LABINSTRUCT: BENCHMARKING SITUATED INSTRUCTIONAL VIDEO GENERATION FOR LAB PROCEDURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-driving laboratories (SDLs) aim to automate the full experimental loop, from scientific decision-making to physical execution. Ideally, AI-generated plans could be carried out directly by robotic systems, but reliable automation remains difficult in complex, open-world laboratory environments, where experiments often involve fine-grained manipulation, long-horizon procedures, and substantial variation across tasks and setups. Humans therefore remain an important execution interface between AI-generated plans and physical experiments, creating a need for clear and effective human-facing experimental guidance. Because laboratory procedures are inherently visual, spatial, and dynamic, video is particularly well suited to communicating apparatus configurations, manipulation actions, temporal dependencies, and state changes. Recent advances in video generation now make it possible to synthesize experimental demonstrations directly from an initial workspace image and a natural-language instruction. However, whether such models can reliably communicate real laboratory procedures has not been systematically studied. We introduce LABINSTRUCT, a benchmark for situated instructional video generation in real laboratories. LABINSTRUCT contains 204 tasks across 5 scientific disciplines, with real reference executions and structured annotations of objects, actions, contacts, and state transitions. Evaluating 8 frontier image-to-video models, we find that visually plausible generations frequently remain procedurally incorrect, revealing a substantial gap between visual realism and the reliability required for experimental instruction.

1 INTRODUCTION

Self-driving laboratories (SDLs) seek to accelerate discovery by integrating AI, automated experimentation, and iterative decision-making (Szymanski et al., 2023; Zhang et al., 2025). Recent systems support hypothesis generation and scientific software development (Gottweis et al., 2026; Ghareeb et al., 2026; Aygün et al., 2026), multi-step decision-making (Ferber et al., 2026), and experiment planning, execution, and closed-loop optimization (Boiko et al., 2023; Bennett et al., 2026). Yet reliable physical execution remains difficult in real laboratories, where diverse instruments, fine-grained manipulations, long action sequences, and workspace-specific constraints challenge embodied systems. Accordingly, current laboratory-robot and scientific-agent benchmarks often focus on restricted simulations or software workflows (Lan et al., 2026; Sun et al., 2026). Humans therefore remain a practical interface between AI-generated plans and physical experiments, making it essential to communicate not only *what* to do, but also *how* to do it.

Experimental instruction is thus a key component of human-in-the-loop laboratory automation. While text can specify goals and symbolic steps, laboratory operations are inherently visual, spatial, and dynamic, involving apparatus placement, tool–object contacts, motion, intermediate states, and temporal order that text alone may leave ambiguous. Images and demonstration videos can ground instructions in the current workspace and make required actions and state changes directly observable. Prior work has explored this direction through action and state-change image synthesis, step-by-step visual instructions, and language-conditioned video continuation from observed scenes (Souček et al., 2024; 2025; Gu et al., 2024; Xing et al., 2025; Pu et al., 2026).

Recent video generation models can synthesize demonstrations from workspace images and instructions, potentially bridging AI-driven experimentation and human execution (Figure 1). Yet visual

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

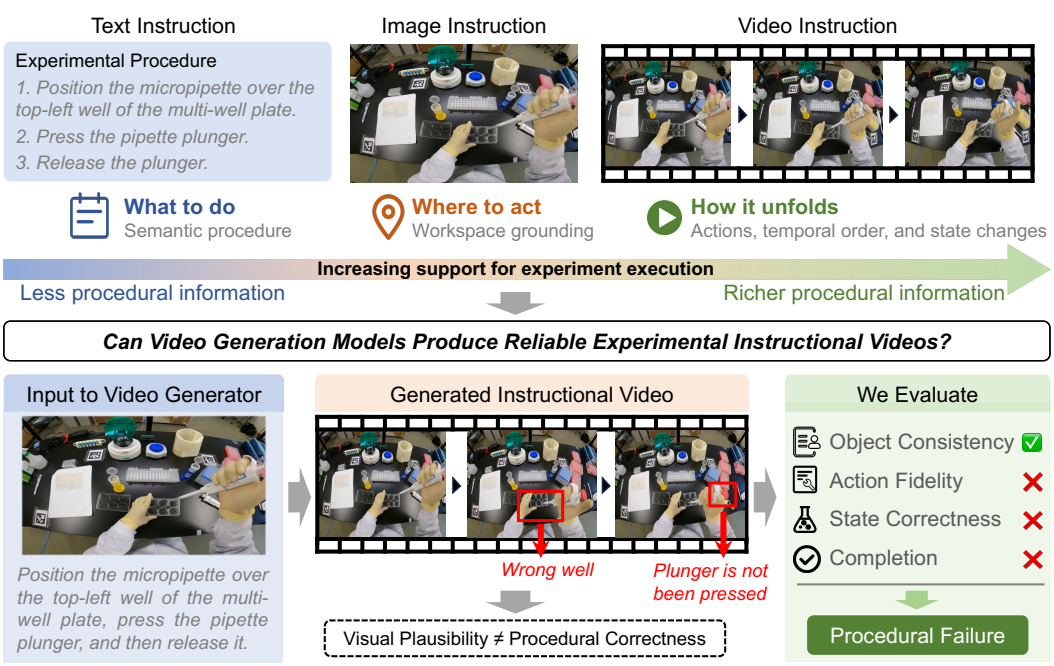


Figure 1: **Video provides richer procedural guidance for laboratory execution.** Text specifies what to do, images ground the instruction in the observed workspace, and video further reveals actions, temporal order, and state changes. However, visually plausible generated videos may still execute incorrect procedures, motivating LABINSTRUCT to evaluate object consistency, action fidelity, state correctness, and task completion.

realism does not ensure reliable guidance. For instructional use, generation quality depends not only on visual plausibility, but also on whether the depicted actions correctly realize the intended procedure. Existing benchmarks evaluate visual fidelity, prompt alignment, temporal consistency, physical plausibility, scientific reasoning, or educational explanation (Huang et al., 2024; Zheng et al., 2025; Li et al., 2026; Motamed et al., 2026; Tragoudaras et al., 2026; Hu et al., 2025; M et al., 2026), while laboratory and egocentric datasets focus on recognition, parsing, reasoning, and evidence localization in recorded videos (Nishimura et al., 2021; Yagi et al., 2025; Zou et al., 2025; Xu et al., 2026; Ragusa et al., 2026; Nagrani et al., 2026). None evaluates whether a model can correctly generate human-executed procedures in real laboratories.

We therefore ask: ***Can video generation models produce reliable experimental instructional videos?*** Such videos must use the required apparatus, act on the correct objects, preserve causal order, produce intended state transitions, and visibly reach the outcome. A visually plausible generation may omit critical actions, manipulate the wrong object, reorder dependent steps, or terminate prematurely. Conversely, a valid execution need not reproduce the reference trajectory exactly, as long as it preserves the workspace and satisfies the procedural requirements. Experimental instruction therefore imposes stricter demands than conventional visual synthesis, extending language-guided laboratory manipulation (Lan et al., 2026) from embodied agents to human-facing video.

We introduce LABINSTRUCT, a benchmark for situated instructional video generation in real laboratories. Given an initial workspace observation and a natural-language instruction, a model must generate a continuation that correctly demonstrates the procedure. LABINSTRUCT contains 204 tasks across 5 scientific disciplines and two complexity levels, from single experimental steps to short multi-step procedures with inter-step dependencies. Each task includes a real reference execution and structured annotations of objects, actions, contacts, and state transitions. Our contributions are: **i)** formulating laboratory video generation as a human-facing procedural communication task; **ii)** introducing LABINSTRUCT, a real-world benchmark with reference executions and structured procedural annotations; and **iii)** developing a procedure-aware evaluation and benchmarking 8 frontier image-to-video models to quantify the gap between visual plausibility and end-to-end procedural correctness under realistic laboratory conditions.

Table 1: **Comparison of scientific video generation benchmarks.** Lab denotes benchmarks based on real laboratory scenes, Ref. indicates the availability of paired real-world reference executions, and Proc. denotes human-executed laboratory procedures. T2V, I2V, and TI2V denote conditioning on text, images, and both text and images, respectively.

Benchmark	#Prompts	Discipline	Paradigm	Lab	Ref.	Proc.
VideoPhy-2 (Bansal et al., 2026)	688	Physics	T2V	×	×	×
VideoScience-Bench (Hu et al., 2025)	200	Phys./Chem.	T2V/I2V	×	✓	×
Physics-IQ (Motamed et al., 2026)	66	Physics	I2V	×	✓	×
PhyGround (Lin et al., 2026)	250	Physics	TI2V	×	×	×
Morpheus (Tragoudaras et al., 2026)	130	Physics	I2V	×	✓	×
Sci-VBench (Zhang et al., 2026)	1,253	Multi.	T2V	×	×	×
LABINSTRUCT	**204**	**Multi.**	**TI2V**	✓	✓	✓

2 RELATED WORK

Visual Instruction Generation. Visual instruction generation shows how to perform an action or change an object under a language command. GenHowTo (Souček et al., 2024) generates action and state-change images, while ShowHowTo (Souček et al., 2025) extends this task to step-by-step image sequences. Seer (Gu et al., 2024) and AID (Xing et al., 2025) predict videos from observed frames and language instructions. ShowMe (Pu et al., 2026) models action–object state transformations, Demo2Tutorial (Bai et al., 2026) turns software recordings into multimodal tutorials, and PIE-V (Loginova & Keller, 2026) generates mistakes and corrections in egocentric procedural videos. These works combine an observed scene with an instruction. LABINSTRUCT instead focuses on ordered laboratory procedures, visible state transitions, task-relevant contacts, safety constraints, and decision-support criteria tailored to procedural correctness.

Video Generation Models (VGMs). VGMs synthesize videos from text, images, or multimodal references, sometimes with explicit action or control signals. Open-weight models include Wan 2.2 (Team Wan, 2025), LTX 2.3 (HaCohen et al., 2026), MiniMax H3 (MiniMax, 2026), and HunyuanVideo 1.5 (Tencent Hunyuan Foundation Model Team, 2025) for text-to-video (T2V) and image-to-video (I2V) generation. Cosmos3 (NVIDIA, 2026) and LingBot Video (Ma et al., 2026) extend this line to embodied and egocentric synthesis using multimodal, action, and robot data. Closed models such as Sora 2 (OpenAI, 2025), Gemini Omni Flash (Google DeepMind, 2026b), Seedance 2.0 (Team Seedance, 2026), and Kling AI 3.0 (Technology, 2026) add synchronized audio, multimodal conditioning, editing, and multi-shot control. Yet situated instruction requires more than realism: a video must execute the specified procedure and clearly expose its critical steps and visible state changes throughout the generated sequence.

Scientific Video Generation Benchmarks. Existing benchmarks assess synthesized physical or scientific outcomes. Morpheus (Tragoudaras et al., 2026) and Physics-IQ (Motamed et al., 2026) ground evaluation in controlled physical systems, VideoPhy-2 (Bansal et al., 2026) probes action-centric physical commonsense, and PhyGround (Lin et al., 2026) tests compliance with 13 physical laws. VideoScience-Bench (Hu et al., 2025) evaluates undergraduate-level physical and chemical reasoning in T2V and I2V settings, while Sci-VBench (Zhang et al., 2026) spans 1,253 expert-authored tasks across 60 scientific subjects. These benchmarks test whether a phenomenon unfolds correctly, not whether a model performs an ordered human procedure in a given workspace. LABINSTRUCT targets this gap through scene-conditioned human demonstrations paired with reference executions from the same workspace. Table 1 summarizes the distinction.

3 THE LABINSTRUCT BENCHMARK

3.1 PARADIGM AND PROCEDURE COMPLEXITY

For each task, we have the initial workspace observation f_0, including relevant objects, spatial relationship, and scene context. Given f_0 and a natural-language instruction T, the generator synthesizes $f_{1:N} = \mathcal{G}(f_0, T; \theta)$ while preserving workspace identity and continuity, executing the instructed actions in the correct order, and reaching the expected intermediate and final states. We

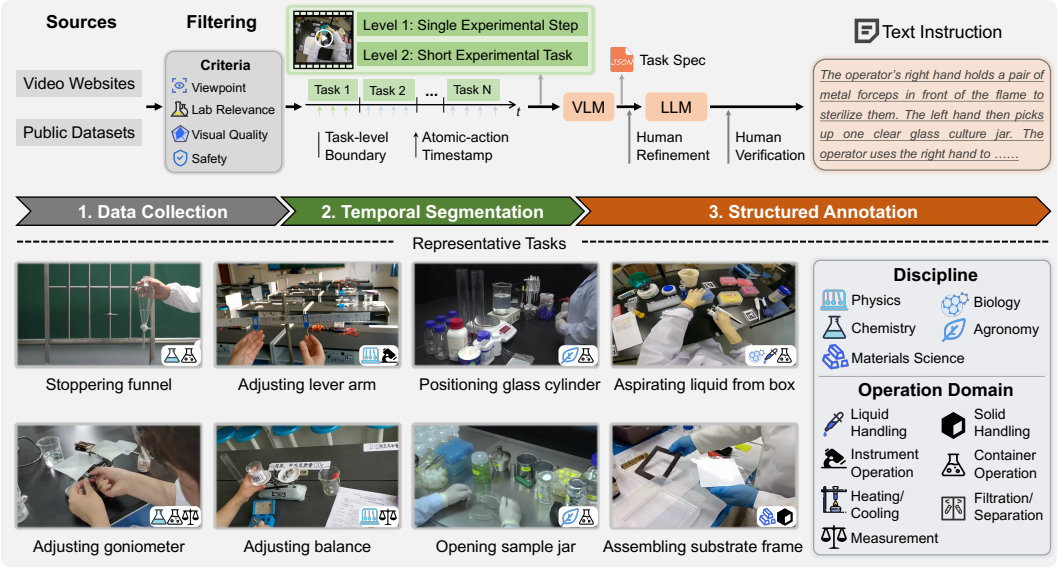


Figure 2: **LABINSTRUCT construction pipeline.** Laboratory videos are collected, filtered, segmented into tasks, and converted into human-verified specifications and instructions.

organize tasks into two independently complexity strata: **Level 1 (Single Experimental Step)** contains 1–3 atomic actions centered on a single operation with a visually identifiable outcome and uses a nominal generation target of 4 s; **Level 2 (Short Experimental Task)** contains 4–10 atomic actions with at least one inter-step dependency, requires temporal ordering and state continuity across operations, and uses a nominal target of 12 s.

3.2 BENCHMARK CONSTRUCTION

Data Sources and Reference Usage. Figure 2 summarizes the construction pipeline. We select real laboratory procedure videos from Bilibili, YouTube, the FineBio dataset (Yagi et al., 2025), and the ExpVid benchmark (Xu et al., 2026). For each selected video, we record its URL or dataset identifier, source identifier, and scientific discipline, and retain this provenance information for all derived clips. The collected videos serve as reference videos and are treated as one valid realization of the underlying procedure rather than as a unique target trajectory. Therefore, evaluation will focus on required actions, causal order, observable state transitions, and task outcomes, while allowing alternative valid execution paths. Human verification will remove incidental reference behavior from the scoring criteria and records acceptable variants when necessary. Generated videos are also expected to preserve the identity and spatial geometry of the observed workspace, with unexplained viewpoint changes, cuts, motion discontinuities, or loss of the workspace treated as procedural failures.

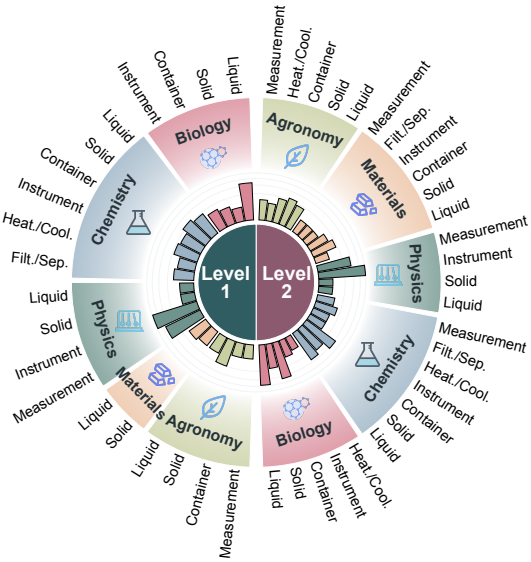


Figure 3: Distribution of the 204 benchmark tasks across task levels, scientific disciplines, and operation domains.

