DIRECTORBENCH: DIAGNOSING LONG-FORM VIDEO GENERATION WITH PERSONALIZED MULTI-AGENT EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form video generation is rapidly moving from short, single-scene synthesis toward minute-long, multi-shot creation with narrative structure, cinematic control, audio, and cross-modal synchronization. However, evaluating such videos remains challenging, since existing benchmarks largely focus on local visual quality, short-horizon temporal consistency, or generic prompt alignment, and provide limited diagnosis of workflow failures and user-dependent preferences. We introduce DirectorBench, a personalized multi-agent diagnostic benchmark for long-form video generation. DirectorBench evaluates generated videos with respect to 80 structured metadata entries, 7 user profiles, and 40 checkpoint criteria across 5 dimensions: script, visual, audio, cross-modal, and stability. Instead of reducing quality to a single aggregate score, DirectorBench localizes checkpoint-level bottlenecks and supports profile-aware evaluation. We evaluate 4 long-form video generation workflows, 6 base LLMs, and 7 user profiles. Across workflows, DirectorBench reveals a between-unit bottleneck: transition quality averages only 0.256 and reaches 0.356 for the best workflow, while prompt-level user demand fulfillment averages 0.71. A study with 14 annotators yields case-level Spearman correlation 0.739 and 84.3% pairwise ranking accuracy; the human-rated bottleneck appears in the automatic top-three list in 96.5% of cases. Additional experiments show that profile-conditioned prompts change speech density and video characteristics. These findings highlight the importance of diagnostic and profile-aware benchmarking for agentic long-form video generation. Our code is available at `https://anonymous.4open.science/r/DirectorBench-835E/`.

1 INTRODUCTION

Recent advances in text-to-video generation (Singer et al., 2023b) have significantly improved the visual fidelity, motion realism, and prompt controllability of short video clips. Modern systems can generate visually appealing scenes from natural-language prompts, and a growing ecosystem of agentic workflows has begun to decompose video creation into stages (Wu et al., 2025; Huang et al., 2026a; Meng et al., 2026). These developments mark a shift from isolated clip generation (Brooks et al., 2024; Peng et al., 2025a) toward long-form, multi-shot video creation (Meng et al., 2026; Wu et al., 2025), where the goal is no longer merely to synthesize a plausible short scene, but to produce a coherent video with narrative structure, cinematic control, audio-visual alignment, and long-term consistency. As a result, long-form video quality depends not only on local frame fidelity, but also on the structure that connects shots, modalities, and user intent over time.

Existing video-generation benchmarks (Huang et al., 2024; Zheng et al., 2025; Han et al., 2025) have made important progress in measuring visual quality, temporal smoothness, subject consistency, and prompt alignment. However, they remain insufficient for evaluating long-form, agentically generated videos. Most existing metrics are designed for short clips or frame-level properties, and therefore under-represent between-shot failures such as transition quality, long-term continuity, and story-level coherence (Huang et al., 2024; Zheng et al., 2025). In addition, current evaluations focus primarily on visual consistency, while audio quality, narration, and cross-modal coordination are less systematically diagnosed (Han et al., 2025). Furthermore, most benchmarks collapse quality into a single aggregate score, ignoring that video creation is inherently user-dependent (Zheng et al.,

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.

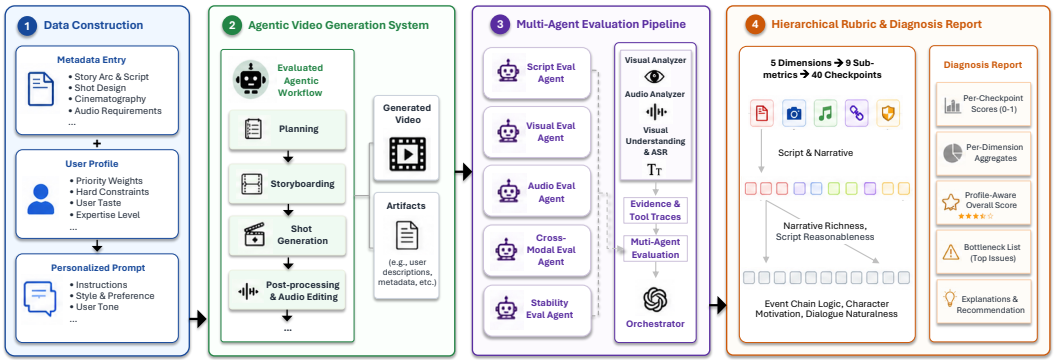

Figure 1: DirectorBench separates metadata-defined intent from user preferences. Profile-conditioned prompts drive generation; specialist evaluation, tool evidence, and diagnostic synthesis assess the resulting videos.

2025; Ma et al., 2026; Peng et al., 2025b). At the same time, modern video-generation systems are increasingly agentic, involving base LLMs, planning modules, storyboard generation, tool calls, and iterative refinement (Wu et al., 2025; Huang et al., 2026a; Meng et al., 2026; Wang et al., 2026), where failures are distributed across components rather than attributable to a single model. As a result, a useful benchmark should go beyond final-score comparison and provide diagnostic signals that localize bottlenecks and explain how different parts of the workflow contribute to overall quality.

To address these challenges, we propose DirectorBench, a personalized multi-agent diagnostic benchmark for long-form video generation. DirectorBench evaluates agentic long-form video generation systems under structured metadata specifications and personalized user profiles, analyzing the generated outputs to produce diagnostic evaluation signals. The key design principle of DirectorBench is to separate objective intent from personalized preference: metadata specifies what the video should contain (e.g., story arcs, shot design, camera control, consistency constraints, and audio requirements), while user profiles determine how different dimensions are prioritized (Ma et al., 2026; Peng et al., 2025b; Wu et al., 2024; Xing et al., 2026). This enables DirectorBench to jointly assess whether a video satisfies concrete requirements and how its perceived quality varies under different user preferences (Ma et al., 2026; Peng et al., 2025b; Han et al., 2025; Zheng et al., 2025).

We instantiate DirectorBench with 80 structured metadata entries, 7 user profiles, and 40 checkpoint criteria across five dimensions: script, visual, audio, cross-modal, and stability. Instead of collapsing quality into a single aggregate score, DirectorBench produces checkpoint-level diagnosis reports that localize bottlenecks across dimensions. Using DirectorBench, we evaluate long-form video generation across various settings to analyze how workflow design, model choice, and personalization affect generation quality. Across workflows, DirectorBench reveals a between-unit bottleneck: transition quality averages only 0.256 and reaches 0.356 for the best workflow, while prompt-level user demand fulfillment averages 0.71. We further conduct human evaluation with multiple annotators and show that DirectorBench aligns with human judgment at the dimension level. Furthermore, we also distinguish rescoring the same video from generating a new video with a different profile. A controlled prompting experiment on two systems demonstrates changes in output content, without equating those changes with improved preference satisfaction. These results demonstrate that diagnostic and profile-aware evaluation reveals workflow failures and user-dependent quality variation that are hidden by aggregate metrics.

Our contributions are threefold:

- We introduce DirectorBench, a personalized multi-agent diagnostic benchmark for long-form video generation, instantiated with 80 structured metadata entries, 7 user profiles, and 40 checkpoint criteria across script, visual, audio, cross-modal, and stability dimensions.
- We propose a multi-agent evaluation pipeline with dynamic checkpoint activation, tool-based evidence, confidence-aware aggregation, and profile-weighted scoring, producing checkpoint-level diagnosis reports rather than only aggregate scores.
- We evaluate long-form video generation across agentic workflows, base LLMs, and user profiles, showing that DirectorBench reveals between-unit bottlenecks, workflow-specific failure signatures, and user-dependent quality variation, with validation from human evaluation.

