

Benchmarking Scientific Understanding and Reasoning for Video Generation using VideoScience-Bench

Lanxiang Hu^{*}, Abhilash Shankarampeta^{*}, Yixin Huang, Zilin Dai,
Haoyang Yu, Yujie Zhao, Haoqiang Kang, Daniel Zhao, Tajana
Rosing, and Hao Zhang

University of California San Diego
{lah003, ashankarampeta, yih119, z6dai}@ucsd.edu

Abstract. The next frontier for video generation lies in developing models capable of zero-shot reasoning, where understanding real-world scientific laws is crucial for accurate modeling of physical outcomes under diverse conditions. However, existing video benchmarks are physical commonsense-based, offering limited insight into video models’ scientific reasoning capability. We introduce VideoScience-Bench¹, a benchmark for evaluating undergraduate-level scientific understanding in video models. Each prompt encodes a composite scientific scenario that requires understanding and reasoning across multiple scientific concepts to generate the correct phenomenon. The benchmark comprises 200 evaluation items built from 160 expert-curated test cases (160 text-to-video prompts; 40 also include image conditioning) spanning 14 topics and 103 concepts. We conduct expert-annotated evaluations across seven state-of-the-art video models in T2V and I2V settings, along five dimensions: Prompt Consistency, Phenomenon Congruency, Dynamism, Immutability, and Spatio-Temporal Coherence. Using a VLM-as-a-Judge to assess video generations, we observe strong correlation with human assessments. To the best of our knowledge, VideoScience-Bench is the first benchmark to evaluate video models not only as generators but also as reasoners, requiring their generations to demonstrate scientific understanding consistent with expected physical and chemical phenomena.

Keywords: Scientific Reasoning · Video Evaluation · Auto Evaluation

1 Introduction

The launches of Sora-2 [30] and Veo-3.1 [10] mark a transformation in the video-generation landscape: video generation models [5, 25] now deliver cinematic-level visual fidelity and exhibit unprecedented world-modeling capability. As video

^{*} Equal contribution.

¹ Code and Data: <https://github.com/hao-ai-lab/VideoScience>

models evolve, the community’s notion of what constitutes “good” video generation is shifting beyond distributional quality metrics such as Fréchet Video Distance (FVD) [39] and Fréchet Video Motion Distance (FVMD) [22] toward evaluations that probe a model’s world-modeling capability [8, 12, 13, 15] and assess whether the generated videos follow commonsense physical regularities [2, 3, 19–21, 28]. Recent models have even begun to exhibit zero-shot reasoning abilities, tackling tasks involving scientific, mathematical, and spatial reasoning [38, 40].

Physical Commonsense:

A transparent tube carrying a steady stream of water pours out of the tube.



Scientific Reasoning:

A transparent tube carrying a steady stream of water is taped to the front of a speaker playing a low-frequency sound.

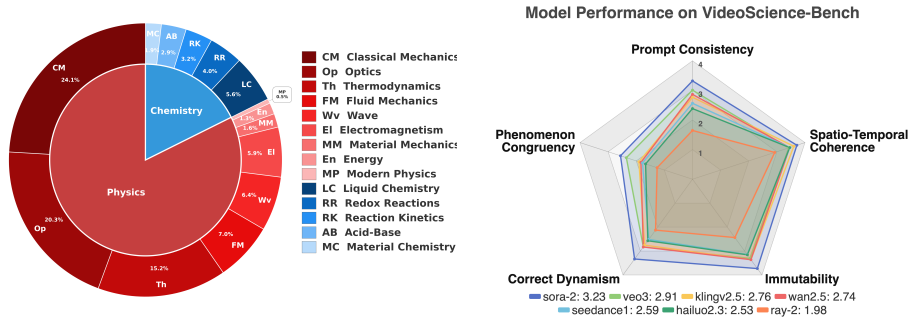


Fig. 1: VideoScience-Bench Overview. **Top:** Comparison of physical-commonsense versus scientific-reasoning generations using Sora-2. The first scientific reasoning scenario hinges on recognizing that water’s high specific heat capacity lets it act as a heat sink, thereby preventing immediate balloon rupture. The second example requires reasoning over physical vibration, wave, and material properties. **Bottom-left:** Subcategory frequency of questions From VideoScience-Bench. Note that we are using a two-letter code in the pie chart that maps the subcategory provided in the legend. **Bottom-right:** Expert annotated model performance of seven video models on VideoScience-Bench. The scores serve as the ground truth for our quantitative correlational analysis.

Human reasoners trained in science model the world using a rich set of scientific principles and anticipate the resulting phenomena, like whether citric acid removes iron stains and what will happen if an LED is powered with or without a resistor. This motivates our central question: *beyond visual fidelity and physical commonsense, can video models correctly portray scientific phenomena and demonstrate scientific reasoning about the real world?* World modeling grounded on scientific principles plays a key role in advancing discovery by allowing models to predict outcomes of real-world processes. Unlike commonsense-based world modeling, true scientific world modeling requires reasoning across *multiple interacting concepts*, for example, Figure 1 shows that integrating specific-heat and heat-transfer principles predicts water will absorb the heat and prevent immediate balloon rupture.

However, existing video model benchmarks mostly focus on physical commonsense, restricting evaluations to everyday scenarios that only require basic understanding of physical laws [2,3,20,28,35]. We introduce *VideoScience-Bench*, a benchmark designed to assess the scientific understanding and reasoning of video models. Each instance involves multiple scientific concepts and requires up to undergraduate-level knowledge, spanning 103 concepts across physics and chemistry. We evaluate whether video models can serve as zero-shot scientific reasoners, capable of understanding complex experimental setups and generating expected phenomena. The benchmark measures performance along five key dimensions: *Prompt Consistency*, *Phenomenon Congruency*, *Dynamism*, *Immutability*, and *Spatio-Temporal Coherence*.

