

FilmBench: A Film-Grade Benchmark for Cinematic Video Generation

Shengyi Wang^{†1,2} Niantong Li^{†1} Guangzheng Hu^{†1} Hong Qi^{*3} Fei Ding^{1,2} Weixu Qiao¹
 Jinlin Wang¹ Xiaotong Lv¹ Peng Han¹ Zimeng Li¹ Fanshu Ding¹ Yushu Wang¹
 Han Wu¹ Jingjing Chen¹ Chongxiao Wang^{1,2} Yanhao Wu^{1,2} Chenglong Huang^{1,2}
 Xiaoqian Zhu^{1,2} Jie Tian^{1,2} Hua Li^{1,2} Jingjing Fan^{1,2} Mingshuang Tang³
 Zhong Li³ Hengxia Qiang³ Weibin Chen¹ Jinyang Zhen¹
 Bing Zhao¹ Lin Qu¹ Jing Li^{*1,2} Hu Wei^{*1}

¹Alibaba Group

²Moku Lab, Hujing Digital Media & Entertainment Group

³Beijing Film Academy

[†] Equal contribution. ^{*} Corresponding authors.

qihong@bfa.edu.cn, kongwang@alibaba-inc.com, lj225205@alibaba-inc.com



<https://huggingface.co/datasets/skylenage/FilmBench>



<https://github.com/Neo-yk/FilmOps>

Abstract

Progress in video generation keeps narrowing the visual gap between AI-generated and professionally produced footage, yet most benchmarks still draw prompts from web sources or LLM templates and score them with untrained, generic multimodal models. More fundamentally, their evaluation taxonomies remain rudimentary (overall visual quality, coarse text alignment and temporal smoothness) rather than the professional Cinematic Language criteria by which films are actually made and judged, so they assess basic video plausibility rather than film-grade craft. We introduce **FilmBench**, a text-to-video (T2V) and reference-to-video (R2V) benchmark grounded in the professional Cinematic Language of the film-academy tradition and co-developed with directors and faculty from the Beijing Film Academy and the Hujing Digital Media & Entertainment Group film studio. It rests on three choices. First, prompts are *reverse-engineered* from clips of award-winning films spanning 20 cinematic genres and chosen by professional directors, so every prompt is anchored to a verified live-action reference; the prompts follow real shot lists, and most script multiple shots (1,056 of the 1,169 prompts are multi-shot), unlike prior single-clip benchmarks. Second, evaluation follows a three-level Cinematic taxonomy of **3 axes, 12 components and 35 (T2V) +3 (R2V-only) sub-metrics**. Third, we develop an in-house expert-grade automatic evaluation agent and open-source its core suite of Cinematic Language operators (FilmOps). Benchmarking leading video generation models (9 for T2V, 7 for R2V), the evaluator reproduces the human model ranking at model-level Spearman $\rho = 0.95$ (T2V) and 0.96 (R2V). Scores fall well below prior web-style benchmarks, with two consistent gaps in *dynamic aesthetics* and a marked *single-to multi-shot* performance drop that widens for weaker models.

1 Introduction

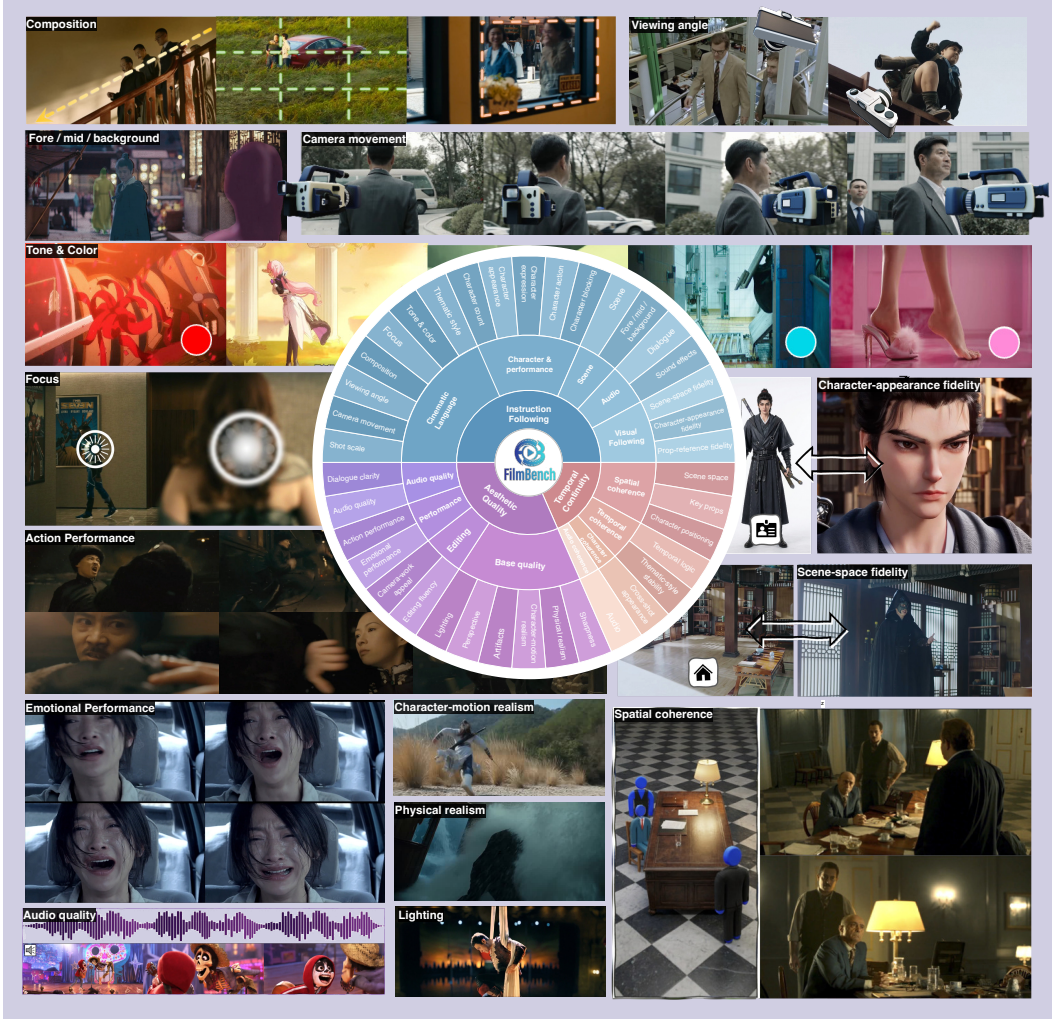


Figure 1: FilmBench evaluation taxonomy: 3 L1 axes, 12 L2 components, 35+3 (R2V-only) L3 sub-metrics, built from clips across 20 film genres curated by expert directors.

Video generation has progressed at an unprecedented pace. Diffusion- and DiT-based models (Veo [4, 24], Kling [12], Seedance [20], Hailuo [17], HappyHorse [6, 7] and others) now produce ~15-second photorealistic clips with controllable camera and even synchronized audio, blurring the line between AI-generated footage and professionally produced cinema. A natural next question is whether the capability of these models has reached the pass mark of *professional cinematic creation*. Modern film production follows a rigorous Cinematic Language (shot scale, camera movement, shot perspective, lighting, scene blocking, composition and film-grade audio design) refined over more than a century of practice. Existing benchmarks were not designed with this professional, academy-taught Cinematic Language in mind, and so far they paint only a coarse, optimistic picture of model capability.

The current benchmarking landscape, and what is missing. The community has produced a rich body of evaluation suites in the past three years. Foundational multi-dimensional benchmarks such as VBench [9], VBench-2.0 [29] and Video-Bench [5] dissect quality into 9–18 generic dimensions, and learned scorers such as VideoScore [8] push automatic evaluation closer to human preference. Compositional and fine-grained benchmarks (T2V-CompBench [22], FETV [15]) target attribute, motion



Figure 2: The director-driven reverse-engineering pipeline that turns award-winning film clips into film-grade prompts; see §3.1 for details.

and interaction binding; temporal, motion and multi-shot benchmarks (ChronoMagic-Bench [27], VMBench [14], SLVMEval [16], MSVBench [21]) probe long-horizon, motion and cross-shot fidelity; reasoning, physical and social benchmarks (TiViBench [2], SVBench [18], WorldJen [10], RBench [3]) stress higher-order or embodied capabilities; the I2V line (ConsistI2V/I2V-Bench [19], UI2V-Bench [28], IP-Bench [13]) focuses on image-conditioned generation; and audio–video and aesthetic benchmarks (AVGen-Bench [30], VGA-Bench [11]) address audio quality and visual aesthetics, respectively.

Despite this breadth, none of these benchmarks evaluates models against the standards used in real cinematic creation. They share three structural blind spots. First, prompts are sourced from web users, captioners or LLM templates rather than from verified professional shots, so the input distribution drifts away from cinematic content. Second, the implicit content distribution skews toward generic, English, web-style scenes, with no balanced coverage of cinematic genres or culturally diverse film traditions. Third, the evaluation taxonomies leave out cinematic axes such as shot scale, camera movement, lighting, visual style, character performance and audio quality, or collapse them into a single “aesthetic” score. As a consequence, current leaderboards saturate while professional cinematic creators readily reject the same outputs as not film-grade.

FilmBench. We address this gap with FilmBench, an evaluation benchmark for cinematic text-to-video (T2V) and reference-to-video (R2V) generation, built jointly with directors and faculty from the Beijing Film Academy and the film studio of Hujing Digital Media & Entertainment Group. FilmBench is grounded in three design principles.

Figure 1 gives an overview of the FilmBench evaluation taxonomy. Expert directors from the Beijing Film Academy selected clips spanning 20 film categories—covering the breadth of professional cinematic genres—and used them as the foundation for both the prompt set and the fine-grained evaluation hierarchy. The taxonomy radiates from the three L1 axes through progressively finer L2 components and L3 sub-metrics, while the surrounding film examples illustrate the kinds of cinematic-language challenges each dimension targets. We next describe how these expert-curated clips are turned into structured, film-grade prompts.



Figure 3: Fine-grained, per-dimension FilmBench evaluation on a multi-shot reference-scene (R2V) example reverse-engineered from *La La Land*: Seedance 2.0 (94.97) vs. Vidu Q2 Pro (73.61). Discussed in the text.

(P1) *Reverse-engineered from real films* (Figure 2): instead of authoring prompts in the abstract, professional directors select clips from award-winning films that stress Cinematic Language and visual expression, reverse-infer draft prompts with a strong multimodal model (Gemini 3.1 Pro), and refine them into structured shot-level prompts that explicitly encode scene, role, prop, shot scale, camera movement, composition, lighting, dialogue and performance, so that every prompt is anchored to a verified professional reference and every fine-grained criterion in the taxonomy is covered. Because these prompts follow real shot lists, most are multi-shot (1,056 of the 1,169 prompts), unlike the single-clip prompts of prior benchmarks. (P2) *Academy-aligned cinematic taxonomy*: our evaluation dimensions follow the Cinematic Language teaching system of the Beijing Film Academy, organized as *three* first-level axes (L1): *Instruction Following*, *Temporal Continuity* and *Aesthetic Quality*. These decompose into 12 second-level components (L2) and 35 third-level sub-metrics (L3). T2V and R2V share this main framework; the R2V task additionally adds a *Visual Following* L2 component of three L3 sub-metrics (scene-space, character-appearance and prop-reference fidelity), each evaluated against the conditioning reference image, giving R2V 13 L2 and 35+3 L3 (per prompt, only the sub-metric matching its reference type is scored). (P3) *Expert-grade automatic evaluator*: every sub-metric is scored by an in-house, professional film-grade evaluation agent whose core Cinematic Language operator suite (FilmOps) we open-source; its model-level ranking reproduces expert film-industry rankings at Spearman $\rho=0.95$ (T2V) / 0.96 (R2V) (§3, §4).

A worked example. Figure 3 shows FilmBench scoring a multi-shot reference-scene (R2V) example, a four-shot clip reverse-engineered from the musical *La La Land*, on Seedance 2.0 (94.97) and Vidu Q2 Pro (73.61). Because every L3 sub-metric is scored independently, the evaluation exposes behavior that an aggregate score hides. Although Seedance 2.0 leads overall by a wide margin, its *scene-space fidelity* is slightly lower (87.5 vs. 100): Vidu Q2 Pro over-adheres to the reference background image and scores lower on nearly every other dimension. It drops markedly on Cinematic Language following (shot scale 75 vs. 25, viewing angle 100 vs. 50, camera movement 100 vs. 50, composition 75 vs. 43.75) and, having sacrificed the shot/reverse-shot staging to lock onto the scene, exhibits character-positioning drift that drives its *character blocking* to 0. FilmBench can thus objectively credit a single dimension (reference fidelity) while still penalizing the broader loss of professional Cinematic Language; Appendix A discusses further per-dimension case studies.