2 DIRECTORBENCH: STRUCTURED AND PROFILE-AWARE DIAGNOSIS

2.1 PROBLEM FORMULATION

For metadata $\mathbf{m}$, profile $\mathbf{u}$, and generation system $\mathcal{G}$, we construct a prompt and evaluate its output:

$$\mathbf{p} = g(\mathbf{m}, \mathbf{u}), \qquad \mathbf{v} = \mathcal{G}(\mathbf{p}), \qquad \mathcal{R} = f(\mathbf{m}, \mathbf{u}, \mathbf{v}). \tag{1}$$

The diagnosis report $\mathcal{R}$ contains rubric results, dimension scores, an overall score, bottlenecks, and evidence traces. This separation matters experimentally: changing $\mathbf{u}$ in g can change the generated content, whereas changing only the evaluation weights for a fixed $\mathbf{v}$ measures sensitivity to the scoring perspective. These are distinct interventions and should not be interpreted interchangeably.

Metadata provides a reference for requirement fulfillment, not a unique ideal video. A profile can prioritize one quality dimension without changing the objective facts requested in the task. We therefore retain dimension-level results and constraint violations alongside the aggregate score; different failure patterns can lead to the same scalar value. Figure 1 illustrates the overall framework.

2.2 BENCHMARK CONSTRUCTION

Metadata. The 80 structured entries span narrative/story, cinematic, commercial/marketing, action, sci-fi, educational, and daily-life/vlog tasks (Figure 2). They are curated and restructured from online creation data or hand-crafted to cover diverse long-form generation challenges. Each JSON entry specifies duration, genre, and creative intent, followed by script, visual, and audio requirements. Script fields encode story structure, per-shot dialogue, narration, and timing; visual fields describe shots, camera motion, lighting, and identity consistency; audio fields specify dialogue, lip-sync, music, sound effects, and language. Appendix D provides an abbreviated example. For cases that do not depend on culture-specific knowledge or context, multilingual and cross-lingual versions are also provided and verified by professional translators.

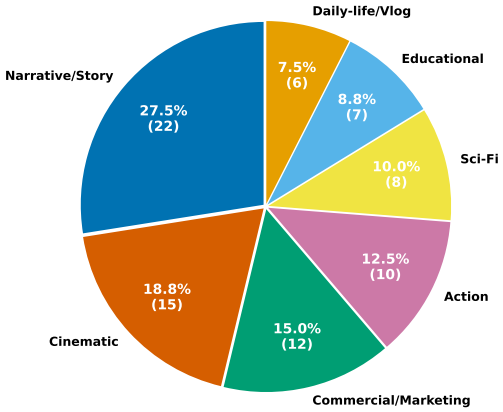

Figure 2: Distribution of the metadata entries across the various video categories.

Profile Provenance and Scope. Seven default profiles were constructed from recurring patterns in Dreamina video-creation sessions. LLM-based intent analysis extracted creation goals, quality priorities, and stylistic tendencies; professional video creators reviewed and consolidated the patterns into profile definitions. These profiles are not only from practical archetypes, but also a statistically representative survey of creators. Keeping profiles separate from metadata lets users substitute a different population's preferences without redefining the task set.

Each profile contains four priority weights for story, visual presentation, audio/emotion, and synchronization, together with taste descriptors, constraints, expertise, and expression style. Stability is reported separately as a dimension; it is not excluded from the overall score. Appendix E gives the weights and profile descriptions.

Profile-Conditioned Prompts. An LLM translates each metadata–profile pair into a concrete creation directive, using the metadata to specify required content and the profile to guide its presentation. Higher-priority dimensions receive greater elaboration, taste descriptors shape aesthetic emphasis, and expertise and expression style guide wording. All required facts must remain consistent across profiles. Prompt checks reject directives with missing task statements, omitted requirements, excessive length, or mismatches between the specified expertise and language. Profile conditioning thus influences video generation in addition to evaluation-time weighting.

In DirectorBench, personalization is expressed through both priority weights and profile-specific taste descriptors. Priority weights determine the contribution of each evaluation dimension, while taste descriptors guide creation directives and preference-sensitive judgments. Profiles with identical

weights can therefore express distinct stylistic goals, such as precise cinematic control or imaginative visual effects. An audio-focused profile can emphasize emotional delivery and musical expression, giving personalization a concrete role in shaping content. By connecting these preferences to prompt construction and evaluation criteria, DirectorBench provides a consistent mechanism for incorporating user priorities into both video generation and quality assessment.

2.3 Dynamic Checkpoint Rubric

A fundamental challenge in evaluating diverse long-form videos with a unified rubric is *applicability*: a lip-sync checkpoint is meaningless for a landscape video without characters, while a story-arc checkpoint is irrelevant for a music visualization. To solve this, we introduce a dynamic checkpoint rubric in DirectorBench with the following mechanisms.

Content-Adaptive Checkpoint Selection. Long-form videos vary substantially in their content and modalities, making many evaluation criteria inapplicable to a given video. Applying all criteria indiscriminately would introduce irrelevant judgments and distort the resulting scores. To address this, DirectorBench dynamically selects applicable checkpoints from a fixed registry, where each checkpoint defines an atomic rubric-scored criterion (Liu et al., 2026; Jia et al., 2026; Deng et al., 2024). Importantly, "dynamic" refers only to checkpoint applicability: the checkpoint definitions and scoring anchors remain fixed across videos. The registry contains 40 checkpoints organized into 9 rubric groups and 5 dimensions, including 35 ordinal (1–5) and 5 binary criteria, with all results normalized to $[0, 1]$ (Appendix F). To determine which checkpoints apply, DirectorBench first performs content profiling: a vision-language model rapidly analyzes representative frames and produces a structured Content Profile with 18 semantic attributes (Appendix G) describing entities, scene structure, stylistic properties, and motion characteristics. The profile is further enriched with signals from automatic speech recognition. These content attributes, together with the checkpoint registry and weighting scheme, were informed by professional production criteria and refined through pilot annotation and workflow analysis.

Requirement-Preserving Activation. Content-adaptive selection must also account for what the video was required to contain. DirectorBench specifies a metadata-first policy that separates required checkpoints from optional, content-dependent checks. Required checkpoints are determined by the task metadata and remain applicable even when the corresponding content is absent; optional checkpoints are activated according to observed content. For example, if dialogue and lip-sync are requested but no speaking character is generated, the omission is recorded as a requirement failure rather than excluded as irrelevant. Evidence availability is handled separately: a tool failure calls for an uncertainty flag, whereas an absent requested element calls for a requirement-failure finding. This policy prevents missing requirements from being treated as successful checks or automatically discarded as N/A. Appendix B distinguishes this intended policy from the output-based routing in the public implementation, which does not by itself guarantee retention of all metadata-required checks.

2.4 Multi-Agent Evaluation and Aggregation

The evaluation pipeline is implemented as a directed acyclic graph (DAG) and decomposed into four sequential phases. **Preprocessing** extracts shot boundaries, representative frames, transition signals, audio tracks, and transcripts, recording tool success, fallback, failure, and latency. **Specialist evaluation** runs script, video, audio, and stability agents in parallel. Each receives the task, relevant content, and available tool traces. **Cross-modal evaluation** then examines text–video, video–audio, and text–audio consistency, using specialist findings as evidence candidates alongside original inputs. **Synthesis** aggregates results and produces a bottleneck list, low-confidence flags, and an evidence-linked narrative report.

Tool measurements are auxiliary evidence, not direct ground truth for narrative quality or synchronization. For example, mouth-motion/audio-energy correlation is only a lightweight synchronization proxy. Tool failures and fallback states inform confidence, but these confidence values are heuristics rather than calibrated probabilities. Stable evaluator outputs can still be systematically wrong; exposing traces makes those errors inspectable.

Table 1: Human alignment and evaluator reliability. These are point estimates on the studied samples; different rows measure different units of agreement.