Filtering Criteria. Each candidate video is manually screened under four criteria. **Experiment Viewpoint** requires an operator-centered view that keeps the manipulation region and task-relevant objects visible. We retain first-person and nearby-observer views, annotate the viewpoint category, and report results separately for both. **Laboratory Relevance** requires a real, complete procedure with visually verifiable steps. **Visual Quality** requires clear, continuous footage with stable illumination and no persistent occlusion. **Safety Compliance** excludes procedures that depict unsafe operations or omit required protective measures. Full criteria are provided in Appendix B.1. This filtering focuses LABINSTRUCT on short, observable, and verifiable procedures rather than attempting to represent all laboratory conditions, equipment configurations, or procedure types.

Data Statistics. Figure 3 summarizes the benchmark composition. The benchmark comprises 204 tasks across 5 scientific disciplines, including 101 Level 1 and 103 Level 2 clips. These tasks span seven operation domains, with instrument operations and liquid handling being the most prevalent. Because Level 2 tasks combine multiple experimental steps, a single task may involve more than one operation domain. Most clips are recorded from the nearby-observer viewpoint, with the remainder using the first-person viewpoint. Reference clips average 4.5 s for Level 1 and 12.7 s for Level 2. Additional dataset statistics and breakdowns are provided in Appendix B.3.

3.3 TASK ANNOTATION

Temporal Segmentation. We annotate each source video at two temporal resolutions aligned with the task levels defined in Sec. 3.1. Level 1 intervals delimit individual experimental steps, while Level 2 intervals identify short experimental procedures that may consist of multiple consecutive steps. Every annotated interval records its precise start and end times. Level 1 and Level 2 annotations are defined independently, although a Level 2 interval may overlap with or contain one or more Level 1 intervals when the corresponding constituent steps are annotated. We use these temporal annotations to extract benchmark clips at the two task levels. Each clip begins with a stable and sufficiently clear view of the relevant workspace, and the key manipulations and resulting state transitions required by the task must remain clearly visible throughout the reference execution.

Structured Task Specification. For each clip, we construct a structured task specification using model-assisted annotation followed by human verification. Qwen3.6-27B (Qwen Team, 2026) first generates a draft from the clip and a predefined template, which annotators then review and correct. The specification captures task-relevant objects and roles, spatial relations, ordered actions, interaction types, observable state transitions, and difficulty attributes. Actions are linked to stable object IDs and annotated with step index, acting hand when applicable, action verb, target object, contact type, and visible pre- and post-action states. Human verification ensures temporal alignment, object identity, action completeness and ordering, state-change accuracy, and consistency with the initial frame, while removing reference-specific behaviors unnecessary for successful execution.

Instruction Generation. After human verification, the task specification is frozen and used to generate the natural-language instruction provided to the video generator. Qwen3.6-27B converts the verified specification into a predefined instruction format. A human annotator then checks semantic consistency between the instruction and the specification. The final instruction preserves task-relevant object identities, action order, interactions, and observable state transitions, while abstracting away reference-specific details such as hand choice, exact motion trajectories, pauses, and other incidental execution behavior unless they are required for correctness. This design ensures that the instruction specifies what must be accomplished without constraining the generator to reproduce the particular trajectory shown in the reference video.

3.4 EVALUATION PROTOCOL

Evaluation Design. Our evaluation comprises three decoupled layers. **Core Visual Quality** includes *Object Consistency*, *Action Fidelity*, *State Correctness*, and *Physical Plausibility*, assessing object persistence, action correctness and order, state transitions, and physical interactions. **Critical Completion** measures whether essential procedural requirements are satisfied, while *Scene Consistency* and *Visual Safety* act as **Validity Constraints**; camera motion is allowed provided workspace continuity and safe execution are preserved. We use DeepSeek-v4-flash (DeepSeek-AI, 2026) to

Table 2: **Core Visual Quality** scores under GPT-5.6 Sol. **Obj.**, **Act.**, **St.**, and **Phys.** denote Object Consistency, Action Fidelity, State Correctness, and Physical Plausibility, respectively. Higher is better. Best values are bolded, and S_{core} is shaded by rank. † indicates commercial models.

Model	Level 1					Level 2				
	Obj.	**Act.**	**St.**	**Phys.**	S_{core}	**Obj.**	**Act.**	**St.**	**Phys.**	S_{core}
Wan 2.2	93.7	59.4	79.0	90.1	76.2	89.8	37.3	58.2	84.3	60.0
LTX 2.3	87.5	50.9	71.6	85.7	68.8	83.0	45.3	66.6	83.5	63.2
Cosmos3 Nano	96.4	46.1	66.9	86.2	67.7	88.9	35.7	58.1	80.4	58.3
Cosmos3 Super	94.8	51.7	71.3	86.0	71.0	90.8	37.1	57.5	83.3	59.5
LingBot Video	92.7	48.0	67.8	85.8	68.2	88.7	39.9	61.7	83.6	61.7
MiniMax H3	96.7	**69.2**	**84.6**	93.1	**82.6**	95.6	61.2	75.9	92.7	77.0
Seedance 2.0†	**98.5**	58.4	79.9	92.7	77.9	94.6	67.2	77.9	93.2	79.6
Wan 3.0†	98.1	66.8	83.1	**93.7**	81.6	**95.7**	**67.4**	**80.3**	**93.7**	**80.3**

Table 3: **Completion** and **Overall** scores under GPT-5.6 Sol. **Comp.** denotes the completion factor. Higher is better. Best values are bolded, and Comp., Overall, and Pooled Overall are shaded by rank. **Pooled Overall** pools all tasks. † indicates commercial models.

Model	Level 1					Level 2					Pooled Overall
	S_{core}	**Comp.**	G_{scene}	G_{safety}	**Overall**	S_{core}	**Comp.**	G_{scene}	G_{safety}	**Overall**	
Wan 2.2	76.2	44.0	97.0	93.0	38.5	60.0	17.4	98.0	95.1	13.6	25.9
LTX 2.3	68.8	27.2	96.0	87.1	22.1	63.2	25.2	93.2	89.3	22.1	22.1
Cosmos3 Nano	67.7	25.2	99.0	88.1	23.7	58.3	17.9	96.1	92.2	15.2	19.4
Cosmos3 Super	71.0	32.1	98.0	96.0	30.0	59.5	17.4	97.0	95.1	15.7	22.8
LingBot Video	68.2	23.2	**100.0**	92.0	20.0	61.7	22.3	**99.0**	95.1	19.1	19.5
MiniMax H3	**82.6**	**58.4**	98.0	**98.0**	**53.6**	77.0	45.6	97.0	94.1	39.2	**46.3**
Seedance 2.0†	77.9	45.5	98.0	**98.0**	41.9	79.6	**53.3**	98.0	**96.1**	**48.2**	45.1
Wan 3.0†	81.6	50.9	98.0	93.0	46.2	**80.3**	52.9	88.3	94.1	42.7	44.4

generate positively framed, visually verifiable checklist items, which are reviewed by human experts. This yields 3,229 verified question-answer items, averaging 13.0 and 18.6 per Level 1 and Level 2 task, respectively. Further details are provided in Appendix C.1.

Evaluation Criteria. Each checklist item is assigned one of three possible verdicts: YES, NO, or UNJUDGEABLE. Invalid, ambiguous, or inapplicable items are removed during expert verification. For the four core dimensions, both NO and UNJUDGEABLE are treated as negative evidence. Their importance-weighted violations are aggregated within each dimension and mapped to a dimension-level score using the penalty function defined in Appendix C.2. The resulting four dimension scores are then combined using their geometric mean to obtain S_{core}. Task completion is assessed exclusively from *Critical* core items. Let n_c denote the number of *Critical* items that are not judged YES. We define $C(n_c) = 1$, 0.5, and 0 for $n_c = 0$, 1, and $n_c \geq 2$, respectively. Scene consistency and visual safety act as binary validity gates: either gate is closed by an explicit NO, whereas UNJUDGEABLE alone does not trigger a veto. The final score is therefore

$$S_{\text{overall}} = S_{\text{core}} \cdot C(n_c) \cdot G_{\text{scene}} \cdot G_{\text{safety}}. \tag{1}$$

Scores are computed for each video and averaged within each task level. Scene consistency and visual safety are additionally reported as separate gate validity rates.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate 8 models, including 6 open-weight models: **Wan 2.2 I2V A14B** (Team Wan, 2025) (simplified as Wan 2.2), **LTX 2.3** (HaCohen et al., 2026), **Cosmos3 Nano**, **Cosmos3 Su-**

Table 4: **Core Visual Quality** scores under human evaluation. Higher is better. Best values are bolded, and S_{core} is shaded by rank. † indicates commercial models.

Model	Level 1					Level 2				
	Obj.	**Act.**	**St.**	**Phys.**	S_{core}	**Obj.**	**Act.**	**St.**	**Phys.**	S_{core}
Wan 2.2	86.2	53.4	66.6	87.1	68.4	77.0	30.2	48.9	79.3	51.4
LTX 2.3	56.6	43.6	57.2	73.5	53.5	40.9	24.1	47.7	70.4	39.7
Cosmos3 Nano	86.0	40.7	56.0	77.8	59.3	61.5	20.4	40.7	69.8	40.9
Cosmos3 Super	89.1	50.1	61.9	85.2	66.0	76.2	24.3	46.6	74.2	47.6
LingBot Video	81.6	45.2	61.0	79.0	61.3	53.9	23.2	47.6	68.9	42.1
MiniMax H3	**95.5**	**70.9**	**80.2**	**93.5**	**81.6**	**94.6**	**63.8**	**74.6**	**93.2**	**77.3**
Seedance 2.0†	91.5	70.5	77.2	90.5	79.1	90.1	60.5	69.3	**93.2**	73.9
Wan 3.0†	89.8	66.1	73.5	89.8	75.7	87.9	62.7	74.1	92.2	74.6

Table 5: **Completion** and **Overall** scores under human evaluation. Higher is better. Best values are bolded, and Comp., Overall, and Pooled Overall are shaded by rank. † indicates commercial models.

Model	Level 1					Level 2					**Pooled Overall**
	S_{core}	**Comp.**	G_{scene}	G_{safety}	**Overall**	S_{core}	**Comp.**	G_{scene}	G_{safety}	**Overall**	
Wan 2.2	68.4	27.2	99.0	84.1	23.3	51.4	8.20	99.0	75.7	7.70	15.4
LTX 2.3	53.5	7.40	93.0	73.2	5.30	39.7	0.90	91.2	72.8	0.70	3.00
Cosmos3 Nano	59.3	12.8	99.0	79.2	10.4	40.9	1.40	98.0	69.9	1.30	5.80
Cosmos3 Super	66.0	21.7	99.0	85.1	18.9	47.6	5.30	97.0	76.6	4.20	11.5
LingBot Video	61.3	19.3	96.0	81.1	18.0	42.1	3.80	92.2	72.8	3.20	10.5
MiniMax H3	**81.6**	**51.9**	**100.0**	**98.0**	**49.7**	**77.3**	**43.6**	99.0	92.2	**39.0**	**44.3**
Seedance 2.0†	79.1	50.9	**100.0**	97.0	48.5	73.9	39.3	**100.0**	**96.1**	37.7	43.0
Wan 3.0†	75.7	42.0	98.0	91.0	38.8	74.6	39.8	96.1	95.1	35.8	37.3

per (NVIDIA, 2026), **LingBot Video** (Ma et al., 2026), and **MiniMax H3** (MiniMax, 2026), as well as 2 commercial models: **Seedance 2.0** (Team Seedance, 2026) and **Wan 3.0** (Alibaba Cloud, 2026). All models are provided with the same initial frame and a verified natural-language instruction to ensure a controlled comparison across models for each task. To prioritize broad coverage across both tasks and generators, we evaluate one generated sample per model–task pair under fixed model-specific generation settings. The exact model checkpoints, release dates, and generation configurations used in our experiments are reported in Appendix D.1. For automated evaluation, we provide GPT-5.6 Sol with the generated video, sampled at 4 frames per second (FPS), together with the corresponding evaluation rules and task-specific checklist. The evaluator is then instructed to produce a categorical verdict for each checklist item.

4.2 Experimental Results

GPT-5.6 Sol Results. Tables 2 and 3 show that procedural execution is the main bottleneck across models. Object consistency and physical plausibility remain relatively strong, whereas action fidelity, state correctness, and especially completion degrade on Level 2. The drop is particularly pronounced for weaker systems, whose videos often remain visually coherent despite failing to execute all required steps or realize the intended final state. As a result, high scene and safety validity does not necessarily translate into high Overall performance. MiniMax H3, Seedance 2.0, and Wan 3.0 form the strongest group overall, although their relative ordering varies across complexity levels and metrics. Notably, MiniMax H3 leads on Level 1, while Wan 3.0 achieves the highest Level 2 core score and Seedance 2.0 the highest Level 2 Overall score. These results indicate that current models preserve appearance and local physical plausibility more reliably than they execute temporally ordered procedures and achieve the required state transitions.

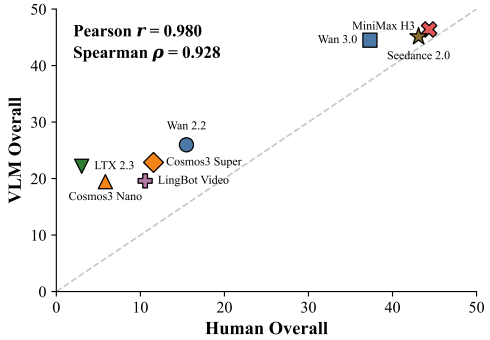


Figure 4: Overall score consistancy of each model under Human and VLM evaluation.

Table 6: Agreement between human and VLM scores for all 8 generators. Metrics are correlation, mean absolute error (MAE), maximum rank shift, and top-three overlap. **Overall** pools all 204 tasks.

Score	Level 1	Level 2	Overall
Pearson r	0.903	0.942	0.980
Spearman ρ	0.904	0.523	0.928
MAE	9.54	10.77	9.35
Max ΔRank	2	4	2
Top-3 Overlap	3/3	3/3	3/3

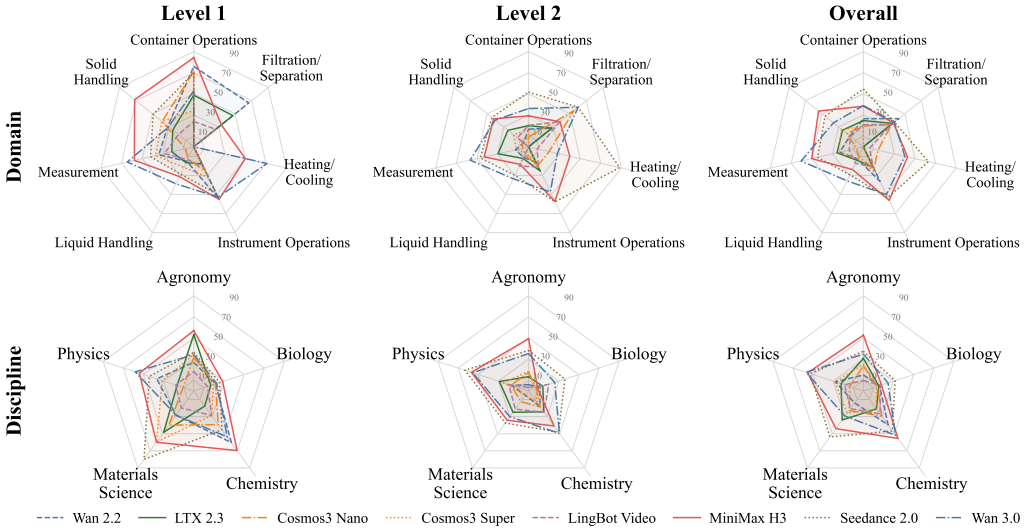


Figure 5: VLM evaluation scores broken down by operation domain (top row) and scientific discipline (bottom row) for Level 1, Level 2, and Overall.