Findings. We benchmark leading video-generation models (9 for T2V, 7 for R2V), including Seedance 2.0, HappyHorse 1.1/1.0, Kling 3.0, Kling 3.0 Omni, Veo 3.1, Grok Imagine Video [26], Vidu [1], and Hailuo 2.3, and distill five field-wide findings. **(i) No saturation:** no model approaches the ceiling on either task (top scores of 88.93 for T2V and 86.66 for R2V), in sharp contrast with the near-ceiling numbers on prior web-style benchmarks; and crucially, what separates the leaders is not generic quality but the professional Cinematic Language sub-metrics of Instruction Following (camera movement, shot scale, viewing angle, focus and composition) together with the Aesthetic Quality axis, which is exactly where FilmBench resolves differences that web-style benchmarks miss. **(ii) A field-wide dynamic-aesthetics bottleneck:** across all models the lowest-scoring sub-metrics are action performance, camera-work appeal, character-motion realism and emotional performance, whereas static image-quality sub-metrics (sharpness, lighting) score markedly higher: today’s models render clean frames but still struggle with believable motion and performance. **(iii) Multi-shot is the hardest regime:** on T2V, moving from single- to multi-shot prompts costs 7.9 points on average and up to 22.8 for the weakest model, concentrated in the Cinematic Language and editing demands of Instruction Following (framing, camera work and cut fluency); it is multi-shot staging, rather than single-take quality, where current models fall short. **(iv) Reference conditioning stresses rather than reorders the field:** across scene, prop and character references the R2V ranking stays stable, yet conditioning lowers scores to varying degrees and penalizes the weakest model most, and the visual-following leader is not necessarily the overall leader, so reference behavior is best read through ranking shifts rather than absolute per-type scores. **(v) No model wins on all L3 sub-metrics:** Seedance 2.0, the overall leader, claims only 18/35 (T2V) or 20/38 (R2V) L3 championships, concentrated on Cinematic Language sub-metrics (Cinematic Language, editing appeal and performance), while HappyHorse 1.1 dominates 11/35 (T2V) or 10/38 (R2V) sub-metrics on character, audio and scene dimensions; notably the three R2V-specific visual-following championships all go to the HappyHorse family; 4 (T2V) and 2 (R2V) models do not claim any championship. This “championship mismatch” reveals structural complementarity concealed by a single leaderboard number.

Contributions. We make three contributions. (i) We release **FilmBench**, a cinematic-grade T2V/R2V benchmark whose prompts are reverse-engineered from award-winning real films and curated jointly with directors and faculty from the Beijing Film Academy and a professional film studio, ensuring coverage of every fine-grained cinematic criterion. (ii) We release **FilmOps**, a modular Cinematic Language operator suite with trained weights and inference scripts, enabling the community to build customizable, expert-knowledge-driven judge agents for video evaluation. (iii) We propose an **academy-aligned evaluation taxonomy** that translates professional Cinematic Language into a three-level hierarchy of 3 L1 axes, 12 L2 components and 35 L3 sub-metrics (35+3 with the R2V-specific Visual-Following extension), scored by an expert-grade automatic evaluator with an open-source Cinematic Language operator suite. (iv) We conduct a comprehensive cinematic evaluation over leading T2V/R2V models, show the automatic evaluator reproduces expert rankings at $\rho=0.95$ (T2V) / 0.96 (R2V), and expose dynamic aesthetics as the shared bottleneck, discussing directions for future research.

2 Related Work

We trace how video-generation evaluation has evolved through three stages (from generic quality scoring, to capability-specific stress tests, to increasingly film-like assessment) and show that each stage, in closing one gap, exposes the next, until the accumulated gaps motivate FilmBench precisely.

From distribution metrics to multi-dimensional, human-aligned evaluation. Early evaluation repurposed action-recognition datasets (UCF-101, MSR-VTT, Kinetics) and reported single scalars such as FVD/IS, which neither localize *which* aspect of a video fails nor align well with human perception. As diffusion- and DiT-based generators—including the systems we evaluate (Veo-3.1, Kling-v3/Omni, Seedance-2.0, Hailuo-2.3, Vidu-Q3-Pro, Grok-Imagine-Video and HappyHorse-1.x)—raised the ceiling to controllable, audio-equipped clips, the community moved to disentangled, human-aligned protocols. VBench [9] and VBench-2.0 [29] decompose quality into 16 then 18 dimensions and push from *superficial* to *intrinsic* faithfulness; Video-Bench [5] enlists MLLMs as scalable judges; VideoScore [8] learns a human-aligned regressor. This stage established the grammar of modern evaluation—quality as a structured set of interpretable dimensions—but its axes are deliberately generic and its prompts are drawn from web users, reflecting what is easy to source

rather than what a film production demands. The first gap is thus one of *content and taxonomy*: neither the prompts nor the dimensions are anchored in a professional production grammar.

Specializing to emerging capabilities. With this foundation in place and baseline visual quality saturating, benchmarks specialized to isolate the *hard* capabilities that generic scores had masked. Compositional suites T2V-CompBench [22] and FETV [15] expose brittle attribute, motion and spatial-relation binding; the temporal–motion line ChronoMagic-Bench [27] and VMBench [14] target metamorphic amplitude and perception-aligned motion, while SLVMEval [16] extends reliability testing to hour-scale footage. As generators began stitching shots into a story, MSVBench [21] introduced multi-shot evaluation with a hybrid LMM-plus-expert scorer, distilling this stage’s lesson: current systems behave as *visual interpolators rather than world models*. A parallel reasoning wave then raised the bar from rendering to understanding, probing structural (TiViBench [2]), social (SVBench [18]), broad multi-axis (WorldJen [10]) and embodied-robotic (RBench [3]) reasoning. Yet even MSVBench scores its shots against generic quality criteria rather than a director’s intended shot list, so evaluation still stops short of the *film*: cross-shot narrative structure and director intent are not judged end-to-end against a verified cinematic reference. This is the second gap.

Toward film: reference, audio, aesthetics and Cinematic Language. The strand closest to our goal relaxes that assumption toward controllable, film-like generation, but does so one aspect at a time. Reference/image-conditioned suites (ConsistI2V/I2V-Bench [19], UI2V-Bench [28] and IP-Bench [13]) evaluate consistency, semantic understanding and protection from a conditioning image, raising *reference fidelity* as a concern. As generators acquired joint audio, AVGen-Bench [30] evaluates text-to-audio-video generation, adding a *cross-modal audio* axis, while VGA-Bench [11], co-developed with the Beijing Film Academy, brings aesthetic tagging into view. In parallel, a cinematic-understanding literature (MovieNet, shot-type taxonomies, CineScale/ShotBench) and storyboard-to-video pipelines treat film as structured language, and MovieBench [25] pushes to feature-length identity consistency—sustained in at most 53% of cases at multi-scene scope. Collectively these efforts touch nearly every ingredient of film, yet no single benchmark unifies them: aesthetics and audio are not anchored in a production grammar, reference fidelity and cross-shot audio–dialogue continuity are evaluated in isolation if at all, the target is typically real or arbitrary footage rather than *generated* films, and the cinematic-understanding line asks only whether models can *read* Cinematic Language, never whether a generator can *speak* it at a director’s granularity. Closing this third gap—a unified, production-grounded evaluation of *generated* film—is exactly the remit of FilmBench.

Positioning of FilmBench. These three accumulated gaps (a generic taxonomy, a clip-level scope, and fragmented coverage of individual film aspects) jointly define what a film-grade benchmark must satisfy at once. To the best of our knowledge, no existing benchmark simultaneously (i) reverse-engineers prompts from genuine, award-winning films selected by directors; (ii) spans a broad set of cinematic genres and both text-to-video and reference-to-video tasks, with prompts organized as real, mostly multi-shot shot lists (1,056 of 1,169) that evaluate the film rather than the isolated clip; (iii) grounds its taxonomy in the Cinematic Language system taught in professional film schools: 3 first-level axes, 12 second-level components and 35 third-level sub-metrics with an R2V-specific Visual-Following extension, co-designed with directors, Beijing Film Academy faculty and a professional film studio; and (iv) treats audio–dialogue continuity (including dialogue intelligibility and per-character timbre stability across shots) as a first-class axis. FilmBench thus asks not whether a model can *read* Cinematic Language but whether it can *speak* it, and because every prompt is reverse-engineered from a verified shot, we always retain both a structured grammar and a ground-truth reference. §3 and §4 detail this taxonomy and the automatic-plus-human evaluator stack, with a head-to-head comparison table in §3 and the appendix.

3 FilmBench: Construction and Dimensions

This section describes how FilmBench is built. We first present the reverse-engineering pipeline (Figure 2) that turns genuine cinematic clips into structured prompts (§3.1), then introduce our three-level Cinematic Language taxonomy of 3 first-level axes (L1), 12 second-level components (L2) and 35 third-level sub-metrics (L3), co-designed with faculty from the Beijing Film Academy and the Hujing Digital Media & Entertainment Group film studio (§3.2).

3.1 Reverse-Engineered Prompt Construction

FilmBench prompts are produced by a director-driven *reverse-engineering* pipeline with three stages (Figure 2). **Clip selection:** professional directors from the film studio of Hujing and the Beijing Film Academy prioritize *award-winning films* (e.g., winners of the Academy Awards, the Golden Horse Awards and the Hundred Flowers Awards) across our 20 cinematic genres, and select *multi-shot* clips (mostly ≥ 4 shots) that specifically stress Cinematic Language and visual-expression skill. **Draft prompt inference:** a FilmOps operator suite together with a strong multimodal model (Gemini 3.1 Pro) reverse-extracts the original narrative script, Cinematic Language elements and on-screen visual elements from each clip, which Gemini 3.1 Pro then assembles into an initial structured prompt. **Expert refinement:** the Beijing Film Academy directing team screens, quality-checks, rewrites and standardizes every draft into professional film-grade Cinematic Language. Throughout clip selection and final curation, the team ensures that *every L1 axis and L2 component in the taxonomy is covered*, so that the benchmark faithfully probes a model’s film-grade generation ability rather than generic visual quality. Each prompt uses slotted cinematic tags (@scene, @role, @prop), e.g., “Shot 1 (medium-shot, eye-level fixed, rule-of-thirds): in a brightly lit ICU @scene1, a young female doctor @role1 ...”.

Tasks and scale. FilmBench covers two tasks: **T2V** (prompt $\rightarrow$ video) and **R2V** (reference + prompt $\rightarrow$ video). The T2V suite contains **515 prompts** spanning **20 base cinematic genres**, balanced across Chinese Films and International Films markets (243 Chinese Films / 272 International Films). The R2V suite contains **654 prompts** split across three reference types (232 scenes / 209 props / 213 characters) and drawn from the same 20-genre pool across both markets (239 Chinese Films / 415 International Films). Because prompts follow real shot lists, most are *multi-shot*: 402 of the 515 T2V prompts and all 654 R2V prompts script multiple shots (1,056 of 1,169 overall), unlike the single-clip prompts of prior benchmarks. Each prompt is reverse-engineered from a distinct award-winning film clip, so the prompt count also reflects the number of source clips.

3.2 Academy-Aligned Evaluation Taxonomy

Our evaluation dimensions follow the academic Cinematic Language system taught in professional film schools. The taxonomy is *co-designed by faculty from the Beijing Film Academy and the film studio of Hujing together with an analysis of current video-generation models’ capabilities*, and is deliberately organized as a multi-level hierarchy so that conclusions and insights can be drawn at coarse (L1), medium (L2) and fine (L3) granularity.

Table 1: FilmBench three-level taxonomy (3 L1 axes, 12 L2 components, 35 L3 sub-metrics). R2V adds a *Visual Following* component (+3 L3). Full definitions in Appendix B.

L1 Axis	L2 Component
Instruction Following (16+3)	Cinematic Language
	Character & performance
	Scene
	Audio
Temporal Continuity (7)	<i>Visual Following (R2V only)</i>
	Spatial coherence
	Temporal coherence
	Character coherence
Aesthetic Quality (12)	Audio coherence
	Base quality
	Editing
	Performance
	Audio quality

Three L1 axes. FilmBench evaluates cinematic competence along three first-level axes, decomposed into 12 L2 components and 35+3 L3 sub-metrics (Table 1); below we enumerate the Cinematic Language sub-metrics and leave the remaining, more generic components to Table 1 and Appendix B. **Instruction Following** captures whether the video faithfully realizes the prompt’s cinematic in-

tent, over four components: *Cinematic Language* (shot scale, camera movement, viewing angle, composition, focus, tone & color, thematic style), *Character & performance*, *Scene*, and *Audio*. **Temporal Continuity** captures whether the output behaves as a film rather than a single clip, over four components: *spatial coherence*, *temporal coherence*, *character coherence*, and *audio coherence*. **Aesthetic Quality** captures production-grade quality, over four components: *base quality*, *editing* (editing fluency, camera-work appeal), *performance* (emotional performance, action performance), and *audio quality*.

Task-specific extension. T2V and R2V share this main framework. R2V additionally introduces one L2 component, **Visual Following**, with three L3 sub-metrics (*scene-space*, *character-appearance* and *prop-reference* fidelity), each measuring adherence to the conditioning reference image. Every R2V prompt conditions on one of three reference types (*scene*, *prop* or *character*), and only the matching fidelity sub-metric is scored for that prompt. Visual Following is scored only for R2V, so R2V has 13 L2 and 35+3 L3 dimensions, while all other dimensions are identical across tasks. For every L3 sub-metric we provide a 5-point anchor rubric; the complete definition table is in Appendix B.