Comparison	Statistic	Result
Human criterion ranking, cohort level	Spearman ρ	0.806
Human overall-quality ranking, case level	Spearman ρ / Kendall τ	0.739 / 0.705
Human case comparisons	Pairwise ranking accuracy	84.3%
Within-case bottleneck profiles	Mean correlation	0.640
Human weakest criterion	Top-1 accuracy / top-3 coverage	64.1% / 96.5%
Blind human cross-rating	ICC	0.871
Repeated automatic evaluations	Checkpoint agreement α	0.828
Repeated checkpoint activation	Jaccard similarity	0.994

Let $\mathcal{R}_d$ be the returned report-metric results for dimension d, with score s_r and confidence c_r. The implementation computes

$$S_d = \frac{\sum_{r \in \mathcal{R}_d} c_r s_r}{\sum_{r \in \mathcal{R}_d} c_r}, \qquad S_{\text{overall}} = \frac{\sum_{d \in \mathcal{D}_{\text{obs}}} \widetilde{w}_d S_d}{\sum_{d \in \mathcal{D}_{\text{obs}}} \widetilde{w}_d}, \tag{2}$$

where $\mathcal{D}_{\text{obs}}$ contains dimensions with returned scores. If all confidences within a dimension are zero, its arithmetic mean is used. For profile weights (w_s, w_v, w_a, w_c), the five effective weights are $(w_s, w_v, w_a, w_c, w_v)$ for script, video, audio, cross-modal, and stability. Renormalization accounts for the repeated visual weight and absent dimensions. This weighting convention is explicit so that the overall score is reproducible; the visual weight is not its final percentage contribution when stability is present. Constraint violations are reflected in relevant requirement results and exposed in the report, rather than implemented as a universal binary rejection gate.

Diagnostic Report. The overall score summarizes a declared evaluation perspective, while the dimension vector and evidence identify what to inspect. For example, high requirement fulfillment with weak transitions calls for inspecting shot boundaries; high visual quality with weak video–audio consistency calls for inspecting synchronization. These are interpretations of observed output properties, not proof of an internal failure cause. Two systems can have similar overall scores while requiring different corrective actions. A comparison should therefore retain the profile, active metric set, and low-confidence flags with the scalar score, especially when content-dependent routing changes which dimensions contribute.

3 Benchmark Validation and Experimental Setup

Before using DirectorBench to compare generation systems, we examine whether its diagnostic outputs reflect human-perceived video quality. We then describe the generation systems, evaluation tools, and computational cost to establish the scope and conditions of the subsequent experiments.

3.1 Human Alignment and Repeatability

Fourteen annotators rated generated videos on a 1–5 scale using 16 perceptual criteria covering consistency, plausibility, instruction following, plot logic, continuity, camera/action quality, timing, and structural stability. Human criteria are compared with the evaluator at the dimension or report-metric level, rather than requiring a one-to-one reproduction of the 40 automatic checkpoints. Appendix A lists the criteria and separates human agreement, automatic–human alignment, and evaluator repeatability.

Table 1 provides quantitative support for ranking and bottleneck localization. Case-level Spearman correlation is 0.739, while pairwise ordering agrees with human judgments in 84.3% of comparisons. The automatic top-one bottleneck matches the human weakest criterion in 64.1% of cases, increasing to 96.5% when the report is used as a top-three shortlist. Thus, the diagnostic output is better supported as a set of candidates for inspection than as an infallible single explanation. Cohort-level agreement ($\rho = 0.806$) describes the ranking of perceptual criteria and should not be substituted for case-level agreement. Blind cross-rating yields ICC 0.871. Repeated evaluator runs produce checkpoint agreement $\alpha = 0.828$ and activation Jaccard 0.994.

Taken together, the human study provides empirical support for the accuracy and diagnostic effectiveness of DirectorBench on the evaluated corpus. Agreement with human quality rankings supports its ability to distinguish stronger from weaker outputs, while the high top-three bottleneck coverage shows that its diagnostic reports capture weaknesses identified by human viewers. Evaluator repeatability further supports the consistency of these measurements. These complementary findings justify using DirectorBench for comparative evaluation and targeted inspection of long-form video generation within the studied setting.

3.2 Systems, Tools, and Cost

Evaluation Configuration. The experimental assessor uses GPT-5.4 (OpenAI, 2026) across evaluator calls. Preprocessing combines ffmpeg/ffprobe, PySceneDetect (Castellano, 2014), OpenCV (Bradski, 2000), SSIM (Wang et al., 2004), color-histogram differences, and Farnebäck optical flow (Farnebäck, 2003). MobileViCLIP-Small (Yang et al., 2025) supplies text–video similarity, Sentence-BERT (Reimers & Gurevych, 2019) supports transcript similarity, and Librosa (McFee et al., 2015) supports audio features. These tools do not independently verify higher-level judgments. Saved traces record available evidence and degraded execution. Appendix C gives implementation details and limitations of the released snapshot.

We compare MovieAgent (Wu et al., 2025), ViMax (Huang et al., 2026a), Dreamina Creation Agent[1], and Kling Canvas Agent[2]. These cover character-oriented planning, staged script/storyboard/shot synthesis, edit-action orchestration, and canvas-based creation. This is an observational comparison of complete systems: their renderers, tools, and other components also differ. It does not isolate workflow architecture as the causal source of every score difference. In a separate experiment, we vary six base LLMs within the same Dreamina workflow; model details are in Appendix H.

Evaluation Cost. The profiled pipeline averages 131.5K input tokens, 11.4K output tokens, 7.2 minutes, and an estimated USD 0.44 per video, with 36.0 of 40 checkpoints activated. These are corpus averages from the reported configuration, not fixed costs or current-price guarantees. Longer videos and more complex traces increase context and preprocessing work. We do not report a matched latency or human-alignment comparison with lightweight metrics; the additional expense buys structured evidence and multimodal diagnosis.

4 Experimental Results and Diagnostic Analyses

In this section we use DirectorBench to study three orthogonal questions about long-form video generation. *RQ1* asks how the design of the agentic workflow that drives a generator affects the quality of its output, and what bottlenecks survive current workflows. *RQ2* holds the workflow fixed and asks how much of the remaining headroom the base LLM that orchestrates the agent can recover. *RQ3* holds the generated videos themselves fixed and asks whether personalized evaluation reveals quality structure that a generic single-profile metric collapses.

4.1 RQ1: Workflow Signatures and Shared Bottlenecks

Figure 3 reveals different strength–weakness patterns. Dreamina scores 0.76 on user-demand fulfillment and 0.59 on stability, but only 0.22 on transition quality. Kling has stronger text–video consistency (0.71) and temporal coherence (0.49), with lower stability (0.36). ViMax's temporal coherence is 0.18, while MovieAgent's audio-related strengths coexist with low user-requirement consistency (0.08). These observations motivate examining how a workflow carries requirements and global state between stages. They do not prove that one architectural component caused a particular failure, since the systems differ in multiple respects.