We also introduce *VideoScience-Judge*, a checklist-based, automated VLM-as-a-Judge evaluation framework [11, 24]. VideoScience-Judge first generates a prompt-specific checklist to guide the evaluation, then selects key frames to identify causally salient moments in each video. These components jointly support the VLM-based scoring process, and the resulting scores are aggregated into a single weighted metric reflecting the model’s overall reasoning quality. For validation, we validate the automated VLM judgments significantly better align with domain-expert annotations in comparison with other physical-commonsense-based and visual fidelity benchmarks.

In summary, this paper investigates whether video generation models can reason over multiple interacting scientific concepts. To this end, we make three key contributions:

1. We introduce VideoScience-Bench, a high-quality benchmark targeting zero-shot scientific reasoning in video generation models, encompassing real-world phenomena across 103 physics and chemistry concepts.
2. We develop VideoScience-Judge, a checklist-based evaluation framework that leverages VLM-as-a-Judge, integrates key-frame selection and computer-vision tools for detecting physical cues (e.g., motion, object interactions), and outputs quantitative scores of scientific validity alongside perceptual quality.
3. We conduct a comprehensive empirical study on VideoScience-Bench. While current video models produce high-quality, photorealistic and temporally coherent videos, VideoScience-Bench remains challenging: many systems still

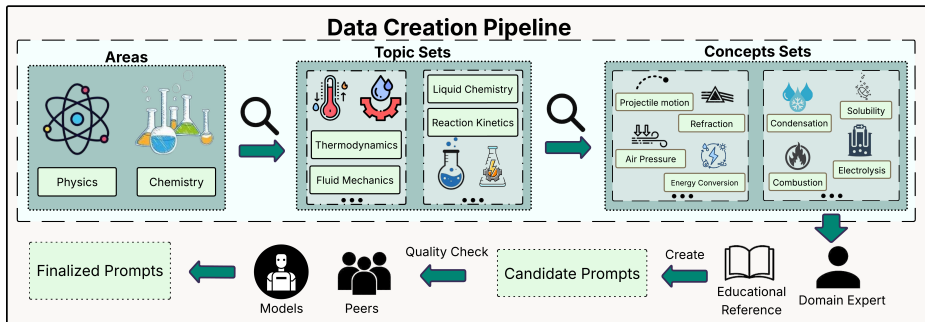


Fig. 2: Overview of the data creation pipeline. Each researcher selects two or more scientific concepts and references relevant educational materials or videos to design a prompt. Prompts undergo peer and model review, followed by model-based quality checking, before being finalized for dataset inclusion.

lack the capacity to model complex scientific phenomena and often violate basic physical laws. Among the evaluated models, Sora-2 and Veo-3 perform best and show early signs of scientific reasoning.

2 VideoScience-Bench

VideoScience-Bench is designed to evaluate video models’ capability of generating correct physical and chemical outcomes for complex scientific events that require an understanding of multiple concepts.

2.1 Benchmark Overview

Categorization. Leveraging established topic categorizations from undergraduate physics and chemistry, VideoScience-Bench comprises 160 test cases (Table 1 covering 14 topics (9 related to physics and 5 related to chemistry) and 103 concepts (with 79 associated with physics and 24 with chemistry). Example topics include physics-related areas like Classical Mechanics and Optics, as well as chemistry-related topics like Liquid Chemistry and Reaction Kinetics. The selected topics aim to reflect commonly encountered course titles in undergraduate science education, while the concepts provide a more detailed categorization corresponding to specific chapters or sections within these courses. The category-frequency distribution for all prompts is illustrated in Figure 1. See Supplementary Material for the complete topic mapping and an extended discussion of the generalizability and robustness of our dataset size relative to existing benchmarks.

Distinction with Other Scientific Benchmarks. Current scientific benchmarks for video models primarily assess physical common sense. These benchmarks typically include daily scenes such as mirror reflections, bird flight, falling water, and

Table 1: VideoScience-Bench statistics. The benchmark consists of 160 expert-authored *test cases* (each a *T2V item*); 40 of these additionally provide a first-frame image and a paraphrased prompt as an *I2V item*, for 200 total evaluation items.

	Physics	Chemistry	Total
Topics	9	5	14
Concepts	79	24	103
Test cases (T2V items)	135	25	160
of which I2V items	35	5	40
Total evaluation items	170	30	200

bouncing balls, which can be understood with a basic or high-school level of scientific knowledge [2, 3, 20, 28]. In contrast, each challenge in VideoScience-Bench requires the understanding of at least two undergraduate-level scientific principles. For instance, analyzing the bending of a laser beam as it passes through a sugar solution with a concentration gradient requires understanding and applying Snell’s Law in optics, along with principles of diffusion physics (See the Supplementary Material for the example figure). This compositional scientific-reasoning requirement defines the difficulty of VideoScience-Bench. Rather than relying on isolated single-principle recall or everyday physical commonsense, generating the correct outcome relies on synthesizing multiple interacting concepts.

Data Splits. The VideoScience-Bench consists of two parts: the T2V split, which includes 160 question prompts, and the I2V split, which features 40 paraphrased question prompts from the T2V split. The I2V split also includes first-frame images for the initial experimental setups. A detailed description of the curation process can be found in Section 2.2. We provide several examples of each split in the Supplementary Material.