4 Evaluation Method

FilmBench is scored by an expert-grade automatic evaluation agent which integrates a suite of Cinematic Language *operators*, expert annotation models, and a judge model. For every dimension in the taxonomy the agent produces a 1–5 score, achieving high agreement with human experts (§4.2). Below we describe the open-sourced operator suite (§4.1) that grounds the Cinematic Language dimensions and the score-aggregation rule (§4.2).

4.1 FilmOps: A Cinematic Language Operator Suite

A key obstacle to professional cinematic evaluation is that neither generic MLLM-as-judge approaches nor existing domain expert models reliably encode film-industry visual rules: general multimodal models are trained on open-domain understanding and misjudge professional categories such as shot scale, composition and camera movement [23], while existing aesthetic/shot models are trained mostly on everyday photography or live-action footage and do not transfer across genres (3D/2D animation, stylized content). We therefore build **FilmOps**, an open-source operator suite that maps a generated video into structured cinematic labels for the Cinematic Language dimensions.

Industry-aligned taxonomy. FilmOps defines a professional classification system, aligned with classical production references and vetted by front-line practitioners, over six core dimensions: *shot scale*, *composition*, *viewing angle*, *tone & color*, *character layout*, and *camera movement*. Five are frame-level and one (camera movement) is shot-level; together they span 55 sub-categories (character layout is an open-ended natural-language field). Categories are defined by narrative/visual function rather than imaging mechanism, so the same standard applies across live-action, 3D and 2D genres.

Multi-genre training data. To match the cross-genre, multi-style distribution of generated video, FilmOps is trained on a pool of 5,000+ real film/TV works spanning live-action, 3D animation, 2D animation and stylized/VFX content, and across narrative types (dialogue-driven, action, group scenes, and long-tail shot forms). Each operator is trained on ~40K–60K annotated samples with a shot-based test set, under a strict annotator-training and quality-control protocol led by professional practitioners.

Task-matched operator design. Operators are built to match each dimension’s characteristics: frame-level visual dimensions (shot scale, composition, angle, color) use vision backbones (DINO, BEiT and ResNet-18), while dimensions requiring relational reasoning or temporal modeling (character layout, camera movement) use a multimodal model (InternVL3-14B, LoRA-SFT). Classification operators are evaluated by macro-F1 and the natural-language layout operator by precision. Against four strong zero-shot general-MLLM baselines, the trained operators improve markedly on every dimension, confirming the value of domain-specialized operators; the full per-operator backbones and macro-F1 figures tested over 400 images/shots against these baselines are reported in the appendix (Table 12).

To support customizable, personalized self-evaluation, we **open-source the classification standard, model weights and inference scripts** of FilmOps. The modular design is not tied to a single model, so users can reuse individual operators or embed specific dimensions into their own evaluation pipelines.

4.2 Score Aggregation

Each L3 sub-metric receives a 1–5 score, which is **linearly mapped to a 0–100 scale before aggregation** ($1 \rightarrow 0, 2 \rightarrow 25, 3 \rightarrow 50, 4 \rightarrow 75, 5 \rightarrow 100$). Aggregation is **sample-level**: each video is first aggregated across its dimensions, and per-model scores are the mean over that model’s videos, preserving sample variance for significance analysis. We aggregate **L3 $\rightarrow$ L1 directly** (each L1 axis is the mean of its constituent L3 sub-metrics), treating every L3 sub-metric as equally important, and the Overall score is the equal-weighted mean of the three L1 axes; L2 components are computed as a side branch for fine-grained reporting only. Not-applicable scores (N/A or null) are excluded from aggregation; for models without audio capability, audio dimensions are handled symmetrically as not-applicable.

Qualitative case studies. Fine-grained per-dimension scoring exposes trade-offs that aggregate scores hide. As a worked example, Figure 4 scores a multi-shot T2V sci-fi mech battle, a case that stresses the dual challenge of intense action combined with multi-shot cutting. Seedance 2.0 (86.11) clearly outperforms Grok Imagine Video (55.97), and the gap is diagnostic rather than uniform: Seedance is markedly stronger on tone & color, character & performance and scene, and on camera-work appeal; in the action passages its character positioning stays consistent across shots and its temporal logic shows no visible errors. Grok, by contrast, exhibits scene drift and drifting cross-shot appearance and positioning, which surfaces directly in the per-dimension gaps on character-motion realism and action performance. The introduction (Figure 3) shows a complementary multi-shot R2V reference-scene case, and Appendix A provides two further worked examples (reference-character and reference-prop) that together cover the major scoring scenarios in FilmBench.

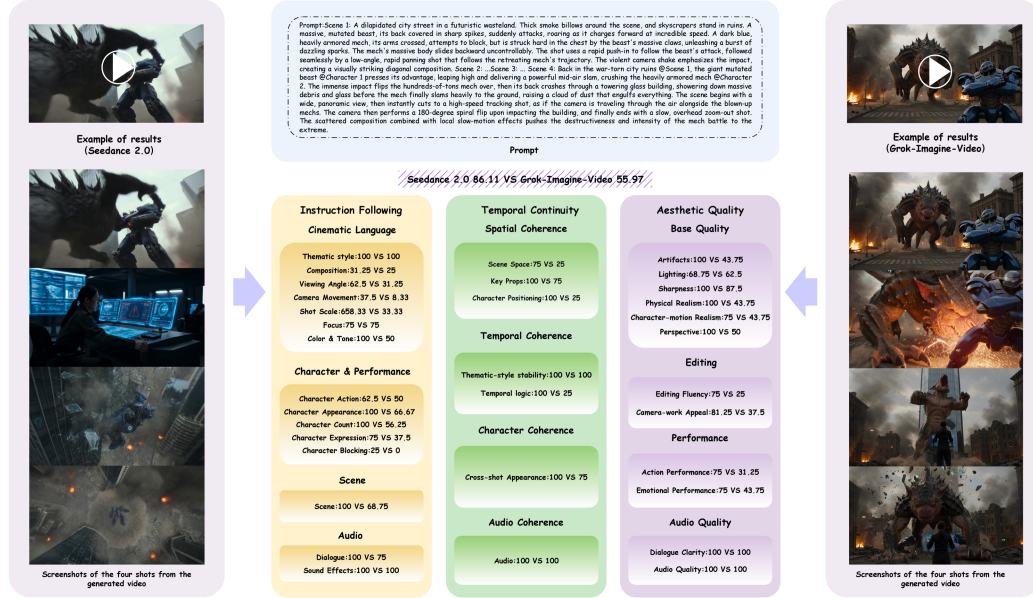


Figure 4: Multi-shot T2V example (a sci-fi mech battle): Seedance 2.0 (86.11) vs. Grok Imagine Video (55.97). Discussed in the text.

5 Experiments

5.1 Models and Setup

We benchmark leading video-generation models under each task. For **T2V** we evaluate 9 models: Seedance 2.0 [20], HappyHorse 1.1 [7], HappyHorse 1.0 [6], Kling 3.0 [12], Kling 3.0 Omni [12], Veo 3.1 [4, 24], Grok Imagine Video [26], Vidu Q3 Pro [1] and Hailuo 2.3 [17]. For **R2V** we evaluate 7 models: Seedance 2.0, HappyHorse 1.1, HappyHorse 1.0, Kling 3.0 Omni, Veo 3.1, Grok Imagine Video, and Vidu Q2 Pro [1]. All models are run under their official APIs / public checkpoints with default settings, generating one video per prompt. This yields $9 \times 515 = 4,635$ machine-scored T2V records and $7 \times 654 = 4,578$ R2V records. Throughout, T2V and R2V are reported with the *same* analysis format: overall, per-axis (L1), per-component (L2), per-sub-metric (L3), discriminability, and capability profiles, so that the two tasks are directly comparable; R2V additionally carries a reference-type analysis (scene / character / prop).

5.2 Human Agreement

We validate the automatic evaluator against 90 professional human raters majoring in film, television, directing, or digital media from the Beijing Film Academy at the *model level*: over 300 prompts ($\sim 30\%$ of the benchmark) carry both machine and human judgments, yielding 2,878 model–prompt pairs. We compute per-model machine and human means at each L2 component and correlate the two rankings (Table 2). The overall model-level Spearman ρ reaches 0.95 (T2V) and 0.96 (R2V), confirming that the automatic leaderboard closely reproduces expert model rankings.

Table 2: L2 component-level human agreement (Spearman ρ), averaged across T2V and R2V. 10 of 13 components reach $\rho \geq 0.78$; the three lower-agreement components (audio quality, audio coherence, editing) are separated by a mid-rule.

	Perform.	Vis. Follow.	Cinem. Lang.	Base qual.	Audio	Char. & perf.	Temp. coh.	Spat. coh.	Scene	Char. coh.	Audio q.	Audio coh.	Editing
ρ	0.99	0.96	0.95	0.95	0.93	0.92	0.85	0.83	0.82	0.78	0.68	0.68	0.60

Drilling into the component level, 10 of the 13 L2 dimensions achieve $\rho \geq 0.78$; the three lower-agreement components—*audio quality* (0.68), *audio coherence* (0.68) and *editing* (0.60)—are those where subjective temporal behaviour (sound fidelity, cut rhythm) is hardest to pin down, and we acknowledge that human and automatic judgments diverge more on these axes. The overall-level agreement ($\rho \geq 0.95$ on both tasks) nonetheless indicates that the aggregate ranking is a reliable proxy for expert judgment, and we flag the three lower-agreement dimensions when interpreting fine-grained results.

5.3 Overall Leaderboard

Tables 7–8 and Figure 5 report the overall FilmBench score (0–100, higher is better). On T2V, Seedance 2.0 leads (88.93), closely followed by HappyHorse 1.1 (87.42) and HappyHorse 1.0 (87.02); the two Kling 3.0 variants form a second tier (~ 86), Vidu Q3 Pro, Grok and Veo 3.1 a third tier (~ 81), with Hailuo 2.3 at 68.94. On R2V, the top group is Seedance 2.0 (86.66), HappyHorse 1.1 (85.51) and HappyHorse 1.0 (84.87). Crucially, no model approaches saturation, in sharp contrast with the near-ceiling scores reported on prior web-style benchmarks, and the top group is consistent across both tasks, indicating that reference conditioning does not reshuffle the leaders.

5.4 Dimension Variance

To quantify where models differ most at fine-grained levels, we compute the cross-model variance of each dimension across the hierarchy (Figure 6; Hailuo is excluded from the variance computation to avoid its audio outliers dominating). At the axis level, *Instruction Following* shows by far the largest variance (99.4), nearly eight times that of Aesthetic Quality (12.6) and seventeen times that of Temporal Continuity (5.6), indicating that models differ primarily in their ability to execute professional cinematic language instructions.

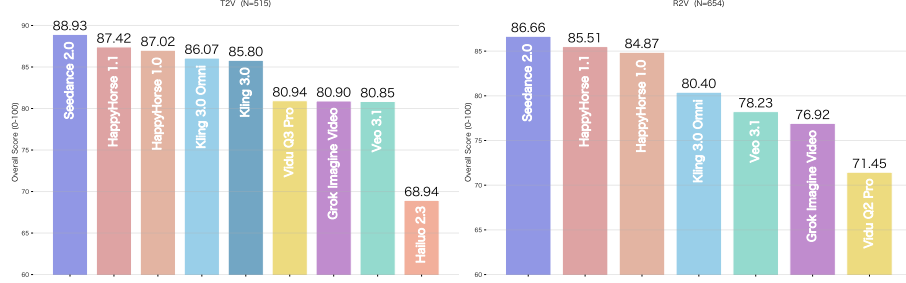


Figure 5: Overall FilmBench scores (0–100) per model; model names are printed inside each bar and the dashed line is the cross-model average. Left: T2V ($N = 515$); right: R2V ($N = 654$).

Drilling into the component level, *Cinematic Language* dominates with a variance of 187.9, far exceeding Audio (148.4), Scene (71.3) and all other components. At the sub-metric level, the top five highest-variance dimensions are all Cinematic Language or scene-related: camera movement (357.6), focus (269.6), shot scale (209.0), viewing angle (200.5) and fore/mid/background (169.8). This concentration confirms that the primary differentiator among current generative video models lies in Cinematic Language instruction following—whether a model understands and executes professional camera-language conventions. Per-task breakdowns are provided in Appendix C.

