Sci-VBench: Evaluating Knowledge- and Reasoning-Intensive Video Generation in Science Domains

Diandian Zhang[1*] **Tingyu Song**[2*] **Lin Fu**[1*] **Zheyuan Yang**[3] **Yilun Zhao**[4]

[1] Zhejiang University [2] UCAS [3] Tongji University [4] Yale University

Abstract

We introduce Sci-VBench, a comprehensive benchmark for evaluating knowledge- and reasoning-intensive video generation across scientific domains. It contains 1,253 expert-annotated examples spanning 60 subjects across four core disciplines: Natural Science, Healthcare, Humanities & Social Sciences, and Engineering. Each example requires models to generate temporally rich videos that demand scientific reasoning and knowledge-grounded synthesis, going beyond surface-level visual plausibility. We further establish a rubric-based evaluation protocol. Our analysis shows that, under this protocol, both non-expert human evaluators and MLLM-as-Judge systems can achieve relatively high agreement with expert judgments, supporting reproducible evaluation at scale. We benchmark 16 frontier proprietary and open-source models and find that, while automatic perceptual-quality scores cluster tightly across systems, performance on Prompt Grounding and Scientific and Causal Correctness varies substantially, with a pronounced proprietary–open-source gap. These findings show that advances in visual realism have not yet translated into reliable modeling of scientific and causal dynamics.

🤗 Sci-VBench Sci-VBench

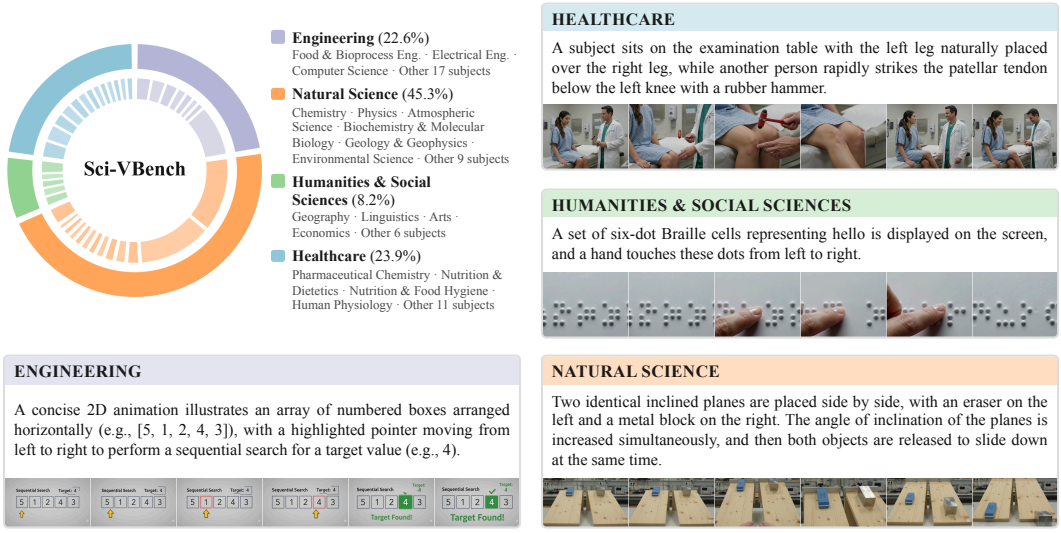


Figure 1: Overview of the Sci-VBench dataset. **Top left:** distribution of the 1,253 expert-annotated examples over 60 subjects across four core disciplines. **Remaining panels:** representative prompt–video examples from each discipline, generated by Gemini-Omni-Flash, where faithful generation requires grounding in the underlying scientific mechanism.

*Equal contributions. Correspondence to: Tingyu Song (songtingyu23@mails.ucas.ac.cn), Yilun Zhao (yilun.zhao@yale.edu).

1 Introduction

Video generative models have advanced rapidly in visual fidelity, motion coherence, and controllability (Ma et al., 2025; The Sora Team, 2025; Google DeepMind, 2025; Wan AI Team, 2025). These gains shift the central question from whether a generated video looks plausible to whether it faithfully realizes the events, constraints, and temporal dependencies specified by the prompt. This question is particularly consequential in science-domain settings, where perceptual plausibility can mask fundamental errors in the underlying process. A video may look convincing while violating a conservation law, reversing a causal relation, or depicting an impossible state transition (Bansal et al., 2025a; Meng et al., 2025). Reliable science-domain video generation therefore requires more than visual realism: it requires knowledge-grounded, temporally consistent synthesis that preserves mechanistic fidelity.

Evaluation, however, has not kept pace with this shift. Existing benchmarks provide increasingly broad coverage of perceptual quality, prompt alignment, and compositionality (Huang et al., 2024; Zheng et al., 2025; Huang et al., 2026; Sun et al., 2025), while reasoning-oriented suites largely focus on generic physical commonsense (Meng et al., 2025; Bansal et al., 2025a;b; Li et al., 2025a). Meanwhile, knowledge- and reasoning-intensive evaluation has primarily been studied for video understanding (Zhao et al., 2025c; Hu et al., 2025a; Deng et al., 2025; Fu et al., 2026; Shangguan et al., 2025). Recent work has begun to examine scientific reasoning in video generation (Hu et al., 2025b), but a multidisciplinary benchmark that tests mechanisms across scientific domains and makes expert-defined evaluation criteria portable across raters is still missing.

To address this gap, we introduce Sci-VBench, a benchmark for knowledge- and reasoning-intensive, visually verifiable video generation across scientific domains. It comprises 1,253 expert-authored and independently reviewed generation tasks spanning 60 subjects across four core disciplines: Natural Science, Healthcare, Humanities & Social Sciences, and Engineering. Each task pairs a minimal generation prompt with an expert-authored evaluation specification consisting of (i) a reference guide that identifies the target concept and expected phase-based phenomena and (ii) a detailed rubric with observable 1–5 scoring anchors. This design requires models to infer and render the underlying mechanism while providing evaluators with a consistent expert-defined standard.

We establish and validate a scalable rubric-based evaluation protocol. Expert human ratings serve as the reference labels. For automatic benchmarking, we combine VBench-based Vision Tools (VT) for Low-level Perceptual Fidelity with rubric-conditioned MLLM-as-judge scoring for Prompt Grounding, Scientific and Causal Correctness, and Spatiotemporal Consistency. A controlled human study shows that providing the evaluation specification substantially improves non-expert agreement with experts, while independent expert re-rating confirms the stability of the reference judgments (Cohen's $\kappa = 0.842$). Across multiple MLLM evaluators, our rubric-conditioned protocol also aligns more closely with expert judgments than prior video scoring methods (He et al., 2024; 2025; Guan et al., 2025).

We evaluate 16 frontier proprietary and open-source text-to-video models on Sci-VBench. Both expert and automatic evaluations reveal a pronounced proprietary–open-source gap concentrated in Prompt Grounding and Scientific and Causal Correctness, whereas the automatic perceptual-quality proxy is nearly flat across systems. Qualitative analysis further shows that even leading models produce systematic mechanistic errors despite visually convincing outputs. Prompt rewriting improves Prompt Grounding and Scientific and Causal Correctness but has limited effect on Spatiotemporal Consistency, suggesting that many temporal failures arise from limitations of the underlying generators rather than underspecified instructions.

2 Related Work

2.1 Video Generative Models

Recent progress in video generation (Ma et al., 2025) has been propelled by the synergy of diffusion models and large-scale transformer architectures, carrying text-to-video

Benchmark	Domain	Focus	Examples	Expert	Eval Spec
VBench	General	Video quality	946	✗	✗
PhyGenBench	Physics	Physical commonsense	160	✗	✗
VideoPhy	Physics	Physical commonsense	688	✗	✗
VideoPhy-2	Physics	Action-centric physical commonsense	3,940	✗	✗
WorldModelBench	General	World modeling	350	✗	✗
V-ReasonBench	General	Structured, spatial, pattern, physical	326	✗	✗
VideoScience-Bench	Phys./Chem.	Scientific reasoning	200	✓	✗
Sci-VBench	60 sci. subjects	Knowledge-, Reasoning-intensive	1,253	✓	✓

Table 1: Comparing existing video-generation benchmarks with Sci-VBench. "Expert" indicates whether domain experts are involved in benchmark construction or annotation. "Eval Spec" denotes released, reusable per-example evaluation guides or rubrics.

(T2V) systems from the short, low-fidelity clips of early models (Li et al., 2025b; Henschel et al., 2024) to the high-fidelity, temporally coherent outputs of current video foundation models (The Sora Team, 2025; Google DeepMind, 2025). Consequently, a significant focus of current research (Xue et al., 2025; Gillman et al., 2025; Hu et al., 2025c) is on enhancing their adherence to physical laws and commonsense reasoning. Adjacent work has studied scientific multimodal reasoning and static scientific visual generation (Wang et al., 2025a; Chen et al., 2026; Zhao et al., 2025b; Wang et al., 2025b).