Metrics. VideoScience-Bench evaluates generated videos across five dimensions that reflect both perceptual quality and physical fidelity. The perceptual aspects include *Immutability* and *Spatio-temporal Coherence*, while the physical aspects are assessed through two levels: *Phenomenon Congruency* and *Correct Dynamism*. For the assessment to be meaningful, *Prompt Consistency* is essential, which means that the experimental setup and procedure must align with the prompt. These evaluation criteria are consistent with current practices in video-generation benchmarking, and a correlational analysis is presented in Section 4.2. Each dimension is scored on a four-point ordinal scale: 1 (absent or contradictory), 2 (weak or partially incorrect), 3 (mostly correct), and 4 (clearly correct). Below, we elaborate on the rating criteria for each of the five dimensions:

- **Prompt Consistency:** The experimental setup and procedures align with the descriptions provided in the prompt.
- **Phenomenon Congruency:** The observed outcomes are consistent with expected scientific principles.

- **Correct Dynamism:** All other fundamental physical laws are followed, including those governing motion and interactions among the objects involved.
- **Immutability:** Objects remain unchanged when no transformation is anticipated.
- **Spatio-Temporal Coherence:** Transitions between frames are smooth, natural, and temporally consistent.

While focused on physical and scientific correctness, these dimensions inherently capture critical perceptual and geometric errors. For example, object deformation is penalized by Immutability, and multi-object viewpoint inconsistency is captured by Spatio-Temporal Coherence (see Supplementary Material for a detailed analysis of how these dimensions comprehensively cover perceptual and geometric qualities, including lighting and shading).

2.2 Benchmark Building

In VideoScience-Bench, we adopt a three-stage data construction pipeline as shown in Figure 2.

Question Making. Following the categorization outlined in Section 2.1, eight domain-specific graduate student experts selected subsets of categories and authored 20 prompts each. Each prompt was designed to require reasoning involving at least two scientific concepts, ensuring a sufficient level of difficulty. The prompts were based on public educational experiment examples [1, 7, 32] to guarantee scientific validity and to ensure that each target phenomenon could be feasibly represented in a video of no more than 10 seconds. Among these contributors, two graduate students also curated 40 I2V questions, each paired with a reference video that depicted the target experiment as the ground truth. The reference videos were sourced from public datasets and online resources [41, 44], with a key frame extracted from the initial segment used as the conditional input image to illustrate the complete experimental setup.

Panel Review and Quality Check. After drafting the full test suite, all participating graduate students convened as a review panel. The panel cross-checked each expert’s set of 20 prompts to verify specificity, procedural accuracy, and clarity of the expected outcomes. As a final quality assurance step, we employed capable video generation models, Sora-2 and Wan-2.5-T2V-Preview [36], to synthesize the opening seconds of each prompt, confirming that the key experimental components and setups can be captured.

3 VideoScience-Judge

VLM-as-a-Judge. Recent studies have shown that powerful VLMs can produce human-aligned judgments when carefully prompted and calibrated [11, 24]. In VideoScience-Bench, we adopt a VLM-as-a-Judge framework and examine the

performance gap between reasoning-oriented VLMs and non-reasoning models such as GPT-4o. We find that directly prompting VLMs, particularly non-reasoning ones, to evaluate video generations often leads to unjustified assessments and false-positive ratings (see Supplementary Material). This observation motivates the design of a quantitative auto-eval harness featuring a checklist-based grading scheme, where each rating must be supported by salient video frames and an evidence table aggregated from computer-vision analysis of the videos as shown in Figure 3. The full prompt template used in our VLM-as-a-Judge framework is provided in the Supplementary Material.

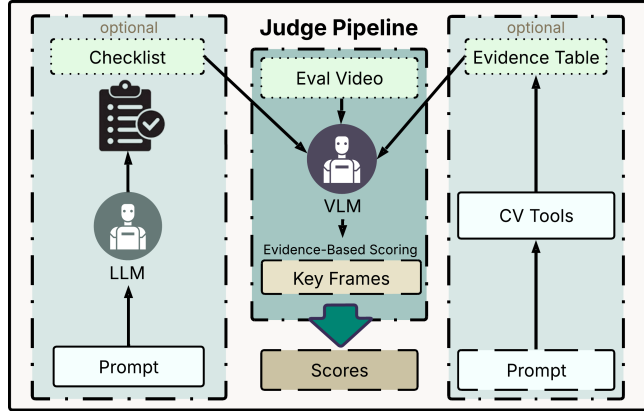


Fig. 3: Overview of our VLM-as-a-Judge pipeline, which combines checklist-based scoring, key-frame extraction, and evidence aggregation to ensure grounded and traceable video evaluations.

Checklist Generation and Key Frame Extraction. To improve interpretability and reduce false positives, we incorporate prompt-specific checklists generated by an LLM agent, and the fraction of checklist items satisfied within each evaluation category is used to derive an ordinal rating on a four-point scale. Each checklist is scored in a reductive, itemized manner, where deductions are only made when supported by concrete evidence. In VideoScience-Judge, we leverage the ground-truth reference phenomenon and reference video (in the I2V split) to guide the VLM in identifying key frames where checklist violations occur, thereby providing evidence-grounded deductions. The prompt template used for checklist generation is provided in the Supplementary Material.

To ensure the reliability of the Checklist Generation step, human validation of the generated checklists yields an accuracy of 95.6% against ground-truth phenomena (detailed in the Supplementary Material).