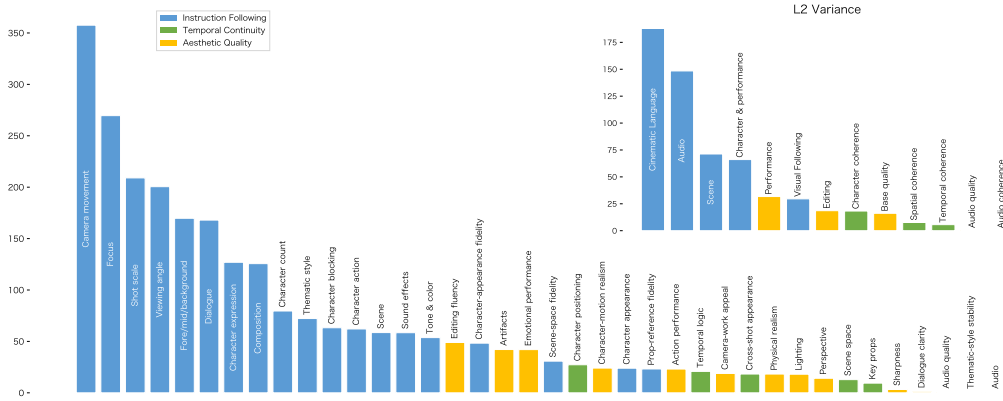


Figure 6: Cross-model variance per L3 sub-metric (main bars, colored by axis), with L2 inset (upper right), computed over the merged T2V+R2V model means (Hailuo excluded).

5.5 Per-Axis Landscape (L1)

Across the three axes, models are most differentiated on *Instruction Following*, which encompasses the Cinematic Language sub-metrics that drive the high variance reported above. On T2V (Figure 7, left), Seedance 2.0 leads all three axes: Instruction Following (91.69, ahead of HappyHorse 1.1 at 90.70), Temporal Continuity (96.06) and Aesthetic Quality (79.03). On R2V (Figure 7, right), Seedance 2.0 likewise leads every axis (Instruction Following 87.45, Temporal Continuity 94.09, Aesthetic Quality 78.44). The Instruction Following axis shows the widest inter-model gap under both tasks, as the Cinematic Language and scene-related sub-metrics amplify the spread in professional-language execution.

5.6 Per-Component Rankings (L2)

Descending one level, we rank models within the L2 components most closely tied to *Cinematic Language* together with the adjacent *Scene* and *Character & performance* components (Figure 8); the full 12/13-panel rankings are deferred to the appendix. Cinematic Language is placed at the centre

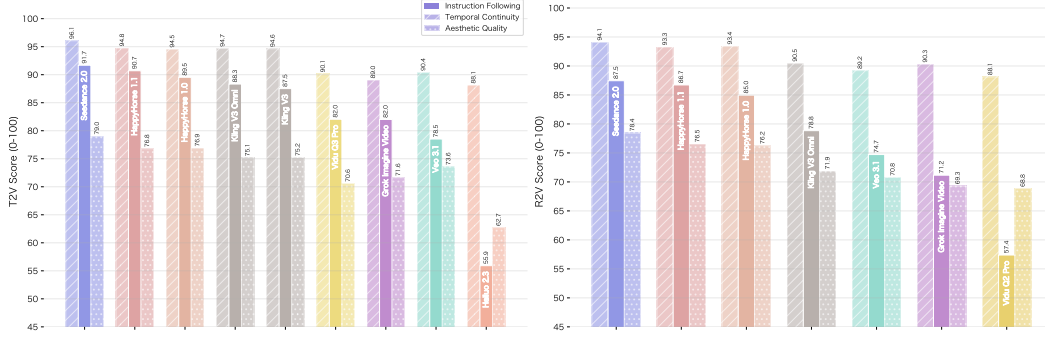


Figure 7: Per-axis rankings for T2V (left, 9 models) and R2V (right, 7 models). Each model is scored on three L1 axes: Instruction Following (solid), Temporal Continuity (hatched), and Aesthetic Quality (dotted). Models are ranked by overall score; R2V Instruction Following includes the visual-following component.

and used as the sort key, and each component isolates a different craft skill, revealing where the head models specialize:

Cinematic Language (camera work, framing, transitions) is the sharpest discriminator of the three, with the field fanning out over a 17–42-point range and even the strongest model reaching only 84.6. This is the one component that hinges on *temporal* decision-making—when to move the camera, how to reframe, when to cut—rather than per-frame rendering, which is why models diverge so widely here. Seedance 2.0 leads decisively and is the only model whose camera competence holds up when a reference image is imposed, marking dynamic camera control as its signature strength.

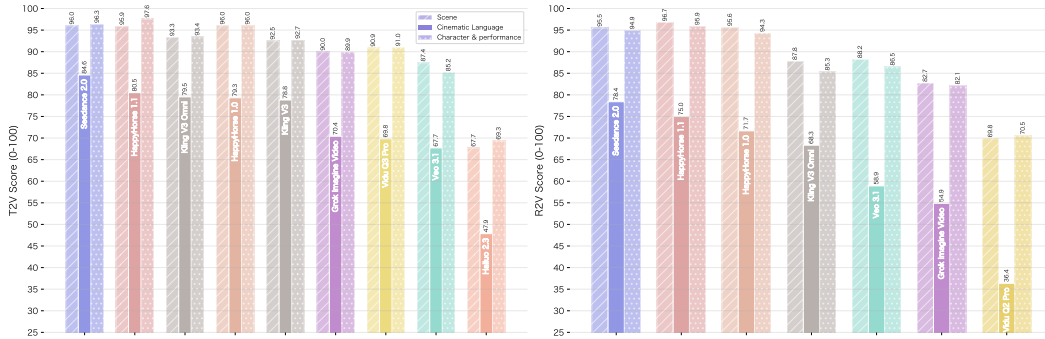


Figure 8: Per-model rankings within the L2 components most related to audiovisual language: *Cinematic Language* (solid, centre, model names inside; also the sort key), flanked by *Scene* (hatched) and *Character & performance* (dotted). T2V (left, 9 models) and R2V (right, 7 models). The full L2 rankings are in the appendix.

Scene (environment consistency, lighting, background layout) also spreads the field, though the leading cluster reaches the mid-90s (HappyHorse 1.0 at 96.0, HappyHorse 1.1 at 96.7). The gap opens further down the ranking, where weaker models struggle to keep a coherent, consistently lit environment across shots. The separation here is owned by the HappyHorse family, whose strength lies specifically in preserving spatial layout—a capability distinct from camera craft, since the models that lead scene rendering are not the ones that lead Cinematic Language.

Character & performance spreads the field the least of the three but still meaningfully, with the leaders reaching the high-90s (HappyHorse 1.1 at 97.6, Seedance 2.0 at 96.3). Here differentiation comes from sustaining believable characters and acting across a full scene, and HappyHorse 1.1 holds a slim but consistent edge, making it the character specialist just as Seedance 2.0 is the camera specialist. Read together, the three components show that the overall leaderboard masks a division of

labor—camera craft, scene rendering and character performance are led by different models—so no single system dominates every axis of film craft.

5.7 Fine-Grained Rankings (L3)

At the finest level, FilmBench resolves 35 (T2V) / 38 (R2V) L3 sub-metrics, each with its own ranking; Figure 9 shows the three sub-metrics selected by highest merged T2V+R2V variance. All three are Cinematic Language sub-metrics, indicating that **camera craft is where the leading models pull apart most**:

Camera movement shows the largest spread among the head models, with scores ranging from Seedance 2.0’s 86.5 down toward the low 50s. Executing a motivated camera move—a push-in, a pan that tracks action, a crane reveal—requires the model to sustain a coherent scene through time, and models diverge considerably in how convincingly they do so. Seedance 2.0 leads by a clear margin, suggesting a training emphasis on dynamic camera choreography that sets it apart from the rest of the field.

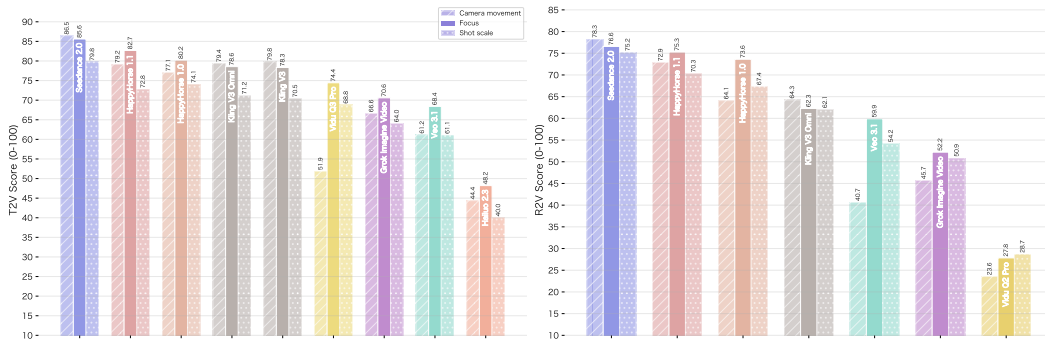


Figure 9: Per-model rankings for the three highest-variance L3 sub-metrics (computed over merged T2V+R2V data): *Camera movement* (hatched), *Focus* (solid, model names inside), and *Shot scale* (dotted). T2V (left, 9 models) and R2V (right, 7 models). The full per-metric panels are in the appendix.

Focus (depth-of-field control, subject isolation) sees the leaders cluster in the low-to-mid 80s (Seedance 2.0 at 85.6, HappyHorse 1.1 at 82.7), with a wider tail below. The metric separates models because it rewards *intentional* focus that directs attention to the dramatically relevant subject—a semantic judgement, beyond merely producing a shallow-depth look, that the stronger models handle more reliably.

Shot scale (close-up, medium, wide framing decisions) tests whether a model chooses the framing a scene calls for—an intimate close-up for a line of dialogue, a wide for an establishing beat. Seedance 2.0 again leads (79.8) with HappyHorse 1.1 the runner-up, and the head ordering across all three sub-metrics is consistent—Seedance 2.0 first, the HappyHorse family close behind. Because these are the dimensions on which models differ most, holding a high standard across them is what distinguishes a film-grade leader: Seedance 2.0’s overall lead comes not from a single capability but from staying near the top on the most challenging Cinematic Language dimensions at once, while the HappyHorse family remains competitive across them.

Crucially, **no model wins on all L3 sub-metrics**: Seedance 2.0, the overall leader, claims only 18/35 (T2V) or 20/38 (R2V) championships, concentrated on Cinematic Language sub-metrics—Cinematic Language (camera movement, shot scale, focus, viewing angle), camera-work appeal, and performance (action and emotional performance)—in both tasks. HappyHorse 1.1, the runner-up, dominates 11/35 (T2V) or 10/38 (R2V) sub-metrics on character instruction-following, audio, and scene dimensions. Notably, on the three R2V-specific visual-following sub-metrics (scene-space fidelity, character-appearance fidelity and prop-reference fidelity), the HappyHorse family takes all three championships—HappyHorse 1.0 leads scene-space and character-appearance fidelity, HappyHorse 1.1 leads prop-reference fidelity. This “championship mismatch” is the key value of fine-grained evaluation: a single leaderboard number conceals the structural complementarity

between models. Beyond the top two, the remaining championships are scattered across several models, with 4 (T2V) or 2 (R2V) models winning no championships at all. This granularity turns a single leaderboard into an actionable diagnostic: a mid-tier model can still lead on individual craft dimensions. The complete per-model score matrices (L3 and L2 heatmaps) are in the appendix (Figures 20–23).

FilmBench spans 20 cinematic genres across Chinese Films and International Films markets (T2V $n = 243/272$; R2V $n = 239/415$) to ensure broad diversity and coverage. Figure 10 splits each task by market: the top group (Seedance 2.0, HappyHorse 1.1/1.0) is identical for Chinese Films and International Films prompts on both tasks, with only minor mid-pack reordering. Figure 11 shows the score heatmap across the top 4 genres by sample size: rankings remain stable across genres, and each model’s per-genre spread is small ($\pm 2\text{--}3$ points), confirming that both leaderboards reflect model capability rather than genre mix. The action subset, often cited as the hardest, shows the same ranking with only a slightly lower ceiling.

Figure 10: Overall ranking split by market, Chinese Films vs. International Films; left: T2V ($n = 243/272$), right: R2V ($n = 239/415$).

Figure 11: Overall score heatmap across the top 4 genres by sample size. Y-axis: genres (with sample sizes T2V/R2V); X-axis: models split into T2V (left) and R2V (right) sections. Each cell shows the genre-specific mean score.

Beyond market and genre, we probe three prompt-level factors that stress generation differently: narrative structure (single- vs multi-shot prompts), content type (action vs dialogue scenes), and rendering style (live-action vs animation). In T2V, 113 of 515 prompts are single-shot and 402 are multi-shot; every R2V prompt is multi-shot, so the shot contrast is T2V only. We tag 117 of 515 T2V prompts and 190 of 654 R2V prompts as action, and label all 654 R2V prompts as live-action (486) or animation (168).