2.2 Evaluation of Video Generative Models

Current evaluation benchmarks (Huang et al., 2024; He et al., 2024; Li et al., 2025a) for video generation primarily assess foundational capabilities such as perceptual fidelity and motion quality. Some evaluation frameworks (Huang et al., 2026; Zheng et al., 2025; Sun et al., 2025) further probe compositionality, object interaction, and higher-level prompt following. These benchmarks are not designed to test whether a video remains faithful to domain-specific mechanisms and rubric-verifiable scientific outcomes. Several recent benchmarks (Bansal et al., 2025a;b; Li et al., 2025a) move toward reasoning-focused evaluation by testing whether generated videos obey physical laws, but they remain grounded in generic physical commonsense rather than science-domain knowledge. In parallel, prior work finds persistent reliability gaps in model-based evaluation for AI-generated videos and complex scientific tasks (Song et al., 2025; Zhao et al., 2025a;d), underscoring the need for structured evaluation specifications such as the per-example rubrics used in Sci-VBench. The closest concurrent benchmark to ours is VideoScience-Bench (Hu et al., 2025b), which evaluates scientific phenomena in physics and chemistry through 200 expert-annotated prompts. In contrast, Sci-VBench spans 60 subjects, provides reusable per-example reference guides and rubrics, and studies evaluation portability across expert, non-expert, and MLLM-based raters. Table 1 summarizes these differences, comparing Sci-VBench with existing video-generation benchmarks in terms of domain coverage, evaluation focus, expert involvement, and released evaluation specifications.

3 Sci-VBench Benchmark

To ensure high data quality and rigorous assessment, Sci-VBench is constructed around four core desiderata: **(1) Breadth of domain knowledge:** 1,253 examples spanning 60 subjects across four disciplines, from astrophysics to public policy. **(2) Depth of expert-level reasoning:** every prompt is authored by a domain expert, so generating it correctly demands grounded scientific understanding and multi-step causal reasoning. **(3) Completeness of spatiotemporal reasoning:** the target phenomenon must unfold over time, so no single well-chosen frame can satisfy the prompt. **(4) Reliable and reproducible evaluation:** a rubric-based protocol with per-example scoring anchors, released so that other groups can apply the same standard without recruiting domain experts.

In the following subsections, we first describe the evaluation dimensions and then detail the benchmark construction pipeline, with an overview shown in Figure 2.

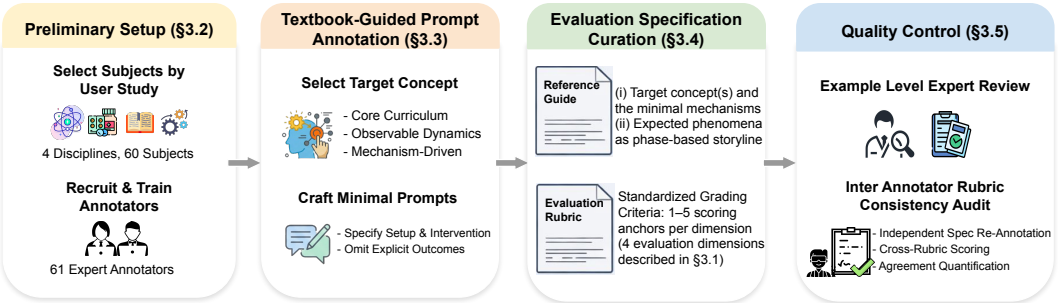


Figure 2: Overview of the Sci-VBench benchmark construction process.

3.1 Evaluation Dimensions

Low-level Perceptual Fidelity and *Prompt Grounding* follow established concerns in prior video-generation evaluation, while *Scientific and Causal Correctness* and *Spatiotemporal Consistency* isolate mechanism-level failures that generic quality or prompt-alignment criteria do not capture: **(1) Low-level Perceptual Fidelity:** the perceptual quality of the synthesized video, including temporal coherence, motion dynamics, and the visual quality of individual frames. **(2) Prompt Grounding:** whether the video faithfully instantiates the explicit conditions stated in the prompt, such as the presence and identity of key entities, their initial states, spatial arrangement, and any required instruments. **(3) Scientific and Causal Correctness:** whether the generated dynamics are consistent with the domain knowledge and causal mechanisms targeted by the prompt, as specified by the per-example reference guide. **(4) Spatiotemporal Consistency:** whether the video maintains the coherent temporal evolution and stable spatial relationships needed for the intended mechanism to be interpretable.

3.2 Preliminary Setup for Benchmark Construction

Subject Selection. To ensure broad, faithful coverage of knowledge- and reasoning-intensive video generation across diverse disciplines, we conducted a user study with 133 undergraduate and graduate students to inform subject selection. Participants are asked to curate two video prompts requiring expert-level reasoning on topics related to their field of study and to provide feedback on their experiences during the curation process. The authors then manually analyzed the collected prompts together with the corresponding videos generated by Sora (The Sora Team, 2025) and Wan 2.2 (Team Wan et al., 2025), and selected the 60 subjects (listed in Table 5 in Appendix A.1) across four disciplines whose core concepts are both expert-level and verifiable from video evidence.

Expert Annotator Recruitment and Training. Each subject is assigned to annotators with matching expertise, and every example is authored by one annotator and independently reviewed by another (Section 3.5). We include a total of 61 expert annotators (detailed biographies are presented in Appendix A.4); based on their current academic status, this pool comprises 11 undergraduate students, 45 graduate students, and five of the authors. All the annotators also participated in our initial user study. Each annotator completes a training session on the annotation protocol before contributing examples.

3.3 Video Prompt Annotation

We annotate video prompts through a textbook-guided pipeline. Specifically, for each subject, expert annotators first select a target concept (or tightly coupled set of concepts) from canonical textbooks and course materials that is representative of the subject's core curriculum and naturally expressed through observable spatiotemporal dynamics. Accordingly, Sci-VBench excludes expert concepts whose correctness cannot be verified from video evidence alone. We require that selected concepts have mechanism-governed visual realizations (*e.g.*, reaction dynamics in chemistry, conservation-driven interactions in engineering, and

intervention response in healthcare), so that videos may look superficially plausible yet still produce clear and systematic deviations when the underlying principles are violated. For each selected concept, annotators craft a minimal prompt that specifies only the observable initial setup and any explicit intervention or task objective, omitting the expected mechanistic trajectory and the key phenomena to be generated, so that a model must infer the mechanism from the setup rather than reproduce an outcome the prompt already describes.

3.4 Evaluation Specification Curation

Constructing Sci-VBench requires domain experts, but future users of the benchmark cannot be expected to recruit them at scale. We therefore release an evaluation specification with each prompt, pairing a **high-level reference guide** that fixes how the task should be interpreted with a detailed **evaluation rubric** that operationalizes scoring under that interpretation, using 1–5 anchors per dimension. This externalizes the expert knowledge evaluation requires, so that non-experts or MLLM judges can score generated videos under the same standard; we test that portability empirically in Section 4.3.

High-level Reference Guide Annotation. We instruct annotators to write a high-level reference guide for each prompt that (i) specifies the target concept(s) and the minimal mechanistic assumptions required for the scenario to be well-defined, and (ii) summarizes the expected phenomena as a concise phase-based storyline, including key causal transitions and any visibility/viewpoint constraints needed for verification. Moreover, to reduce terminology barriers for downstream evaluators applying the released specification, we instruct annotators to anticipate the knowledge gaps of a non-expert verifier and selectively provide brief clarifications.

Evaluation Rubric Annotation. We instruct annotators to produce a detailed evaluation rubric aligned with our evaluation dimensions (discussed in Section 3.1), defining 1–5 scoring anchors per dimension and tying each anchor to observable evidence. The rubric specifies what evidence is sufficient for full credit, what constitutes partial correctness, and which violations or omissions warrant low scores. Prompt Grounding, Scientific and Causal Correctness, and Spatiotemporal Consistency receive per-example rubrics, since what counts as evidence depends on the mechanism the example targets. Low-level Perceptual Fidelity instead uses a single rubric shared by every example: its anchors describe generic video-quality properties that do not depend on the scientific content.

3.5 Data Quality Control

Table 2 reports per-discipline example counts together with the average length of each prompt, reference guide, and rubric. Every example is reviewed in a second pass by an independent domain expert, who checks that the prompt is clear and fully specified, that the intended mechanism is visually testable in the described scene, and that the prompt, reference guide, and rubric are mutually consistent. The reviewer revises any example that fails these checks, and an author verifies the revision before the example is finalized. We further audit 200 randomly sampled examples by having a second annotator write an independent specification for the same prompt and two independent scorers grade the same video under both. Agreement is high on every reasoning dimension (quadratic weighted $\kappa = 0.79$ for Prompt Grounding, 0.75 for Scientific and Causal Correctness, and 0.73 for Spatiotemporal Consistency), indicating that a score is determined by the released specification rather than by who wrote it. Appendix A.2 details the design. Sci-VBench is released with two evaluation splits: **full**, containing all 1,253 examples, and ***testmini***, a fixed

Feature	Eng.	Health.	Soc.	Nat.
# Example	283	300	103	567
Length				
Video Prompt	42.6	40.5	35.8	33.3
Ref. Guide	163.3	161.3	164.5	160.1
LPF Rubric	245.0	245.0	245.0	245.0
PG Rubric	288.7	300.3	295.3	280.6
SCC Rubric	294.7	292.2	315.2	283.7
SC Rubric	327.6	344.9	321.5	313.5

Table 2: Sci-VBench statistics by discipline.

subset of 150 examples (37 Engineering, 16 Healthcare, 78 Natural Science, and 19 Humanities & Social Sciences) that supports rapid iteration and cost-constrained evaluation; in our experiments, open-source models are evaluated on both splits, while proprietary systems are evaluated on *testmini* only, as generating the full benchmark through commercial APIs is prohibitively expensive (Section 5.1).