CV-Augmented Reasoning. VideoScience-Judge further incorporates deterministic CV modules that produce quantitative evidence for the judge. The toolbox

includes: Grounding DINO for open-vocabulary-conditioned detection to verify entity presence and attributes in experiment setups [23], ByteTrack for identity association and to quantify spatio-temporal coherence [43], RAFT optical flow to capture and quantify motion direction and magnitude [37], as well as CLIP4Clip for text-video alignment [27]. When a reference video is available, we additionally report LPIPS similarity with a VGG backbone [33, 42]. Each module outputs compact JSON records (`frame_second`, `type`, `region`, `score`, `attributes`). Depending on the evaluation setting, the prompt aggregator optionally attaches a checklist, an evidence table, both, or neither to the prompt, and instructs the model to convert the fraction of satisfied items per category into a four-point rating. Detailed qualitative case studies demonstrating how specific CV components, such as RAFT for tracking physical interactions and Grounding DINO with ByteTrack for tracing motion trajectories, ground the VLM’s judgment are provided in the Supplementary Material.

4 Experiments

4.1 Model Evaluation.

Experiment Settings. We evaluate seven state-of-the-art video models on VideoScience-Bench: Sora-2 [30], Veo-3 [10], Seedance 1.0 Pro [6, 9], Kling-v2.5-Turbo-Pro [18], Hailuo 2.3 [29], Ray2 [26], and Wan-2.5-T2V-Preview [36]. We generate three videos for each model in all of our experiments.

Human Annotation Results. We asked domain experts to rate video outputs of all the models based on the five-dimensional rubrics (defined in section 2.1). Using the human annotations, we calculated both per-criterion and aggregate model scores. This human-grounded evaluation serves as a reference for analyzing the scientific reasoning performance of video models and the effectiveness of VideoScience-Judge. It allows us to calibrate automatic VLM-as-a-Judge metrics and correlate judge predictions with expert assessments.

The performance of the annotated models is summarized in Figure 1, and the weighted average scores across the five dimensions are presented in the legend. The weights assigned are as follows: *Phenomenon Congruency* (0.3), *Prompt Consistency* (0.2), *Spatio-temporal Coherence* (0.2), *Correct Dynamism* (0.15), and *Immutability* (0.15). The weighting prioritizes the causal target of each experiment: Phenomenon Congruency carries the largest weight (0.3) as the ultimate scientific outcome; Prompt Consistency and Spatio-Temporal Coherence (0.2 each) gate whether the correct setup and a temporally valid video are produced at all; and Correct Dynamism and Immutability (0.15 each) capture supporting physical constraints.

The results from human annotations are summarized in the Supplementary Material. These results show that while most models perform well in maintaining *Spatio-temporal Coherence* and *Immutability*, qualities commonly emphasized in existing video benchmarks [11, 15, 16], many still struggle to generate physically

correct phenomena (*Phenomenon Congruency*) or adhere to other fundamental physical laws (*Correct Dynamism*). On *Phenomenon Congruency*, even the strongest closed-source systems, such as Sora-2 and Veo-3, achieve only approximately 64% and 58.7% on the Likert scale, respectively, while the strongest open-source model, Wan-2.5, reaches only around 46%. Achieving high *Prompt Consistency* also remains challenging: even the best-performing models, such as Sora-2 and Veo-3, with average scores above 3, oftentimes fail to produce accurate experimental setups and procedures, frequently fail to produce accurate experimental setups and procedures, particularly in more complex scenarios involving interactions among multiple key objects.

VideoScience-Judge Results. We evaluate different configurations of VideoScience-Judge on the same set of video generations, with results summarized in Table 2. For each method including the baselines, we compute ranking correlations with human annotations using Kendall’s τ [17] and Spearman’s ρ [34]. To enable a controlled comparison, we evaluate all automatic methods using a unified judge model (GPT-5 mini), except for LMArena-T2V, whose ELO scores are derived from human voting rather than model-based judging. Under this unified setting, the three VSci-Judge variants consistently exhibit higher alignment with human evaluations than representative alternatives such as VideoScore2 ($\tau = 0.45$, $\rho = 0.53$) and PhyGenEval ($\tau = 0.62$, $\rho = 0.82$), demonstrating their effectiveness in capturing human-judged quality across our test suite.

Table 2: We report our VLM-as-a-judge evaluation scores (using GPT-5-mini) on all baselines (except for LMArena-T2V, where ELO scores are computed based on human voting rather than judge scores) and their ranking correlations to human annotations. “VSci-Judge” refers to VideoScience-Judge, “Base” is standard evaluation, “CL” refers to Checklist, and “CV” to CV-augmented reasoning. For each model, min-max normalization is applied to annotation scores and scores from other benchmarks before averaging across all test cases and runs (except for LMArena-T2V). For each evaluation, the highest score is **bolded** and the second is underlined.

Model	VSci-Judge			T2V-CompBench	VideoScore2	PhyGenEval	LMArena-T2V
	Base	CL	CL+CV				
Sora-2	0.76	0.75	0.76	0.80	0.90	<u>0.40</u>	<u>1319</u>
Veo-3	0.67	<u>0.65</u>	<u>0.65</u>	<u>0.78</u>	0.94	0.65	1361
Kling-v2.5-Turbo-Pro	0.63	0.59	0.59	0.70	0.89	0.26	1226
Wan-2.5-T2V-Preview	<u>0.68</u>	0.60	0.59	0.80	0.94	0.31	1128
Seedance 1.0 Pro	0.63	0.56	0.54	0.74	<u>0.93</u>	0.28	1191
Hailuo 2.3	0.58	0.52	0.50	0.69	0.89	0.24	1253
Ray2	0.54	0.39	0.34	0.62	0.84	0.25	1062
$\tau(\uparrow)$	0.78	0.90	0.90	0.62	0.45	0.62	0.52
$\rho(\uparrow)$	0.86	0.96	0.96	0.71	0.53	0.82	0.71

4.2 Quantitative Analysis

VideoScience-Judge Design Choice

Model Choice. We evaluate the performance of VideoScience-Judge using both non-reasoning and reasoning models as the backbone VLM judge. As shown in the Supplementary Material, non-reasoning models such as GPT-4o exhibit substantial deviations from human annotations, with a strong tendency to assign inflated scores, leading to a high false-positive rate. Replacing GPT-4o with a reasoning model like GPT-5 pro mitigates inflation by up to 57% in *prompt consistency*, yet the discrepancy remains noticeable, leading to ranking misalignments with human judgments (detailed in the Supplementary Material).