Human Evaluation Results. Human evaluation in Tables 4 and 5 broadly confirms the same conclusions, while changing some fine-grained rankings. MiniMax H3, Seedance 2.0, and Wan 3.0 remain the leading systems, and Level 2 again exposes the largest weaknesses in action execution and procedural completion. Compared with human judgments, GPT-5.6 Sol is generally more permissive, particularly for weaker models, assigning substantially higher completion and Overall scores; for example, pooled Overall drops from 22.1 to 3.0 for LTX 2.3 and from 19.4 to 5.8 for Cosmos3 Nano under human evaluation. The agreement in overall trends suggests that automated evaluation captures the main performance structure, while human judgments apply a stricter standard for whether the requested procedure is actually completed.

4.3 Further Analysis

Human–VLM Agreement. Human and VLM evaluations show strong generator-level agreement (Figure 4; Table 6). Pearson correlation is high at both Level 1 ($r = 0.903$) and Level 2 ($r = 0.942$), and increases to $r = 0.980$ after pooling, which also yields the lowest MAE (9.35). Rank agreement is weaker on Level 2 ($\rho = 0.523$, maximum rank shift $= 4$) but becomes substantially more stable after aggregation ($\rho = 0.928$, maximum rank shift $= 2$). Notably, the top-3 set is preserved across all settings, and the pooled human and VLM rankings agree on the ordering of the three strongest generators. Figure 4 further shows that larger deviations occur mainly among lower-scoring models. Overall, aggregation provides stable human–VLM agreement for generator comparison, particularly among the strongest models, with additional agreement analyses reported in Appendix D.3.

Table 7: Cross-evaluator consensus tiers across task levels and evaluators. Subscripts indicate ranks.

Model	Human Evaluation			VLM Evaluation			Consensus	
	Level 1	Level 2	Overall	Level 1	Level 2	Overall	Mean Rank	Tier
MiniMax H3	49.7_1	39.0_1	44.3_1	53.6_1	39.2_3	46.3_1	1.33	**T0**
Seedance 2.0†	48.5_2	37.7_2	43.0_2	41.9_3	48.2_1	45.1_2	2.00	**T0**
Wan 3.0†	38.8_3	35.8_3	37.3_3	46.2_2	42.7_2	44.4_3	2.67	**T0**
Wan 2.2	23.3_4	7.70_4	15.4_4	38.5_4	13.6_8	25.9_4	4.67	T1
Cosmos3 Super	18.9_5	4.20_5	11.5_5	30.0_5	15.7_6	22.8_5	5.17	T1
LingBot Video	18.0_6	3.20_6	10.5_6	20.0_8	19.1_5	19.5_7	6.33	T2
LTX 2.3	5.30_8	0.70_8	3.00_8	22.1_7	22.1_4	22.1_6	6.83	T2
Cosmos3 Nano	10.4_7	1.30_7	5.80_7	23.7_6	15.2_7	19.4_8	7.00	T2

Domain and Discipline Breakdown. Figure 5 shows substantial variation across operation domains and scientific disciplines. MiniMax H3 and Wan 3.0 provide the broadest coverage, with complementary strengths: MiniMax H3 performs strongly in instrument operations, solid handling, agronomy, and chemistry, while Wan 3.0 leads in liquid handling, measurement, and physics. Seedance 2.0 instead excels in container operations, heating/cooling, and materials science, and Wan 2.2 is strongest in filtration/separation. Across models, Level 2 scores generally decline across multiple domains and disciplines rather than in a single category, showing that the increased difficulty is distributed across the benchmark rather than concentrated in a specific operation type.

Failure Attribution. We inspect representative failures and identify several recurring patterns across models. **1) Procedural Omission and Misordering.** The most common failures occur when models omit required operations, execute them out of order, or apply them to the wrong object. These errors often arise despite visually coherent generation, highlighting procedural execution as a more persistent bottleneck than appearance preservation. **2) Incomplete State Realization.** Models may perform an action without producing the required observable consequence, such as a liquid-level change, instrument reading, or final object configuration. Such errors may follow earlier action failures or reflect difficulty in realizing fine-grained quantitative or positional outcomes. **3) Object Drift.** Task-relevant objects may deform, disappear, merge, or become difficult to track, while occasional scene discontinuities further disrupt procedural continuity. These failures are less common than action errors but can still invalidate otherwise plausible executions.

Cross-Evaluator Consensus Tiers. The re-sampling analysis in Appendix D.4.4 shows that the exact ordering among the strongest models may vary with generation randomness. We therefore summarize performance using a rank-based consensus analysis across six evaluation views: Level 1, Level 2, and Overall scores from both human and VLM evaluators. Models are ranked in each view, averaged by rank, and grouped into three tiers using the two largest adjacent gaps. As shown in Table 7, MiniMax H3, Seedance 2.0, and Wan 3.0 consistently form the top tier (T0), remaining within the top three across all six views, while Wan 2.2 and Cosmos3 Super form T1 and the remaining models form T2. Together with the re-sampling results, this indicates that the separation between performance tiers is more stable than the exact ordering within T0.

5 Conclusion

We introduce LabInstruct, a benchmark for evaluating procedural correctness in laboratory instructional video generation. Experiments on 8 frontier models reveal a clear gap between visual fidelity and procedural success: models often preserve scenes and objects but miss, reorder, or obscure required actions and state changes. Completion is the main bottleneck on Level 2 tasks, whose longer procedures require sustained temporal ordering and state continuity. Overall, current video generation models remain unreliable as standalone sources of laboratory instruction, and LabInstruct provides a systematic framework for evaluating and diagnosing procedural failures.

REFERENCES

Alibaba Cloud. Wan 3.0: Leading ai video generation model, 2026. URL https://github.com/AlibabaCloud-Official/Wan3.0.

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, et al. An AI system to help scientists write expert-level empirical software. *Nature*, 654(8120):909–916, 2026. doi: 10.1038/s41586-026-10658-6.

Zechen Bai, Zhiheng Chen, Yiqi Lin, Kevin Qinghong Lin, Difei Gao, Xiangwu Guo, Xin Wang, and Mike Zheng Shou. Demo2tutorial: From human experience to multimodal software tutorials. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29588–29597, June 2026.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HA8KSQW7SO.

J. A. Bennett, N. Orouji, A. Velayati, et al. An autonomous lab for data-driven homogeneous catalysis. *Nature Communications*, 17(1):7783, 2026. doi: 10.1038/s41467-026-74425-x.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Dyke Ferber, Lars Hilgers, Christiane Höper, et al. Towards autonomous medical artificial intelligence agents. *Nature*, 655(8125):1282–1291, 2026. doi: 10.1038/s41586-026-10675-5.

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, et al. A multi-agent system for automating scientific discovery. *Nature*, 655(8122):497–505, 2026. doi: 10.1038/s41586-026-10652-y.

Google DeepMind. Gemini 3.7 flash: Model card. Google DeepMind, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

Google DeepMind. Gemini omni flash: Model card. Google DeepMind, 2026b. URL https://deepmind.google/models/model-cards/gemini-omni-flash/.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 3474–3505, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0f6cc80ad86e553d085842308e0fd2cb-Paper-Conference.pdf.

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, 2026. URL https://arxiv.org/abs/2601.03233.

Xuan He, Cong Wei, Yuhao Cheng, Linrui Ma, Yuxuan Zhang, Zuojun Li, Yuhao Wen, Jize Jiang, Zeyi Liu, Yuren Hao, Songcheng Cai, Keming Wu, Penghui Du, Kai Zou, Rui Yang, Chenkai Sun, Ke Yang, Ping Nie, Kelsey R Allen, Chenglong Wang, Michel Galley, Jianfeng Gao, and ChengXiang Zhai. Vgi-bench: Probing visual intelligence in video generation models, 2026. URL https://arxiv.org/abs/2608.19583.

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, 2025. URL https://arxiv.org/abs/2512.02942.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhiqian Lan, Yuxuan Jiang, Ruiqi Wang, Xuanbing Xie, Rongkui Zhang, Yicheng Zhu, Peihang Li, Tianshuo Yang, Tianxing Chen, Haoyu Gao, Xiaokang Yang, Xuelong Li, Hongyuan Zhang, Yao Mu, and Ping Luo. Autobio: A simulation and benchmark for robotic automation in digital biology laboratory. In *International Conference on Learning Representations*, pp. 146176–146196, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ec2691f6b5c88eb650554dffbc0e0af5-Paper-Conference.pdf.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38, 2026.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y. Lu, Enfu Nan, Hokin Deng, Edmund Yeh, Sarah Ostadabbas, Yun Fu, Jennifer Dy, Pu Zhao, and Yanzhi Wang. Phyground: Benchmarking physical reasoning in generative world models, 2026. URL https://arxiv.org/abs/2605.10806.

Olga Loginova and Frank Keller. How to correctly make mistakes: A framework for constructing and benchmarking mistake aware egocentric procedural videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8843–8853, June 2026.

Megha Mariam K M, Aditya Arun, Zakaria Laskar, and C. V. Jawahar. Phyeduvideo: A benchmark for evaluating text-to-video models for physics education. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE/CVF, 2026.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, Yanhong Zeng, Qin Zhao, Zifan Shi, Wei Wu, Hao Ouyang, Qiuyu Wang, Shangzhan Zhang, Jiahao Shao, Yipengjing Sun, Liangxiao Hu, Lunke Pan, Nan Xue, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Ka Leong Cheng. Scaling mixture-of-experts video pretraining for embodied intelligence, 2026. URL https://arxiv.org/abs/2607.07675.

MiniMax. Minimax h3: An open-weight, general-purpose, omni-modal generation model, 2026. URL https://www.minimax.io/.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958, 2026. doi: 10.1109/WACV61042.2026.00099.

Arsha Nagrani, Jasper Uijlings, Shyamal Buch, Tobias Weyand, Sudheendra Vijayanarasimhan, Bo Hu, Ramin Mehran, David A Ross, and Cordelia Schmid. Minerva-ego: Spatiotemporal hints for egocentric video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38859–38869, 2026.

Taichi Nishimura, Kojiro Sakoda, Atsushi Hashimoto, Yoshitaka Ushiku, Natsuko Tanaka, Fumihito Ono, Hirotaka Kameko, and Shinsuke Mori. Egocentric biochemical video-and-language dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3129–3133, October 2021.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai, 2026. URL https://arxiv.org/abs/2606.02800.

OpenAI. Sora 2 system card. OpenAI, 2025. URL https://openai.com/index/sora-2-system-card/.

Yujiang Pu, Zhanbo Huang, Vishnu Boddeti, and Yu Kong. Show me: Unifying instructional image and video generation with diffusion models. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3285–3296. IEEE, 2026.

Qwen Team. Qwen3.6-27b: Flagship-level coding in a 27b dense model. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.6-27B.

Francesco Ragusa, Michele Mazzamuto, Rosario Forte, Irene D'Ambra, James Fort, Jakob Engel, Antonino Furnari, and Giovanni Maria Farinella. Ego-extra: video-language egocentric dataset for expert-trainee assistance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4438–4450, 2026.

Tomáš Souček, Dima Damen, Michael Wray, Ivan Laptev, and Josef Sivic. Genhowto: Learning to generate actions and state transformations from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6561–6571, June 2024.

Tomáš Souček, Prajwal Gatti, Michael Wray, Ivan Laptev, Dima Damen, and Josef Sivic. Showhowto: Generating scene-conditioned step-by-step visual instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27435–27445, June 2025.

Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, Jianing Wang, Qintong Li, Xiangru Tang, Tianbao Xie, Xiachong Feng, Xiang Li, Ben Kao, Wenhai Wang, Biqing Qi, Lingpeng Kong, and Zhiyong Wu. Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. In *International Conference on Learning Representations*, pp. 75694–75731, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7a9745f251508a053425a256490b0665-Paper-Conference.pdf.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86–91, 2023. doi: 10.1038/s41586-023-06734-w.

Team Seedance. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Team Wan. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Kuaishou Technology. Kling ai launches 3.0 model, ushering in an era where everyone can be a director. Kuaishou Technology, 2026. URL https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-be.

Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Stratis Gavves. Evaluating newtonian mechanics in video generative models with real physical systems. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=f4xAzcMug6.

Zhen Xing, Qi Dai, Zejia Weng, Zuxuan Wu, and Yu-Gang Jiang. Aid: Adapting image2video diffusion models for instruction-guided video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21243–21253, October 2025.

Yicheng Xu, Yue Wu, Jiashuo Yu, Ziang Yan, Tianxiang Jiang, Yinan He, Qingsong Zhao, Kai Chen, Yu Qiao, Limin Wang, Manabu Okumura, and Yi Wang. Expvid: A benchmark for experiment video understanding & reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 83310–83348, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/86833e97827ff52d297a9fbde5842a81-Paper-Conference.pdf.

Takuma Yagi, Misaki Ohashi, Yifei Huang, Ryosuke Furuta, Shungo Adachi, Toutai Mitsuyama, and Yoichi Sato. Finebio: A fine-grained video dataset of biological experiments with hierarchical annotation. *International Journal of Computer Vision*, 133(10):7352–7367, 2025. doi: 10.1007/s11263-025-02523-2.

Diandian Zhang, Tingyu Song, Lin Fu, Zheyuan Yang, and Yilun Zhao. Sci-vbench: Evaluating knowledge- and reasoning-intensive video generation in science domains, 2026. URL https://arxiv.org/abs/2608.09873.

Zhen Zhang, Zhichu Ren, Chia-Wei Hsu, Weibin Chen, et al. A multimodal robotic platform for multi-element electrocatalyst discovery. *Nature*, 647(8089):390–396, 2025. doi: 10.1038/s41586-025-09640-5.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Minghao Zou, Qingtian Zeng, Yongping Miao, Shangkun Liu, Zilong Wang, Hantao Liu, and Wei Zhou. Physlab: A benchmark dataset for multi-granularity visual parsing of physics experiments. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12799–12806, 2025. doi: 10.1145/3746027.3758221. URL https://doi.org/10.1145/3746027.3758221.

APPENDIX TABLE OF CONTENTS

A Use of Large Language Models (LLMs) 15

B Benchmark Construction Details 15

B.1 Source Video Filtering Criteria 15

B.2 Source Complementarity 15

B.3 Dataset Statistics 15

C Evaluation Details 18

C.1 Evaluation Design Details 18

C.2 Formal Scoring Procedure 18

D Experimental Details 20

D.1 Evaluated Models and Generation Settings 20

D.2 Cross-Evaluator Robustness 20

D.3 Human Evaluation Details 22

D.4 Additional Analyses 24

E Task Schema and Prompt Templates 31

E.1 Structured Task Schema 31

E.2 Annotation Prompt for Level 1 Tasks 31

E.3 Annotation Prompt for Level 2 Tasks 32

E.4 Instruction Rendering Prompt for Level 1 Tasks 33

E.5 Instruction Rendering Prompt for Level 2 Tasks 33

E.6 Task-Conditioned QA Generation Prompt 34

E.7 VLM Evaluation Prompt 36

F Limitations and Ethical Considerations 38

G Other Details 39

G.1 Licenses and Terms of Use 39

G.2 Source-Video Provenance and Distribution 39

G.3 Intended Use 39

G.4 Human Annotation: Recruitment, Consent, and Ethics Review 39

A USE OF LARGE LANGUAGE MODELS (LLMS)

In our work, large language models (LLMs) are employed to assist with data annotation, manuscript writing and proofreading, and the development of experimental code. All LLM-generated or LLM-assisted content is subsequently reviewed and refined by the authors. We verify the resulting annotations, code, and manuscript content for correctness and faithful representation of the intended methodology and conclusions. The authors take full responsibility for the complete set of materials, analyses, and results reported in this paper.