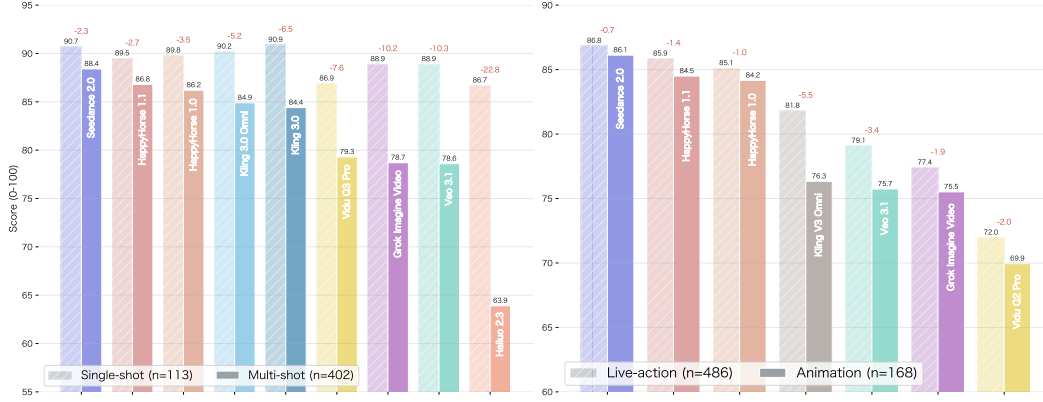


Figure 12: Left: T2V single-shot (hatched) vs multi-shot (solid) mean scores (0–100) per model, ordered by multi-shot score; red numbers give the drop Δ ; $n=113/402$. Right: R2V live-action (hatched) vs animation (solid) mean scores per model; $n=486/168$.

top models lose little (Seedance 2.0 -2.3 , HappyHorse 1.1 -2.7), whereas lower-ranked models show much larger drops (Grok Imagine Video -10.2 , Veo 3.1 -10.3 , Hailuo 2.3 -22.8). The degradation concentrates on shot-craft dimensions (Table 5 in the appendix): Cinematic Language (composition, viewing angle, camera movement, focus), editing fluency, and cross-shot spatial consistency, while audio and temporal-coherence dimensions move little. A single-shot prompt only tests intra-shot rendering, whereas a multi-shot prompt additionally requires planning distinct shots and cutting between them inside one clip; the wider inter-model spread on multi-shot prompts reflects a compositional competence that current generators have not yet mastered, and we therefore treat shot structure as a first-class evaluation axis rather than folding it into a single leaderboard number.

Action scenes uniformly lower scores across all models. Every model scores lower on action than on dialogue prompts (Figure 13), and, as with multi-shot prompts, the drop is larger for lower-ranked models: the top group loses least (T2V Seedance 2.0 -5.3 , HappyHorse 1.1 -6.7) while lower-ranked models show larger drops (Vidu Q3 Pro -12.3 , Hailuo 2.3 -15.5). **Seedance 2.0 maintains the lead on action-specific L3 sub-metrics:** on T2V action prompts it tops 16/35 sub-metrics, notably character action (99.3), character expression (96.7), fore/mid/background (95.0) and viewing angle (91.1); on R2V action prompts it tops 11/38 sub-metrics, notably tone & color (100.0), camera movement (84.9), composition (83.1) and shot scale (78.2). Its action drops on these dimensions remain moderate (T2V ≤ 3.6 , R2V ≤ 10.7), indicating that its Cinematic Language and character-performance strengths carry over to action content. **The action performance degradation concentrates in Instruction Following and Aesthetic Quality, with two distinct L3 clusters** (Table 6 in the appendix). In Instruction Following the drop is dominated by camera-work sub-metrics—camera movement (-31.1), focus (-28.6), shot scale (-26.6) and composition (-21.0)—indicating that models show larger drops in camera framing and motion once the scene becomes fast and multi-agent. In Aesthetic Quality a second cluster of *physical-realism* sub-metrics degrades sharply: physical plausibility (-16.9), character-motion realism (-15.8), perspective (-15.3), artifacts (-11.9) and clarity (-11.8). Action scenes thus reveal two concurrent degradation patterns that an aggregate score hides—models not only shift camera framing, but also show reduced physical plausibility and increased artifacts under vigorous motion. The same two clusters lead the R2V action drop, at a smaller magnitude (physical plausibility -9.9 , artifacts -3.5); we report this contrast descriptively, as the two tasks differ in prompt and sample distribution.

Top models already generalize across visual style. On R2V, the leading group scores almost identically on live-action and animation (Figure 12, right; Seedance 2.0 86.8 vs. 86.1, gap < 1.5), indicating that rendering style is no longer a constraint at the frontier. Style generalization instead shows wider variation in the mid-tier: several mid-ranked models score markedly lower on animation than live-action (Kling 3.0 Omni $81.8 \rightarrow 76.3$, -5.5), and the loss again concentrates in Instruction Following (Kling 3.0 Omni $80.9 \rightarrow 72.8$) and Aesthetic Quality, echoing the action analysis: the two harder distributions, action and animation, both challenge models on the same two axes.

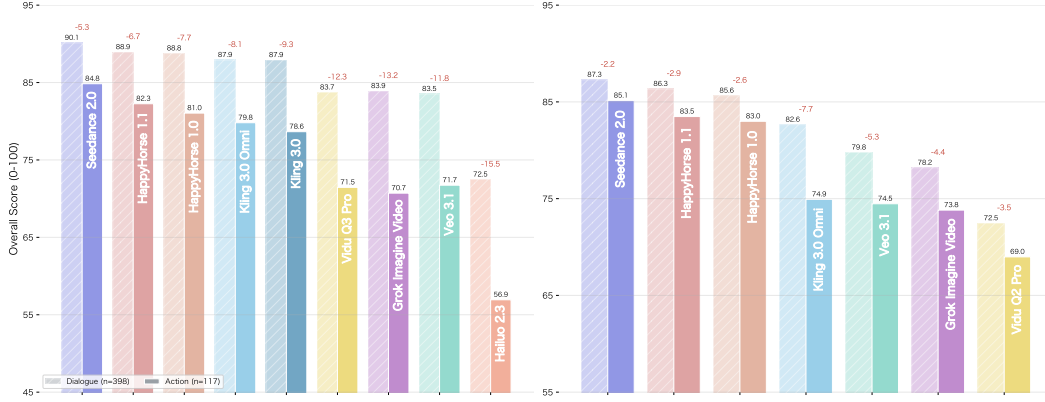


Figure 13: T2V (left, 9 models) and R2V (right, 7 models): dialogue (hatched) vs. action (solid) mean scores (0–100) per model, ordered by dialogue score; red numbers give the drop Δ (action – dialogue). T2V: action $n=117$, dialogue $n=398$. R2V: action $n=190$, dialogue $n=464$.

5.10 R2V: Reference Types

R2V prompts condition on one of three reference types: scene (232 prompts), prop (209) and character (213). Figure 14 reports the reference-fidelity (visual-following) ranking for each type. The visual-following leader differs from the overall leader: HappyHorse 1.0 tops scene- and character-fidelity, while HappyHorse 1.1 leads on prop-fidelity, indicating that raw reference fidelity and overall film-quality are distinct capabilities. Across the three types, scene-space and character-appearance fidelity show a similar ranking order, whereas prop-reference fidelity reveals a notably different model ordering and a much tighter spread, confirming that current models still struggle to faithfully preserve a referenced prop even when scene and character references are respected.

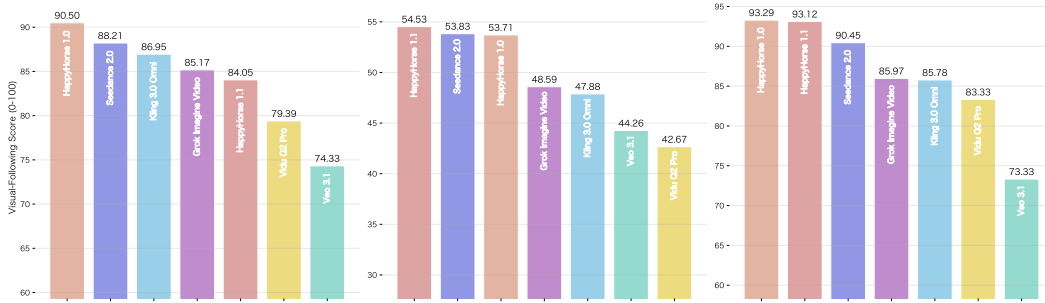


Figure 14: R2V visual-following (reference-fidelity) ranking by reference type (columns: scene / prop / character).

5.11 Cinematic Language Related L3 Landscape

Finally, we zoom in on the L3 sub-metrics most directly tied to film production craft. Figure 15 collects 10 sub-metrics from three L2 groups across the two tasks: Cinematic Language (camera movement, shot scale, focus, viewing angle, tone & color, composition), editing (camera-work appeal, editing fluency), and performance (action performance, emotional performance). Across both tasks, models score highest on tone & color (T2V 95.0, R2V 91.2) and lowest on action performance (50.8 / 46.8) and camera-work appeal (54.8 / 57.5). The performance group is uniformly the lowest-scoring L2 cluster in both tasks, while Cinematic Language dimensions show the widest T2V–R2V gap (e.g., camera movement 69.5 vs. 55.6), indicating that reference conditioning amplifies the spread on camera-work execution.

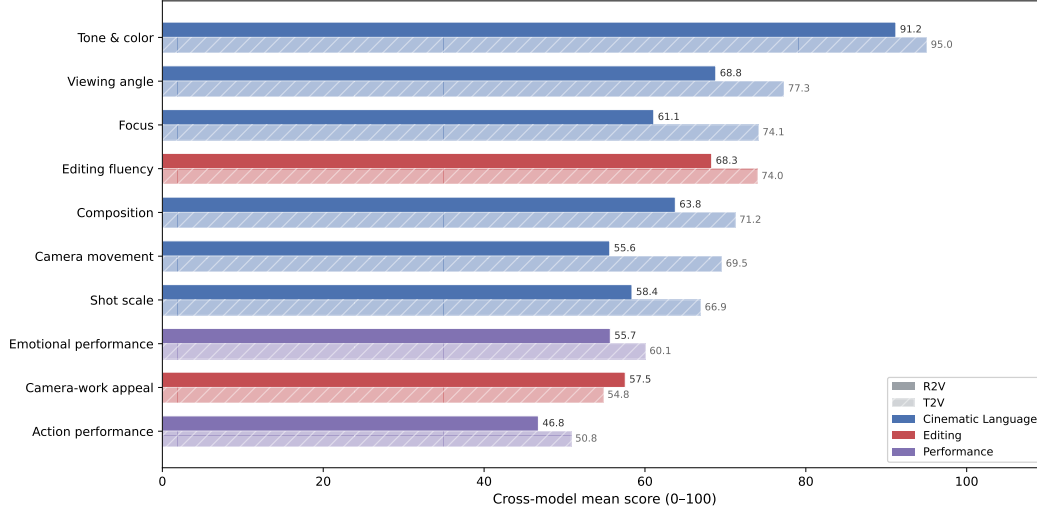


Figure 15: Cinematic Language L3 sub-metric cross-model means (0–100), T2V (hatched) vs. R2V (solid), sorted by descending score. Bars are colored by L2 group: Cinematic Language (blue), editing (red), performance (purple).

5.12 Summary of Findings

Taken together, the analyses above distill into five findings that hold across the field. **(i) No saturation; models differ most on Cinematic Language sub-metrics:** no model nears the ceiling on either task (tops of 88.93 / 86.66), and camera movement, focus, shot scale and viewing angle show the largest inter-model spread, far more so under R2V (camera-movement variance 384.7 vs. 135.7), while temporal and audio dimensions show near-uniform scores. **(ii) Dynamic aesthetics are the lowest-scoring sub-metrics across the field:** action performance, camera-work appeal, character-motion realism and emotional performance are the lowest sub-metrics in both tasks, while static image-quality sub-metrics (sharpness, lighting) score markedly higher. **(iii) Multi-shot prompts are uniformly harder and show wider model spread:** every model scores lower on multi-shot prompts (average drop 7.9 points), with the drop larger for lower-ranked models; the drop concentrates on shot-craft dimensions (Cinematic Language, editing fluency, scene space). **(iv) Reference conditioning stresses rather than reorders:** it leaves the top group intact but compresses the field, and the visual-following leader differs from the overall leader, indicating that raw reference fidelity and overall film-quality are distinct capabilities. **(v) No model wins on all L3 sub-metrics:** Seedance 2.0, the overall leader, claims only 18/35 (T2V) or 20/38 (R2V) championships, concentrated on Cinematic Language sub-metrics (Cinematic Language, editing appeal and performance), while HappyHorse 1.1 dominates 11/35 (T2V) or 10/38 (R2V) sub-metrics on character, audio and scene dimensions; notably the three R2V-specific visual-following championships all go to the HappyHorse family, not to Seedance; 4 (T2V) and 2 (R2V) models do not claim any championship. This “championship mismatch” reveals structural complementarity concealed by a single leaderboard number.

6 Conclusion

We presented FilmBench, an evaluation benchmark for text-to-video (T2V) and reference-to-video (R2V) generation grounded in the Cinematic Language system used in professional film production. Built in collaboration with directors and faculty from the Beijing Film Academy and a professional film studio, FilmBench features prompts reverse-engineered from award-winning real films, an academy-aligned three-level evaluation taxonomy (3 L1 axes, 12 L2 components, 35 L3 sub-metrics), and an expert-grade automatic evaluator suite (FilmOps) whose model-level ranking reproduces expert rankings at Spearman $\rho = 0.95$ (T2V) / 0.96 (R2V). We evaluated 9 T2V and 7 R2V leading systems and released the benchmark to support research toward film-grade generative systems.