4 Sci-VBench Evaluation Protocol

In this section, we first describe our human and automated evaluation protocols. We further present a detailed analysis of the reliability of our automated evaluation protocols.

4.1 Human Evaluation Protocols

Expert human ratings serve as the primary reference labels for each evaluation dimension. Specifically, for each generated video, an expert annotator in the corresponding discipline is provided with the text prompt, the generated video, and the per-example evaluation specification, and then assigns 1–5 integer scores for all four dimensions. To ensure consistent interpretation of the rubric, annotators are trained with a brief calibration set and guided to map concrete, observable cues (*e.g.,* key state transitions, causal consistency, and verifiable evidence in the video) to the corresponding score anchors.

4.2 Automated Evaluation Protocols

Expert ratings are labor-intensive and do not scale to the full benchmark or to broad model comparisons. We therefore introduce an automated protocol that scores at scale while remaining anchored to the same per-example specifications used in human evaluation. For model benchmarking, LPF is reported via VBench-based Vision Tools (VT) and the remaining three dimensions via the rubric-conditioned judge.

Automatic Metrics for Low-level Perceptual Fidelity. We assess low-level perceptual fidelity with the established automatic metrics used throughout video-generation evaluation, so that this dimension stays directly comparable to prior work. Following VBench (Huang et al., 2024), we adopt the six metrics under its *Video Quality* dimension (with definitions and implementations detailed in Appendix B.1). We use the official VBench evaluation protocol to compute these metrics, then normalize and average them into a single composite score, linearly mapped onto the same 1–5 range as the rubric-based dimensions, which we report as the Vision Tools (VT) score.

MLLM-as-Judge for All Other Dimensions. We evaluate Prompt Grounding, Scientific and Causal Correctness, and Spatiotemporal Consistency using a rubric-conditioned MLLM-as-judge pipeline, instantiated with the open-source Qwen3.5-397B-A17B for reproducibility and native video input. Specifically, for a given video–prompt pair and target dimension, MLLM-as-Judge is provided with (1) the original text prompt, (2) the generated video, (3) the per-example high-level reference guide, and (4) the 1–5 anchored rubric for that dimension only, which keeps evidence for one dimension from bleeding into another. MLLM-as-Judge is required to state a brief justification grounded in the video before assigning a single 1–5 integer score for that dimension. The single-dimension evaluation prompt is shown in Figure 5 in Appendix B.3. To reduce run-to-run variance, every video–dimension pair is scored in three independent judge runs, and we report the mean of the three scores.

4.3 Measuring Reliability of Evaluation Protocol

We assess the reliability of our evaluation protocol by measuring agreement between human experts and (1) human non-experts, with and without evaluation specifications, and (2) MLLM judges provided with evaluation specifications.

Collecting Reference Expert Ratings. Every *testmini* video is rated by an expert, and Table 4 reports those ratings for all 16 models. For the reliability analysis in this section we use the 1,500 videos generated by the ten systems released on or before January 2026. For each video, we ask an expert annotator from the same discipline, who did not author the corresponding prompt or evaluation specification, to rate video quality on each dimension using a 1–5 scale under the standardized interface described above. These scores are the reference labels for every correlation reported in Table 3. To assess the stability of the reference labels, we randomly sample 300 of them and ask a second independent expert from the same discipline, again not an author of the example, to re-rate them under the same protocol. Across the paired expert ratings, we obtain Cohen's $\kappa = 0.842$, indicating strong inter-expert agreement.

Analyzing Non-Expert Ratings With and Without Evaluation Specifications. Two separate non-expert cohorts score the same expert-scored videos: in the *without evaluation specification* condition evaluators see only the text prompt and the generated video, while in the *with evaluation specification* condition a different cohort additionally receives the full per-example specification (*i.e.,* the high-level reference guide and scoring rubric) and is asked to follow it strictly. As shown in Table 3, providing the evaluation specification increases instance-level Pearson correlation in all four dimensions. The improvement is modest for Low-level Perceptual Fidelity (84.7 vs. 83.2) but substantially larger for Prompt Grounding (83.3 vs. 74.1), Scientific and Causal Correctness (82.5 vs. 68.2), and Spatiotemporal Consistency (79.6 vs. 70.3). With the specification, non-experts agree with experts more closely than any MLLM-as-Judge instantiation we tested, on every dimension.

	LPF	PG	SCC	SC	Avg.
Measuring Reliability of Our Evaluation Protocol					
Human Non-expert					
with Eval Spec.	84.7	83.3	82.5	79.6	82.5
without Eval Spec.	83.2	74.1	68.2	70.3	74.0
Qwen3.5-397B-A17B	53.5	73.9	71.7	54.5	63.4
Gemma-4-31B	51.2	68.5	67.2	54.6	60.4
Qwen3.5-9B	45.9	64.0	64.4	48.3	55.7
Comparison with Prior Auto-Eval Methods					
VideoScore	44.8	58.1	40.6	46.3	47.5
VideoScore2	47.9	60.4	49.8	48.2	51.6
VideoReward	50.7	55.6	38.9	44.5	47.4
ETVA	31.2	62.7	64.1	33.6	47.9

Table 3: Instance-level Pearson correlation ($\times 100$) with expert human ratings. **LPF**: Low-level Perceptual Fidelity, **PG**: Prompt Grounding, **SCC**: Scientific and Causal Correctness, **SC**: Spatiotemporal Consistency.

Analyzing MLLM-as-Judge Ratings Across Base Evaluators. We next assess how closely MLLM-as-Judge scores match those of human experts, evaluating multiple MLLM-as-Judge instantiations that swap the underlying evaluator among open-weight video understanding models of varying scale (*i.e.,* Qwen3.5-397B-A17B, Gemma-4-31B, and Qwen3.5-9B). As shown in Table 3, Qwen3.5-397B-A17B attains the highest correlations overall, and agreement broadly increases with evaluator scale. For every evaluator, alignment is markedly stronger on the reasoning-centric dimensions (Prompt Grounding and Scientific and Causal Correctness) than on Low-level Perceptual Fidelity and Spatiotemporal Consistency, whose fine-grained visual and temporal artifacts remain difficult for general-purpose MLLM judges. We further compare our automated evaluation protocol against representative prior paradigms, including VideoScore (He et al., 2024), VideoScore2 (He et al., 2025), VideoReward (Liu et al., 2025), and ETVA (Guan et al., 2025). As Table 3 shows, the strongest rubric-conditioned MLLM-as-Judge correlates more closely with expert ratings than any of them on every dimension; Appendix B.2 reports the full setup.

5 Experiment

5.1 Evaluated Models

We benchmark 16 frontier text-to-video models on Sci-VBench, spanning eight **proprietary** systems: Sora-2 (The Sora Team, 2025), Veo-3.1-Fast and Veo-3.1 (Google DeepMind, 2025), Kling-2.6 (Kuaishou Technology, 2025), Wan-2.6 (Wan AI Team, 2025), Seedance-2.0 (ByteDance Seed, 2026), HappyHorse-1.1 (Alibaba Cloud, 2026), and Gemini-Omni-Flash (Google, 2026), and eight **open-source** models: HunyuanVideo-

Models	Release	Full Avg.	Automatic Eval					Human Eval				
			VT	SC	PG	SCC	Avg.	LPF	SC	PG	SCC	Avg.
Proprietary Models												
Gemini-Omni-Flash (Google, 2026)	2026-06	–	**4.12**	2.53	**3.52**	**3.34**	**3.38**	**3.76**	2.53	**3.37**	**3.06**	**3.18**
HappyHorse-1.1 (Alibaba Cloud, 2026)	2026-06	–	3.96	2.66	3.21	2.76	3.15	3.75	2.85	3.09	2.57	3.06
Seedance-2.0 (ByteDance Seed, 2026)	2026-02	–	3.86	2.46	3.17	3.06	3.14	3.73	2.32	3.01	2.68	2.94
Veo-3.1 (Google DeepMind, 2025)	2025-10	–	4.10	2.26	2.83	2.75	2.98	3.53	2.66	2.65	2.37	2.80
Veo-3.1-Fast (Google DeepMind, 2025)	2025-10	–	4.03	2.34	2.92	2.70	3.00	3.45	2.71	2.67	2.34	2.79
Sora-2 (The Sora Team, 2025)	2025-09	–	3.82	2.29	3.12	2.72	2.98	3.40	2.29	2.77	2.24	2.68
Kling-2.6 (Kuaishou Technology, 2025)	2025-12	–	3.92	2.59	2.60	1.71	2.70	3.49	**2.91**	2.41	1.65	2.62
Wan-2.6 (Wan AI Team, 2025)	2025-12	–	3.89	2.61	2.43	1.79	2.68	3.42	2.82	2.27	1.58	2.52
Open-source Models												
MiniMax-H3 (MiniMax, 2026)	2026-08	**2.61**	3.94	2.62	2.52	1.63	2.68	3.64	2.83	2.29	1.54	2.58
HunyuanVideo-1.5 (Tencent, 2025)	2025-11	2.43	3.79	2.66	1.87	1.45	2.44	3.45	2.90	1.80	1.21	2.34
LongCat-Video (Meituan LongCat Team et al., 2025)	2025-10	2.34	3.92	2.46	1.76	1.24	2.34	3.45	2.76	1.75	1.13	2.27
Cosmos3-Nano (NVIDIA, 2026)	2026-06	2.43	4.05	2.73	1.76	1.27	2.45	3.37	2.73	1.67	1.14	2.23
Wan2.2-5B (Team Wan et al., 2025)	2025-07	2.43	3.87	**2.79**	1.85	1.33	2.46	3.05	2.82	1.79	1.21	2.22
LTX-2 (HaCohen et al., 2026)	2026-01	2.31	4.07	2.36	1.51	1.26	2.30	3.13	2.74	1.48	1.12	2.12
CogVideoX1.5-5B (Yang et al., 2025)	2024-11	2.30	3.98	2.11	1.92	1.40	2.35	2.68	2.11	1.75	1.19	1.93
LTX-2.3 (HaCohen et al., 2026)	2026-03	2.27	3.96	2.16	1.80	1.28	2.30	2.66	2.15	1.70	1.18	1.92