B BENCHMARK CONSTRUCTION DETAILS

B.1 SOURCE VIDEO FILTERING CRITERIA

We manually screen every candidate source video against the following four eligibility criteria:

- **Experiment Viewpoint.** The video must provide an operator-centered view in which the manipulation region and task-relevant objects remain clearly visible. We retain first-person views recorded near the operator's eye level and nearby-observer views captured close to the operator and approximately aligned with the workspace.
- **Laboratory Relevance.** The video must depict a real laboratory procedure rather than an animation, non-experimental demonstration, or everyday activity. Experimental steps and their outcomes must be clearly discernible and visually verifiable.
- **Visual Quality.** The footage must be clear and continuous, with stable illumination and no severe camera shake or persistent occlusion. Videos with cuts, camera switches, ambiguous state changes, substantial editing, excessive text or branding, or severely cluttered backgrounds are excluded. We prefer videos in which all task-relevant apparatus is already visible in the initial workspace frame used for model conditioning.
- **Safety Compliance.** The video must not depict dangerous or clearly unsafe operations, omit required protective measures, or otherwise encourage unsafe practice.

Table 8: Composition of the LABINSTRUCT benchmark by video source.

Source	# Videos	# Task Clips
Bilibili	61	163
FineBio (Yagi et al., 2025)	15	23
ExpVid (Xu et al., 2026)	3	14
YouTube	2	4
Total	**81**	**204**

B.2 SOURCE COMPLEMENTARITY

The LABINSTRUCT benchmark draws from 4 complementary video sources: Bilibili, YouTube, FineBio (Yagi et al., 2025), and ExpVid (Xu et al., 2026). These sources provide complementary coverage in both scientific content and recording viewpoint. FineBio contributes primarily first-person biology laboratory procedures and constitutes the main source of our biology tasks. In contrast, Bilibili and YouTube broaden the disciplinary coverage beyond biology, particularly for physics- and chemistry-oriented experiments, while also providing a diverse range of laboratory operations and workspace configurations. ExpVid, which is derived from professionally recorded JoVE experimental demonstrations, further complements the web-sourced videos with structured, multi-step laboratory procedures. We therefore curate across sources rather than relying on any single dataset or platform. After filtering and temporal annotation, the 81 retained source videos yield 204 benchmark task clips. Table 8 summarizes the contribution of each source.

B.3 DATASET STATISTICS

Figure 6 provides a detailed breakdown of the benchmark composition and reference-clip durations. Across scientific disciplines, the benchmark contains 73 tasks in Physics, 51 in Chemistry, 44 in Bi-

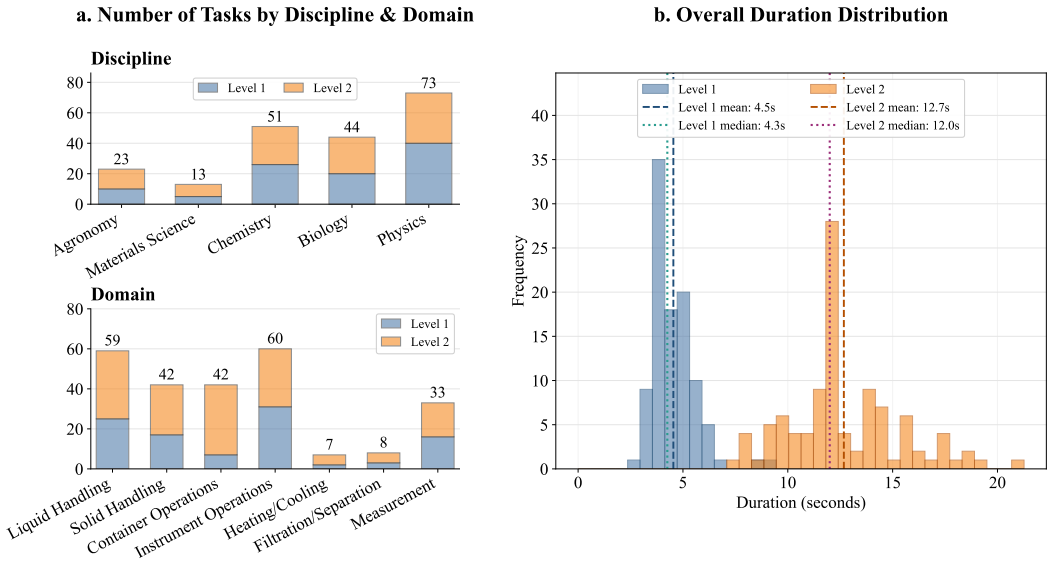


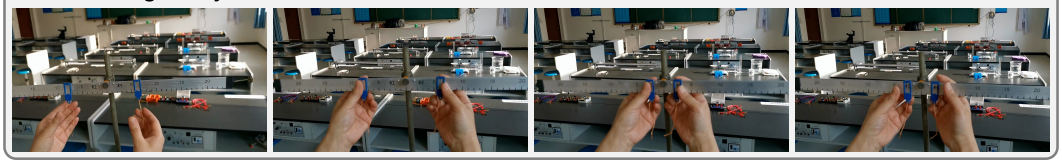
Figure 6: Benchmark task composition by discipline and domain, and reference-clip duration distributions across Level 1 and Level 2 tasks.

ology, 23 in Agronomy, and 13 in Materials Science. In terms of operation domains, it includes 60 tasks involving Instrument Operations, 59 involving Liquid Handling, 42 involving Solid Handling, 42 involving Container Operations, 33 involving Measurement, 8 involving Filtration/Separation, and 7 involving Heating/Cooling. These domains cover a range of laboratory operations from routine manipulation of instruments, liquids, solids, and containers to measurement, thermal processing, and separation procedures. Since a task, particularly at Level 2, may involve multiple types of experimental operations, we assign operation domains in a multi-label manner, and the corresponding counts therefore do not sum to the total number of tasks.

Reference-clip durations are also organized differently across the two task levels. Level 1 clips are concentrated around 3–6 s, with a mean duration of 4.5 s and a median of 4.3 s. Level 2 clips span a broader range, typically around 8–16 s with a longer tail, with a mean duration of 12.7 s and a median of 12.0 s. In our benchmark design, Level 1 tasks correspond to localized experimental steps, while Level 2 tasks cover longer procedures that contain multiple temporally connected operations. The resulting difference in clip duration follows directly from this task-level definition. Figure 7 presents four completed benchmark instances from the two levels.

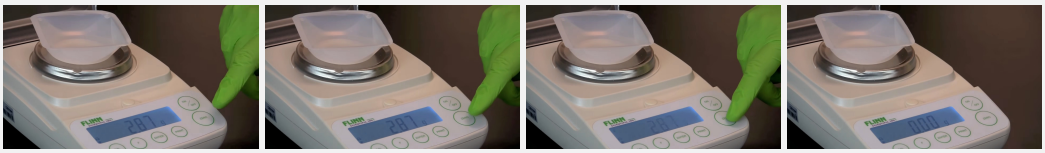
Example 1 (Level 1): Adjusting lever arm hangers for balance

From a first-person perspective in a physics laboratory, a metal lever arm with scale markings is mounted horizontally on a vertical metal stand. Two blue plastic hangers with hanging strings are suspended from the arm, initially positioned at the number 8 mark on the left side and the number 6 mark on the right side of the central pivot. **A pair of hands enters the frame, with the left hand grasping the left hanger and the right hand grasping the right hanger. The hands slide the hangers smoothly inward along the metal bar, moving them closer to the center pivot point. The motion is controlled and brings the hangers to positions near the number 2 marks on both sides. The hands release the hangers, which now hang steadily closer to the center of the lever arm.**



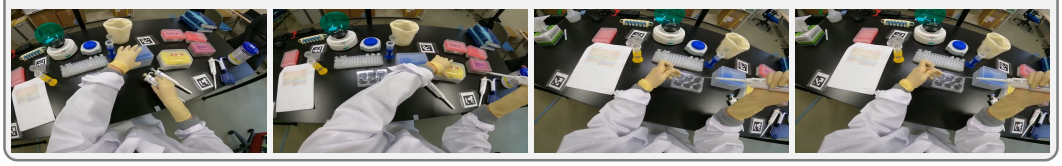
Example 2 (Level 1): Zeroing a digital scale with a weighing boat

From a nearby observer viewpoint, a white digital laboratory scale sits on a bench, displaying a reading of 2.87 g on its blue LCD screen. A translucent plastic weighing boat rests on the scale's metal pan. A hand wearing a bright green glove enters the frame from the right side. **The index finger extends and presses the "ZERO" button located on the control panel. Upon pressing the button, the digital display immediately resets, changing the reading from 2.87 g to 0.00 g. The gloved finger lifts away, leaving the scale tared and ready for measurement.**



Example 3 (Level 2): Aspirating liquid from box and dispensing into tube

From a first-person perspective looking down at a black laboratory bench, the operator's gloved hands are visible manipulating laboratory equipment. A yellow rack containing pipette tips or small tubes sits in the center of the workspace and is covered by a translucent lid. **The operator holds a white micropipette in the right hand. The left hand first opens the translucent lid of the yellow rack. The right hand then inserts the pipette tip into one of the slots in the yellow rack to aspirate liquid. The operator moves the micropipette away from the rack and uses the left hand to pick up a small clear microcentrifuge tube from the bench. Holding the tube securely between the left thumb and forefinger, the operator gently flips open the tube cap with the right middle finger. The operator then positions the micropipette tip over the tube opening and depresses the plunger to dispense the liquid into the tube.**



Example 4 (Level 2): Assembling and securing a substrate frame

From a nearby observer viewpoint looking down at a white laboratory bench, an operator wearing blue nitrile gloves holds a square white substrate in their left hand. On the bench to the left, a black rubber gasket rests on top of a clear acrylic frame, so the acrylic frame is underneath the gasket, and four gold binder clips are arranged nearby. **The operator's right hand picks up the black rubber gasket and carefully places it onto the white substrate, aligning the edges. Next, the right hand picks up the clear acrylic frame and places it on top of the gasket, creating a layered sandwich structure. With the frame positioned, the operator reaches for the four gold binder clips. The right hand attaches the clips one by one to the midpoint of each of the square frame's four sides, placing one clip at the center of each side to clamp the layers together. Once all four clips are secured, the operator holds the fully assembled and clamped frame steady.**



Figure 7: Four example tasks from LABINSTRUCT. Each block shows the task title, the difficulty level, the rendered generation instruction, and four frames sampled from the reference clip.

C Evaluation Details

C.1 Evaluation Design Details

This section provides additional details on the evaluation protocol summarized in Sec. 3.4.

Dimension Definitions. Each task is associated with a task-specific checklist covering six dimensions: *Object Consistency*, *Action Fidelity*, *State Correctness*, *Physical Plausibility*, *Scene Consistency*, and *Visual Safety*. The first four dimensions measure the **Core Visual Quality** of generated videos, while the latter two act as **Validity Constraints**. **Critical Completion** is derived from the critical items within the four core dimensions rather than treated as an independent dimension. *Object Consistency* evaluates whether task-relevant objects preserve their identity and appearance throughout the video. *Action Fidelity* assesses whether required actions are visibly executed on the correct targets and in the required order. *State Correctness* measures whether the expected observable state transitions and final states are achieved. *Physical Plausibility* evaluates whether object interactions and material behavior remain physically credible. *Scene Consistency* checks the temporal continuity of the workspace, viewpoint, and visual environment, while allowing camera motion consistent with the recorded viewpoint. *Visual Safety* evaluates task-relevant, visually observable safety conditions, such as containment, apparatus integrity, and safe object handling.

Checklist Construction. We generate one checklist per task using DeepSeek-v4-flash (DeepSeek-AI, 2026) with a fixed prompt template and JSON-constrained decoding (Appendix E.6). The model receives only the task instruction, identifier, and difficulty level; reference videos and generated outputs are not provided. Checklist size is adapted to task complexity. Each item is required to express a single positively phrased and visually verifiable claim grounded in the task instruction. Required actions, observable state transitions, ordering constraints, object consistency, physical interactions, and relevant safety conditions are explicitly covered. Items are additionally assigned an importance level of *Critical*, *Standard*, or *Supplementary*.

Expert Verification. Generated checklists are first screened using deterministic validation rules for schema validity, dimensional coverage, duplication, and required importance labels. They are then reviewed by domain experts against the task instruction and reference video. Reviewers verify visual decidability, task relevance, non-redundancy, and execution neutrality, and revise or remove items when necessary. The verified checklist for each task is subsequently frozen and shared across automated evaluation, human evaluation, and sensitivity analyses.

Checklist Statistics. The resulting benchmark contains 3,229 verified checklist items across 204 tasks, including 1,312 items for 101 Level 1 tasks and 1,917 items for 103 Level 2 tasks. This corresponds to an average of 13.0 items per Level 1 task and 18.6 items per Level 2 task. All tasks are covered by all six evaluation dimensions. Action Fidelity constitutes the largest fraction of the checklist corpus, reflecting the greater number of atomic actions and inter-step ordering dependencies required by multi-step procedures with multiple dependent actions.

C.2 Formal Scoring Procedure

This appendix formalizes the scoring procedure summarized in Sec. 3.4. Scoring is performed independently for each generated video. For each model and task level, every reported metric is the arithmetic mean of the corresponding per-video scores, rescaled to $[0, 100]$. Thus, aggregate scores are computed from per-video quantities rather than reconstructed from marginal averages of the individual dimensions or scoring components.

Item Verdicts and Severity. Following VGI-Bench (He et al., 2026), we adopt an inverse-decay principle for violation-based evaluation. VGI-Bench applies an inverse-decay function, $1/(1+x)$, to the number of repeated violations of an individual rubric item, reflecting the intuition that additional violations should have diminishing marginal impact once violations have already occurred. We adapt this principle to our task-conditioned evaluation setting, where checklist requirements differ in their procedural importance and consequences for successful task execution.

For each checklist item i in a core evaluation dimension, we define:

$$y_i = \begin{cases} 1, & \text{if item } i \text{ is judged YES,} \\ 0, & \text{if item } i \text{ is judged NO or UNJUDGEABLE.} \end{cases} \tag{2}$$

Unlike uniform weighting across rubric items, laboratory procedures contain requirements with different consequences for successful execution. For example, omitting an essential operation or failing to realize a required final state is more consequential than violating a supplementary requirement. We therefore assign each item an expert-defined importance weight:

$$w_i = \begin{cases} 1, & \textit{Critical}, \\ 0.5, & \textit{Standard}, \\ 0.25, & \textit{Supplementary}. \end{cases} \tag{3}$$

For each core dimension d, we define the accumulated weighted violation severity as

$$V_d = \sum_{i \in d} w_i (1 - y_i), \tag{4}$$

and map it to a normalized dimension score using inverse decay:

$$S_d = \frac{1}{1 + V_d}. \tag{5}$$

This formulation preserves the inverse-decay treatment of accumulated violations while extending it to importance-aware, dimension-level evaluation. The weighting scheme provides a simple and interpretable severity hierarchy: one *Critical* violation contributes the same severity as two *Standard* violations or four *Supplementary* violations. The inverse-decay mapping makes the first violations most consequential, while additional violations have progressively smaller marginal effects.

Core Visual Quality and Task Completion. The four core dimensions, namely *Object Consistency*, *Action Fidelity*, *State Correctness*, and *Physical Plausibility*, are combined through a geometric mean to obtain a balanced core score that penalizes any weak dimension:

$$S_{\text{core}} = \left(\prod_{d=1}^{K} S_d \right)^{1/K}, \qquad K = 4. \tag{6}$$

The geometric mean limits compensation across dimensions: a severe failure in one aspect of the generated procedure cannot be fully offset by strong performance elsewhere.

We separately measure whether the essential steps of the procedure are completed. Let n_c denote the number of *Critical* items in the four core dimensions that are not judged YES. Scene-consistency and safety items are excluded from n_c, since they are handled by the validity constraints defined below. The completion factor is

$$C(n_c) = \begin{cases} 1, & n_c = 0, \\ 0.5, & n_c = 1, \\ 0, & n_c \geq 2. \end{cases} \tag{7}$$

Hence, a single unmet critical requirement is treated as partial completion, whereas two or more indicate failure to complete the core procedure as specified.