To validate the robustness of our framework across different vendors, we also evaluated Claude-4.5-Sonnet and Gemini-3-Pro (see Supplementary Material). We found their performance to be broadly comparable to GPT-5-Pro, with Gemini-3-Pro achieving the smallest absolute gaps to human annotations overall.

Configuration Choice. Table 3 compares evaluation scores from different configurations of VideoScience-Judge against human annotations. With GPT-5 pro, the score inflation is a less pronounced issue, but notable gaps remain in dimensions such as *Dynamism*, *Coherence*, and *Immutability*, where the VLM still tends to overestimate performance. Incorporating a checklist as hint and requiring the model to reason each rating using evidence from key frames leads to a noticeable reduction in overestimation across multiple dimensions. Further integrating CV tools as evidence sources, particularly ByteTrack, which enables reliable identity association across frames and provides concrete evidence for penalizing frame jumps or unexpected object transformations. This substantially narrows the scoring gaps, by 0.07 on *Coherence* and 0.09 on *Immutability* respectively, compared with the vanilla VideoScience-Judge without the checklist.

A detailed ablation of each CV component’s contribution to the final scoring is provided in the Supplementary Material. For instance, LPIPS yields the largest calibration gain on Immutability, while RAFT is highly effective for motion-related dimensions like Correct Dynamism and Expected Phenomenon.

Table 3: Gap to human annotations (Δ) (mean $\Delta \pm$ std) for each score dimension and VSci-Judge variants using GPT-5 pro. **Larger positive Δ indicates stronger over-estimation relative to annotators.** (see human expert annotation scores in the Supplementary Material)

Dimension	w/o CL	w/ CL	w/ (CL +CV)
Prompt Consis.	-0.10 ± 0.10	-0.06 ± 0.08	-0.05 ± 0.09
Exp. Phenomenon	-0.22 ± 0.10	-0.20 ± 0.11	-0.20 ± 0.11
Dynamism	$+0.32 \pm 0.13$	$+0.27 \pm 0.10$	$+0.29 \pm 0.13$
Coherence	$+0.28 \pm 0.08$	$+0.26 \pm 0.09$	$+0.21 \pm 0.09$
Immutability	$+0.27 \pm 0.12$	$+0.24 \pm 0.14$	$+0.18 \pm 0.17$

Comparison with Baselines Detailed quantitative results and per-dimension breakdowns for all baseline frameworks, including PhyGenEval, VideoScore2, and T2V-CompBench, are provided in the Supplementary Material.

PhyGenEval. We adopt the PhyGenEval framework from PhyGenBench [28] to assess physical commonsense in VideoScience videos, employing GPT-4o as our VLM for video understanding and physical commonsense assessment, and CLIP to compute visual-semantic alignment between generated videos and reference prompts, with scores averaged across all prompts within each domain to obtain category-specific and total performance metrics (provided in the Supplementary Material). However, the significant correlation gap between VSci-Judge ($\rho = 0.96$, CL+CV) and PhyGenEval ($\rho = 0.82$) reveals fundamental limitations in the applicability of PhyGenEval to complex real-world video generation tasks. The three-tier hierarchical framework (key phenomena detection, physics order verification, overall naturalness) of PhyGenEval enforces rigid binary judgments at each stage, leading to score quantization and reduced discrimination among similar models, while its rule-based approach of checking explicit violations of discrete physical laws (e.g., "Does the ball fall downward?", "Does water flow correctly?") cannot capture holistic aspects of physical realism such as consistent motion trajectories, causal relationships across time, acceleration patterns, momentum conservation, and nuances of motion smoothness that humans naturally evaluate. These limitations are particularly pronounced for VideoScience-Bench, which differs fundamentally from PhyGenBench’s simplified, single-principle scenarios. VideoScience-Bench involves multiple scientific concepts (e.g., multi-object interactions, fluid dynamics) where outcomes arise from combinations of laws rather than single principles, featuring cascading effects where second-order dynamics matter.

VideoScore2. Although VideoScore2 [13] represents a comprehensive multi-dimensional evaluation framework covering visual quality, text-to-video alignment and physical consistency, it exhibits limited effectiveness when evaluating complex scientific reasoning in video generation models. VideoScore2’s correlations with human annotations ($\tau=0.45$, $\rho=0.53$) are noticeably lower than the proposed VSci-Judge ($\tau=0.78$, $\rho=0.86$) and its checklist-augmented variants ($\tau=0.90$, $\rho=0.96$) (detailed in the Supplementary Material). This gap can be attributed to the fact that VideoScore2’s physical consistency dimension was designed to capture broad common-sense violations, rather than performing deeper scientific reasoning or evaluating physical dynamics. Our dataset, by contrast, emphasizes non-trivial physical processes and scientific constraints. Consequently, explicit physics decomposition through domain-specific checklists yields substantially more reliable rankings than general-purpose rationales, validating the need for specialized evaluation frameworks when assessing nuanced physical reasoning in video generation. VideoScore2 is therefore not well suited for our benchmark, which requires strict, scientifically grounded physical evaluation, resulting in weaker alignment with human judgments on this dataset.