Table 4: Model performance on the Sci-VBench *testmini* split (150 examples), with **Full Avg.** reporting the automatic average on the full benchmark for open-source models. **VT** = VBench-based Vision Tools, **SC** = Spatiotemporal Consistency, **PG** = Prompt Grounding, **SCC** = Scientific and Causal Correctness, and **LPF** = Low-level Perceptual Fidelity. Both **Avg.** columns are unweighted means over the four dimensions of their block; VT is the automatic proxy for LPF. Provider filters rejected 6 Gemini-Omni-Flash prompts and 2 Seedance-2.0 prompts; affected averages use successful generations. Rows are sorted by Human Eval **Avg.** within each model group; **bold** and underline mark the best and second-best results.

1.5-480P-T2V (Tencent, 2025), LTX-2.0-19B-distilled and LTX-2.3 (HaCohen et al., 2026), LongCat-Video (Meituan LongCat Team et al., 2025), Wan2.2-5B-T2V (Team Wan et al., 2025), CogVideoX1.5-5B (Yang et al., 2025), Cosmos3-Nano (NVIDIA, 2026), and MiniMax-H3 (MiniMax, 2026). All videos are generated from the verbatim benchmark prompts under each model's default configuration, with the per-model version, resolution, frame rate, and clip duration listed in Appendix A.3.

5.2 Main Results

Table 4 reports *testmini* results for all 16 models; complete per-dimension results on the full benchmark are provided in Appendix C.1 (Table 9). For the open-source models, which we run on both splits, per-model full-benchmark averages differ from *testmini* by at most 0.07, confirming *testmini* as a faithful low-cost proxy. We highlight three main findings below. **What separates current systems is mechanism, not appearance.** VT is nearly flat across all 16 models (3.79–4.12); human *Low-level Perceptual Fidelity* ratings of the same construct spread wider (2.66–3.76), so perceptual quality is saturated as far as the VBench-based proxy can resolve it rather than in absolute terms. The reasoning-centric dimensions separate models far more sharply, and do so under both protocols: automatic *Scientific and Causal Correctness* ranges from 1.24 to 3.34, and human SCC from 1.12 to 3.06. What distinguishes current systems on Sci-VBench is less whether videos look right than whether they get the underlying mechanism right. **The proprietary–open-source gap is concentrated on reasoning, not on consistency.** Gemini-Omni-Flash attains the best automatic (3.38) and human (3.18) averages, followed by HappyHorse-1.1

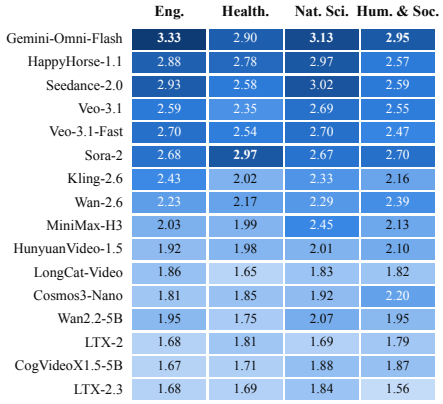


Figure 3: Per-discipline mean of the SC/PG/SCC judge scores on *testmini*.

and Seedance-2.0; all three surpass the strongest earlier systems, Veo-3.1-Fast, Veo-3.1, and Sora-2 (2.98–3.00), and human evaluation reproduces the automatic ordering at the top of the table. Open-source systems fall behind on exactly the reasoning dimensions: MiniMax-H3 leads that group but reaches only 1.63 on SCC, half the proprietary best. On *Spatiotemporal Consistency* they are not behind at all, with Wan2.2-5B attaining the highest automatic score of any model (2.79). **No model is uniformly strong across disciplines.** As shown in Figure 3, no system is strongest everywhere: Gemini-Omni-Flash leads three of the four disciplines (3.33 on Engineering, 3.13 on Natural Science, and 2.95 on Humanities & Social Sciences) but Sora-2 takes Healthcare (2.97), and open-source profiles are similarly uneven. Models whose overall averages nearly coincide can therefore differ substantially in which domain mechanisms they preserve.

5.3 Error Analysis

We classify the observed errors into three major categories. **(1) Poor Adherence to Instructions:** models misinterpret core concepts and miss the fine-grained details specified in the prompt. **(2) Inaccurate Simulation of Scientific Principles:** models prioritize visual aesthetics over physical realism, producing factually incorrect dynamics. **(3) Deficiencies in Temporal Coherence and Visual Quality:** videos suffer from temporal inconsistencies, such as objects changing illogically over time. Appendix C.2 illustrates each category with frames from the evaluated models.

5.4 Effect of Prompt Rewriting

Our main results use verbatim prompts (Section 5.1); as an ablation, we ask how much of the gap more explicit prompting recovers. We rewrite each *testmini* prompt with Gemini-3-Flash, instructing it to restate the scene and the requested dynamics in more visually concrete terms without changing the scenario, and regenerate with Wan2.2-5B and HunyuanVideo-1.5 under unchanged generation and evaluation settings. The gains are ordered consistently across both models (Figure 4): *Scientific and Causal Correctness* (SCC) improves most (+23.3% and +51.7%), then *Prompt Grounding* (+12.4% and +26.7%), while *Spatiotemporal Consistency* gains least (+7.5% and +16.5%). The gap narrows without closing: HunyuanVideo-1.5 reaches 2.20 on SCC, above the best verbatim open-source score in Table 4 (1.63) yet far below the strongest proprietary system (3.34). Explicit wording helps where the prompt left the mechanism implicit, but it cannot supply the mechanistic fidelity the generator lacks.

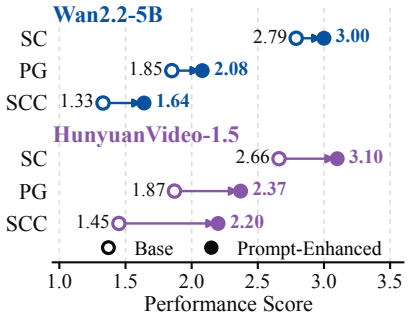


Figure 4: Performance enhancement via prompt rewriting on *testmini*.

6 Conclusion

Sci-VBench fills a key gap in evaluating text-to-video models as expert-domain "world simulators" by providing 1,253 expert-authored prompts spanning 60 subjects across four disciplines, where success hinges on mechanistic reasoning and temporally coherent causal dynamics rather than surface-level realism. Benchmarking 16 frontier proprietary and open-source models on Sci-VBench reveals a persistent gap between perceptual realism and scientific validity, together with a substantial proprietary/open-source performance gap on the reasoning-centric dimensions. The observed failure modes, in turn, point to concrete opportunities to improve instruction adherence, mechanistic consistency, and spatiotemporal stability in expert-domain generation.

References

Alibaba Cloud. Alibaba cloud model studio: Happyhorse text-to-video api reference, 2026. URL https://www.alibabacloud.com/help/en/model-studio/happyhorse-text-to-video-api-reference. Model happyhorse-1.1-t2v, accessed via DashScope.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=9D2QvO1uWj.

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

ByteDance Seed. Seedance 2.0: Advancing video generation for world complexity. *CoRR*, abs/2604.14148, 2026. doi: 10.48550/ARXIV.2604.14148. URL https://doi.org/10.48550/arXiv.2604.14148.

Ziyu Chen, Yilun Zhao, Chengye Wang, Rilyn R. Han, Manasi Patwardhan, and Arman Cohan. Scimdr: Advancing scientific multimodal document reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 44718–44742. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.2070. URL https://doi.org/10.18653/v1/2026.acl-long.2070.

Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *CoRR*, abs/2510.08559, 2025. doi: 10.48550/ARXIV.2510.08559. URL https://doi.org/10.48550/arXiv.2510.08559.