The lowest pooled metrics concern relationships between units: transition quality (0.256), temporal coherence (0.405), and video–audio consistency (0.416). Even Kling, the strongest transition result, reaches only about 0.36. In comparison, user-demand fulfillment is 0.70, text–video consistency 0.63, and stability 0.57. An acceptable aggregate score can therefore coexist with weak cross-shot

[1] https://jimeng.jianying.com/ai-tool/generate
[2] https://kling.ai/canvas/home

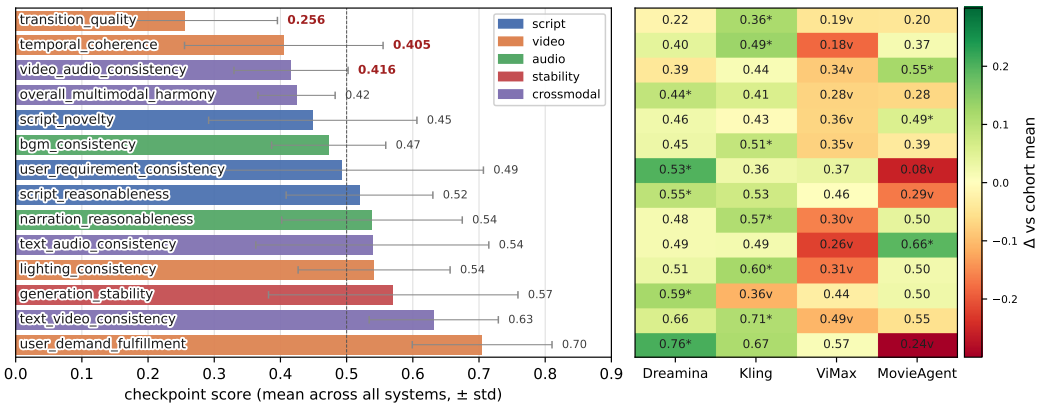

Figure 3: RQ1: Workflow signatures and cohort-wide bottlenecks in long-form video generation. **Left**: Mean checkpoint scores across all four workflows, sorted ascending, with population standard deviation error bars. **Right**: Heatmap of each workflow's deviation from the four-workflow cohort mean (Δ). Asterisks mark the cohort-best score per row.

assembly or audiovisual coordination. Genre-level summaries also differ: Action averages 0.455 overall, Cinematic/Shot-Design 0.509, Commercial/Marketing 0.484, and Education/News 0.489. These descriptive averages motivate targeted analysis of action-heavy continuity

4.2 RQ2: Base Agent LLMs within a Fixed Workflow

RQ2 isolates the effect of the base LLM that orchestrates the agent. We use the model-variation bucket of DirectorBench: the Dreamina edit-action workflow is held fixed, the same input prompts are shared across all variants, and only the base LLM that drives the skills and MCP tool-chain is swapped. Each model is evaluated under the equal-weight neutral profile. We evaluate six representative proprietary chat-oriented LLMs (ByteDance Seed, 2026; Z.ai, 2026; Anthropic, 2026; OpenAI, 2026; Moonshot AI, 2026; MiniMax, 2026). These models are treated as interchangeable controllers of the same agentic workflow, allowing us to isolate the effect of the base LLM from workflow-level components. Detailed descriptions are provided in Appendix H.

Narrative dimensions dominate the variation across base LLMs. Base-LLM choice primarily affects reasoning-intensive stages, while leaving system-level components unchanged. Figure 4 shows a column-normalized heatmap of dimension-level performance. The visual, audio, and stability dimensions exhibit low contrast: most models appear in similar color bands, indicating minimal differentiation. This confirms that these dimensions are governed by shared workflow components and are largely insensitive to the choice of base LLM. In contrast, the script and cross-modal dimensions display substantially higher contrast, with clear separation between models. This indicates that the primary variation across base LLMs lies in narrative reasoning and multi-modal coordination, which are directly handled by the LLM.

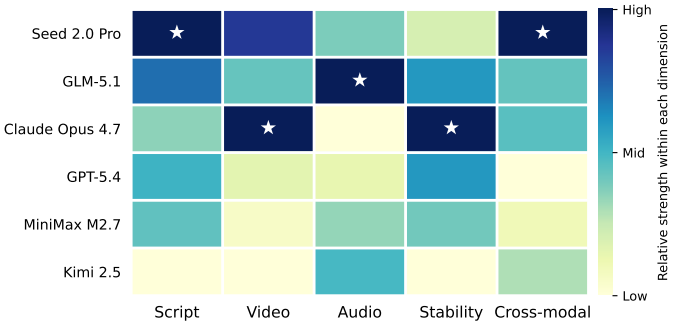

Figure 4: Relative performance across dimensions (column-normalized). Darker indicates stronger performance; stars mark per-dimension best.

Base LLMs exhibit complementary specialization rather than a strict ranking. No single model consistently dominates across all dimensions. Instead, models show complementary strengths across different dimensions (Figure 4). For example, *Seed 2.0 Pro* (ByteDance Seed, 2026) leads on script and cross-modal reasoning, *GLM-5.1* (Z.ai, 2026) on audio, *Claude Opus 4.7* (Anthropic, 2026) on

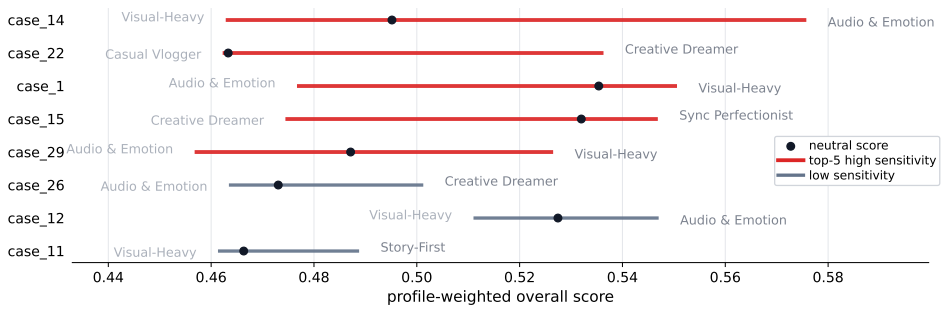

Figure 5: Personalized benchmarking changes the score assigned to the same metadata intent. Each segment shows the score range obtained under seven user profiles; dots indicate the neutral equal-weight score. We show representative high- and low-sensitivity cases, demonstrating that generic evaluation hides substantial user-dependent variation.

video and stability, and *GPT-5.4* (OpenAI, 2026) on stability robustness. The pattern demonstrates that base LLMs specialize in different aspects of the pipeline rather than forming a strict global ranking. From a practical perspective, this implies that model selection can be task-dependent: applications prioritizing narrative quality favor script-strong models, while synchronization-sensitive tasks benefit from cross-modal specialists.

The above analysis reveals that the base LLM acts primarily as a *narrative and coordination modulator* within a fixed agentic workflow . Its impact is concentrated on script generation and cross-modal alignment, while leaving system-level dimensions (visual quality, audio fidelity, and stability) largely unchanged. Compared to the substantial variation induced by workflow design in RQ1, the effect of base-LLM substitution is secondary. This suggests a clear optimization priority: improvements should focus on workflow architecture before model selection, unless the application is explicitly narrative-sensitive.

4.3 RQ3: Scoring Sensitivity and Generation-Side Effects

RQ3 asks whether such a generic protocol is sufficient when users express different creation preferences. Under such setting, we hold the metadata entry, Dreamina agentic workflow, and base LLM fixed, **but only** vary the user profile.

Evaluation-Time Sensitivity. We first evaluate the same generated videos under seven user profiles. Figure 5 shows selected high- and low-sensitivity cases, with a neutral equal-priority score as reference. Some profile ranges exceed ten percentage points. This variation demonstrates that user profile choice alone can substantially change evaluation outcome with even the same task intent. Meanwhile, under equal-weight aggregation, each case is assigned a single neutral score, implicitly assuming that all users value dimensions equally. However, the wide intervals in Figure 5 show that this assumption hides important profile-specific behavior. A profile that emphasizes audio and emotional tone may reward a case differently from a profile that prioritizes visual control or narrative structure. Generic evaluation therefore averages away the very preference-dependent quality signals that personalized video creation is meant to capture.

Table 2: Profile-conditioned generation: ASR characters/minute on two systems. Speech density measures behavioral change.

Profile	**Dreamina**	**Kling**
Visual-Heavy	206.7	178.7
Audio & Emotion	92.7	165.7
Creative Dreamer	88.2	102.0
Sync Perfectionist	86.1	136.0
Casual Vlogger	83.2	126.0
Story-First	24.4	67.3

Generation-Time Intervention. To test whether profiles affect output content, we additionally keep metadata and generation configuration fixed within Dreamina and Kling and change only the profile-conditioned prompt. Table 2 reports ASR-transcribed characters per minute, a duration-normalized speech-content measure with the six profiles presented.