Limitations. FilmBench is designed for evaluating professional film and cinematic content creation capabilities. The evaluation dimensions and model rankings presented in this work do not necessarily reflect model performance on non-cinematic video generation tasks such as vertical short-form video creation, casual user-generated content, or other non-cinematic video production scenarios.

Broader Impact. We do not anticipate any negative societal impacts arising from this work. FilmBench is an evaluation benchmark and does not itself generate or deploy content; its intended use is to support research toward higher-quality video generation systems.

References

- [1] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models, 2024. URL <https://arxiv.org/abs/2405.04233>.
- [2] Harold Haodong Chen, Disen Lan, Wen-Jie Shu, Qingyang Liu, Zihan Wang, Sirui Chen, Wenkai Cheng, Kanghao Chen, Hongfei Zhang, Zixin Zhang, Rongjin Guo, Yu Cheng, and Ying-Cong Chen. TiViBench: Benchmarking think-in-video reasoning for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11403–11413, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_TiViBench_Benchmarking_Think-in-Video_Reasoning_for_Video_Generation_CVPR_2026_paper.html.
- [3] Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking video generation model for the embodied world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL <https://openreview.net/forum?id=p5QSlwume>.
- [4] Google DeepMind. Veo: A text-to-video generation system. Technical report, Google, 2025. URL <https://deepmind.google/discover/blog/veo-2/>.
- [5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18858–18868, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Han_Video-Bench_Human-Aligned_Video_Generation_Benchmark_CVPR_2025_paper.html.
- [6] HappyHorse AI Team. Happyhorse 1.0: Core capabilities and structural features for sota video generation. Technical report, HappyHorse AI, 2026. URL <https://happy-horse.art/features>.
- [7] HappyHorse AI Team. Happyhorse 1.1: Advancing multimodal integration and ai-driven video production. Technical report, HappyHorse AI, 2026. URL <https://happy-horse.art/happy-horse-1-1-ai>.
- [8] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhuranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhui Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2105–2123, 2024. doi: 10.18653/v1/2024.emnlp-main.127. URL <https://aclanthology.org/2024.emnlp-main.127/>.
- [9] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.
- [10] Karthik Inbasekar, Guy Rom, and Omer Shlomovits. WorldJen: An end-to-end multi-dimensional benchmark for generative video models. *arXiv preprint arXiv:2605.03475*, 2026. doi: 10.48550/arXiv.2605.03475. URL <https://arxiv.org/abs/2605.03475>.
- [11] Longteng Jiang, DanDan Zheng, Qianqian Qiao, Heng Huang, Huaye Wang, Yihang Bo, Bao Peng, Jingdong Chen, Jun Zhou, and Xin Jin. VGA-Bench: A unified benchmark and multi-model framework for video aesthetics and generation quality evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30457–30466, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jiang_VGA-Bench_A_Unified_Benchmark_and_Multi-Model_Framework_for_Video_Aesthetics_CVPR_2026_paper.html.

- [12] Kling Team and Kuaishou Technology AI Team. Kling-omni technical report: A unified framework for video generation, editing, and reasoning. Technical Report arXiv:2512.16776, Kuaishou Technology, 2025. URL <https://arxiv.org/abs/2512.16776>.
- [13] Xiaofeng Li, Leyi Sheng, Zhen Sun, Zongmin Zhang, Jiaheng Wei, and Xinlei He. IP-Bench: Benchmark for image protection methods in image-to-video generation scenarios. *arXiv preprint arXiv:2603.26154*, 2026. doi: 10.48550/arXiv.2603.26154. URL <https://arxiv.org/abs/2603.26154>.
- [14] Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. VMBench: A benchmark for perception-aligned video motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13087–13098, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ling_VMBench_A_Benchmark_for_Perception-Aligned_Video_Motion_Generation_ICCV_2025_paper.html.
- [15] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c481049f7410f38e788f67c171c64ad5-Abstract-Datasets_and_Benchmarks.html.
- [16] Ryosuke Matsuda, Keito Kudo, Haruto Yoshida, Nobuyuki Shimizu, and Jun Suzuki. SLVMEval: Synthetic meta evaluation benchmark for text-to-long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7794, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Matsuda_SLVMEval_Synthetic_Meta_Evaluation_Benchmark_for_Text-to-Long_Video_Generation_CVPR_2026_paper.html.
- [17] MiniMax AI Team. Minimax hailuo 2.3: A new level of complex video performance & media agent. MiniMax News Release, 2025. URL <https://www.minimax.io/news/minimax-hailuo-23>.
- [18] Wenshuo Peng, Gongxuan Wang, Tianmeng Yang, Chuanhao Li, Xiaojie Xu, Hui He, and Kaipeng Zhang. SVBench: Evaluation of video generation models on social reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 32872–32881, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Peng_SVBench_Evaluation_of_Video_Generation_Models_on_Social_Reasoning_CVPR_2026_paper.html.
- [19] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhui Chen. ConsistI2V: Enhancing visual consistency for image-to-video generation. *Transactions on Machine Learning Research*, 2024. URL <https://openreview.net/forum?id=vqniLmUDvj>.
- [20] Team Seedance, De Chen, Liyang Chen, et al. Seedance 2.0: Advancing video generation for world complexity, 2026. URL <https://arxiv.org/abs/2604.14148>.
- [21] 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. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 24034–24058, 2026. doi: 10.18653/v1/2026.findings-acl.1203. URL <https://aclanthology.org/2026.findings-acl.1203/>.
- [22] 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 (CVPR)*, pages 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.
- [23] 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, 2025. URL <https://arxiv.org/abs/2505.15145>.
- [24] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.
- [25] Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. MovieBench: A hierarchical movie level dataset for long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28984–28994, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_MovieBench_A_Hierarchical_Movie_Level_Dataset_for_Long_Video_Generation_CVPR_2025_paper.html.

- [26] xAI Team. Grok imagine video 1.5: Native audio generation and sota image-to-video workflow. xAI Official Blog, 2026. URL <https://x.ai/news/grok-imagine-video-1-5>.
- [27] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. ChronoMagic-Bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, pages 21236–21270, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/25b9960c8a5bd887eb5476c951260403-Abstract-Datasets_and_Benchmarks_Track.html.
- [28] Ailing Zhang, Lina Lei, Dehong Kong, Zhixin Wang, Jiaqi Xu, Fenglong Song, Chun-Le Guo, Chang Liu, Fan Li, and Jie Chen. UI2V-Bench: An understanding-based image-to-video generation benchmark. *arXiv preprint arXiv:2509.24427*, 2025. URL <https://arxiv.org/abs/2509.24427>.
- [29] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL <https://arxiv.org/abs/2503.21755>.
- [30] Ziwei Zhou, Zeyuan Lai, Rui Wang, Yifan Yang, Yuqing Yang, Qi Dai, Lili Qiu, and Chong Luo. AVGen-Bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL <https://openreview.net/forum?id=aJdgt8xDMy>.

A Qualitative Evaluation Examples

Figures 16–17 present two R2V cases from 3D animated film production, illustrating how Film-Bench’s fine-grained scoring reveals dimension-level trade-offs that an aggregate score alone would conceal. In both cases, the primary score gaps concentrate on Cinematic Language and character performance within Instruction Following, rather than on temporal or aesthetic dimensions where models tend to cluster.

The reference-character case (Figure 16) shows that character-appearance fidelity scores are relatively high across models, as the reference identity is generally well preserved. By contrast, the reference-prop case (Figure 17) exposes a field-wide prop-reference bottleneck: both models score 0 on prop-reference fidelity because the prompt describes a golden energy aura around the character’s body, yet the generated effect bleeds onto the reference prop and alters its color, illustrating that current models struggle to isolate stylized visual effects from unrelated reference objects.

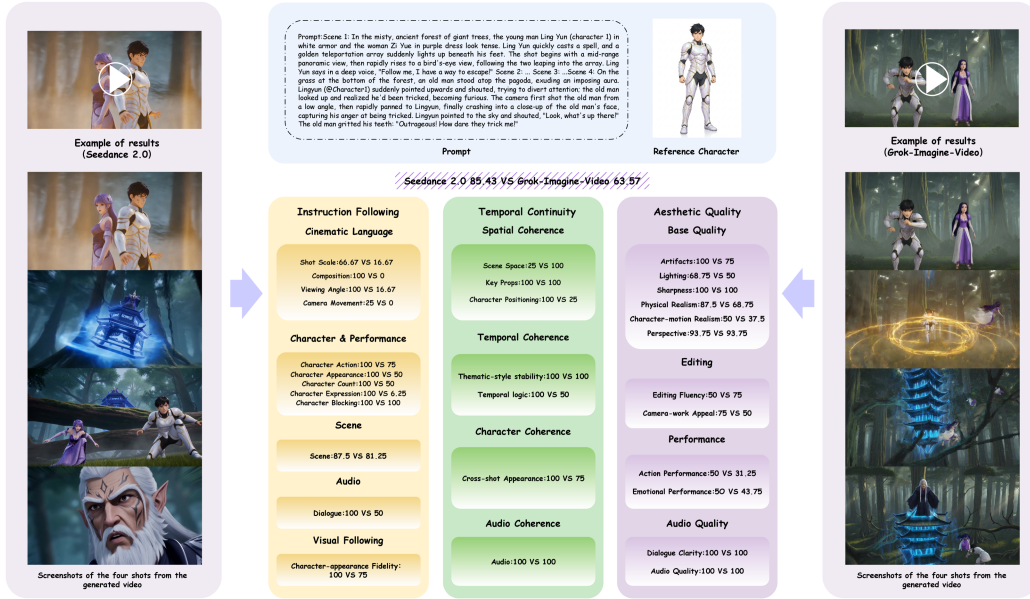


Figure 16: R2V reference-character example in 3D animation: Seedance 2.0 (85.43) vs. Grok Imagine Video (63.57). Character-appearance fidelity is high, but the score gap is driven by Cinematic Language and character performance.

B L3 Dimension Definitions

Tables 3 and 4 list all 35 L3 sub-metrics scored in the T2V task, grouped by their L1 axis and L2 component. The R2V task adds three further L3 sub-metrics (*scene-space*, *character-appearance* and *prop-reference* fidelity) under a Visual-Following L2 component of Instruction Following, each scored against the corresponding reference type, giving 38 in total. Each sub-metric is rated on a 1–5 scale against an anchor rubric.

C Per-Task Variance Breakdown

Figures 18 and 19 give the per-task cross-model variance for each L3 sub-metric. The merged view combining both tasks is shown in the main text (Figure 6).

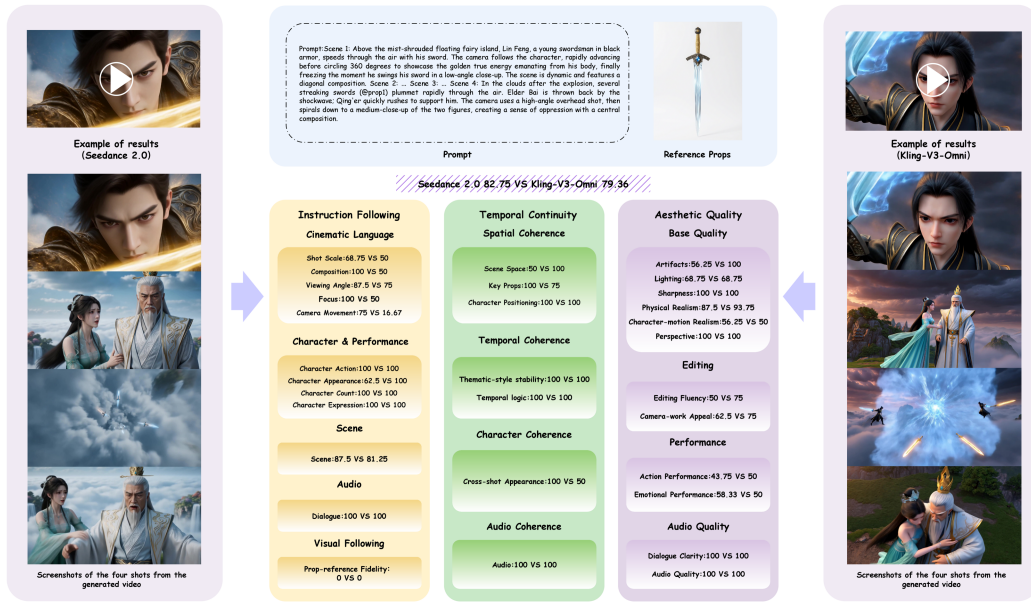


Figure 17: R2V reference-prop example in 3D animation: Seedance 2.0 (82.75) vs. Kling 3.0 Omni (79.36). Both score 0 on prop-reference fidelity because the prop’s color is altered by a golden energy aura intended for the character; the overall gap is again driven by Cinematic Language.