Lin Fu, Zheyuan Yang, Yang Wang, Tingyu Song, Arman Cohan, and Yilun Zhao. Videokr: Towards knowledge- and reasoning-intensive video understanding. *CoRR*, abs/2606.05259, 2026. doi: 10.48550/ARXIV.2606.05259. URL https://doi.org/10.48550/arXiv.2606.05259.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. In Danielle Belgrave, Cheng Zhang, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, and Nancy Chen (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, volume 38, pp. 103174–103201. Curran Associates, Inc., 2025. doi: 10.52202/085713-3448. URL http://papers.nips.cc/paper_files/paper/2025/hash/953dcf20cbd275465066ad63ea111f13-Abstract-Conference.html.

Google. Generate and edit videos with gemini omni flash, 2026. URL https://ai.google.dev/gemini-api/docs/omni. Public preview model gemini-omni-flash-preview, June 2026.

Google DeepMind. Veo: a text-to-video generation system. https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf, 2025. Latent diffusion audio-video model.

Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. ETVA: evaluation of text-to-video alignment via fine-grained question generation and answering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-23, 2025*, pp. 21299–21309. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01978. URL https://doi.org/10.1109/ICCV51701.2025.01978.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: efficient joint audio-visual foundation model. *CoRR*, abs/2601.03233, 2026. doi: 10.48550/ARXIV.2601.03233. URL https://doi.org/10.48550/arXiv.2601.03233.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 2105–2123. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.127. URL https://doi.org/10.18653/v1/2024.emnlp-main.127.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. Videoscore2: Think before you score in generative video evaluation. *CoRR*, abs/2509.22799, 2025. doi: 10.48550/ARXIV.2509.22799. URL https://doi.org/10.48550/arXiv.2509.22799.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *CoRR*, abs/2403.14773, 2024. doi: 10.48550/ARXIV.2403.14773. URL https://doi.org/10.48550/arXiv.2403.14773.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *CoRR*, abs/2501.13826, 2025a. doi: 10.48550/ARXIV.2501.13826. URL https://doi.org/10.48550/arXiv.2501.13826.

Lanxiang Hu, Abhilash Shankarampeta, Yixin Huang, Zilin Dai, Haoyang Yu, Yujie Zhao, Haoqiang Kang, Daniel Zhao, Tajana Rosing, and Hao Zhang. Benchmarking scientific understanding and reasoning for video generation using videoscience-bench. *CoRR*, abs/2512.02942, 2025b. doi: 10.48550/ARXIV.2512.02942. URL https://doi.org/10.48550/arXiv.2512.02942.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *CoRR*, abs/2505.04512, 2025c. doi: 10.48550/ARXIV.2505.04512. URL https://doi.org/10.48550/arXiv.2505.04512.

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 17-21, 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, 2026. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Kuaishou Technology. Kling AI launches video 2.6 model with "simultaneous audio-visual generation" capability, redefining AI video creation workflow, December 2025. URL https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-video-26-model-simultaneous-audio-visual. Accessed: 2025-12-16.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. *CoRR*, abs/2502.20694, 2025a. doi: 10.48550/ARXIV.2502.20694. URL https://doi.org/10.48550/arXiv.2502.20694.

Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video model post-training through data, reward, and conditional guidance design. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=BZwXMqu4zG.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In Danielle Belgrave, Cheng Zhang, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, and Nancy Chen (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, volume 38, pp. 82155–82192. Curran Associates, Inc., 2025. doi: 10.52202/085713-2750. URL http://papers.nips.cc/paper_files/paper/2025/hash/76227feb18ea0ee40bd15cf02c33e18e-Abstract-Conference.html.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Bingyuan Wang, Qinghe Wang, Xuanhua He, Hongfa Wang, Chenyang Zhu, Hongyu Liu, Yingqing He, Zeyu Wang, Zhifeng Li, Xiu Li, Sirui Han, Yike Guo, Wei Liu, Dan Xu, Linfeng Zhang, and Qifeng Chen. Controllable video generation: A survey. *CoRR*, abs/2507.16869, 2025. doi: 10.48550/ARXIV.2507.16869. URL https://doi.org/10.48550/arXiv.2507.16869.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report. *CoRR*, abs/2510.22200, 2025. doi: 10.48550/ARXIV.2510.22200. URL https://doi.org/10.48550/arXiv.2510.22200.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 2025. URL https://proceedings.mlr.press/v267/meng25c.html.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities, 2026. URL https://www.minimax.io/blog/minimax-h3. Open weights: https://huggingface.co/MiniMaxAI/MiniMax-H3.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf. Open weights: https://huggingface.co/nvidia/Cosmos3-Nano.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: assessing visual temporal reasoning capabilities in multimodal foundation models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=fCi4o83Mfs.

Tingyu Song, Tongyan Hu, Guo Gan, and Yilun Zhao. Vf-eval: Evaluating multimodal llms for generating feedback on AIGC videos. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 21126–21146. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1027. URL https://doi.org/10.18653/v1/2025.acl-long.1027.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 8406–8416. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00787. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-video_Generation_CVPR_2025_paper.html.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025. doi: 10.48550/ARXIV.2503.20314. URL https://doi.org/10.48550/arXiv.2503.20314.

Tencent. Hunyuanvideo 1.5 technical report. *CoRR*, abs/2511.18870, 2025. doi: 10.48550/ARXIV.2511.18870. URL https://doi.org/10.48550/arXiv.2511.18870.

The Sora Team. Sora 2 is here, September 2025. URL https://openai.com/index/sora-2/.

Wan AI Team. Introducing wan 2.6. https://wan.video/introduction/wan2.6, 2025.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. Sciver: Evaluating foundation models for multimodal scientific claim verification. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 8562–8579. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.ACL-LONG.420. URL https://doi.org/10.18653/v1/2025.acl-long.420.

Zihang Wang, Yilun Zhao, Kaiyan Zhang, Chen Zhao, Manasi Patwardhan, and Arman Cohan. Scisketch: An open-source framework for automated schematic diagram generation in scientific papers. In Ivan Habernal, Peter Schulam, and Jörg Tiedemann (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - System Demonstrations, Suzhou, China, November 4-9, 2025*, pp. 403–417. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.EMNLP-DEMOS.28. URL https://doi.org/10.18653/v1/2025.emnlp-demos.28.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18826–18836. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01754. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xue_PhyT2V_LLM-Guided_Iterative_Self-Refinement_for_Physics-Grounded_Text-to-Video_Generation_CVPR_2025_paper.html.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang,

Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Yilun Zhao, Weiyuan Chen, Zhijian Xu, Manasi Patwardhan, Chengye Wang, Yixin Liu, Lovekesh Vig, and Arman Cohan. Abgen: Evaluating large language models in ablation study design and evaluation for scientific research. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 12479–12491. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.ACL-LONG.611. URL https://doi.org/10.18653/v1/2025.acl-long.611.

Yilun Zhao, Chengye Wang, Chuhan Li, and Arman Cohan. Can multimodal foundation models understand schematic diagrams? an empirical study on information-seeking QA over scientific papers. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 18598–18631. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.FINDINGS-ACL.957. URL https://doi.org/10.18653/v1/2025.findings-acl.957.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: measuring expert-level multi-discipline video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 8475–8489. Computer Vision Foundation / IEEE, 2025c. doi: 10.1109/CVPR52734.2025.00793. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhao_MMVU_Measuring_Expert-Level_Multi-Discipline_Video_Understanding_CVPR_2025_paper.html.

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Yixin Liu, Robert Tang, Joseph Chee Chang, Jesse Dodge, Jonathan Bragg, Chen Zhao, Hanna Hajishirzi, Doug Downey, and Arman Cohan. Sciarena: An open evaluation platform for non-verifiable scientific literature-grounded tasks. In Danielle Belgrave, Cheng Zhang, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, and Nancy Chen (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, volume 38. Curran Associates, Inc., 2025d. doi: 10.52202/085713-3532. URL http://papers.nips.cc/paper_files/paper/2025/hash/9811fe727c94e7ff79f701c94ed1d938-Abstract-Datasets_and_Benchmarks_Track.html.

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. *CoRR*, abs/2503.21755, 2025. doi: 10.48550/ARXIV.2503.21755. URL https://doi.org/10.48550/arXiv.2503.21755.

A Sci-VBench Dataset

A.1 Subject Selection

Natural Sciences (15)	Engineering (20)	Healthcare (15)	Humanities & Social Sciences (10)
Astrophysics & Astronomy	Aerospace Engineering	Anthropotomy	Archaeology
Atmospheric Science	Agricultural Engineering	Basic Immunology	Architecture & Urban Planning
Biochemistry & Molecular Biology	Biomedical Engineering	Clinical Laboratory Science	Arts
Biophysics	Chemical Engineering	Digestive Physiology	Business Administration
Cell & Developmental Biology	Civil & Environmental Engineering	Human Physiology	Economics
Chemistry	Communications Engineering	Imaging Medicine	Education
Earth Science	Computer Science	Medicine	Geography
Ecology & Evolutionary Biology	Electrical Engineering	Microbiology-Immunology	Linguistics
Environmental Science	Energy Engineering	Nutrition & Dietetics	Public Policy Administration
Geology & Geophysics	Food and Bioprocess Engineering	Nutrition and Food Hygiene	Sport
Marine Science	Geotechnical Engineering	Pathology	
Microbiology	Hydraulic Engineering	Pathophysiology	
Morphology	Instrument and Meter Engineering	Pharmaceutical Chemistry	
Physics	Materials Science & Engineering	Pharmacy	
Statistics & Data Science	Mechanical Engineering	Skin Physiology	
	Network Engineering		
	Nuclear Engineering		
	Optical Engineering		
	Scientific Computing		
	Structural Engineering		

Table 5: Complete subject list by major disciplines. Columns list subfields under each of the four major disciplines.