Speech density varies from 24.4 to 206.7 characters/minute on Dreamina and 67.3 to 178.7 on Kling. Story-First is lowest and Visual-Heavy highest on both systems. The latter observation is

particularly instructive: visual emphasis does not imply low speech density, so this metric cannot stand in for profile fulfillment. The experiment supports the claim that profiles are behaviorally active during generation, beyond post-hoc rescoring. A complementary video-language-model inspection found multi-shot narrative structure for Story-First, dialogue-present live-action content for Audio & Emotion, stylized effects with little dialogue for Creative Dreamer, and fewer but more deliberately designed camera shots for Visual-Heavy. These qualitative observations connect prompt emphasis to different video characteristics.

5 Related Work

Existing work evaluates generative quality through complementary metrics and benchmarks. Inception Score (Salimans et al., 2016) and Fréchet Inception Distance (Heusel et al., 2017) assess image-generation quality, while Fréchet Video Distance (Unterthiner et al., 2018) compares video distributions using spatiotemporal representations. These aggregate measurements provide useful indicators of generative performance, but offer limited information about where specific requirements fail within an individual video. VBench (Huang et al., 2024) and its extensions (Zheng et al., 2025; Huang et al., 2026b) provide more fine-grained assessments of subject consistency, motion, visual quality, prompt alignment, and faithfulness. Alongside these metric-based approaches, LLM-as-judge methods (Gu et al., 2026) enable evaluation through explicit criteria and structured reasoning. VideoBench (Han et al., 2025) uses multimodal language models with few-shot scoring and chain-of-query evaluation to improve agreement with human judgments. EvalCrafter (Liu et al., 2024) combines 700 prompts with 17 metrics spanning visual, content, motion, and text–video properties, and fits aggregation coefficients to human opinions. T2V-CompBench (Sun et al., 2025) further examines compositional requirements through 1,400 prompts across seven categories, using model-, detection-, and tracking-based measurements. Together, these efforts establish complementary perspectives on perceptual quality, human alignment, and requirement fulfillment. Meanwhile, advances in text-to-video synthesis (Singer et al., 2023a; Villegas et al., 2023; Kondratyuk et al., 2024), multi-shot generation (Meng et al., 2026), and multi-agent workflows (Wu et al., 2025; Huang et al., 2026a; Wang et al., 2026; Wei et al., 2025) increasingly support the creation of coordinated video sequences. Evaluating such outputs requires connecting individual-shot quality with narrative continuity, cross-shot transitions, and audiovisual coordination, as well as identifying how failures in these aspects affect the assembled video.

Positioning of Our Work. DirectorBench complements existing benchmarks with structured, profile-aware long-form diagnosis. Compared with single holistic LLM judgments, it combines explicit checkpoints, specialist agents, and tool-supported evidence, aiming to improve reliability and transferability across generation workflows. Our human study supports the perceptual validity of quality rankings and bottleneck reports, while repeated evaluations demonstrate consistency on the studied corpus. This combination extends established quality measurements with interpretable diagnosis rather than replacing them.

6 Conclusion

DirectorBench provides a structured, profile-aware framework for diagnosing long-form video generation. It separates metadata-defined task requirements from user preferences and combines a fixed checkpoint rubric with specialist agents and tool-supported evidence. The resulting reports connect overall scores with dimension-level assessments and localized bottlenecks, making evaluation outcomes easier to interpret and inspect. Human validation supports the framework's ability to rank video quality and identify candidate weaknesses that agree with human observations, while repeated evaluations demonstrate consistency on the studied corpus.

Our experiments reveal recurring weaknesses in cross-shot transitions and cross-modal coordination, showing that strong individual components do not necessarily yield a coherent assembled video. Profile analyses further distinguish sensitivity to evaluation perspectives from changes introduced during generation: rescoring exposes preference-dependent quality variation, while controlled prompting on two systems demonstrates that profile information can influence generated content. Together, these findings support evaluating long-form outputs through both explicit task requirements and clearly specified user priorities.

AI USE STATEMENT

AI-assisted tools were only used for language polishing. The authors retain full responsibility for the final manuscript. Meanwhile, LLMs also form part of the research pipeline, including profile construction, prompt generation, automated evaluation, and qualitative video analysis, as described in the paper.

ETHICS STATEMENT

DirectorBench is intended to support research on the quality and limitations of generated videos. Its scores do not establish content authenticity, safety, or suitability for deployment. Automated judgments may inherit biases from underlying models and tools, while the curated metadata and user profiles do not represent all cultural contexts or individual preferences. Evaluation results should therefore be interpreted within the documented study setting and accompanied by human inspection where appropriate. Responsible use of the benchmark requires attention to data provenance, privacy, applicable content licenses, and the potential misuse of synthetic media.

REPRODUCIBILITY STATEMENT

The evaluation framework and experimental configuration are described in Sections 2 and 3. The appendices provide the human evaluation criteria, profile definitions, and implementation details, including the scope and limitations of the released snapshot. Source code and released metadata and profile files are available at `https://anonymous.4open.science/r/DirectorBench-835E/`. These materials support inspection and reimplementation of the evaluation procedure. Exact reproduction of numerical results additionally depends on the corresponding study-version data, human annotations, and external model versions; the current availability and version constraints are documented in Appendix C.

REFERENCES

Anthropic. Introducing Claude Opus 4.7, 2026. URL `https://www.anthropic.com/news/claude-opus-4-7`. Accessed 2026-09-26.

Gary Bradski. The OpenCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI technical report, 2024. URL `https://openai.com/index/video-generation-models-as-world-simulators/`.

ByteDance Seed. Seed 2.0 official launch, 2026. URL `https://seed.bytedance.com/en/blog/seed-2-0-official-launch`. Accessed 2026-09-26.

Brandon Castellano. PySceneDetect: Intelligent scene cut detection and video splitting tool, 2014. URL `https://github.com/Breakthrough/PySceneDetect`. Software; accessed 2026-09-26.

Shihan Deng, Weikai Xu, Hongda Sun, Wei Liu, Tao Tan, Jianfeng Liu, Ang Li, Jian Luan, Bin Wang, Rui Yan, and Shuo Shang. Mobile-bench: An evaluation benchmark for llm-based mobile agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 8813–8831. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.478. URL `https://doi.org/10.18653/v1/2024.acl-long.478`.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, volume 2749 of *Lecture Notes in Computer Science*, pp. 363–370. Springer, 2003. URL `https://www.ida.liu.se/ext/WITAS-ev/Computer_Vision_Technologies/PaperInfo/farneback03.html`.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. A survey on LLM-as-a-Judge. *The Innovation*, 7 (6):101253, 2026. doi: 10.1016/j.xinn.2025.101253. URL https://doi.org/10.1016/j.xinn.2025.101253.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-bench: Human-aligned video generation benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18858–18868. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01757. URL https://openaccess.thecvf.com/content/CVPR2025/html/Han_Video-Bench_Human-Aligned_Video_Generation_Benchmark_CVPR_2025_paper.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 21807–21818. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

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

Mengzhao Jia, Zhihan Zhang, Ignacio Cases, Zheyuan Liu, Meng Jiang, and Peng Qi. Autorubric: Rubric-based generative rewards for faithful multimodal reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 25707–25724. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.1282. URL https://doi.org/10.18653/v1/2026.findings-acl.1282.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25105–25124. PMLR, 2024. URL https://proceedings.mlr.press/v235/kondratyuk24a.html.