T2V-CompBench. While T2V-CompBench [35] provides a comprehensive evaluation of compositional text-to-video generation across seven categories (attribute binding, spatial relationships, motion binding, action binding, and object interactions), VSci-Judge focuses specifically on complex physics reasoning with a methodologically distinct approach. Using Llava-1.5-7B as the judge for the attribute binding category of T2V-CompBench, it achieves correlations $\tau = 0.62$ and $\rho = 0.71$. In contrast, VSci-Judge achieves substantially higher human agreement ($\tau=0.78$, $\rho=0.86$). This critical architectural difference demonstrates that T2V-CompBench’s VLM component struggles to capture nuanced physics violations, whereas VSci-Judge’s evidence-grounded VLM approach can integrate multiple signals to assess physical plausibility holistically.

4.3 Qualitative Analysis

We provide qualitative examples illustrating how different models perform. As shown in the top portion of Figure 4, Sora-2 generates the expected physical phenomenon with strong prompt consistency and accurate motion. The video follows the described setup closely, exhibits coherent temporal evolution, and captures the ignition and color dynamics characteristic of the reaction. In contrast, Hailuo-2.3, despite maintaining reasonable dynamism, immutability and spatial coherence, fails to respect key elements of the prompt, resulting in a static video where the expected phenomenon never emerges. These differences underscore that correct experimental setup and prompt grounding are essential for achieving high scores on our benchmark, even when other dimensions appear satisfactory.

In contrast to the successful cases, simpler physical scenarios still pose significant challenges for current models. The bottom portion of Figure 4 reveals that even Sora-2 and Veo-3, the two highest-ranked models in our evaluation, struggle significantly with a scenario where text models can readily infer the expected phenomenon. Sora-2 generates a setup that partially aligns with the prompt, but contains significant errors and fails to capture the expected phenomenon. Despite this, Sora-2 maintains relatively good dynamism, immutability, and spatio-temporal coherence. Veo-3 depicts the setup poorly, though it momentarily suggests the expected outward movement of the balls.

Beyond these specific examples, an extended analysis of model vulnerabilities (detailed in the Supplementary Material) reveals several systematic shortcomings in scientific reasoning. First, models consistently score higher on Chemistry than Physics (e.g., Sora-2 overall scores of 3.37 vs. 3.12), because generative priors better mimic continuous appearance changes than strict spatial geometry. Second, they suffer from a "setup vs. execution" bottleneck, accurately rendering initial setups (e.g., Sora-2 achieving 3.53 *Prompt Consistency* in Energy) but failing to simulate causal outcomes (dropping to 2.67 in *Phenomenon Congruency*). Furthermore, models present an "illusion of coherence" by prioritizing artifact-free visuals over physical accuracy, such as Kling-v2.5, achieving perfect *Coherence* (4.00) in Physics despite a dismal *Phenomenon Congruency* (1.04). Finally, failure modes follow distinct archetypes: models predominantly generate

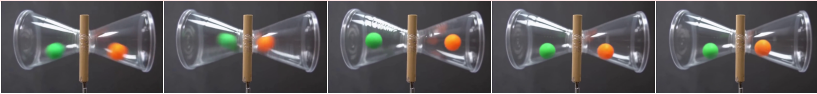
Aluminum-Iodine Reaction	
Prompt	A few <u>drops of water</u> are added to a small pile of <u>powdered aluminum</u> and <u>iodine crystals</u> in a shallow dish.
Expected	No reaction occurs before water is added. After water is added, the mixture spontaneously <u>ignites</u> with a <u>violet-blue flash</u> and smoke.
Sora-2: Successfully Generated Expected Phenomenon	
	
PCS: 4/4 PCG: 4/4 CDN: 4/4 IMB: 4/4 STC: 4/4	
Hailuo-2.3: Failed to Generate Expected Phenomenon	
	
PCS: 2/4 PCG: 1/4 CDN: 4/4 IMB: 4/4 STC: 4/4	
Rotating Cups with Balls	
Prompt	Two <u>plastic cups</u> are joined mouth-to-mouth, with a <u>wooden stick</u> fixed along the outside connecting them. One <u>small ball</u> is placed inside each cup, and the stick is <u>spun rapidly</u> around its center.
Expected	As the system spins, both <u>balls move outward</u> and <u>press against the sides</u> of their cups. They <u>remain</u> in that outward position while the rotation continues.
Sora-2: Failed to Generate Expected Phenomenon	
	
PCS: 2/4 PCG: 1/4 CDN: 3/4 IMB: 4/4 STC: 4/4	
Veo-3: Failed to Generate Correct Experiment Setup	
	
PCS: 1/4 PCG: 2/4 CDN: 1/4 IMB: 2/4 STC: 2/4	

Fig. 4: Comparison of video generation models on VideoScience. **Top:** Aluminum-iodine reaction testing chemical dynamics. Sora-2 correctly depicts the expected ignition phenomenon, while Hailuo-2.3 fails to generate the reaction. **Bottom:** Rotating cups with balls testing centrifugal force understanding. Both Sora-2 and Veo-3, the two highest-ranked models in our evaluation, fail to conduct correct experimental procedure and simulate the expected phenomenon accurately. Human Annotation Rating: Prompt Consistency (PCS), Phenomenon Congruency (PCG), Correct Dynamism (CDN), Immutability (IMB), and Spatio-Temporal Coherence (STC).

"Motion with Wrong Physics" (Sora-2 at 69%), while others default to a "Static Collapse" where the state freezes (Seedance at 34%) or suffer "Catastrophic Hallucinations" involving structural melting (Ray-2 at 53%).