A.2 Data Quality Control

To test whether independently authored evaluation specifications support consistent judgments across independent evaluators, we conduct a targeted audit on a random sample of 200 examples. For each example, we recruit a second annotator from the same subject area to construct the full evaluation specification independently, including the high-level reference guide and the 1–5 anchored rubrics for all dimensions, without access to the first annotator's specification. This produces two independently authored specifications for the same prompt. To quantify consistency, for each sampled example we randomly select one generated video from a randomly chosen model and recruit two independent scorers from the same subject area, neither of whom authored either specification. Each scorer evaluates the same video under both independently authored specifications in counterbalanced order, yielding a crossed scorer–rubric design. We then compute dimension-wise agreement across the resulting paired score sets using quadratic weighted Cohen's kappa: Prompt Grounding (κ = 0.79), Scientific and Causal Correctness (κ = 0.75), and Spatiotemporal Consistency (κ = 0.73). These results indicate that independently authored specifications yield strongly consistent judgments.

A.3 Video Generation Models

We provide the detailed settings of the video generation models in Table 6.

Organization	Model	Release	Version	Access	Width	Height	FPS	Duration
			Proprietary Models					
OpenAI	Sora-2	2025-09	`sora-2`	API	1280	720	30	15s
Google	Veo-3.1	2025-10	`veo-3.1-generate-preview`	API	1280	720	24	8s
	Veo-3.1-Fast	2025-10	`veo-3.1-fast-generate-preview`	API	1280	720	24	8s
	Gemini-Omni-Flash	2026-06	`gemini-omni-flash-preview`	API	1280	720	24	10s
Kuaishou	Kling-2.6	2025-12	`kling-v2-6`	API	1440	1440	24	10s
Alibaba	Wan-2.6	2025-12	`wan2.6-t2v`	API	1280	720	30	15s
	HappyHorse-1.1	2026-06	`happyhorse-1.1-t2v`	API	1280	720	24	15s
ByteDance Seed	Seedance-2.0	2026-02	`doubao-seedance-2-0-260128`	API	1280	720	24	15s
			Open-source Models					
Tencent	HunyuanVideo-1.5	2025-11	`tencent/HunyuanVideo-1.5`	Open	848	480	10	15s
Lightricks	LTX-2	2026-01	`Lightricks/LTX-2` (19B-distilled)	Open	1536	1024	24	15s
	LTX-2.3	2026-03	`Lightricks/LTX-2.3`	Open	1536	1024	24	15s
Meituan	LongCat-Video	2025-10	`meituan-longcat/LongCat-Video`	Open	832	480	10	15s
Alibaba	Wan2.2-5B	2025-07	`Wan-AI/Wan2.2-TI2V-5B`	Open	1280	704	16	15s
Zhipu AI	CogVideoX1.5-5B	2024-11	`THUDM/CogVideoX1.5-5B`	Open	1360	768	16	5s
NVIDIA	Cosmos3-Nano	2026-06	`nvidia/Cosmos3-Nano`	Open	1280	720	24	15s
MiniMax	MiniMax-H3	2026-08	`MiniMaxAI/MiniMax-H3`	Open	1344	768	24	15s

Table 6: Details of the evaluated video generation models. All videos are generated from the verbatim benchmark prompts under each model's default configuration; **Duration** is the supported setting closest to our 15-second target (Veo-3.1 caps at 8s, Gemini-Omni-Flash at 10s, Kling-2.6 at 10s, and CogVideoX1.5-5B at 5s). **Version** gives the exact API model identifier for proprietary systems and the HuggingFace repository for open-source models; resolution and frame rate are measured from the generated videos. HappyHorse-1.1 clips carry a provider watermark (present in our generations even with the documented `watermark: false` request parameter).

A.4 Annotator Information

We list annotator biographies in Table 7 and Table 8; entries for annotators who are also authors are withheld to preserve anonymity.

ID	Year	Major	Assigned Subject(s)	Author?	Validator?
1	4th yr Undergraduate	Physics	Physics	✗	✗
2	4th yr Undergraduate	Chemistry	Chemistry	✗	✗
3	1st yr Master	Mathematics	Statistics & Data Science	✗	✗
4	1st yr Master	Earth Science	Earth Science	✗	✗
5	-	-	-	✓	✓
6	1st yr Master	Biology	Cell & Developmental Biology	✗	✗
7	1st yr Master	Statistics	Statistics & Data Science	✗	✗
8	1st yr Master	Environmental Science	Environmental Science	✗	✗
9	1st yr Master	Geology	Geology & Geophysics	✗	✗
10	1st yr Master	Physics	Astrophysics & Astronomy	✗	✓
11	1st yr Master	Chemistry	Biochemistry & Molecular Biology	✗	✓
12	1st yr Master	Mathematics	Statistics & Data Science	✗	✓
13	2nd yr Master	Biology	Ecology & Evolutionary Biology	✗	✓
14	1st yr Master	Environmental Science	Atmospheric Science	✗	✓
15	1st yr Master	Physics	Physics	✗	✓
16	1st yr Master	Chemistry	Chemistry	✗	✓
17	1st yr Master	Mathematics	Statistics & Data Science	✗	✓
18	-	-	-	✓	✓
19	2nd yr PhD	Geophysics	Geology & Geophysics	✗	✓
20	1st yr PhD	Neuroscience	Human Physiology	✗	✓
21	2nd yr PhD	Genetics	Biochemistry & Molecular Biology	✗	✓
22	3rd yr PhD	Atmospheric Science	Atmospheric Science	✗	✓
23	2nd yr PhD	Paleontology	Ecology & Evolutionary Biology	✗	✓
24	4th yr Undergraduate	Computer Science	Computer Science	✗	✗
25	1st yr Master	Electrical Engineering	Electrical Engineering	✗	✗
26	4th yr Undergraduate	Mechanical Engineering	Mechanical Engineering	✗	✗
27	1st yr Master	Chemical Engineering	Chemical Engineering	✗	✗
28	1st yr Master	Software Engineering	Computer Science	✗	✓
29	2nd yr Master	Materials Science	Materials Science & Engineering	✗	✓
30	1st yr Master	Biomedical Engineering	Biomedical Engineering	✗	✓
31	2nd yr Master	Industrial Engineering	Mechanical Engineering	✗	✓

Table 7: Biographies of 61 annotators involved in Sci-VBench construction (Author biographies are hidden to protect identity confidentiality).

ID	Year	Major	Assigned Subject(s)	Author?	Validator?
32	1st yr PhD	Electrical Engineering	Electrical Engineering	✗	✓
33	-	-	-	✓	✓
34	2nd yr PhD	Computer Engineering	Scientific Computing	✗	✓
35	3rd yr PhD	Mechanical Engineering	Mechanical Engineering	✗	✓
36	2nd yr PhD	Chemical Engineering	Chemical Engineering	✗	✓
37	4th yr Undergraduate	Economics	Economics	✗	✗
38	1st yr Master	Sociology	Public Policy Administration	✗	✗
39	4th yr Undergraduate	Psychology	Education	✗	✗
40	4th yr Undergraduate	Geography	Geography	✗	✗
41	1st yr Master	Economics	Economics	✗	✓
42	1st yr Master	Political Science	Public Policy Administration	✗	✓
43	1st yr Master	Anthropology	Archaeology	✗	✓
44	2nd yr Master	Education	Education	✗	✓
45	1st yr Master	Economics	Economics	✗	✓
46	1st yr Master	Psychology	Education	✗	✓
47	-	-	-	✓	✓
48	3rd yr PhD	Urban Studies	Architecture & Urban Planning	✗	✓
49	2nd yr PhD	Linguistics	Linguistics	✗	✓
50	4th yr Undergraduate	Nursing	Medicine	✗	✗
51	4th yr Undergraduate	Public Health	Human Physiology	✗	✗
52	4th yr Undergraduate	Pharmacy	Pharmacy	✗	✗
53	4th yr Undergraduate	Nutrition	Nutrition & Dietetics	✗	✗
54	1st yr Master	Clinical Medicine	Medicine	✗	✓
55	1st yr Master	Pharmacy	Pharmaceutical Chemistry	✗	✓
56	1st yr Master	Nutrition Science	Nutrition and Food Hygiene	✗	✓
57	2nd yr Master	Epidemiology	Pathology	✗	✓
58	1st yr Master	Public Health	Digestive Physiology	✗	✓
59	-	-	-	✓	✓
60	3rd yr PhD	Biomedical Science	Microbiology-Immunology	✗	✓
61	2nd yr PhD	Health Informatics	Imaging Medicine	✗	✓

Table 8: Biographies of 61 annotators involved in Sci-VBench construction (Author biographies are hidden to protect identity confidentiality).

