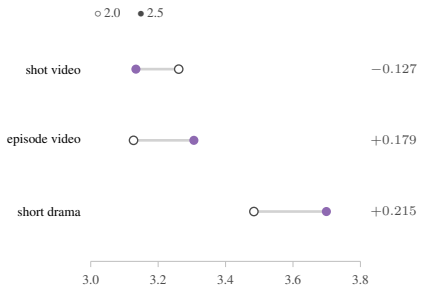
Dimension	Axis	hy4-preview	hy3	doubao-seed-2.1-pro
S-A1 event coverage	**F**	**4.53**	4.31	4.27
S-A2 dialogue fidelity	**F**	**4.84**	4.42	4.25
S-B1 executability	**P**	3.73	3.82	**3.90**
S-B2 restraint in additions	**P**	3.38	**3.48**	3.40
S-C1 narrative flow	**E**	4.33	4.26	**4.40**
S-C2 shooting rhythm	**E**	**3.05**	2.91	2.58
S-C3 emotional expression	**E**	4.23	4.05	**4.33**
S-C4 audiovisual style	**E**	**4.03**	3.89	3.99
Overall		**4.02**	3.89	3.89

Bold marks the highest mean in each row; Overall averages the eight leaf dimensions.

A separate qualitative comparison uses the script for episode 2 of the animated *I Insist on Taking You* (2,019 Chinese characters, 20 dialogue turns). In table-column order, the outputs contain $10/10/7$ shots and $2{,}596/1{,}306/2{,}586$ Chinese characters, preserving $20/20/3$ turns verbatim. hy4-preview retains the dialogue and adds camera movement and blocking, but chains multiple actions within one shot; hy3 mainly uses "rapid cut" and "close-up", whereas doubao-seed-2.1-pro adds visual detail but rewrites dialogue.

G.7 SEEDANCE-2.5: STRONGER EPISODE COHERENCE, WEAKER CLIP QUALITY

On paired items, the newer version scores lower at shot-video level but higher at both episode and short-drama levels (Fig. 20). The declines include audio-visual sync (-0.658), motion smoothness and camera plausibility, while cross-shot style consistency rises by $+1.143$. The case in Fig. 20b suggests that more literal prompt adherence can reproduce a deficiency in the prompt rather than resolve it. Duration also differs: single shots increase from 15 s to 30 s, with four internal cuts in one shot still summarised by a single score. These observations illustrate why stronger episode coherence need not imply better continuous behaviour within a clip.



(a) Paired means at the three video granularities.

Multi-reference · animated · EP1 shot 2 · 21.1 s

V−E2 audio-visual sync: 1 (seedance−2.0 on the same item: 5)

Prompt, excerpt. shot 3 — she lowers her eyes, still facing right, { Come clean, leave clean. I owe nobody. } shot 4 — **Fu Jingxing in profile** [Image 2] stands in the doorway, facing the open suitcase, his gaze passing over **Qiao Wan from behind** [Image 4]; then, voice lowered, { You are packing now? }

Reference sheets (10 for this shot, 4 named). Qiao Wan · front | Qiao Wan · **back** | Fu Jingxing · **profile** | master bedroom

Both lines have a named speaker, and the prompt pins both speakers to a back view or a profile silhouette; one of the ten sheets is a dedicated rear view. There is no mouth for the dialogue to come out of.

(b) A case in which following the prompt preserves its conflicting requirements.

Figure 20: seedance−2.5 and seedance−2.0 on paired animated, multi-reference items, matched by drama, episode and shot (39/14/10 pairs). The newer version scores higher at episode and short-drama granularity but slightly lower on individual clips.

G.8 Automated scoring without an external reference

In the item shown in Fig. 21, reference-based scores are close to human ratings, but a purely perceptual dimension is overestimated by 3.67 points. The contrast illustrates, at item level, the reference-related limitation observed in metric validation (Sec. D.3, Tab. 18).



Dimension	human	judge	err	
I−B1 subject attributes	2.00	2	0.00	correct
I−C2 appearance, costume	2.33	2	0.33	correct
I−B3 scene and layout	2.67	3	0.33	correct
I−A1 technical quality	1.33	5	3.67	**overestimated**

The three annotators scored technical quality 1/1/2 and tagged it consistently: structural distortion, visible noise, blur, ghosting.

Figure 21: All four scored dimensions for one nano−banana−pro panel. Reference-based judgements are close to human scores, but technical quality is overestimated by almost four points.