4.4 Limitations

Externally verifiable correctness, not latent reasoning. VideoScience-Bench scores whether a generated video *satisfies* an externally verifiable scientific phenomenon under a controlled prompt; it does not claim to recover a model’s latent reasoning. Some chemical phenomena, characterized by continuous appearance changes (color gradients, bubbling, fizzing), can therefore be partially mimicked from learned visual priors, a property of current models rather than a flaw of the benchmark. We mitigate single-prior recall by construction: every test case couples at least two interacting scientific concepts (Section 2). Strong per-case behavior from Veo-3 and Sora-2 on physics scenarios (e.g., the curved-refraction gradient in Appendix ??, Fig. ??) indicates genuine causal competence rather than pure retrieval.

Robustness to catastrophic hallucinations. Off-the-shelf CV tools are brittle when applied naively to heavily distorted generations. In VideoScience-Judge, catastrophic failures, deformation, melting, identity loss, teleportation, and unstable geometry are penalized by *Immutability* and *Spatio-Temporal Coherence before* any phenomenon-correctness credit is assigned, so a structurally broken video cannot score highly on scientific outcome. Appendix ?? (Table ??) quantifies these archetypes (e.g., catastrophic hallucination accounts for 53% of Ray-2’s failures).

5 Related Work

5.1 Video Generation Models

Recent breakthroughs in T2V generation have been driven by large-scale diffusion and autoregressive architectures capable of synthesizing high-resolution, temporally coherent videos from prompts. State-of-the-art systems such as Sora-2 [30] and Veo-3.1 [10], demonstrate impressive visual quality and scene diversity. However, despite the rapid advances, these models often struggle with temporal consistency and physical plausibility in generated videos.

To address these challenges, many efforts have focused on establishing comprehensive evaluation suites for video generation. *Huang et al.* introduce VBench [14], which decomposes video generation quality into multiple hierarchical dimensions containing subject identity consistency, motion smoothness, temporal flicker, and spatial relationships. VBench++ [16] extends this framework by incorporating additional dimensions such as compositionality, physical realism, and instruction following, offering broader coverage of both perceptual and reasoning-oriented aspects of video generation. Other works, such as WorldModelBench [20], VideoPhy [2], and Physion [4], further emphasize the need to measure dynamic coherence, physics adherence, and material interactions.

5.2 VLM-as-a Judge for Video Model Evaluation

The use of vision-language models (VLMs) and multimodal large language models as automated judges for video generation has become increasingly prevalent. [12, 13]. Complementary to these judges, frameworks like T2V-CompBench [35] propose a hybrid evaluation pipeline that integrates both VLM-based and computer vision-based metrics for compositional video understanding. Its VLM-as-a-Judge framework employs Grid-LLaVA to reason over sampled frames combined with detection-based and tracking-based modules (like GroundingDINO, SAM, Depth-Anything, and DOT) to create a structured rubric that aligns semantic, spatial and motion dimensions.

5.3 Physical Commonsense Benchmarks for Video Models

Much of our inspiration for this work comes from past benchmarks designed to evaluate VLM’s physical reasoning capabilities. Benchmarks like PhyGen-Bench [28], VideoPhy [2], VideoPhy-2 [3], Physion [4], and IntPhys [31], all attempt to evaluate basic physical laws of video generation models, like mechanics, object interactions, and gravity.

However, there is currently a large gap in the space of video reasoning benchmarks that both require a synthesis of multiple concepts across multiple physics domains at a college undergraduate level and provide an auto-eval framework for evaluation. Our work, VideoScience, fills that need by providing complex examples that require a fundamental understanding of the physical world. Our benchmark reveals that advancement is still needed for SOTA video models to truly understand physical phenomena.

6 Conclusion

We introduce VideoScience-Bench, the first benchmark designed to evaluate scientific reasoning in video models via the generation of scientific phenomena. Unlike prior benchmarks focused on physical commonsense, VideoScience-Bench emphasizes understanding and reasoning across undergraduate-level topics and concepts in physics and chemistry. We ask expert annotators to rate video generations from the seven latest models along five dimensions. The results show that while the latest video models achieve high visual quality and temporal coherence, they still struggle with following instructions and adhering to physical laws. To further support automatic evaluation, we develop VideoScience-Judge, a checklist-based VLM-as-a-Judge framework, and requires the VLM judge to reason over evidence extracted from key frames from the video and a CV-grounded evidence table. Correlation analysis between our auto-evaluation scores and human annotations shows that VideoScience-Judge achieves the strongest alignment with expert-rated rankings and best captures a video model’s scientific reasoning capability in comparison with existing benchmarks. We hope VideoScience-Bench will foster progress toward video models that not only produce visually compelling results, but also reason faithfully about the underlying scientific principles.