B Sci-VBench Evaluation Protocol

B.1 Low-level Video Quality: Definitions and Implementation Details (Adapted from VBench)

Subject Consistency: Measures whether the main subject(s) in the video remain stable and coherent across frames. It evaluates whether the subject's identity, shape, appearance, and key attributes (e.g., color, size, structure) are preserved throughout the video, without unexpected changes, distortions, or disappearance during temporal progression.

Background Consistency: Assesses whether the scene background remains temporally stable across frames. It focuses on the continuity of environmental elements, such as layout, lighting, and spatial structure, ensuring that the background does not flicker, shift unnaturally, or change inconsistently when no scene transition is intended.

Motion Smoothness: Motion Smoothness evaluates the temporal continuity and physical plausibility of motion in the video. Measure whether object movements, camera motion, and transitions between frames are smooth, continuous, and free from jitter, abrupt jumps, or unnatural temporal artifacts.

Dynamic Degree: Dynamic Degree reflects the intensity and richness of motion present in the video. Assesses whether the video contains an appropriate level of dynamic variation, such as movement, deformation, or interaction of an object, rather than being overly static or motionless. This metric does not judge correctness, but rather the amount of motion activity.

Aesthetic Quality: Aesthetic Quality evaluates the overall visual appeal and artistic quality of the video. It considers factors such as composition, color harmony, lighting, visual balance, and stylistic coherence, measuring how pleasing and well-structured the video appears from a human perceptual perspective.

Imaging Quality: Imaging Quality measures the low-level visual fidelity of the video frames. It focuses on technical aspects including sharpness, resolution, noise level, compression artifacts, blur, and rendering clarity, reflecting how clean and realistic the generated images appear at the pixel level.

B.2 Comparing with Prior Video Scoring Methods

We adopt four methods: **(1) VideoScore** (He et al., 2024) trains a learned automatic evaluator on VideoFeedback by fine-tuning a video-capable VLM to regress five aspect scores: Visual Quality, Temporal Consistency, Dynamic Degree, Text-to-Video Alignment, and Factual Consistency. We run the released evaluator on each generated video to obtain the five aspect scores, then compute their unweighted average as the overall VideoScore. **(2) VideoScore2** (He et al., 2025) is a "think-before-scoring" judge trained on VideoFeedback2 with human scores plus reasoning traces, using a two-stage SFT+RL pipeline, and it outputs three scores: visual quality, text alignment, and physical/common-sense consistency. We apply the released model to each video and take the unweighted average of the three reported dimension scores as the overall VideoScore2. **(3) VideoReward** (Liu et al., 2025) is a VLM-based reward model trained from a large-scale human preference dataset over three dimensions: Visual Quality (VQ), Motion Quality (MQ), and Text Alignment (TA). We compute per-video VQ/MQ/TA with VideoReward and use their unweighted average as the overall VideoReward score. **(4) ETVA** (Guan et al., 2025) parses video prompts into semantic scene graphs, generating fine-grained atomic questions, and scoring videos via knowledge-augmented, multi-stage question answering. We follow the official protocol to compute ETVA's per-video alignment score (aggregated over atomic questions) and use it as the ETVA score.

As shown in Table 3, our rubric-conditioned MLLM-as-Judge aligns more closely with expert ratings than the compared prior auto-eval methods on the reasoning-centric dimensions, achieving the highest instance-level Pearson correlation among the compared automatic methods on Prompt Grounding, Scientific and Causal Correctness, and Spatiotemporal Con-

sistency. Among prior methods, ETVA is comparatively competitive on Prompt Grounding and Scientific and Causal Correctness but degrades sharply on Low-level Perceptual Fidelity and Spatiotemporal Consistency.

B.3 Evaluation Prompt Templates

Figure 5 shows the template that conditions MLLM-as-Judge on one evaluation dimension. Figure 6 shows the evaluation specification it is conditioned on, as released with the benchmark: the verbatim generation prompt, followed by the 1–5 anchored rubric for each judged dimension.

MLLM As Judge

You are a professional video quality assessment expert. You are given a video generated by a Text-to-Video model. Your task is to score this video on 1 dimension(s) based on the provided rubric.

Generation Prompt

<generation prompt, verbatim from the benchmark>

Reference Guide

<per-example high-level reference guide: target concept, minimum mechanism, and the expected phase-based storyline>

Scoring Rubric

1. <target dimension>

<dimension definition>

- Score 1: <anchor for score 1>
- Score 2: <anchor for score 2>
- Score 3: <anchor for score 3>
- Score 4: <anchor for score 4>
- Score 5: <anchor for score 5>

Instructions

Watch the video carefully, then score the video on each of the 1 dimension(s). For each dimension, provide:

1. A brief justification (2-3 sentences)
2. A score from 1 to 5

Output your response in the following JSON format (use double quotes):

```
{
"<target dimension>": {"justification": "...", "score": N}
}
```

Figure 5: Prompt template for rubric-conditioned MLLM-as-judge scoring on a single evaluation dimension, reproduced from the scoring code. Angle-bracketed fields are filled per example; the rubric block carries only the anchors for the dimension being scored.

Rubric Example

We present a rubric example to demonstrate rubric-based scoring in Sci-VBench. The prompt below is the verbatim benchmark prompt; the anchors are the expert-authored evaluation specification released with it.

Prompt: A subject sits on the examination table with the left leg naturally placed over the right leg, while another person rapidly strikes the patellar tendon below the left knee with a rubber hammer.

pg_specification: {
1 point: "No seated subject appears; the left-over-right leg crossing is absent; no reflex hammer is visible anywhere in frame; the region just below the left kneecap is never shown or is fully covered by clothing or hands."
2 points: "A seated subject is shown, but the legs are not crossed left-over-right (e.g., legs parallel, right-over-left, or uncrossed); or a hammer appears but is never brought near the left knee; or the region just below the left kneecap is not visible when the hammer approaches."
3 points: "Subject sits with the left leg crossed over the right leg and the region just below the left kneecap exposed; a rubber-headed reflex hammer is held and brought down to strike that exact region (not the kneecap itself, not the thigh muscle, not the shin)."
4 points: "All elements of 3-score are present AND the subject's crossed-leg posture and upper body stay motionless in the moments before impact (no shifting, flinching, or leg repositioning as the hammer approaches); the left knee stays extended or nearly so, not bent more than about 45°; the hammer's head meets the tendon region square-on rather than at a shallow glancing angle."
5 points: "All elements of 4-score are present AND the seating, leg-crossing, hammer approach, and moment of contact play out as one continuous, uninterrupted shot with the left knee, hammer, and lower leg all kept in frame and unobstructed by hands, clothing, or camera angle from start through the strike; whether the leg subsequently moves is not scored at this level."
}

scc_specification: {
1 point: "No leg movement follows the strike (the leg stays completely still); or the leg movement shown is physiologically wrong for this reflex, e.g., the hip flexes, the ankle flexes or dorsiflexes, or the leg withdraws rather than extends."
2 points: "The leg moves after the strike, but the response is inconsistent with a knee-jerk reflex: onset is delayed by roughly a second or more rather than immediate, both legs move, or the knee extension is accompanied by visible hip or ankle motion."
3 points: "Within the same continuous shot as the strike, the left lower leg extends forward at the knee with no perceptible delay after contact; the movement is confined to the knee joint, with the hip and ankle visibly uninvolved."
4 points: "All elements of 3-score are present AND the extension is a single brief kick rather than a slow push or an exaggerated fling, and the leg returns toward its resting position afterward without additional jerks, bounces, or oscillations."
5 points: "All elements of 4-score are present AND nothing suggests the subject is voluntarily assisting or resisting the movement (no visible muscle bracing, grimace, or hand moving toward the leg); the kick begins and ends cleanly with no tremor, sustained contraction, or rebound afterward, consistent with a brief involuntary contraction rather than a deliberate motion."
}

sc_specification: {
1 point: "The subject, hammer, or leg jumps to a different position or shape between frames without continuous motion connecting them; the scene's lighting flips or changes abruptly partway through the shot; the left knee or lower leg bends or bulges in a way a real joint cannot."
2 points: "A visible cut or gap interrupts the hammer's path between its approach and the moment it reaches the knee; the direction of the light and shadows on the leg or hammer changes partway through the clip with no light source moving to explain it; part of the leg or hammer flickers in and out of view where nothing should be blocking it."
3 points: "The subject, hammer, and both legs stay in consistent relative positions throughout (the crossed-leg posture does not shift on its own); the hammer's path from raised position to contact is shown as one continuous motion; the light source and the resulting shadows on the leg and hammer keep the same direction and strength across the whole clip."
4 points: "All elements of 3-score are present AND the hammer's swing speeds up and slows down the way a real swung object would (not a constant-speed glide or a sudden jump into contact); if the leg extends, its arc follows a natural pivot at the knee rather than sliding in a straight line or bending at an impossible point."
5 points: "All elements of 4-score are present AND the camera framing stays fixed with no visible drift, wobble, or unexplained repositioning across the whole shot; the hammer never appears to pass through or behind the leg when it should be in front of it; the subject, hammer, and leg all appear to occupy consistent depth and scale relative to one another throughout."
}

Figure 6: The evaluation specification of the knee-jerk reflex example, reproduced verbatim from the released dataset.