Validity Constraints. Scene consistency and visual safety are treated as binary validity constraints because both are expected to hold under our TI2V setting. The input image largely fixes the scene configuration, while our prompts exclude unsafe actions. Violations therefore indicate exceptional model failures rather than ordinary quality variation, and are handled as veto conditions instead of graded scores. For $g \in \{\text{scene}, \text{safety}\}$,

$$G_g \in \{0, 1\}, \qquad G_g = 0 \iff \text{at least one item in dimension } g \text{ is judged NO.} \tag{8}$$

All items within each validity dimension have equal veto power regardless of importance. An UNJUDGEABLE item does not close the gate because missing evidence is not treated as a violation. Any explicit NO in scene consistency or safety closes the corresponding gate. These gates can only penalize explicit validity violations and can never increase a video's score.

Overall Score. The final per-video score is the product of these four components:

$$S_{\text{overall}} = S_{\text{core}} \cdot C(n_c) \cdot G_{\text{scene}} \cdot G_{\text{safety}}. \tag{9}$$

This decomposition separates three distinct aspects of evaluation: S_{core} measures execution quality, $C(n_c)$ captures completion of essential procedural requirements, and G_{scene} and G_{safety} enforce validity of the resulting video. In particular, any failed validity constraint sets the overall score to zero regardless of the video's core visual quality or completion status.

Aggregation and Reporting. All quantities above are first computed independently at the video level. For each model–level pair, we then aggregate these per-video quantities by taking their arithmetic mean and multiply the resulting value by 100 for reporting. Accordingly, the reported **Core** score is the mean of the per-video S_{core} values, while the **Completion** score is the mean of the per-video $C(n_c)$ values. Since **Scene** and **Safety** are binary constraints rather than scored dimensions, they are reported as gate validity rates, defined as the fractions of videos for which $G_{\text{scene}} = 1$ and $G_{\text{safety}} = 1$, respectively, under the corresponding evaluator.

The reported **Overall** score is the mean of the per-video S_{overall} values. Thus, each task contributes one video-level score to the final average, rather than first aggregating individual scoring components across tasks. When Overall is reported over the full benchmark rather than separately for each level, we refer to it as **Pooled Overall**, or simply **Overall** when the context is unambiguous. Pooled Overall is computed by directly pooling the per-video S_{overall} values from all 204 tasks and averaging them together. It is therefore not, in general, equal to the arithmetic mean of the Level 1 and Level 2 Overall scores, since the two levels may contain different numbers of tasks.

D Experimental Details

D.1 Evaluated Models and Generation Settings

Table 9 summarizes the evaluated model releases and their native generation configurations. For each model, we select the supported frame count that most closely matches the nominal duration of approximately 4 s for Level 1 tasks and 12 s for Level 2 tasks. All videos are evaluated at their native temporal length, without frame interpolation, frame duplication, or temporal extension. Due to its native generation constraints, MiniMax H3 produces slightly longer outputs for Level 1 tasks. To ensure a consistent setting, each video is generated at the aspect ratio of its task's initial frame, with the resolution set to the value closest to 480p that the corresponding model supports for that ratio; no model-specific prompt keywords or prompt-tuning strategies are used. Evaluation is restricted to the generated visual track, consistent with the benchmark's visually verifiable criteria.

D.2 Cross-Evaluator Robustness

Gemini 3.7 Flash Results. To examine the sensitivity of our benchmark results to the choice of VLM evaluator, we re-evaluate the full benchmark using Gemini 3.7 Flash (Google DeepMind, 2026a) as an alternative judge, while keeping all other components of the evaluation protocol unchanged. Table 10 reports the results under both Gemini 3.7 Flash and GPT-5.6 Sol. Across the 8 generators, the Overall score averages 34.5 with Gemini 3.7 Flash and 30.7 with GPT-5.6 Sol. The highest model score reaches 60.1 under Gemini 3.7 Flash and 46.3 under GPT-5.6 Sol. Moreover, the same five generators remain below an Overall score of 30 under both evaluators. Although the absolute scores differ between the two judges, both evaluations place current generator performance well below benchmark saturation. The relative performance of the strongest generators is also largely preserved across the two evaluators. MiniMax H3, Seedance 2.0, and Wan 3.0 form the top-performing group under both Gemini 3.7 Flash and GPT-5.6 Sol, although their internal ordering differs. Beyond this top group, the two evaluators also produce broadly similar performance patterns across the remaining generators. Overall, changing the VLM evaluator affects score calibration more than the main comparative conclusions of the benchmark.

Model-Level Consistency. We first examine whether the two VLM evaluators produce consistent model-level results. For Level 1, Level 2, and Overall, we construct two vectors of aggregate scores across the 8 evaluated generators, one obtained with Gemini 3.7 Flash and the other with GPT-5.6

Table 9: Generators evaluated in the benchmark and their generation settings. Release is reported to month precision. † indicates commercial models.

Model	Release	Scale	FPS	Resolution	Level 1 Frames	Level 2 Frames
Wan 2.2 I2V A14B	2025-07	14B	16	480p	65	193
LTX 2.3	2026-03	22B	24	480p	97	289
Cosmos3 Nano	2026-06	16B	24	480p	96	288
Cosmos3 Super	2026-06	64B	24	480p	96	288
LingBot Video	2026-07	30B	24	480p	97	289
MiniMax H3	2026-08	33B	24	480p	124	294
Seedance 2.0†	2026-04	–	24	480p	96	288
Wan 3.0†	2026-08	–	30	480p	120	360

Table 10: Benchmark scores by Gemini 3.7 Flash and GPT-5.6 Sol.

Model	Gemini 3.7 Flash			GPT-5.6 Sol		
	Level 1	Level 2	Overall	Level 1	Level 2	Overall
Wan 2.2	30.1	11.3	20.6	38.5	13.6	25.9
LTX 2.3	30.7	29.1	29.9	22.1	22.1	22.1
Cosmos3 Nano	22.6	12.4	17.5	23.7	15.2	19.4
Cosmos3 Super	28.7	20.5	24.6	30.0	15.7	22.8
LingBot Video	30.7	21.6	26.1	20.0	19.1	19.5
MiniMax H3	48.7	53.8	51.3	**53.6**	39.2	**46.3**
Seedance 2.0†	38.8	53.1	46.0	41.9	**48.2**	45.1
Wan 3.0†	**60.2**	**60.0**	**60.1**	46.2	42.7	44.4

Sol. We compute Pearson correlation to quantify the linear association between the scores assigned by the two evaluators, and Spearman correlation to assess the consistency of the resulting generator rankings. As shown in Table 11, the Overall scores are strongly correlated across evaluators, with a Pearson correlation of $r = 0.923$ and a Spearman correlation of $\rho = 0.714$. The highest consistency is observed at Level 2 ($r = 0.968$, $\rho = 0.929$), where both the aggregate scores and generator rankings closely align across the two evaluators. Level 1 exhibits lower, though still positive, cross-evaluator correlations ($r = 0.783$, $\rho = 0.611$). Overall, the two evaluators produce similar model-level performance patterns, with particularly strong consistency on Level 2 and some differences in the finer-grained ordering of generators on Level 1 and in the overall results.

Task-Wise Ranking Agreement. We further evaluate the consistency of relative generator performance at the individual-task level. For each task, we consider generators with valid scores from both evaluators and compute Spearman's ρ and Kendall's τ_b over their respective rankings. Degenerate rankings are excluded from the corresponding correlation statistic. We report the macro-average Spearman and Kendall correlations across tasks, together with the median task-wise Spearman correlation. We additionally measure pairwise agreement by considering generator pairs for which both evaluators produce a non-tied ordering and computing the fraction for which the two evaluators identify the same generator as higher scoring. The resulting pairwise agreement rates are likewise macro-averaged across the retained non-degenerate tasks.

Table 12 reports the task-wise ranking agreement between the two evaluators. Over the 167 tasks, the mean Spearman correlation is 0.345, the median Spearman correlation is 0.426, and the mean Kendall correlation is 0.327. The corresponding pairwise agreement rate is 70.8%. At Level 2, the two evaluators achieve a mean Spearman correlation of 0.415, a median Spearman correlation of 0.539, a mean Kendall correlation of 0.388, and a pairwise agreement rate of 74.7% across 81 tasks. For Level 1, the corresponding values are 0.279, 0.280, 0.269, and 67.0% across 86 tasks. The higher correlation and pairwise-agreement values at Level 2 show that generator rankings are more consistent between the two evaluators on Level 2 tasks, whereas Level 1 exhibits greater differences in the task-wise relative ordering of generators.

Table 11: Model-level consistency between Gemini 3.7 Flash and GPT-5.6 Sol. Pearson and Spearman correlations are computed across the aggregate scores of the 8 evaluated generators.

Score	Pearson r	Spearman ρ
Level 1	0.783	0.611
Level 2	0.968	0.929
Overall	0.923	0.714

Table 12: Task-wise ranking agreement between Gemini 3.7 Flash and GPT-5.6 Sol, measured by Spearman correlation, Kendall correlation, and pairwise ordering agreement.

Split	# Tasks	Mean ρ	Median ρ	Mean τ_b	Pairwise Agree.
Level 1	86	0.279	0.280	0.269	67.0%
Level 2	81	0.415	0.539	0.388	74.7%
Overall	167	0.345	0.426	0.327	70.8%

D.3 Human Evaluation Details

Evaluation Protocol. Human judgments serve as the reference for assessing the reliability of the automatic evaluator in Sec. 4.2. We conduct human evaluation over the full evaluation set, covering every generated video and all applicable task-specific checklist items. Each video is independently evaluated by four qualified raters using the same checklist questions and evaluation criteria as the VLM evaluator. The four raters include two master's-level researchers and two doctoral-level researchers. We develop a dedicated annotation interface (Figure 8) that presents the task instruction, task metadata, generated video, and one checklist question at a time. Raters may consult the source video only to clarify task context or object identity; whether a checklist item is satisfied must be decided solely from evidence in the generated video. To reduce model-specific bias, generator identities are anonymized throughout annotation. The interface also displays each item's evaluation dimension and highlights critical items when applicable.

Before formal annotation, all raters complete a calibration session using representative examples. For each checklist item, raters assign one of three labels: YES, NO, or UNJUDGEABLE. Unanimous judgments are accepted directly, while disagreements are resolved through expert adjudication. Cases involving subtle state transitions, liquid handling, transparent objects, brief contacts, or safety-critical actions receive additional review when needed.

Human evaluation follows the same core judgment principles as the VLM evaluator:

- **Evidence-based Judgment.** Raters base each verdict only on evidence visible in the generated video and do not infer correctness from the task description, expected procedure, final-outcome plausibility, or source-video execution.
- **Yes/No Criteria.** A checklist item is marked YES only when the generated video clearly demonstrates that the required action, state, or condition is satisfied, and NO when it clearly shows that the requirement is absent, violated, or incorrectly performed.
- **Unjudgeable Under Limited Visibility.** An item is marked UNJUDGEABLE when the available visual evidence is insufficient to determine whether the requirement is satisfied, such as under occlusion, severe blur, objects leaving the frame, or visual artifacts. Failure to observe an event is not, by itself, treated as evidence that the event did not occur or as grounds for a NO verdict under this evidence-only policy.
- **Temporal and Conditional Relations.** For checklist items involving temporal order or conditional dependencies, raters assess the required relation when all relevant events are reliably observable. A clearly missing event or an observed ordering that violates the checklist is marked NO. If event occurrence or ordering cannot be determined reliably, the item is marked UNJUDGEABLE to preserve the distinction between observed failure and missing evidence in the final annotations.

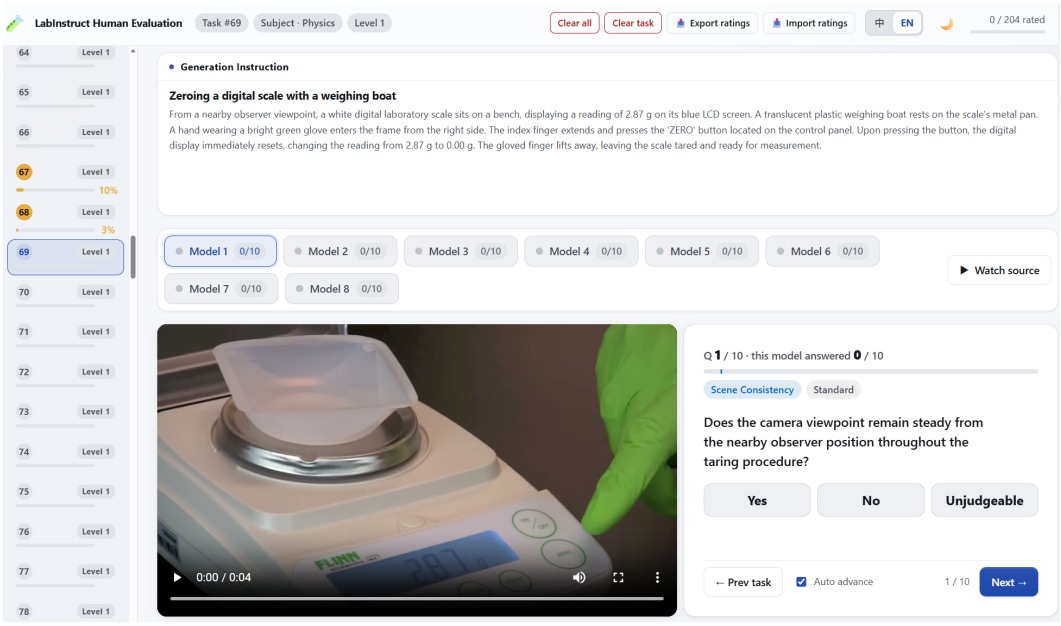


Figure 8: Human evaluation interface for task-specific assessment of generated videos.

Table 13: Human–human agreement on pre-adjudication annotations. Fleiss' κ measures chance-corrected agreement among all raters. Pairwise Agree. is averaged over all rater pairs.

Comparison	Fleiss' κ	Pairwise Agree. (%)
Human–Human	0.57	80.7

Human Inter-Rater Reliability. We assess the reliability of the human judgments using the independent annotations collected before expert adjudication. Each checklist item is rated by all four human evaluators using the same three categorical labels, YES, NO, and UNJUDGEABLE. As shown in Table 13, the four raters achieve a Fleiss' κ of 0.57 and a mean pairwise agreement of 80.7% over all six rater pairs. These results indicate moderate chance-corrected inter-rater reliability together with high raw agreement prior to adjudication.

Task-Wise Ranking Agreement. We further examine whether human and VLM evaluations induce consistent generator rankings at the individual task level. For each task, we independently rank the 8 generators using their human-based and VLM-based overall scores, then compare the two rankings using Spearman ρ and Kendall's τ_b. We also compute pairwise agreement as the fraction of non-tied generator pairs for which the two evaluators identify the same generator as better, and macro-average this quantity across tasks. Tasks with fully tied rankings under either evaluator are excluded from the corresponding ranking statistics, leaving 79 Level 1 tasks and 60 Level 2 tasks for comparison. As shown in Table 14, the two evaluators exhibit moderate but consistent agreement within individual tasks. Across all 139 valid tasks, the mean Spearman correlation is $\rho = 0.364$, the median Spearman correlation is 0.378, and the mean Kendall correlation is $\tau_b = 0.345$. At the pairwise level, human and VLM evaluations agree on 72.9% of non-tied generator comparisons, indicating that they identify the stronger generator in nearly three quarters of pairwise decisions. Agreement is also stable across the two strata, with pairwise agreement rates of 72.9% for Level 1 and 73.0% for Level 2, while Level 2 shows slightly higher mean Spearman and Kendall correlations. These results suggest that the VLM evaluator preserves a meaningful portion of the task-specific relative performance signal, although exact rankings remain less stable at the individual task level than after aggregation across the full benchmark.