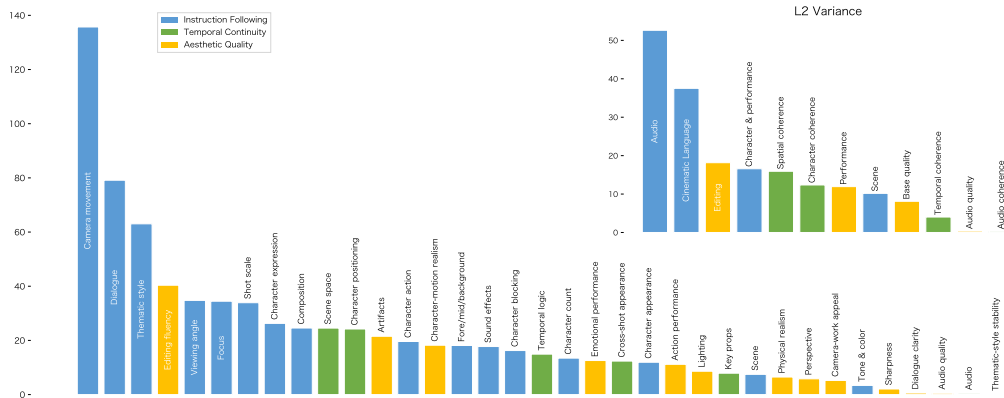


Figure 18: T2V cross-model variance per L3 sub-metric, with L2 inset (upper right).

Table 3: FilmBench L3 sub-metric definitions, part 1: the Instruction Following (IF) axis, including the three R2V-only Visual-Following sub-metrics. “IF” = Instruction Following, “TC” = Temporal Continuity, “AQ” = Aesthetic Quality.

L1	L2 component	L3 sub-metric	Definition (5-point rubric anchor)
IF	Cinematic Language	Shot scale	Adherence to requested shot scale (extreme close-up → extreme long shot).
		Camera movement	Adherence to requested camera move (fixed, push, pull, pan, track, orbit, follow, zoom, handheld, strong move).
		Viewing angle	Adherence to requested horizontal/vertical angle (eye-level, high, bird’s-eye, extreme low/high, Dutch, over-shoulder, POV).
		Composition	Adherence to requested composition rule (center, symmetry, rule-of-thirds, diagonal, leading-line, framing, depth, etc.).
		Focus	Whether in/out-of-focus and focus changes match the prompt.
		Tone & color Thematic style	Adherence to requested hue / temperature / saturation. Adherence to requested style (realism, cyberpunk, horror; or a director’s style).
	Character & performance	Character count	Whether the number of characters matches the prompt.
		Character appearance	Whether facial features, hairstyle, accessories, age match the prompt.
		Character expression	Whether expression / emotion / gaze match, with no spurious expressions.
		Character action Character blocking	Whether actions match, with no spurious actions. Whether blocking, entrance/exit and orientation match the prompt.
	Scene	Scene Fore/mid/background	Whether the scene matches the prompt. Whether foreground/middle/background match the prompt.
	Audio	Dialogue	Dialogue completeness, emotional accuracy, lip-sync, no ID jumps.
		Sound effects	Whether sound effects / BGM match the prompt.
	Visual Following (R2V)	Scene-space fidelity	<i>Whether the generated space/scene stays consistent with the reference scene image (scene reference).</i>
		Character-appearance fidelity	<i>Whether character appearance stays consistent with the reference character image (character reference).</i>
		Prop-reference fidelity	<i>Whether key props stay consistent with the reference prop image (prop reference).</i>

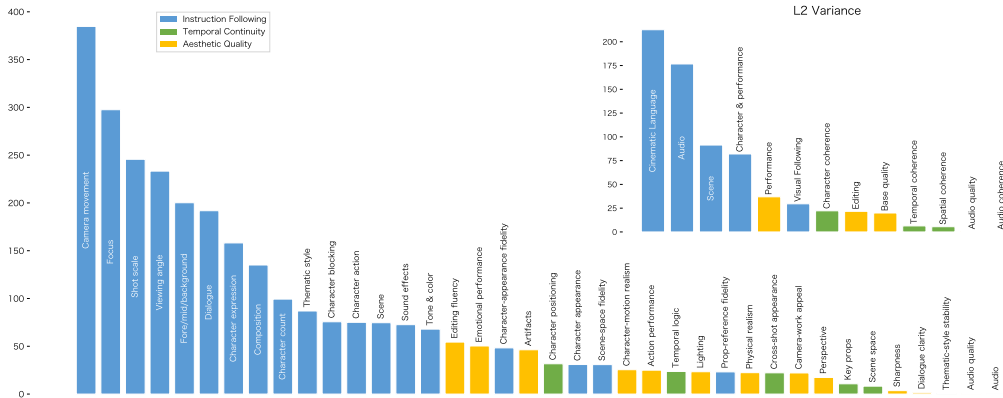


Figure 19: R2V cross-model variance per L3 sub-metric, with L2 inset (upper right).

Table 4: FilmBench L3 sub-metric definitions, part 2: the Temporal Continuity (TC) and Aesthetic Quality (AQ) axes.

L1	L2 component	L3 sub-metric	Definition (5-point rubric anchor)
TC	Spatial coherence	Scene space	Cross-shot spatial/layout consistency; no unexplained jumps or appearing/vanishing elements.
		Key props	Key props keep count, form, material and color; no flicker or vanishing.
		Character positioning	Character positioning/orientation stays temporally consistent.
	Temporal coherence	Temporal logic	Events unfold with coherent causal/temporal logic, no fragmentation.
		Thematic-style stability	Style / tone / grading stay unified over time, no drift.
	Character coherence	Cross-shot appearance	Same character keeps face and overall appearance throughout, no ID drift.
AQ	Audio coherence	Audio	Timbre / volume / audio-style stay consistent over time.
	Base quality	Sharpness	Image clarity and detail resolution.
		Physical realism	Props/scene obey real-world physics (excluding character motion).
		Character-motion realism	Character movement is natural and physically plausible; no distortion / clipping.
		Artifacts	Free of AI artifacts (flicker, noise, extra limbs, garbled text, seams).
		Perspective	Correct scaling, geometry and spatial consistency; no perspective conflict.
	Editing	Lighting	Correct exposure, consistent light/shadow, atmosphere and dimensionality.
		Editing fluency	Smooth multi-shot editing: subject continuity, shot-scale/axis consistency, eyeline match, pacing.
	Performance	Camera-work appeal	Whether camera work is compelling, purposeful and serves the narrative.
		Emotional performance	Expressiveness and believability of emotional acting.
	Audio quality	Action performance	Quality and believability of physical/action performance.
		Audio quality Dialogue clarity	Overall audio production quality. Intelligibility and clarity of dialogue.

D Performance Degradation Table

Tables 5 and 6 detail the performance drops observed in single- to multi-shot and dialogue-to-action transitions, supporting the analysis in Section 5.9.

E Per-axis Scores

This appendix reports per-model breakdowns for T2V and R2V under the sample-level *stand* aggregation used throughout the paper (machine scores over the full prompt set; T2V $N = 515$, R2V $N = 654$). Model abbreviations: Seed (Seedance 2.0), HH1.1/HH1.0 (HappyHorse 1.1/1.0), KV3O (Kling 3.0 Omni), KV3 (Kling 3.0), Grok (Grok Imagine Video), Veo (Veo 3.1), ViduQ3/ViduQ2 (Vidu Q3-Pro/Q2-Pro), Hailuo (Hailuo 2.3).

Tables 7–9 give the per-axis breakdown for both tasks, complementing the overall rankings in the main text.

Table 10 summarizes T2V robustness across markets: the top ranks are stable across Chinese and international source titles (Seedance 2.0 is #1 in both).

Table 5: T2V multi-shot performance degradation by taxonomy level: the worst-hit L1 axis, the three sharpest L2 components and the five sharpest L3 sub-metrics. Values are mean scores (0–100) on single- vs multi-shot prompts, macro-averaged over the nine models; Δ is the multi-shot drop.

Level	Dimension	Single	Multi	Δ
L1	Instruction Following	91.9	80.4	−11.5
	Temporal Continuity	97.8	91.0	−6.8
	Aesthetic Quality	77.7	72.4	−5.3
L2	Cinematic Language	86.6	69.4	−17.2
	Editing	75.0	61.5	−13.5
	Spatial coherence	96.0	84.5	−11.5
L3	Editing fluency	97.3	67.5	−29.8
	Focus	91.7	68.7	−23.0
	Camera movement	87.0	65.1	−21.9
	Viewing angle	88.6	74.3	−14.3
	Scene space	93.4	80.1	−13.3

Table 6: T2V action performance degradation by taxonomy level: the worst-hit L1 axis and the sharpest L3 sub-metrics, split into a camera-work cluster (Instruction Following) and a physical-realism cluster (Aesthetic Quality). Values are mean scores (0–100) on dialogue vs. action prompts, macro-averaged over the nine models; Δ is the action drop.

Level	Dimension	Dialogue	Action	Δ
L1	Instruction Following	85.7	73.5	−12.2
	Aesthetic Quality	75.9	65.5	−10.4
L3	Camera movement	76.5	45.4	−31.1
	Focus	81.5	52.9	−28.6
	Shot scale	72.9	46.3	−26.6
	Physical plausibility	86.3	69.3	−16.9
	Character-motion realism	61.8	45.9	−15.8
	Perspective	94.5	79.2	−15.3
	Artifacts	67.6	55.7	−11.9

Table 7: Per-axis scores (0–100) by model, T2V ($N = 515$). Overall is the equal-weighted mean of the three axes.

Model	Overall	Instruction	Temporal	Aesthetic
Seedance 2.0	88.9	91.7	96.1	79.0
HappyHorse 1.1	87.4	90.7	94.8	76.8
HappyHorse 1.0	87.0	89.5	94.5	76.9
Kling 3.0 Omni	86.1	88.3	94.7	75.1
Kling 3.0	85.8	87.5	94.6	75.2
Vidu Q3 Pro	80.9	82.0	90.1	70.6
Grok Imagine Video	80.9	82.0	89.0	71.6
Veo 3.1	80.8	78.5	90.4	73.6
Hailuo 2.3	68.9	55.9	88.1	62.7

Table 8: Per-axis scores (0–100) by model, R2V ($N = 654$).

Model	Overall	Instruction	Temporal	Aesthetic
Seedance 2.0	86.7	87.5	94.1	78.4
HappyHorse 1.1	85.5	86.7	93.3	76.5
HappyHorse 1.0	84.9	85.0	93.4	76.2
Kling 3.0 Omni	80.4	78.8	90.5	71.9
Veo 3.1	78.2	74.7	89.2	70.8
Grok Imagine Video	76.9	71.2	90.3	69.3
Vidu Q2 Pro	71.4	57.4	88.1	68.8

Table 9: R2V Visual-Following fidelity (0–100) by reference type. Each R2V prompt is scored only on the sub-metric matching its reference type (scene $N = 232$, character $N = 213$, prop $N = 209$), so columns report scores on disjoint prompt subsets.

Model	Scene-space	Character-appearance	Prop-reference
Seedance 2.0	88.2	90.5	53.8
HappyHorse 1.1	84.0	93.1	54.5
HappyHorse 1.0	90.5	93.3	53.7
Kling 3.0 Omni	86.9	85.8	47.9
Veo 3.1	74.3	73.3	44.3
Grok Imagine Video	85.2	86.0	48.6
Vidu Q2 Pro	79.4	83.3	42.7

Table 10: T2V overall scores (0–100) on the market-labeled prompt subset. Columns restrict to prompts drawn from Chinese (1,431 records) vs. international (1,764 records) source titles; prompts without a market label are excluded.

Model	CN	INTL
Seedance 2.0	90.3	90.4
HappyHorse 1.1	89.1	89.5
HappyHorse 1.0	89.1	89.1
Kling 3.0 Omni	88.1	88.7
Kling 3.0	87.7	88.7
Vidu Q3 Pro	84.2	84.2
Grok Imagine Video	84.4	85.2
Veo 3.1	84.5	84.6
Hailuo 2.3	73.6	74.3

F Full L3 / L2 Heatmaps

For completeness, Figures 20–23 give the complete per-model score matrices at both granularities. Rows are grouped by L1 axis (and L2 component for the L3 maps); columns are ordered by overall rank. Darker cells are higher scores. These maps make the low-variance rows (temporal/audio) and the discriminative rows (camera/framing, dynamic aesthetics, and, for R2V, prop-reference fidelity) visible at a glance, and corroborate the aggregated rankings in the main text.

G FilmOps Taxonomy Details

FilmOps grounds the Cinematic Language dimensions by mapping each frame or shot into structured cinematic labels. Its taxonomy spans six core dimensions and 55 fixed sub-categories (character layout is an open-ended natural-language field and is not counted), with all category definitions aligned to classical production references and validated by practitioners. Table 11 summarizes the dimensions, and the category inventories are listed below.

Table 11: FilmOps taxonomy: six operators, their granularity, number of label classes and label form. Total: 55 fixed sub-categories (character layout is an open-ended natural-language field, not counted).