Xue Liu, Xin Ma, Yuxin Ma, Yongchang Peng, Duo Wang, Zhoufutu Wen, Ge Zhang, Kaiyuan Zhang, Xinyu Chen, Yida Ding, Tianci He, Jiani Hou, Liang Hu, Ziyun Huang, Yongzhe Hui, Jianpeng Jiao, Chennan Ju, Yingru Kong, Yiran Li, Jiashuo Liu, Mengyun Liu, Luyao Ma, Fei Ni, Yiqing Ni, Pengbo Niu, Yueyan Qiu, Yanle Ren, Xinyu Shen, Zilin Shi, Zaiyuan Wang, Wenjie Yue, Chun Zhang, Shiyu Zhang, Xinyi Zhang, Kaiwen Zhao, Zhenwei Zhu, Shanshan Wu, Qi Zhao, and Wenhao Huang. Xpertbench: Expert level tasks with rubrics-based evaluation. *CoRR*, abs/2604.02368, 2026. doi: 10.48550/ARXIV.2604.02368. URL https://doi.org/10.48550/arXiv.2604.02368.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EvalCrafter_Benchmarking_and_Evaluating_Large_Video_Generation_Models_CVPR_2024_paper.html.

Qiyao Ma, Dechen Gao, Rui Cai, Boqi Zhao, Hanchu Zhou, Junshan Zhang, and Zhe Zhao. Personalized rewardbench: Evaluating reward models with human aligned personalization. *arXiv preprint arXiv:2604.07343*, 2026. URL https://arxiv.org/abs/2604.07343.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In Kathryn D. Huff and James Bergstra (eds.), *Proceedings of the 14th Python in Science Conference, SciPy 2015, Austin, Texas, USA, July 6-12, 2015*, pp. 18–24. scipy.org, 2015. doi: 10.25080/MAJORA-7B98E3ED-003. URL https://doi.org/10.25080/Majora-7b98e3ed-003.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, Yanhong Zeng, Xing Zhu, Yujun Shen, and Huamin Qu. HoloCine: Holistic generation of cinematic multi-shot long video narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 461–471, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Meng_HoloCine_Holistic_Generation_of_Cinematic_Multi-Shot_Long_Video_Narratives_CVPR_2026_paper.html.

MiniMax. MiniMax M2.7: Early echoes of self-evolution, 2026. URL https://www.minimax.io/news/minimax-m27-en. Accessed 2026-09-26.

Moonshot AI. Kimi K2.5: Visual agentic intelligence, 2026. URL https://www.kimi.com/en/blog/kimi-k2-5. Accessed 2026-09-26.

OpenAI. Introducing GPT-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4/. Accessed 2026-09-26.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k. *CoRR*, abs/2503.09642, 2025a. doi: 10.48550/ARXIV.2503.09642. URL https://doi.org/10.48550/arXiv.2503.09642.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=4GSOESJrk6.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1410. URL https://aclanthology.org/D19-1410/.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2226–2234, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/forum?id=nJfylDvgzlq.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/forum?id=nJfylDvgzlq.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-video_Generation_CVPR_2025_paper.html.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=vOEXS39nOF.

Qian Wang, Ziqi Huang, Ruoxi Jia, Paul Debevec, and Ning Yu. Mavis: A multi-agent framework for long-sequence video storytelling. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pp. 2273–2295. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.eacl-long.101/.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Zheng Wei, Mingchen Li, Zeqian Zhang, Ruibin Yuan, Pan Hui, Huamin Qu, James Evans, Maneesh Agrawala, and Anyi Rao. Hollywood town: Long-video generation via cross-modal multi-agent orchestration. *CoRR*, abs/2510.22431, 2025. doi: 10.48550/ARXIV.2510.22431. URL https://doi.org/10.48550/arXiv.2510.22431.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *CoRR*, abs/2503.07314, 2025. doi: 10.48550/ARXIV.2503.07314. URL https://doi.org/10.48550/arXiv.2503.07314.

Yihan Wu, Ruihua Song, Xu Chen, Hao Jiang, Zhao Cao, and Jin Yu. Understanding human preferences: Towards more personalized video to text generation. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee (eds.), *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pp. 3952–3963. ACM, 2024. doi: 10.1145/3589334.3645711. URL https://doi.org/10.1145/3589334.3645711.

Jiazheng Xing, Fei Du, Hangjie Yuan, Pengwei Liu, Hongbin Xu, Hai Ci, Ruigang Niu, Weihua Chen, Fan Wang, and Yong Liu. LumosX: Relate any identities with their attributes for personalized video generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ee6e771ce962badeea34b918d3f962dd-Abstract-Conference.html.

Min Yang, Zihan Jia, Zhilin Dai, Sheng Guo, and Limin Wang. MobileViCLIP: An efficient video-text model for mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20824–20835, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_MobileViCLIP_An_Efficient_Video-Text_Model_for_Mobile_Devices_ICCV_2025_paper.html.

Z.ai. GLM-5.1: Overview, 2026. URL https://docs.z.ai/guides/llm/glm-5.1. Accessed 2026-09-26.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *CoRR*, abs/2503.21755, 2025. doi: 10.48550/ARXIV.2503.21755. URL https://doi.org/10.48550/arXiv.2503.21755.

A Human Validation: Measures and Interpretation

The human study uses 14 annotators and a 1–5 Likert scale. The 16 perceptual criteria are character consistency, scene consistency, object consistency, voice consistency, style consistency, character-setting plausibility, scene plausibility, voice plausibility, instruction following, plot logic, element continuity, segment continuity, camera plausibility, action quality, duration allocation, and structural stability. These perceptual questions are compared with automatic evaluation at a coarser level than the fixed checkpoint registry.

The validation summaries answer different questions. Cohort-level Spearman correlation measures agreement in the ranking of the 16 perceptual criteria. Case-level Spearman and Kendall correlations assess overall-quality ordering between cases, while pairwise accuracy measures the fraction of human case orderings recovered. Within-case bottleneck correlation compares the pattern of strengths and weaknesses for a given case. Top-1 accuracy measures whether the automatic weakest item matches the human weakest item, whereas top-3 coverage measures whether the human bottleneck is among three automatic candidates. The latter should not be described as exact identification accuracy. ICC 0.871 summarizes blind human cross-rating; checkpoint agreement $\alpha = 0.828$ and activation Jaccard 0.994 summarize repeated automatic evaluations. Inter-annotator agreement, automatic repeatability, and automatic–human alignment must therefore remain separate.

B Applicability and Missing Requirements

The intended evaluation protocol distinguishes optional content from required content. A request to include speech, a character, a story arc, or a sound effect remains meaningful when generation omits it. The following policy prevents applicability filtering from rewarding such omissions:

Table 3: Metadata-first evaluation policy. The table specifies the intended behavior; output-only checkpoint routing is insufficient to enforce every row.

Required	Detected	Behavior
Yes	Yes	Activate and evaluate quality
Yes	No	Retain the requirement and penalize omission
No	Yes	Activate if the criterion is applicable
No	No	Mark as not applicable

For example, a metadata entry with dialogue and lip-sync requirements must not receive an implicit pass when no speaking character appears. Requirement fulfillment and applicable signal quality are separate: absence can establish failure to follow a request even when a low-level lip-sync measurement cannot be computed. Missing requirements should remain visible in the diagnostic report rather than being silently removed from its interpretation.

There is an implementation boundary in the public snapshot dated July 28, 2026. Its checkpoint filter selects from detected content attributes; the released filter does not itself accept metadata requirements. Requirement-following metrics can penalize omissions through evaluator instructions, but this is not the same as an explicit metadata override that guarantees retention of each required checkpoint. Accordingly, the metadata-first policy should not be mistaken for a mechanically verified property of the released output-only filter. Analyses using that snapshot must inspect requirement results and missing-checkpoint traces, and cannot assume that all unreturned checks were irrelevant.

C Implementation and Release Scope

Tool Evidence. Video probing and audio extraction use ffmpeg/ffprobe, shot detection uses PySceneDetect, and transition evidence includes SSIM, color-histogram differences, and Farnebäck optical flow (Farnebäck, 2003). These are low-level signals for subsequent semantic assessment, not complete measures of coherence. The lip-sync proxy compares mouth-region motion with audio energy; it is not an implementation of a learned audiovisual synchronization network. Sentence-BERT compares text with ASR transcripts, rather than directly embedding audio. Tool outcomes

and fallback paths should be retained with scores so that judgments supported by degraded evidence remain identifiable.