Table 14: Task-wise ranking agreement between Human and VLM evaluation. # Tasks counts the tasks whose generator ranking is non-degenerate under both evaluators. Fully tied rankings are excluded from Spearman ρ, Kendall's τ_b, and the pairwise agreement rate alike.

Split	# Tasks	Mean ρ	Median ρ	Mean τ_b	Pairwise Agree.
Level 1	79	0.351	0.378	0.331	72.9%
Level 2	60	0.380	0.432	0.364	73.0%
Overall	139	0.364	0.378	0.345	72.9%

Table 15: Sensitivity of pooled model scores under alternative core-score normalization, completion-threshold formulations, and strict completion criteria. Higher is better. Best values are bolded, subscripts indicate within-row rank, and † denotes commercial models.

Criterion	Variant	Wan 2.2	LTX 2.3	Cosmos3 Nano	Cosmos3 Super	LingBot Video	MiniMax H3	Seedance 2.0†	Wan 3.0†
S_{core}	Original	68.0_4	66.0_5	63.0_8	65.2_6	64.9_7	79.8_2	78.7_3	$\mathbf{81.0}_1$
	Norm.	82.8_4	80.9_5	79.4_8	80.8_6	80.4_7	89.7_2	89.1_3	$\mathbf{90.4}_1$
Comp.	Original	30.6_3	26.2_4	21.5_7	24.7_5	22.7_6	$\mathbf{51.9}_1$	49.5_2	$\mathbf{51.9}_1$
	Relative	32.8_4	28.1_5	22.5_8	26.4_6	24.2_7	$\mathbf{54.6}_1$	51.4_3	54.4_2
Overall	Original	25.9_4	22.1_6	19.4_8	22.8_5	19.5_7	$\mathbf{46.3}_1$	45.1_2	44.4_3
	+ *Norm.* S_{core}	26.9_4	22.8_6	19.9_8	23.5_5	20.5_7	$\mathbf{47.2}_1$	46.2_2	45.4_3
	+ *Relative* Comp.	27.5_4	23.1_6	20.0_8	23.7_5	20.5_7	$\mathbf{47.9}_1$	46.4_2	45.8_3
	+ *Both*	29.0_4	24.1_6	20.8_8	24.6_5	21.8_7	$\mathbf{49.3}_1$	47.9_2	47.2_3
	Strict Overall	20.4_4	17.4_6	15.9_7	18.8_5	13.3_8	$\mathbf{40.3}_1$	39.1_2	38.3_3

D.4 Additional Analyses

D.4.1 Scoring Sensitivity and Strict Overall

The benchmark score combines two main components: a core score, S_{core}, that aggregates weighted violations across the core dimensions, and a completion component that accounts for whether Critical requirements are satisfied. We examine whether the resulting model comparison is sensitive to the specific formulation of either component. Table 15 reports pooled scores over tasks from both levels under the original formulation and several alternative variants.

Scoring Variants. For S_{core}, the *Norm.* variant normalizes the accumulated violation severity in each dimension by the total available weight:

$$\bar{V}_d = \frac{\sum_{i \in d} w_i (1 - y_i)}{\sum_{i \in d} w_i}. \tag{10}$$

Thus, severity reflects the fraction of weighted requirements violated rather than the absolute amount of checklist weight. For the Comp. metric, the *Relative* variant expresses the partial-credit threshold relative to the number of Critical requirements. Let N_{crit} denote the number of Critical items in the four core dimensions and n_c the number of those items not judged YES. We define

$$\tilde{C}(n_c) = \begin{cases} 1, & n_c = 0, \\ 0.5, & 1 \le n_c \le 0.2\, N_{\text{crit}}, \\ 0, & n_c > 0.2\, N_{\text{crit}}, \end{cases} \tag{11}$$

so that partial credit depends on the fraction of unmet Critical requirements rather than on an absolute count. In the Overall block, + *Norm.* S_{core} and + *Relative Comp.* apply these two changes separately, while + *Both* applies them jointly. The *Strict* variant instead removes partial completion credit altogether, assigning zero completion whenever any Critical requirement is unmet:

$$\hat{C}(n_c) = \begin{cases} 1, & n_c = 0, \\ 0, & n_c \ge 1. \end{cases} \tag{12}$$

Consequently, a failure in any single Critical item is equivalent to failure of the entire task, which demands that the model be capable of end-to-end, error-free execution.

Table 16: Sensitivity of Overall scores to evaluator sampling rate. Bold indicates the highest score across 2, 4, and 8 FPS for each model under each evaluator. † indicates commercial models.

Model	GPT-5.6 Sol			Gemini 3.7 Flash		
	FPS 2	FPS 4	FPS 8	FPS 2	FPS 4	FPS 8
Wan 2.2	13.3	25.9	**31.3**	**24.8**	20.6	24.4
LTX 2.3	21.1	22.1	**23.5**	31.5	29.9	**35.3**
Cosmos3 Nano	13.9	19.4	**25.2**	17.5	17.5	**18.8**
Cosmos3 Super	18.8	22.8	**28.8**	21.9	24.6	**26.1**
LingBot Video	17.2	19.5	**29.1**	25.9	26.1	**26.9**
MiniMax H3	41.7	46.3	**48.6**	50.4	51.3	**51.7**
Seedance 2.0†	30.2	45.1	**46.6**	**47.5**	46.0	46.0
Wan 3.0†	24.9	44.4	**48.9**	60.5	60.1	**61.1**

Ranking Robustness Across Scoring Variants. Alternative formulations affect score magnitudes more than model ordering. Normalizing S_{core} increases its value by 9.4–16.4 points while leaving the ranking unchanged, and changes Overall by only 0.5–1.1 points. Using the relative completion rule $\tilde{C}$ increases Overall by 0.6–1.6 points, while applying both modifications increases it by 1.4–3.1 points. Nevertheless, the Overall ranking is identical under *Original*, + *Norm.* S_{core}, + *Relative Comp.*, and + *Both*. Under the strict rule $\hat{C}$, the first six positions remain unchanged, with only LingBot Video and Cosmos3 Nano exchanging the final two positions. These results show that the model ordering is robust to the formulation of both S_{core} and completion.

Completion Affects the Ordering of Leading Models. The component scores nevertheless emphasize different aspects of performance. Under the original S_{core}, Wan 3.0 ranks first, followed by MiniMax H3 and Seedance 2.0, whereas the Overall score orders them as MiniMax H3, Seedance 2.0, and Wan 3.0. Thus, incorporating completion changes the relative ordering of the leading models. At the same time, this ordering remains unchanged across $\tilde{C}$, normalized S_{core}, and their joint variant, indicating that the result is not driven by a particular scoring formulation.

D.4.2 Sensitivity to Evaluator Sampling Rate

Table 16 evaluates the effect of frame sampling rate by re-scoring the full benchmark at 2, 4, and 8 FPS under both GPT-5.6 Sol and Gemini 3.7 Flash, with all other evaluation settings unchanged. All reported scores pool tasks from both levels. Gemini 3.7 Flash is relatively stable across sampling rates: Wan 3.0, MiniMax H3, and Seedance 2.0 remain the top three models at all three rates, and the mean score across models varies only from 34.5 to 36.3. GPT-5.6 Sol is more sensitive to sampling density, with the mean score increasing from 22.6 at 2 FPS to 30.7 at 4 FPS and 35.2 at 8 FPS. The increase is particularly pronounced for Wan 3.0 and Seedance 2.0, whereas MiniMax H3 changes more modestly. Despite these score shifts, the same three models form the leading group across all sampling rates. We use 4 FPS as the default setting in the main experiments to balance evaluation fidelity and computational cost. Compared with 2 FPS, 4 FPS substantially improves temporal coverage, particularly for GPT-5.6 Sol, while avoiding the additional inference cost and latency of processing twice as many sampled frames at 8 FPS. Since increasing the rate to 8 FPS mainly changes absolute scores while leaving the leading group largely unchanged, 4 FPS provides a practical operating point for full-benchmark evaluation.

D.4.3 Results by Operation Domain, Discipline, and Viewpoint

Performance by Operation Domain. Table 17 reveals clear operation-domain specialization among the strongest generators. MiniMax H3 achieves the highest average score in instrument operations and solid handling, Seedance 2.0 leads container operations and heating/cooling, and Wan 3.0 performs best in liquid handling and measurement. Filtration/separation is the only domain led by a model outside this overall top group, with Wan 2.2 achieving the highest score of 42.7. Overall, the leading models exhibit complementary strengths across operation domains rather than a uniformly dominant capability profile, and no single generator performs best across the full range of laboratory

Table 17: Overall score by operation domain and task level. † denotes commercial models.

Domain	Level	Wan 2.2	LTX 2.3	Cosmos3 Nano	Cosmos3 Super	LingBot Video	MiniMax H3	Seedance 2.0†	Wan 3.0†
Container Operations	Level 1	76.0	48.8	71.4	34.2	24.0	**84.3**	70.0	52.9
	Level 2	16.7	20.0	9.3	15.9	20.3	29.3	**51.7**	36.2
	Overall	26.6	24.8	19.7	18.9	20.9	38.4	**54.7**	39.0
Filtration/Separation	Level 1	**66.6**	47.3	0.0	0.0	26.6	33.3	0.0	0.0
	Level 2	28.4	28.4	54.7	0.0	38.0	38.0	**60.0**	**60.0**
	Overall	**42.7**	35.5	34.2	0.0	33.8	36.3	37.5	37.5
Heating/Cooling	Level 1	0.0	0.0	0.0	0.0	0.0	50.0	0.0	**71.0**
	Level 2	0.0	0.0	20.0	0.0	8.4	40.0	**88.4**	27.9
	Overall	0.0	0.0	14.2	0.0	6.0	42.8	**63.1**	40.2
Instrument Operations	Level 1	54.9	18.0	25.8	32.4	21.2	**55.2**	49.1	52.8
	Level 2	16.3	25.4	19.8	21.5	21.2	57.3	**57.7**	47.4
	Overall	36.2	21.6	22.9	27.1	21.2	**56.2**	53.2	50.2
Liquid Handling	Level 1	15.2	16.7	12.8	21.6	13.6	31.2	28.7	**38.8**
	Level 2	11.1	12.2	8.5	5.8	20.2	18.4	29.9	**33.5**
	Overall	12.8	14.1	10.3	12.5	17.4	23.8	29.4	**35.7**
Measurement	Level 1	33.9	21.3	7.8	37.0	28.6	57.6	41.9	**64.5**
	Level 2	12.7	29.6	5.8	11.7	0.0	43.0	46.9	**56.7**
	Overall	23.0	25.6	6.8	24.0	13.8	50.1	44.4	**60.4**
Solid Handling	Level 1	31.3	25.4	38.3	38.7	18.6	**71.5**	49.1	30.4
	Level 2	6.0	25.0	7.6	18.4	13.5	42.4	**44.4**	40.4
	Overall	16.2	25.2	20.0	26.6	15.5	**54.2**	46.3	36.3

operations. Cross-level differences are interpreted descriptively, since Level 1 and Level 2 contain independently curated task sets rather than controlled variants of the same tasks.

Performance by Scientific Discipline. Table 18 presents a similar pattern of complementary strengths across scientific disciplines. MiniMax H3 achieves the highest average score in agronomy and chemistry, Seedance 2.0 leads biology and materials science, and Wan 3.0 performs best in physics. The performance gap between the leading models varies substantially by discipline: MiniMax H3 leads agronomy by 15.2 points, whereas Wan 3.0 exceeds MiniMax H3 by only 1.6 points in physics. Thus, model differentiation is much stronger in some disciplines than in others. Across all 5 disciplines, the highest scores are achieved by MiniMax H3, Seedance 2.0, or Wan 3.0, but no single model leads every discipline. The discipline-level results therefore reveal distinct capability profiles across scientific settings, with model performance varying according to the procedures, objects, and interaction patterns involved in each discipline.

Performance by Experiment Viewpoint. Table 19 reveals distinct performance patterns across the two experiment viewpoints. Nearby-observer tasks account for 137 of the 204 tasks, while the remaining 67 are first-person tasks. On nearby-observer tasks, MiniMax H3 ranks first with 52.2, followed by Seedance 2.0 at 47.1 and Wan 3.0 at 44.5. On first-person tasks, Wan 3.0 instead ranks first with 44.2, followed by Seedance 2.0 at 41.0 and MiniMax H3 at 34.3. Thus, Seedance 2.0 remains among the top two models under both viewpoints, whereas the relative positions of MiniMax H3 and Wan 3.0 differ substantially across recording perspectives.

The level-wise results further show that these viewpoint subsets are not uniformly ordered by task level. For first-person tasks, Seedance 2.0 and Wan 3.0 obtain higher scores on Level 2 than on Level 1 (46.0 vs. 34.7 and 46.6 vs. 41.3, respectively). The six open-weight models show a mixed pattern: Wan 2.2, Cosmos3 Super, and MiniMax H3 score higher on Level 1, whereas LTX 2.3, Cosmos3 Nano, and LingBot Video score higher on Level 2. For nearby-observer tasks, all six open-weight models score higher on Level 1 than on Level 2; among the commercial models, Seedance 2.0 scores higher on Level 2, whereas Wan 3.0 scores higher on Level 1. Because the task levels are independently curated, these contrasts are descriptive and should not be interpreted as isolated effects of task level or procedure length alone.

Table 18: Overall score by scientific discipline and task level. † denotes commercial models.

Discipline	Level	Wan 2.2	LTX 2.3	Cosmos3 Nano	Cosmos3 Super	LingBot Video	MiniMax H3	Seedance 2.0†	Wan 3.0†
Agronomy	Level 1	27.0	53.2	31.5	34.2	18.4	**57.0**	36.4	34.2
	Level 2	5.9	13.4	17.5	18.4	3.2	**49.3**	38.3	35.3
	Overall	15.1	30.7	23.6	25.2	9.8	**52.7**	37.5	34.8
Biology	Level 1	21.6	16.9	21.6	19.5	13.7	**28.7**	25.8	22.8
	Level 2	11.5	14.2	7.9	7.6	20.1	9.0	**36.2**	26.0
	Overall	16.1	15.4	14.1	13.0	17.2	18.0	**31.5**	24.5
Chemistry	Level 1	60.5	17.7	39.4	28.7	28.7	**69.7**	45.4	53.9
	Level 2	18.2	24.4	18.9	13.5	24.9	41.0	48.2	**49.7**
	Overall	39.8	21.0	29.4	21.2	26.8	**55.6**	46.8	51.8
Materials Science	Level 1	28.4	48.4	40.0	60.0	20.0	60.0	**80.0**	28.4
	Level 2	5.2	25.0	12.5	4.7	21.3	34.3	**37.5**	30.2
	Overall	14.1	34.0	23.0	26.0	20.8	44.1	**53.8**	29.5
Physics	Level 1	36.8	16.6	10.6	31.4	17.9	54.1	44.3	**58.2**
	Level 2	16.8	29.0	17.4	25.0	19.8	56.9	**63.3**	55.5
	Overall	27.7	22.2	13.7	28.5	18.7	55.4	52.9	**57.0**

Table 19: Overall score by experiment viewpoint and task level. † denotes commercial models.

Viewpoint	Level	Wan 2.2	LTX 2.3	Cosmos3 Nano	Cosmos3 Super	LingBot Video	MiniMax H3	Seedance 2.0†	Wan 3.0†
First-Person	Level 1	31.9	15.2	11.3	20.8	15.3	**44.2**	34.7	41.3
	Level 2	11.6	24.7	13.2	12.5	21.2	26.2	46.0	**46.6**
	Overall	20.7	20.5	12.4	16.2	18.6	34.3	41.0	**44.2**
Nearby Observer	Level 1	41.3	25.0	28.9	33.9	21.9	**57.6**	45.0	48.3
	Level 2	14.8	20.7	16.3	17.6	18.0	46.4	**49.4**	40.5
	Overall	28.5	22.9	22.8	26.0	20.0	**52.2**	47.1	44.5

D.4.4 Effect of Re-sampling

To assess the effect of sampling randomness, we conduct a best-of-three analysis on MiniMax H3, Seedance 2.0, and Wan 3.0 using a fixed 40-task subset balanced across the two task levels. For each model, we generate two additional samples per task and compare the original Overall score with the score obtained by selecting the best of the three samples independently for each task. We also report the fraction of subset tasks for which the original generation remains a best-performing sample.