Dimension	Granularity	# Classes	Label form
Shot scale	Frame	8	single-label
Composition	Frame	12	multi-label
Viewing angle	Frame	7	multi-label
Tone & color	Frame	18	multi-label (3 sub-axes)
Character layout	Frame	–	natural language
Camera movement	Shot	10	multi-label

Category inventories. **Shot scale** (8): extreme close-up, close-up, close shot, medium shot, medium full shot, full shot, long shot, extreme long shot. **Composition** (12): center, rule of thirds,

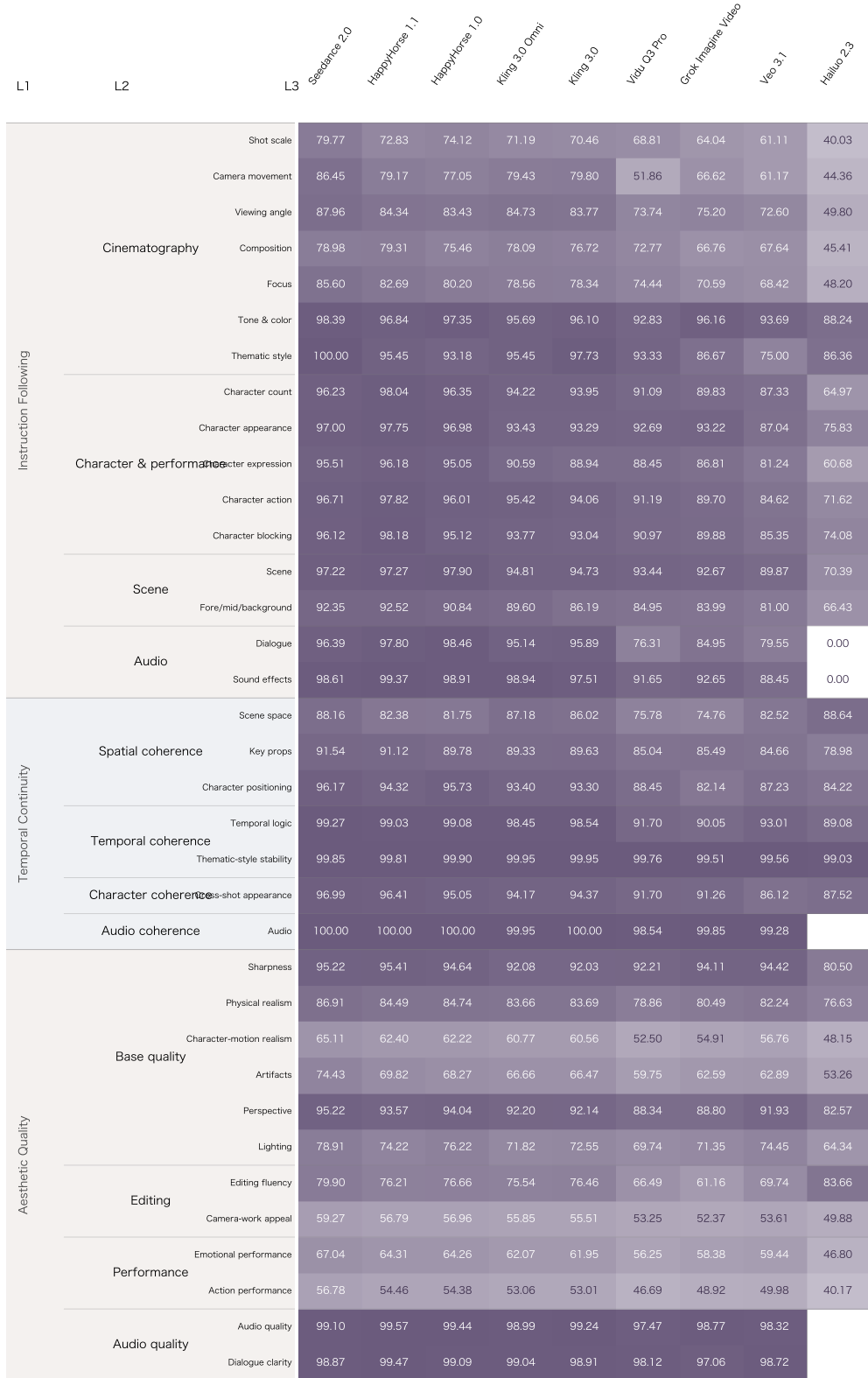


Figure 20: T2V per-model L3 heatmap (35 sub-metrics × 9 models).

L1	L2	L3	Sdance 2.0	Happyhorse 1.1	Happyhorse 1.0	Kling 3.0 Omni	Veo 3.1	Grok Imagine Video	Wukong 2 Pro
Instruction Following	Cinematography	Shot scale	75.22	70.30	67.35	62.06	54.19	50.86	28.71
		Camera movement	78.28	72.88	64.08	64.26	40.65	45.68	23.56
		Viewing angle	83.41	79.39	77.63	75.31	65.51	61.41	39.00
		Composition	73.43	72.58	71.11	70.25	62.09	54.53	42.44
		Focus	76.57	75.27	73.62	62.27	59.90	52.21	27.84
		Tone & color	96.12	96.60	97.92	92.32	91.94	89.75	73.73
	Character & performance	Thematic style	99.29	97.79	94.85	88.64	85.34	92.36	72.06
		Character count	95.48	96.78	95.18	89.11	87.36	80.83	69.05
		Character appearance	95.10	96.97	95.89	88.35	91.35	88.61	81.19
		Character expression	93.15	93.72	91.77	80.74	80.92	76.37	58.39
		Character action	95.35	96.28	94.16	85.05	86.19	82.67	72.13
		Character blocking	94.82	94.63	93.56	82.15	85.72	81.97	71.38
	Scene	Scene	96.87	98.11	97.51	90.95	90.54	85.70	73.86
		Fore/mid/background	88.64	89.71	85.30	72.45	75.33	67.59	49.83
	Audio	Dialogue	95.67	96.97	95.82	91.78	80.21	76.40	59.82
		Sound effects	97.39	97.48	96.72	98.62	87.65	86.88	75.66
	Visual Following fidelity	Scene-space fidelity	88.21	84.05	90.50	86.95	74.32	85.17	79.39
		Character-appearance fidelity	90.45	93.12	93.29	85.78	73.33	85.97	83.33
		Prop-reference fidelity	53.83	54.53	53.71	47.88	44.26	48.59	42.67
Temporal Continuity	Spatial coherence	Scene space	80.39	76.46	80.05	77.64	80.28	82.19	85.21
		Key props	91.31	89.63	88.47	85.44	83.99	86.33	81.99
		Character positioning	93.35	93.31	92.13	87.31	85.21	83.55	78.56
	Temporal coherence	Temporal logic	97.90	98.81	98.20	93.58	92.85	93.31	84.82
		Thematic-style stability	99.81	100.00	99.77	99.08	99.16	98.24	99.77
	Character coherence	Cross-shot appearance	95.64	94.46	94.95	90.14	83.03	89.07	86.85
	Audio coherence	Audio	100.00	100.00	100.00	99.69	99.50	99.57	99.35
Aesthetic Quality	Base quality	Sharpness	95.96	95.70	95.67	92.40	92.96	91.51	92.49
		Physical realism	88.17	85.21	85.11	79.96	78.54	77.87	75.28
		Character-motion realism	61.25	58.03	57.73	54.64	51.40	51.00	46.59
		Artifacts	76.99	73.00	72.54	69.20	61.09	64.09	58.66
		Perspective	94.66	92.92	92.51	87.99	87.27	84.53	84.44
		Lighting	77.01	74.21	75.46	67.00	67.54	67.25	65.22
	Editing	Editing fluency	72.02	70.80	68.92	63.28	62.65	59.21	81.19
		Camera-work appeal	63.73	61.58	61.81	55.23	55.29	53.09	52.12
	Performance	Emotional performance	63.84	61.83	60.82	54.80	53.76	51.20	43.58
		Action performance	53.34	50.91	50.50	45.76	44.60	42.63	39.50
	Audio quality	Audio quality	99.75	99.66	99.61	99.07	98.75	98.88	98.46
		Dialogue clarity	99.46	99.71	99.73	99.61	97.75	97.80	96.98

Figure 21: R2V per-model L3 heatmap (38 sub-metrics $\times$ 7 models; adds the three Visual Following fidelity rows).

L1		L2								
		Seedance 2.0	HappyHorse 1.1	HappyHorse 1.0	Kling 3.0 Omni	Kling 3.0	Vidu Q3 Pro	Grok Imagine Video	Veo 3.1	Haiku 2.3
Instruction Following	Cinematic Language	84.60	80.51	79.32	79.53	78.81	69.83	70.41	67.71	47.85
	Character & performance	96.35	97.59	95.99	93.43	92.70	90.97	89.91	85.17	69.33
	Scene	96.00	95.90	96.01	93.31	92.50	90.87	89.95	87.42	67.70
	Audio	96.94	98.14	98.55	96.32	96.39	81.25	87.10	82.71	0.00
Temporal Continuity	Spatial coherence	92.10	89.26	89.15	90.07	89.85	82.90	80.82	84.90	84.26
	Temporal coherence	99.56	99.42	99.49	99.20	99.25	95.73	94.78	96.29	94.05
	Character coherence	96.99	96.41	95.05	94.17	94.37	91.70	91.26	86.12	87.52
	Audio coherence	100.00	100.00	100.00	99.95	100.00	98.54	99.85	99.28	
Aesthetic Quality	Base quality	82.63	79.99	80.02	77.87	77.90	73.57	75.38	77.11	67.58
	Editing	69.59	66.50	66.85	65.69	65.98	59.87	56.76	61.68	66.77
	Performance	61.86	59.33	59.29	57.27	57.14	51.34	53.61	54.55	42.86
	Audio quality	99.05	99.52	99.35	98.91	99.16	97.85	98.15	98.58	

Figure 22: T2V per-model L2 heatmap (12 components $\times$ 9 models).

horizontal, vertical, symmetric, framing, scattered, leading lines, diagonal, oblique, triangular, depth of field. **Viewing angle** (7): eye level, low angle, high angle, bird’s eye, extreme low angle, extreme high angle, Dutch angle. **Tone & color** (18, three sub-axes): hue (red, orange, yellow, green, cyan, blue, purple, magenta, pink, brown, monochrome, white); temperature (cool, warm, mixed); and saturation (high, medium, low). **Character layout**: an open-ended natural-language description of each main character’s on-screen position and orientation, expressed in image-frame coordinates. **Camera movement** (10): push in, pull out, pan, tracking, static, arc, follow, roll, zoom, dynamic (strong movement).

H FilmOps Operator Evaluation

Table 12 reports, for each FilmOps operator, its backbone and its accuracy against four strong zero-shot general MLLM baselines (GPT-4.1, Qwen3-VL-235B, Gemini 3.5 Flash and Gemini 3.1 Pro). Classification operators are scored by macro-F1 and the natural-language character-layout operator by precision. The trained operators outperform every baseline on all six dimensions, most sharply on the temporally/professionally demanding ones (camera movement, composition, tone & color), corroborating the design discussion in the main text (Section 4.1).

L1		L2						
		Seedance 2.0	HappyHorse 1.1	HappyHorse 1.0	Kling 3.0 Omni	Veo 3.1	Grok Imagine Video	Vidu Q2 Pro
Instruction Following	Cinematic Language	78.40	74.98	71.66	68.34	58.91	54.86	36.36
	Character & performance	94.89	95.86	94.28	85.34	86.49	82.13	70.52
	Scene	95.49	96.75	95.62	87.77	88.20	82.71	69.85
	Audio	95.95	97.00	95.94	92.69	81.18	77.16	61.61
	Visual Following	77.92	77.51	79.65	74.06	64.39	73.69	68.92
Temporal Continuity	Spatial coherence	88.28	86.31	86.79	83.30	83.16	83.91	81.84
	Temporal coherence	98.85	99.41	98.99	96.33	96.01	95.78	92.30
	Character coherence	95.64	94.46	94.95	90.14	83.03	89.07	86.85
	Audio coherence	100.00	100.00	100.00	99.69	99.50	99.57	99.35
Aesthetic Quality	Base quality	82.34	79.85	79.84	75.20	73.15	72.76	70.46
	Editing	67.87	66.19	65.37	59.26	58.97	56.15	66.66
	Performance	58.54	56.32	55.60	50.06	49.00	46.75	41.39
	Audio quality	99.64	99.62	99.66	99.25	98.27	98.46	97.93

Figure 23: R2V per-model L2 heatmap (13 components $\times$ 7 models; adds the Visual Following component).

Table 12: FilmOps operators vs. four zero-shot general MLLM baselines. Classification operators are scored by macro-F1; the natural-language character-layout operator by precision (“–” denotes an unsupported setting). "F" and "P" stand for "Flash" and "Pro".

Operator	Backbone	GPT-4.1	Qwen3-VL	Gemini 3.5 F	Gemini 3.1 P	FilmOps
Shot scale	DINO (ViT-B)	0.483	0.615	0.733	0.725	0.906
Composition	DINO (ViT-B)	0.330	0.289	0.370	0.500	0.859
Viewing angle	BEiT-Base	0.497	0.452	0.534	0.560	0.810
Tone & color	ResNet-18	0.522	0.503	0.512	0.565	0.828
Character layout	InternVL3-14B	0.569	0.556	0.698	0.692	0.881
Camera movement	InternVL3-14B	–	0.277	0.318	0.324	0.795