Rubrics, Confidence, and Aggregation. The registry contains 35 ordinal checkpoints and five binary ones: `object_permanence`, `motion_continuity`, `internal_consistency`, `scene_order`, and `duration_completeness`. Report metrics aggregate checkpoint evidence or return additional task-level judgments. The plotted 14 metrics therefore have a different granularity from the 40 registry entries; reference-script evaluation can additionally return `script_video_fidelity`. Within-dimension aggregation operates on returned metric scores, followed by profile-weighted aggregation across available dimensions as in Equation 2.

Confidence is heuristic. The checkpoint aggregation path uses a checkpoint-count rule, while other judgments can include model-reported confidence informed by tool-availability instructions. These values have not been established as calibrated correctness probabilities. Hard-constraint violations are emphasized in requirement-following evaluation and diagnostic reporting; the synthesis step does not implement a universal pass/fail gate. Applications that require strict rejection should therefore use explicit constraint checks instead of inferring acceptance from the overall score alone.

Cost and reproducibility limits. Per-video averages are 131.5K input tokens, 11.4K output tokens, 7.2 minutes, USD 0.44 estimated API cost, and 36.0 activated checkpoints. Token counts, latency, and cost depend on video length, content, tool execution, and pricing. These summaries do not give a latency distribution or a matched comparison with a plain LLM judge. Likewise, raw human annotations and generation repeat counts are not included in the public snapshot. The reported validation and personalization results should therefore be interpreted within the documented study, rather than as an independently reproduced deployment guarantee.

D METADATA ENTRY

The metadata entries are structured with JSON documents that encode the complete objective task specification of a target video. Each entry serves as the *ground-truth intent* against which generated videos are evaluated.

Below is an abbreviated example for a narrative video; the displayed two shots do not exhaust the requested 65-second story.


```
{
  "meta_id": "narrative_001",
  "duration_sec": 65.0,
  "video_type": "narrative",
  "main_instruction": "Generate a romantic story of a girl chasing a boy in the rain, featuring a complete three-act arc.",
  "// ======= Core Task Specification (Objective Facts) =======": "",
  "modality_details": {
    "text": {
      "story_arc": {
        "act1_setup": "The girl encounters the boy in the rain",
        "act2_conflict": "A misunderstanding erupts and the boy turns to leave",
        "act3_resolution": "They embrace and reconcile in the rain"
      },
      "script": [
        {
          "shot_id": 1,
          "duration_sec": 15,
          "dialogue": "Wait for me!",
          "narration": "Rain blurs her vision as she runs desperately ..."
        },
        {
          "shot_id": 2,
          "duration_sec": 20,
          "dialogue": "",
```

```
            "narration": "The misunderstanding hits like a sudden downpour..."
          }
        ],
        "tone_requirements": "romantic_bittersweet, emotional progression"
      },
      "visual": {
        "shots": [
          {
            "shot_id": 1,
            "description": "wide shot of the girl running in the rain",
            "camera_movement": "tracking",
            "lighting": "natural_rainy_glow"
          },
          {
            "shot_id": 2,
            "description": "close-up of tearful eyes",
            "camera_movement": "push_in",
            "lighting": "soft_dramatic"
          }
        ],
        "camera_requirements": ["tracking", "push_in", "orbiting", "handheld"],
        "consistency_requirements": ["character_identity", "clothing",
        "lighting_shadow", "spatial_layout"]
      },
      "audio": {
        "dialogue": true,
        "lip_sync": true,
        "bgm_style": "soft_piano_orchestral",
        "sound_effects": ["rain_ambient", "footsteps_in_puddle"],
        "tone_control": "emotional_buildup_to_warm_resolution",
        "multi_language": "zh_en_switch"
      }
    }
  }
}
```


Below we briefly explain the meaning and purpose of each key.

- **meta_id**: Unique identifier for this metadata entry (e.g., "narrative_001").
- **duration_sec**: Target duration of the generated video in seconds.
- **video_type**: High-level category of the video (e.g., narrative, action, commercial).
- **main_instruction**: A concise, high-level creative brief that summarizes the overall video concept.
- **modality_details**: The core block containing detailed requirements for the three modalities (text, visual, audio).
- **text**: Narrative and script-related requirements.
  - **story_arc**: High-level three-act story structure (setup, conflict, resolution).
  - **script**: Per-shot script details, including dialogue and narration text.
  - **tone_requirements**: Desired emotional tone and progression of the story.
- **visual**: Visual and cinematic requirements.
  - **shots**: Array of per-shot visual descriptions, each with shot content, camera movement, and lighting.
  - **camera_requirements**: List of required camera movement types that must appear.
  - **consistency_requirements**: List of visual consistency constraints (character identity, clothing, lighting, spatial layout, etc.).
- **audio**: Audio and synchronization requirements.

- **dialogue**: Boolean flag indicating whether spoken dialogue is required.
- **lip_sync**: Boolean flag indicating whether lip-sync is required.
- **bgm_style**: Desired style and mood of the background music.
- **sound_effects**: List of required sound effects.
- **tone_control**: Emotional tone progression for the audio track.
- **multi_language**: Optional language switching requirements.

E User Profile Specifications

DirectorBench defines seven distinct user profiles that represent different real-world preferences in long-form video creation. Each profile $\mathbf{u}$ consists of three core components:

- **Priority weights** $\mathbf{w} \in \mathbb{R}^4$ ($\sum_d w_d = 1$) that control the relative importance of four user-facing evaluation dimensions: narrative quality, visual cinematography, audio emotion, and cross-modal synchronization. The released implementation assigns the visual priority to both video and stability, then renormalizes over dimensions with returned scores (Equation 2). Stability has no independent profile parameter.
- **Hard constraints** that specify non-negotiable requirements. Violations are emphasized in the requirement-consistency judgment and diagnostic report; this is not a universal binary acceptance gate.
- **User taste** descriptors that capture fine-grained aesthetic preferences, expertise level, and expression style, which are used to generate stylistically distinct personalized prompts.

The priority weights for each profile are summarized in Table 4. To improve readability, we abbreviate the four evaluation dimensions as *Story*, *Visual*, *Audio*, and *Sync*, corresponding to text/story arc, visual/camera quality, audio/emotion quality, and cross-modal synchronization respectively.

Table 4: Priority weights of the seven user profiles across the four evaluation dimensions.

Profile	Story	Visual	Audio	Sync
Story-First User	0.55	0.15	0.15	0.15
Visual-Heavy User	0.15	0.50	0.10	0.25
Audio & Emotion	0.20	0.15	0.45	0.20
Sync Perfectionist	0.20	0.20	0.20	0.40
Creative Dreamer	0.25	0.30	0.25	0.20
Casual Vlogger	0.30	0.20	0.30	0.20
Detail Obsessive	0.25	0.30	0.25	0.20

Each profile represents a distinct preference pattern commonly observed in long-form video creation scenarios.

- **Story-First User.** Prioritizes narrative coherence, emotional progression, and causal storytelling. This profile values strong story arcs and logical structure over complex visual effects or cinematic camera movement.
- **Visual-Heavy User.** Emphasizes cinematic presentation, including camera movement, lighting quality, and visual aesthetics. Minor narrative imperfections are tolerated if the visual experience remains compelling.
- **Audio & Emotion User.** Focuses on emotional impact through narration, tone control, and background music. Lip-sync and emotional alignment between audio and visuals are treated as important requirements.
- **Sync Perfectionist.** Strongly prioritizes cross-modal consistency, including lip-sync accuracy, text–visual alignment, and synchronized audio–visual transitions.
- **Creative Dreamer.** Prefers imaginative and stylized content with unusual camera motion, dreamy transitions, and fantasy aesthetics. This profile encourages creative expression over strict realism.