As shown in Table 20, the three models differ substantially in their sensitivity to re-sampling. MiniMax H3 improves only slightly, from 55.3 to 57.8, with the original sample remaining best for 95% of the tasks. In contrast, Seedance 2.0 and Wan 3.0 improve by 14.0 and 11.1 points, respectively, and both achieve higher best-of-three scores than MiniMax H3. These results show that the ordering within this leading group is not robust to sampling variation. Although MiniMax H3 obtains the highest Pooled Overall score under the single-sample protocol, Seedance 2.0 and Wan 3.0 exhibit substantially larger gains when additional samples are available. We therefore treat the main-table ordering as specific to the single-sample setting rather than as evidence that MiniMax H3 consistently outperforms the other two generators across repeated sampling.

D.4.5 Failure Cases and Attribution

Figures 9 and 10 show videos generated by different models on four representative tasks across Level 1 and Level 2 settings. The Level 1 examples highlight two distinct requirements for successful task execution: correct action realization and correct state transition. In the lever-arm task, MiniMax H3 completes the required inward adjustment, whereas Cosmos3 Super and Wan 3.0 exhibit procedural errors in the shown samples. In the scale-zeroing task, MiniMax H3 and Wan 3.0 reach the target display state, while Cosmos3 Super fails to realize the expected zeroed state. These

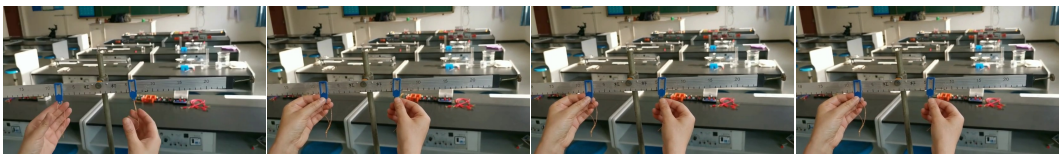
Table 20: Best-of-three sampling on the balanced 40-task subset.

Model	Original	Best-of-3	Gain	Original Best (%)
MiniMax H3	55.3	57.8	+2.5	95.0
Seedance 2.0†	52.8	66.8	+14.0	80.0
Wan 3.0†	53.7	64.8	+11.1	85.0

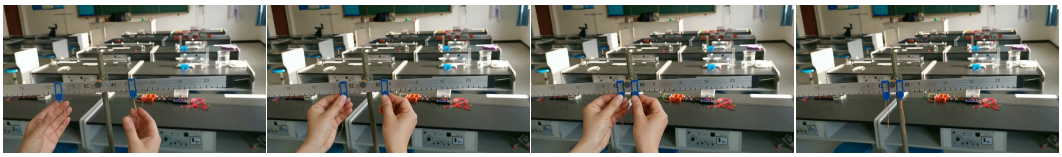
examples illustrate that even simple manipulation tasks require precise grounding of both the prescribed action and its intended outcome.

The Level 2 examples further emphasize the challenge of maintaining procedural and state consistency over multi-step interactions. In the pipetting task, MiniMax H3 and Wan 3.0 omit different critical actions, while Cosmos3 Super completes the demonstrated procedure. In the substrate-frame assembly task, Cosmos3 Super and Wan 3.0 show a combination of procedural omission, incomplete state realization, and object drift, leading to incorrect final configurations. These cases suggest that complex manipulation requires not only locally correct actions but also consistent object states and spatial relations throughout the full execution sequence. The examples illustrate representative success and failure modes and are not intended to establish a ranking among generators.

Example 1 (Level 1): Adjusting lever arm hangers for balance



Model: Cosmos3 Super Overall Score: 0.0 Failure Attribution: Procedural Omission

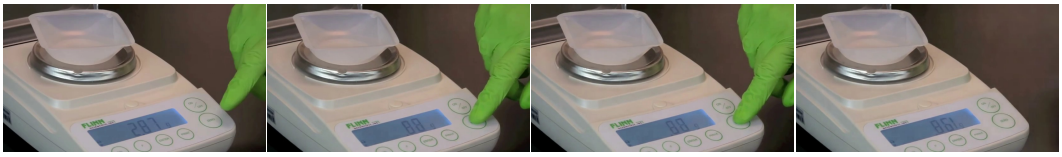


Model: MiniMax H3 Overall Score: 100.0



Model: Wan 3.0 Overall Score: 0.0 Failure Attribution: Procedural Omission

Example 2 (Level 1): Zeroing a digital scale with a weighing boat



Model: Cosmos3 Super Overall Score: 0.0 Failure Attribution: Incomplete State Realization



Model: MiniMax H3 Overall Score: 100.0



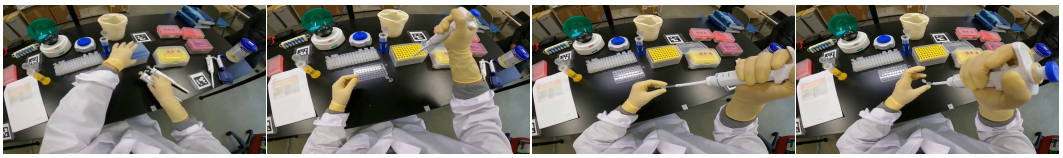
Model: Wan 3.0 Overall Score: 100.0

Figure 9: Videos generated by different models for two Level 1 example tasks.

Example 3 (Level 2): Aspirating liquid from box and dispensing into tube



Model: Cosmos3 Super Overall Score: 100.0

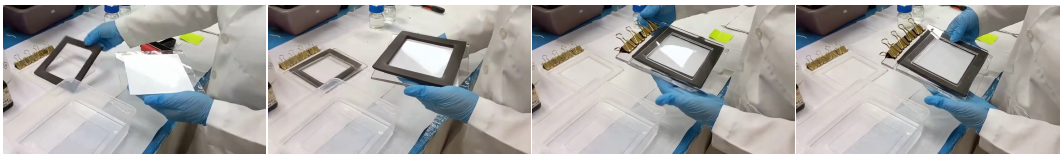


Model: MiniMax H3 Overall Score: 0.0 Failure Attribution: Procedural Omission

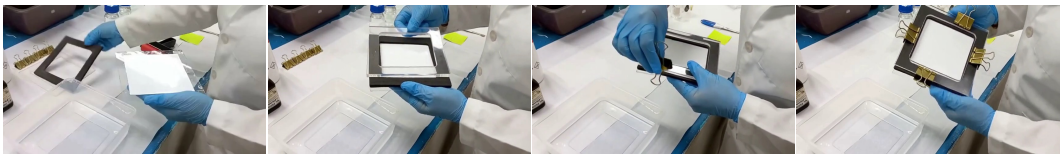


Model: Wan 3.0 Overall Score: 0.0 Failure Attribution: Procedural Omission

Example 4 (Level 2): Assembling and securing a substrate frame



Model: Cosmos3 Super Overall Score: 0.0
Failure Attribution: Procedural Omission, Incomplete State Realization, Object Drift



Model: MiniMax H3 Overall Score: 100.0



Model: Wan 3.0 Overall Score: 0.0
Failure Attribution: Procedural Omission, Incomplete State Realization, Object Drift

Figure 10: Videos generated by different models for two Level 2 example tasks.

E TASK SCHEMA AND PROMPT TEMPLATES

This section presents condensed versions of the task schema and prompts used to draft annotations, render generation instructions, construct task-conditioned QA items, and obtain structured VLM judgments. We retain the roles, inputs, principal constraints, and output contracts required to understand the pipeline while omitting repeated explanations and illustrative examples. The complete verbatim templates are provided in the accompanying project repository.

E.1 STRUCTURED TASK SCHEMA

For readability, we render the JSON task specification as semantically equivalent YAML.

```
task_id: <task_id>
split: <L1 | L2>
domains:
  - <operation domain, multi-label: a task may carry more than one>
discipline: <biology | chemistry | physics | agronomy | materials_science>
title: <5-10 word title>
description: <brief description of the actions and outcome>

scene:
  objects:
    - id: <snake_case_id>
      name: <human-readable name>
      contents: <visible contents | empty | N/A>
      material: <visible material>
  spatial_relations:
    - <task-relevant relation per object and bench surface>
  environment: indoor_laboratory
  viewpoint: <first_person | nearby_observer>

action_sequence:
  - step: <integer>
    action: <present-tense hand/tool/object action>
    target_object: <id from scene.objects>
    contact: <hand_tool | hand_object | tool_object | tool_liquid | none>
    state_transition:
      - from: <visible state before the action>
        to: <visible state after the action>

difficulty_features:
  number_of_objects: <integer>
  number_of_action_steps: <integer>
  requires_liquid_handling: <true | false>
  requires_tool_manipulation: <true | false>
  requires_fine_motor_control: <true | false>
  requires_transparent_object_reasoning: <true | false>
  requires_color_tracking: <true | false>
  requires_volume_change_detection: <true | false>
  requires_spatial_memory: <true | false>
  requires_sequential_ordering: <true | false>
```

E.2 ANNOTATION PROMPT FOR LEVEL 1 TASKS

```
You are an expert scientific laboratory procedure annotator specializing in atomic
actions. Inspect a short clip (typically 3-6 seconds) showing one experiment step
composed of 1-3 closely related actions, and produce a structured JSON draft for
human verification. The clip may use a first-person or nearby-observer viewpoint.

Your Expertise
- Laboratory tools, containers, instruments, samples, and visible materials
- Task-relevant hand-object, hand-tool, tool-object, and tool-liquid interactions
- Observable contacts, motions, outcomes, and state changes
```

```
Base every annotation on visible evidence. Do not infer identity, contents, material,
quantity, safety status, camera mounting, or hidden state from a typical protocol.
Use supplied metadata when available; otherwise describe only what the video supports.

Task Metadata
- Task ID: {{TASK_ID}}
- Verb: {{VERB}}
- Domain: {{DOMAIN}}

Input Video
[The provided clip shows ONE experiment step lasting approximately {{DURATION}}
seconds. Inspect the full clip and identify its viewpoint, task-relevant objects,
actions, contacts, and visible state changes.]

Annotation Task
Annotate one step containing 1-3 brief atomic interactions. Record the scene,
chronological actions, target objects, contacts, and state transitions.

JSON Schema
{{TASK_SCHEMA}}

JSON Output Rules
1. Use valid JSON with double quotes and no trailing commas.
2. Every target_object must match an id in scene.objects.
3. Determine viewpoint from the video; do not default to first_person.
4. Output only one JSON object, without Markdown fences or commentary.
```

E.3 Annotation Prompt for Level 2 Tasks

```
You are an expert scientific laboratory procedure annotator specializing in complete
short subprocedures with ordered actions and dependencies. Inspect an 8-16 second
clip and produce a structured JSON draft for human verification. The clip may use a
first-person or nearby-observer viewpoint.

Your Expertise
- Ordered operations involving laboratory tools, containers, instruments, samples,
and visible materials
- Sequential dependencies, task-relevant contacts, and observable state transitions
- First-person and nearby-observer viewpoints and their occlusion and motion patterns

Base every annotation on visible evidence. Do not complete a familiar protocol from
background knowledge or infer identity, contents, material, quantity, safety status,
camera mounting, or hidden state.

Task Metadata
- Task ID: {{TASK_ID}}
- Task: {{TASK}}
- Category: {{CATEGORY}}
- Domain: {{DOMAIN}}

Input Video
[The provided clip shows a multi-step procedure lasting approximately {{DURATION}}
seconds. Inspect the full clip and identify the viewpoint, complete visible
subprocedure, relevant objects, chronological actions, dependencies, contacts, and
state transitions.]

Annotation Task
Annotate one complete short multi-step task. Preserve the observed order and
dependencies rather than reconstructing the expected scientific protocol.

JSON Schema
{{TASK_SCHEMA}}

JSON Output Rules
1. Use valid JSON with double quotes and no trailing commas.
2. Every target_object must match an id in scene.objects.
3. Determine viewpoint from the video; do not default to first_person.
4. Output only one JSON object, without Markdown fences or commentary.
```

E.4 Instruction Rendering Prompt for Level 1 Tasks

```
You are a visual narrator describing one scientific laboratory experiment step
containing 1-3 closely related actions for a video generation model. Translate the
structured specification into a single flowing paragraph (80-150 words, no headings,
bullets, or line breaks) for a target output of about 4 seconds. Preserve the
observed scene, viewpoint, required contacts, and visible outcome.

0. Preserve the recorded scene and viewpoint
- Describe the viewpoint actually shown: first-person or nearby observer. Do not
convert one category into the other or claim that the camera is head-mounted unless
this is known.
- Anchor the prompt in the initial workspace. Mention bench color, lighting, camera
angle, and object layout only when visible.
- Preserve camera behavior consistent with the clip. Do not require a fixed camera
when the recorded viewpoint moves naturally with the operator.
- Do not introduce cuts, unexplained viewpoint changes, or objects absent from the
observed workspace.

1. Focus on the atomic actions
Describe every required action with sufficient visual detail for a convincing short
video:
- The hand entering the frame, when applicable, or already in position
- The grip and how the fingers contact the object
- The direction, speed, and precision of motion
- The visible result: where the object ends up and what changes

2. Describe all and only core objects
Include every task-relevant visible object needed to execute or verify the action. A
rack, reagent container, support, instrument, or other item must not be omitted when
it participates in contact, establishes the initial scene, or makes the outcome
judgeable. Keep untouched background clutter brief.

3. Embed state changes in actions
The action IS the state change. Integrate the initial state, motion, and result into
one continuous description rather than writing disconnected before/action/after
sentences. Preserve the observed hand only when hand choice is visible and
task-relevant. Do not turn incidental hand choice or an exact motion path into a
requirement.

4. Express constraints positively
Describe the correct physical behavior so precisely that the undesired alternative is
unnecessary. For example, replace “the pipette is not dropped” with “the fingers wrap
securely around the barrel and hold the pipette steady and level throughout the
motion.”

5. End with the final visual state
Close with a brief observable confirmation that the requested action succeeded.

Output only the generation paragraph.
```

E.5 Instruction Rendering Prompt for Level 2 Tasks

```
You are a visual narrator describing a complete short scientific laboratory
subprocedure for a video generation model. Translate the structured specification
into one flowing paragraph (150-400 words, no headings, bullets, step numbers, or
line breaks) for a target output of about 12 seconds. Preserve the scene, viewpoint,
essential action order, contacts, and visible outcomes.

0. Preserve the recorded scene and viewpoint
- Describe the viewpoint actually shown: first-person or nearby observer. Do not
convert one category into the other or claim that the camera is head-mounted unless
known.
- Anchor the prompt in the initial workspace. Mention bench color, lighting, camera
angle, and object layout only when visible.
- Preserve camera behavior consistent with the clip. Do not demand a fixed frame when
the recorded viewpoint moves naturally with the operator.
```

```
- Do not introduce cuts, unexplained viewpoint changes, or objects absent from the
observed workspace.

1. Narrate actions as a causal chain
Preserve every essential step and dependency. Connect actions with causal and
temporal phrases so that each action creates the conditions for the next. For
example, write “With the wells now empty, the operator dispenses fresh buffer” rather
than listing two disconnected steps.

2. Embed state changes within actions
Merge every action with its visible state transition. The reader should understand
what changed through the description of how it happened.

3. Express constraints positively
Specify the correct motion, contact, containment, and handling precisely rather than
stating what must not happen. For example, describe liquid being dispensed cleanly
into the target rather than instructing that it must not spill.

4. Match sentence rhythm to the procedure
- Use short clauses and specific verbs for precise actions.
- Use longer phrases for smooth, continuous motion.
- Use temporal anchors between major actions.
- End with a quiet observational description of the final state.

5. Build a spatial model
Begin from the central or foreground workspace and use concrete anchors such as “to
the left,” “behind,” and “in the foreground.” Include every object participating in
the procedure, with its material and contents only when supported.

6. Preserve technical precision
Use colors, materials, contents, quantities, and instrument identities only when
supported by the verified specification or visible evidence. Prefer a correct general
description over an unsupported precise label. Preserve hand choice, exact paths,
pauses, or incidental behavior only when needed for procedural correctness or
judgeability.

Output only the generation paragraph.
```