C Additional Experimental Results

C.1 Full-Benchmark Results

Table 9 reports the complete per-dimension automatic evaluation results for open-source models on the full Sci-VBench benchmark. The corresponding overall averages are also included in the **Full Avg.** column of Table 4.

Models	Release	VT	SC	PG	SCC	Avg.
Open-source Models						
MiniMax-H3 (MiniMax, 2026)	2026-08	3.91	2.61	**2.34**	**1.60**	**2.61**
Wan2.2-5B (Team Wan et al., 2025)	2025-07	3.81	**2.78**	1.87	1.26	2.43
HunyuanVideo-1.5 (Tencent, 2025)	2025-11	3.78	2.78	1.83	1.32	2.43
Cosmos3-Nano (NVIDIA, 2026)	2026-06	**4.02**	2.70	1.72	1.27	2.43
LongCat-Video (Meituan LongCat Team et al., 2025)	2025-10	3.85	2.52	1.78	1.23	2.34
LTX-2 (HaCohen et al., 2026)	2026-01	4.01	2.49	1.53	1.20	2.31
CogVideoX1.5-5B (Yang et al., 2025)	2024-11	3.89	2.11	1.88	1.35	2.30
LTX-2.3 (HaCohen et al., 2026)	2026-03	3.93	2.07	1.79	1.29	2.27

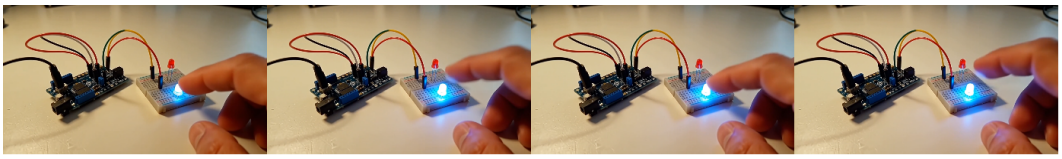
Table 9: Complete automatic evaluation results for open-source models on the full Sci-VBench benchmark (1,253 examples), sorted by the rightmost **Avg.** column. Metric definitions follow Table 4.

C.2 Error Analysis

We conduct a detailed error analysis and classify the errors into three major categories as follows: (1) **Poor Adherence to Instructions**: Our analysis reveals that even leading models struggle with instruction consistency in knowledge-intensive generation tasks. In particular, models demonstrate a lack of detail awareness, failing to generate fine-grained specifics mentioned in the prompt: in Figure 7, the model is unable to render the momentary push-button switch specified in the prompt, showing a bare fingertip touching the board instead.

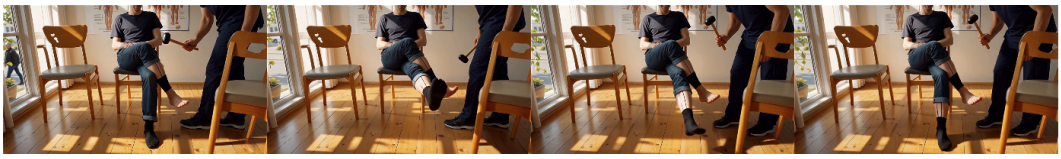
(2) **Inaccurate Simulation of Scientific Principles:** A crucial weakness of current T2V models is their inability to generate content that respects scientific knowledge and the laws of the physical world. Models often fail to reason from preconditions and produce scientifically plausible outcomes. For example, as shown in Figure 8(a), a video meant to depict a knee-jerk reflex incorrectly shows the un-struck leg kicking forward. Similarly, in Figure 8(b), the bag is squeezed directly beside the eye, yet the subject never blinks, demonstrating a broken understanding of the corneal reflex. These errors indicate that models tend to prioritize visual aesthetics over physical and scientific realism.

(3) **Deficiencies in Temporal Coherence and Visual Quality**: Beyond semantic and scientific inaccuracies, T2V models suffer from significant artifacts that break the illusion of realism. These issues primarily concern temporal consistency and overall visual quality. A major problem is the lack of object permanence and consistency; for instance, in Figure 9, the two balls released simultaneously on the two tracks merge into a single ball mid-motion. Similarly, models struggle to maintain a consistent visual style, as seen in Figure 10, where a realistic prism scene jarringly transitions to an animated style for the refracted light. In addition to these consistency failures, models also exhibit other common defects, such as poor image quality with coarse textures (Figure 11(a)) and weak causal links between events, like an LED switching on and off without any relation to the hand waving in front of the motion sensor (Figure 11(b)). Together, these flaws disrupt the visual flow and undermine the generated video's believability.

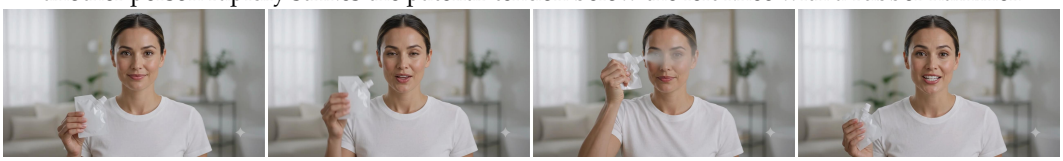


A desktop circuit demonstrates a momentary push-button switch connected to a power source and a single LED with a resistor, while a hand repeatedly presses and releases the button.

Figure 7: Category (1), poor adherence to instructions: the video is inconsistent with the specified setup. The verbatim benchmark prompt is quoted below the example.



(a) A subject sits on the examination table with the left leg naturally placed over the right leg, while another person rapidly strikes the patellar tendon below the left knee with a rubber hammer.

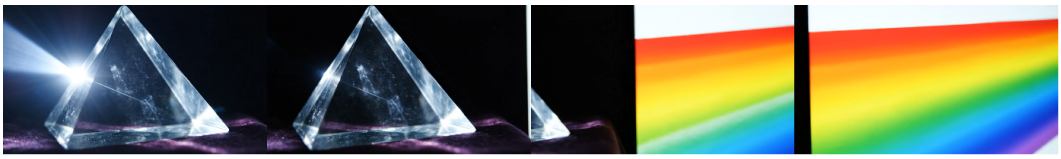


(b) A person is facing the camera, with one hand holding a bag and delivering a brief puff of air near one eye from the side.

Figure 8: Category (2), inaccurate simulation of scientific principles: the required mechanism is not realized. The verbatim benchmark prompt is quoted below each example.

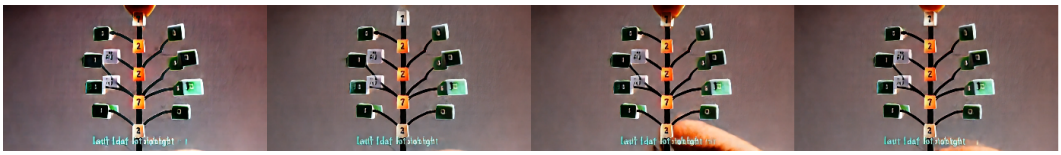
Within the same vertical plane, two fixed tracks are set up: one is a straight line, and the other is a cycloid-shaped curve. The starting and ending heights of both tracks are the same. Two identical small balls are released simultaneously from the starting point of the tracks, and the one that reaches the endpoint first is observed.

Figure 9: Category (3), deficiencies in temporal coherence: objects lose permanence over time. The verbatim benchmark prompt is quoted below the example.

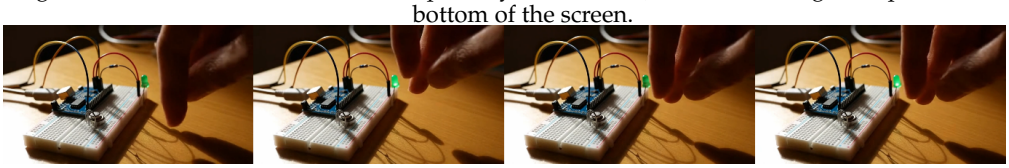


A triangular glass prism is placed in a beam of sunlight, allowing the light to pass through it. Observe what happens to the light as it exits the prism, noting the appearance, distribution, and order of any colors that result.

Figure 10: Category (3), deficiencies in visual quality: the visual style shifts discontinuously. The verbatim benchmark prompt is quoted below the example.



(a) Display a simple binary tree on a light background, with nodes labeled (1–7); highlight the nodes from top to bottom according to their level using bright colors (e.g., red) to demonstrate level order traversal. Briefly illuminate each node when visited, and add a mark for it (its numeral if legible digits can be rendered, otherwise a simple tally mark or dot) to the left-to-right output list at the bottom of the screen.



(b) On the workbench, a microcontroller is connected to a PIR motion sensor and an LED, and is powered on. A hand enters the frame and waves back and forth in front of the sensor for a few seconds before withdrawing.

Figure 11: Further category (3) defects: coarse textures and weak causal links between events. The verbatim benchmark prompt is quoted below each example.