- **Casual Vlogger.** Seeks natural, conversational, and easy-to-watch content. This profile favors balanced storytelling and emotional tone while avoiding overly complex cinematic techniques.
- **Detail Obsessive.** Demands high precision across all dimensions, with particular emphasis on lighting consistency, synchronization accuracy, and structural coherence throughout the video.

F CHECKPOINT TAXONOMY

DirectorBench adopts a hierarchical evaluation taxonomy composed of dimensions, sub-metrics, and checkpoints. The taxonomy is designed to support both high-level quality assessment and fine-grained bottleneck localization for long-form video generation. In total, the benchmark contains 40 checkpoints organized into 9 sub-metrics across 5 evaluation dimensions: *Script & Narrative*, *Visual*, *Audio*, *Cross-Modal*, and *Stability*.

The checkpoint hierarchy was iteratively refined with reference to common long-form video production and cinematic evaluation practices, followed by pilot annotation and workflow analysis to ensure broad coverage of representative failure modes while maintaining practical evaluation cost. Rather than evaluating all checkpoints uniformly, DirectorBench dynamically activates only checkpoints that are applicable to the current video according to detected content attributes; Appendix B distinguishes this implementation from the intended metadata-first protocol. This design avoids penalizing videos for irrelevant criteria while enabling targeted diagnosis across diverse generation scenarios.

Table 5 summarizes the complete checkpoint registry. Each sub-metric corresponds to a coherent aspect of video quality, while individual checkpoints capture finer-grained evaluation criteria within that aspect.

Table 5: Complete checkpoint registry in DirectorBench (40 checkpoints across 9 sub-metrics and 5 dimensions).

Dimension	Sub-metric	Checkpoints
Script & Narrative	Script Reasonableness (5)	event_chain_logic
		character_motivation
		pacing_structure
		internal_consistency
		dialogue_naturalness
	Narrative Richness (3)	premise_originality
		narrative_surprise
		visual_creativity
Visual	Temporal Coherence (7)	char_face_consistency
		char_clothing_consistency
		object_permanence
		background_consistency
		scale_proportion
		motion_continuity
		temporal_logic
	Lighting Consistency (4)	light_direction
		shadow_consistency
		color_temperature
		exposure_stability
Audio	BGM Consistency (4)	bgm_mood_match
		bgm_transition_smoothness
		bgm_tempo_pacing
		bgm_volume_balance
	Narration Quality (3)	speech_timing
		speech_emotion_fit
		speech_intelligibility
Cross-Modal	Text–Video Consistency (5)	scene_presence
		scene_order
		character_matching
		hallucinated_content
		setting_accuracy
	Video–Audio Synchronization (4)	lip_sync_quality
		sound_event_alignment
		audio_continuity
		ambient_sound_match
Stability	Generation Stability (5)	visual_artifact_frequency
		resolution_sharpness
		temporal_degradation
		color_consistency
		duration_completeness
Total		**40 checkpoints**

G CONTENT PROFILING ATTRIBUTES

To support dynamic checkpoint activation across diverse long-form video types, DirectorBench first constructs a lightweight `ContentProfile` describing the high-level semantic properties of the generated video. The content profile is designed to provide sufficient structural information for checkpoint routing while remaining computationally efficient for large-scale evaluation.

The current implementation uses a compact schema consisting of 18 attributes and auxiliary signals grouped into five categories:

- **Entity attributes**
  - `has_characters`
  - `character_count`
  - `has_dialogue`
  - `has_held_objects`
  - `has_animals`
- **Scene-structure attributes**
  - `scene_count`
  - `has_scene_changes`
  - `is_single_shot`
- **Style/content attributes**
  - `has_text_overlay`
  - `has_special_effects`
  - `is_live_action_style`
  - `is_animation_style`
  - `has_background_music`
- **Motion attributes**
  - `has_camera_movement`
  - `has_fast_motion`
  - `has_slow_motion`
- **Auxiliary fusion signals**
  - `vlm_char_count`
  - `asr_speaker_count`

Most visual attributes are inferred from a fast VLM pass over representative shot thumbnails extracted during preprocessing, while dialogue- and music-related cues are derived from ASR outputs and separated audio tracks. To improve robustness in multi-character or multi-speaker scenes, the final `character_count` is conservatively fused as

$$\texttt{character_count} = \max(\texttt{vlm_char_count}, \texttt{asr_speaker_count}). \tag{3}$$

The content profile is subsequently used during checkpoint activation to determine which evaluation criteria are applicable for the current video. For optional content, the output-based filter can activate lip-sync checks when speaking characters and dialogue are detected. Metadata-required but missing content must be distinguished from irrelevant content; Appendix B explains the protocol and the boundary of the released implementation.

The attribute taxonomy was iteratively designed with reference to common long-form video production characteristics and refined through pilot annotation and workflow analysis to ensure broad coverage of practical evaluation scenarios while maintaining lightweight execution cost.

H BASE LLM DETAILS

We include six proprietary large language models (LLMs) as base controllers in our study. All models are used in their chat-oriented configuration and interact with the same toolchain under identical workflow conditions. This setup allows us to isolate the effect of the base LLM while holding all system-level components fixed.

Selection rationale. The selected models cover a diverse set of commercial LLM families, spanning differences in training data, alignment strategies, multimodal capabilities, and inference efficiency. This diversity ensures that the observed behaviours are not specific to a single vendor or model design, but reflect broader patterns in how LLMs interact with agentic workflows.

Model descriptions. We include six representative proprietary LLMs as the base controllers for the agentic workflows evaluated in this study. The selected models cover a diverse range of reasoning styles, planning capabilities, tool-use behaviors, and multimodal interaction characteristics. All models are evaluated under the same workflow setting and evaluation protocol to isolate the effect of base-LLM substitution.

- **Seed 2.0 Pro (ByteDance Seed, 2026).** A proprietary multimodal LLM developed by ByteDance, designed for reasoning, planning, and content generation in multimedia pipelines. It features strong integration with video-oriented tasks and exhibits consistent performance in structured generation and cross-modal coordination.
- **GLM-5.1 (Z.ai, 2026).** A general-purpose LLM from Zhipu AI, optimized for multilingual reasoning and structured output generation. It is particularly competitive in dialogue and tool-use scenarios, with strengths in controllability and format adherence.
- **Claude Opus 4.7 (Anthropic, 2026).** A high-capability LLM from Anthropic, emphasizing reasoning depth, alignment, and stable long-context behavior. It is designed to produce consistent and reliable outputs under complex instructions, making it well-suited for multi-step planning tasks.
- **GPT-5.4 (OpenAI, 2026).** A frontier LLM from OpenAI, optimized for general reasoning, tool use, and robust execution across a wide range of tasks. It exhibits strong capability in instruction following and error recovery within agentic pipelines.
- **MiniMax M2.7 (MiniMax, 2026).** A multimodal-capable LLM from MiniMax, designed for efficient inference and interactive generation. It emphasizes latency-performance trade-offs and performs well in real-time or user-facing generation settings.
- **Kimi K2.5 (Moonshot AI, 2026).** A chat-oriented LLM from Moonshot AI, with strengths in long-context understanding and conversational coherence. It is particularly suited for extended dialogue and context-heavy reasoning tasks.

Experimental control. All models are accessed via their respective APIs using default decoding settings. We do not apply model-specific prompt engineering or tuning, and all models operate under the same prompt templates, tool interfaces, and workflow execution logic. This ensures that observed differences reflect intrinsic model behaviour under a controlled workflow, rather than artefacts of per-model optimisation.

Scope and limitations. We note that all evaluated models are proprietary systems accessed via remote APIs, and their internal architectures and training data are not fully disclosed. As such, our analysis focuses on observable behavioural differences rather than architectural explanations.