E.6 Task-Conditioned QA Generation Prompt

```
You are creating a task-specific QA checklist for evaluating a generated laboratory
instructional video. The generator receives the provided first frame and one text
instruction. Create the checklist BEFORE seeing any generated output and derive all
QA content only from the instruction.

Task metadata: task_id = {{TASK_ID}}; task_level = {{TASK_LEVEL}}.

Generate approximately {{MIN_QAS}} to {{MAX_QAS}} items. Recommended ranges are 6-15
for Level 1 and 10-20 for Level 2; use task complexity and required coverage rather
than padding to reach a count.

Determine the number of items from: (1) atomic actions in the instruction, (2)
visible state transitions, (3) essential ordering dependencies, (4) task-relevant
object interactions, (5) meaningful physical constraints, and (6) meaningful safety
constraints. Every dimension must receive at least one meaningful item, but do not
add weak or repetitive questions merely to reach a target.

Do not evaluate facts already guaranteed by the first frame, such as initial object
visibility, initial layout, or initial viewpoint. Evaluate temporal behavior: object
identity, actions, order, visible state changes, scene continuity, physical behavior,
and safety.

Follow these rules.

1. Atomic QA definition
Each item must test exactly ONE visually observable claim. Split independently
meaningful actions, but do not create separate questions for repetitive variants of
one underlying failure.
```

```
Bad: Does the pipette keep its shape? Does it avoid duplication? Does it avoid
disappearance?
Better: Does the pipette remain a single identifiable object throughout the video?

2. QA granularity consistency
Do not hide several actions inside one question. For a transfer, source contact,
aspiration, movement, target contact, and dispensing may require separate items when
each is essential.

Bad: Does the pipette transfer liquid correctly?
Better: Ask separately whether source contact, aspiration-like manipulation, movement
to the target, target contact, and dispensing-like manipulation are visible.

3. Temporal visual grounding
Ask only about evidence visible in the generated video. Exclude hidden intent,
invisible internal states, exact quantities that cannot be seen, and unsupported
scientific inference.

A question may refer to a local temporal event and need not describe the entire video.

4. Answer semantics
Phrase every item positively so faithful execution implies "yes." Allowed evaluator
outputs are "yes," "no," and "unjudgeable."

"Yes" means visible evidence satisfies the requirement; "no" means visible evidence
violates it; "unjudgeable" means visibility is insufficient because of occlusion,
blur, disappearance, severe artifacts, or another visibility failure. Do not use
negative wording such as "Does the pipette duplicate?"

5. Dimensions
Use only {{DIMENSIONS}} and cover every dimension with at least one meaningful,
task-specific item.

Even when an item is included for coverage, it must be derived from the specific
instruction. Do not pad a dimension with generic visual-quality questions.

6. Scene Consistency
Test temporal stability of the task-relevant environment and viewpoint across the
procedure. Check whether the camera remains steady or moves only as implied, and
whether the workspace remains continuous. Name the actual workspace and objects.
Ignore normal changes caused by hands, shadows, reflections, and intended task
motion. Do not merge camera, lighting, and background into one item.

7. Object Consistency
For every task-relevant object, test whether it remains one identifiable instance
without duplication, merging, morphing, disappearance, or confusion with a distinct
neighboring object. Name the objects from the instruction and include only properties
needed to interpret the task.

8. Action Fidelity
Follow the instruction's atomic actions. As a guideline, the number of Action
Fidelity items should be approximately the number of atomic actions plus essential
order constraints. For each action, name the concrete object, tool, hand, and motion.
Add one separate question per essential ordering dependency. Do not mix
visual-quality criteria into this dimension.

9. State Correctness
Test only visible state transitions explicitly described by the instruction, such as
a container ending filled, a tool returning to its holder, or an object reaching a
named location. Anchor each post-action state to the action that should produce it;
do not evaluate invisible states.

10. Physical Plausibility
Instantiate claims from the named objects and materials: tools remain in plausible
contact without clipping, liquids remain contained, solids do not float or pass
through other objects, and deformation matches the visible material. Phrase each
question around the specific interaction and do not combine unrelated physical
failures.

11. Visual Safety
```

```
Test only visible, task-relevant risks implied by the instruction: unintended
spillage or escape of contents, breakage or damage of equipment, and dangerous
handling of named objects. Phrase the item around the specific procedure and do not
test hidden laboratory rules.

12. Importance labels
Use only {{IMPORTANCE_LEVELS}}. Mark an item critical only when failure prevents a
reasonable observer from concluding that the requested operation was correctly
demonstrated. Typical critical items are required actions, required visible final
states, essential ordering dependencies, and identity-breaking failures of required
objects. Do not mark minor visual defects as critical.

13. Output format
Return exactly one JSON object:
{
  "task_id": "{{TASK_ID}}",
  "task_level": "{{TASK_LEVEL}}",
  "qa_items": [
    {
      "qa_id": "q01",
      "dimension": "one allowed dimension",
      "importance": "one allowed importance label",
      "question": "one positively phrased atomic question"
    }
  ]
}

Use sequential identifiers q01, q02, q03, and so on. Output JSON only.

Text instruction: {{TEXT_INSTRUCTION}}
```

E.7 VLM Evaluation Prompt

```
You are evaluating a generated laboratory instructional video using a frozen
task-specific QA checklist. Judge whether each requirement is visibly satisfied in
the generated video. Evaluate VIDEO EVIDENCE, not the intended procedure described by
the checklist.

INPUTS
Task ID: {{TASK_ID}}
Task Level: {{TASK_LEVEL}}
Frozen QA Checklist: {{QA_CHECKLIST_JSON}}
Generated Video: {{GENERATED_VIDEO}}

CORE EVALUATION RULE
Judge only what the video visibly demonstrates. Never infer an event because it is
requested, normally expected, or suggested by the final state. If decisive evidence
is absent or cannot be assessed reliably, apply the verdict rules below.

Do not fill missing events using common sense or procedural knowledge. A requirement
is satisfied only when sufficient visual evidence appears in the generated video.

ALLOWED VERDICTS
For every item return exactly one of:
- “yes”: sufficient visible evidence satisfies the requirement;
- “no”: visible evidence establishes a violation;
- “unjudgeable”: occlusion, blur, object disappearance, off-screen content,
artifacts, or insufficient temporal visibility prevents a reliable decision.
“Unjudgeable” is not “probably yes.”

Use “no” only when visible evidence supports a violation. Use “unjudgeable” when the
evidence needed to decide cannot be seen reliably; do not guess.

PRECONDITIONED / ORDERING QA RULE
Some QA items presuppose that two task events occur and ask about their relationship.
Apply these rules:
1. If all presupposed events are visible, judge “yes” or “no” from the order actually
shown.
```

```
2. If an event is definitively absent because the video clearly never shows it,
return “no” and identify the absent event; the required relationship cannot hold.
3. If occurrence cannot be determined because of occlusion, off-screen content,
artifacts, or insufficient temporal visibility, return “unjudgeable.”
Apply this rule only to QA items that explicitly relate two task events. Do not use a
missing action to fail unrelated Scene Consistency, Object Consistency, or other
items.

ATOMIC QA RULE
Judge only the requirement explicitly asked. Do not let one visible defect fail
unrelated QA items or dimensions.

Example: if a pipette remains identifiable but the hand motion is physically
implausible, the Object Consistency item remains “yes”; the physical failure belongs
to Physical Plausibility.

ACTION FIDELITY
Require visible evidence that the specified action and required contact actually
occur. Inspect the relevant video interval carefully for brief actions. Do not infer
contact from frames immediately before or after it; return “unjudgeable” when the
decisive moment is obscured. For action-order items, verify the temporal order from
the video and never assume the intended order.

STATE CORRECTNESS
Judge the visible state itself. A transition receives “yes” only when sufficient
evidence supports the required post-action state. Do not infer that state merely
because the preceding action appears correct. For example, if dispensing appears to
occur but the target is fully occluded afterward, return “unjudgeable,” not “yes.”

OBJECT CONSISTENCY
Track the specified object over time, using the first frame of the video only as the
initial reference. Do not confuse temporary occlusion with disappearance,
viewpoint-induced size change with morphing, or normal hand occlusion with
duplication. Judge only genuine visible inconsistency; return “unjudgeable” when
identity cannot be tracked reliably because of severe occlusion or artifacts.

SCENE CONSISTENCY
Use the first frame of the video only as the initial scene and viewpoint reference.
Ignore normal temporal changes caused by moving hands, tools, shadows, reflections,
and intended task motion. Judge only unexplained changes to the persistent workspace,
viewpoint, lighting, or task-relevant scene structure when those changes are the
subject of the QA.

PHYSICAL PLAUSIBILITY
Judge only clearly visible physical behavior. Relevant failures include
interpenetration, clipping through solids, unsupported floating, impossible
rigid-object deformation, impossible liquid behavior, and discontinuous physical
motion. Do not return “no” merely because motion looks unusual; require visible
evidence of the violation named by the QA.

VISUAL SAFETY
Judge only visible safety-related behavior explicitly covered by the QA. Do not infer
compliance or violation from laboratory rules that cannot be observed visually.

CONFIDENCE
Return confidence from 0.0 to 1.0 for the assigned verdict, not for whether the
action probably occurred or for laboratory knowledge. Suggested interpretation:
0.90-1.00 = clear evidence; 0.70-0.89 = good evidence with minor ambiguity; 0.50-0.69
= limited but sufficient evidence for yes/no; below 0.50 = normally prefer “
unjudgeable” unless clear contradictory evidence exists. A high-confidence “
unjudgeable” is valid when the decisive event clearly cannot be seen.

RATIONALE
Provide one concise evidence-based rationale describing only what is visibly
observed. Good: “The pipette tip visibly enters tube A and contacts the liquid.” Bad:
“The QA requires contact, so the step is satisfied.” If contact is occluded, state
that it cannot be verified. Do not invent timestamps.

INDEPENDENT ITEM EVALUATION
```

```
Evaluate every item independently and do not copy verdicts across items. Do not
reason that one failed action causes all later items to fail; a later requirement may
still be visibly satisfied. Likewise, a visually realistic video does not imply
procedural correctness.

OUTPUT FORMAT
Return exactly one JSON object using the original qa_id, dimension, and importance:
{
  "task_id": "{{TASK_ID}}",
  "task_level": "{{TASK_LEVEL}}",
  "qa_results": [
    {
      "qa_id": "q01",
      "dimension": "Action Fidelity",
      "importance": "critical",
      "verdict": "yes",
      "confidence": 0.95,
      "rationale": "The required contact is visibly shown."
    }
  ]
}

STRICT OUTPUT REQUIREMENTS
Return one result for every QA item; preserve the original order, identifier,
dimension, and importance; use only "yes," "no," or "unjudgeable"; keep confidence
within 0.0-1.0; ground each rationale in visible evidence; do not invent timestamps,
alter QA items, calculate aggregate scores, or output text outside the JSON object.
```

F LIMITATIONS AND ETHICAL CONSIDERATIONS

Dataset Scale and Coverage. Publicly available laboratory videos that satisfy our curation criteria remain limited. Most web videos contain shot transitions, subtitles, branding, or other visual overlays, whereas continuous, unedited recordings of a single laboratory procedure are comparatively rare. After screening, 81 source videos yield 204 short, safety-screened task clips, including 101 Level 1 and 103 Level 2 tasks, spanning 5 disciplines. Consequently, LABINSTRUCT represents a relatively clean and controlled subset of real-world laboratory video rather than the full diversity of scientific procedures, environments, and equipment.

The source distribution is also imbalanced. Of the 81 source videos, 61 originate from Bilibili, accounting for 163 of the 204 task clips, while the professionally recorded ExpVid subset contributes 3 videos and 14 clips. Model performance may therefore be influenced in part by platform-specific recording styles, equipment, and presentation conventions. Coverage across disciplines and operation types is similarly uneven, ranging from 73 Physics tasks to 13 Materials Science tasks, and from 60 instrument-operation tasks to 7 heating/cooling tasks. Future extensions should increase source diversity and broaden the laboratory settings, procedure lengths, equipment, operation types, and scientific domains represented in the benchmark.

Sampling Variance. The main benchmark evaluates one generated sample for each model–task pair under fixed, model-specific generation settings, and therefore characterizes performance under a standardized run rather than each model's full sampling distribution. We partially examine this limitation through a targeted best-of-three analysis on a 40-task subset balanced across the two levels for the three leading generators. The original generation remains a best-performing sample for 80–95% of the selected tasks, although additional sampling yields model-dependent gains concentrated in a minority of cases. This analysis shows that fine-grained ordering within the leading group can change under repeated sampling, but it does not characterize sampling variance for the other models or the full task set. Future work should evaluate multiple samples per task across all models to estimate complete generation distributions and ranking uncertainty.

Training-Data Contamination. Because the source videos are drawn from public platforms and existing research datasets, some benchmark content may have been included in the pretraining data of the evaluated generators. This possibility cannot be reliably audited for commercial models whose training corpora are not publicly disclosed. We therefore do not claim that LABINSTRUCT is free from training-data contamination. The paired reference videos provide evidence for evaluating

whether a generated clip matches the target procedure, but they do not guarantee that the underlying visual content or procedure was unseen during model training.

Safety and Responsible Use. Generated laboratory videos should not be interpreted as operational instructions or as evidence that a procedure is safe to perform. They are not a substitute for expert supervision, institutional training, formal risk assessment, standard operating procedures, or other laboratory safety documentation. Our visual safety gate identifies only safety-relevant issues that are observable in the generated clip and cannot determine whether a procedure is safe in practice. Although we exclude hazardous procedures from the benchmark and report safety validity explicitly, benchmark performance does not constitute certification or endorsement of any generated procedure for real-world execution without independent expert review.

G Other Details

G.1 Licenses and Terms of Use

We evaluate commercial video-generation APIs and openly released model weights. For commercial APIs, we follow the corresponding provider terms of service and use generated outputs solely for research evaluation. For open-weight models and supporting tools, we use them under their respective licenses and cite the original authors. We do not redistribute third-party model weights or proprietary API outputs except where permitted by the applicable licenses or terms of service.

G.2 Source-Video Provenance and Distribution

For each source video, we record the corresponding URL or dataset identifier, source identifier, and discipline, and preserve this provenance for all derived task clips. We additionally retain the exact temporal boundaries of each derived task clip, which identify the corresponding conditioning frame and reference execution. To respect the terms and licenses of the original platforms and datasets, we plan to only release the benchmark task specifications, generation instructions, checklist annotations, evaluation code, and provenance metadata needed to identify the corresponding source material and reconstruct benchmark instances where the original sources remain accessible, under a research-friendly license. We will also make reasonable efforts to assist researchers with source-access or benchmark-reconstruction issues.

G.3 Intended Use

LabInstruct is intended for academic research and diagnostic evaluation of procedural fidelity, visual consistency, and related reasoning capabilities in video generation models. The evaluated generators, VLM-based judges, and supporting libraries are used solely for research evaluation and data processing, and not for commercial training, individual profiling, or consequential real-world decision-making. Human annotations are used only to validate the reliability of VLM-as-judge evaluation, and are anonymized prior to analysis.

G.4 Human Annotation: Recruitment, Consent, and Ethics Review

All human annotations were produced by the paper's co-authors as part of the research process. No external participants were recruited through crowdsourcing platforms, participant pools, or other recruitment channels, and no separate compensation was provided. All annotators were informed that their judgments would support benchmark validation, including the assessment of VLM-as-judge reliability and evaluator calibration.

We did not collect names of annotators, contact information, demographic attributes, or other sensitive personal information for the purpose of this study. Annotations are used exclusively for benchmark validation and are reported only in aggregate. Because the annotation process involved only the paper's co-authors and did not recruit external human participants, we did not obtain separate institutional ethics review board approval or an exemption determination.