References

1. Archive, E.: Experiment archive: Fun and easy science experiments. <https://www.experimentarchive.com/> (2025), collection of everyday-item experiments in biology, chemistry, physics, earth science and more.
2. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024), <https://arxiv.org/pdf/2406.03520>
3. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. arXiv preprint arXiv:2503.06800 (2025), <https://arxiv.org/pdf/2503.06800>
4. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y.F., Pramod, R.T., Holdaway, C., Tao, S., Smith, K., Sun, F.Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J.B., Yamins, D.L.K., Fan, J.E.: Physion: Evaluating physical prediction from vision in humans and machines. arXiv preprint arXiv:2106.08261 (2021), <https://arxiv.org/pdf/2106.08261>
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (Nov 2023)
6. ByteDance: Seedance 1.0 pro. Model card (2025), <https://replicate.com/bytedance/seedance-1-pro>, text-to-video and image-to-video model; accessed 2025-11-12
7. of Chemistry, R.S.: Practical videos — resource collections. <https://edu.rsc.org/resources/collections/practical-videos> (2025), high-quality chemistry practical videos supporting classroom experiments.
8. Deng, A., Yang, T., Yu, S., Spencer, L., Bansal, M., Chen, C., Yeung-Levy, S., Wang, X.: Scivideobench: Benchmarking scientific video reasoning in large multi-modal models. arXiv preprint arXiv:2510.08559 (oct 2025). <https://doi.org/10.48550/arXiv.2510.08559>, <https://arxiv.org/abs/2510.08559>
9. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., Li, X., Li, Y., Lin, S., Lin, Z., Liu, J., Liu, S., Nie, X., Qing, Z., Ren, Y., Sun, L., Tian, Z., Wang, R., Wang, S., Wei, G., Wu, G., Wu, J., Xia, R., Xiao, F., Xiao, X., Yan, J., Yang, C., Yang, J., Yang, R., Yang, T., Yang, Y., Ye, Z., Zeng, X., Zeng, Y., Zhang, H., Zhao, Y., Zheng, X., Zhu, P., Zou, J., Zuo, F.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025), <https://arxiv.org/abs/2506.09113>
10. Google, D.: Veo 3: A video generation system with audio from text or image prompt. <https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf> (2025), latent diffusion audio-video model.
11. Han, H., Li, S., Chen, J., Yuan, Y., Wu, Y., Deng, Y., Leong, C.T., Du, H., Fu, J., Li, Y., Zhang, J., Zhang, C., Li, L.j., Ni, Y.: Video-bench: Human-aligned video generation benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18858–18868 (June 2025)
12. He, X., Jiang, D., Nie, P., Liu, M., Jiang, Z., Su, M., Ma, W., Lin, J., Ye, C., Lu, Y., Wu, K., Schneider, B., Do, Q.D., Li, Z., Jia, Y., Zhang, Y., Cheng, G., Wang, H., Zhou, W., Lin, Q., Zhang, Y., Ge, Z., Huang, W., Chen, W.: Videoscore: Evaluating video generation with multi-dimensional criteria. arXiv preprint arXiv:2406.15252 (2024), <https://arxiv.org/pdf/2406.15252>

13. He, X., Jiang, D., Nie, P., Liu, M., Jiang, Z., Su, M., Ma, W., Lin, J., Ye, C., Lu, Y., Wu, K., Schneider, B., Do, Q.D., Li, Z., Jia, Y., Zhang, Y., Cheng, G., Wang, H., Zhou, W., Lin, Q., Zhang, Y., Ge, Z., Huang, W., Chen, W.: Videoscore2: Think before you score in generative video evaluation. arXiv preprint arXiv:2509.22799 (2025), <https://arxiv.org/abs/2509.22799>
14. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. arXiv preprint arXiv:2311.17982 (2023), <https://arxiv.org/pdf/2311.17982>
15. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024), https://openaccess.thecvf.com/content/CVPR2024/papers/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.pdf
16. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., Wang, Y., Chen, X., Chen, Y.C., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench++: Comprehensive and versatile benchmark suite for video generative models. arXiv preprint arXiv:2411.13503 (2024), <https://arxiv.org/pdf/2411.13503>
17. Kendall, M.G.: A new measure of rank correlation. *Biometrika* **30**(1/2), 81–93 (1938). <https://doi.org/10.2307/2332226>
18. Kuaishou Technology: Kling 2.5 turbo pro. Model announcement (9 2025), <https://blog.fal.ai/kling-2-5-turbo-pro-is-now-available-on-fal/>, accessed 2025-11-12
19. Lee, T.W., et al.: VHELM: A holistic evaluation of vision language models. In: NeurIPS Datasets and Benchmarks (2024)
20. Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J.E., Stoica, I., Han, S., Lu, Y.: Worldmodelbench: Judging video generation models as world models. arXiv preprint arXiv:2502.20694 (2025), <https://arxiv.org/pdf/2502.20694>
21. Lin, Z., et al.: Genai-bench: Evaluating and improving compositional text-to-visual generation. arXiv preprint arXiv:2406.13743 (2024)
22. Liu, J., Qu, Y., Yan, Q., Zeng, X., Wang, L., Liao, R.: Fréchet video motion distance: A metric for evaluating motion consistency in videos. arXiv preprint arXiv:2407.16124 (2024), <https://arxiv.org/abs/2407.16124>
23. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
24. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Shan, Y.: Eval-crafter: Benchmarking and evaluating large video generation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22139–22149 (2024), open-access version at CVF Open Access
25. Lu, H., Yang, G., Fei, N., Huo, Y., Lu, Z., Luo, P., Ding, M.: Vdt: General-purpose video diffusion transformers via mask modeling (2023), <https://arxiv.org/abs/2305.13311>
26. Luma AI: Ray2: Large-scale text-to-video generative model. Product page (2025), <https://lumalabs.ai/ray2>, accessed 2025-11-12

27. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021), <https://arxiv.org/abs/2104.08860>
28. Meng, F., Liao, J., Tan, X., Lu, Q., Shao, W., Zhang, K., Cheng, Y., Li, D., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In: Proceedings of the 42nd International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 267, pp. 43781–43806. PMLR (2025), <https://proceedings.mlr.press/v267/meng25c.html>
29. MiniMax: Hailuo 2.3. Official news post (10 2025), <https://www.minimax.io/news/minimax-hailuo-23>, accessed 2025-11-12
30. OpenAI: Sora 2: An advanced video generation model from openai. <https://openai.com/index/sora-2/> (Sep 2025), released September 30 2025.
31. Riochet, J., Lerer, A., Lerer, J., Poggio, T., Tenenbaum, J.B., Ullman, S., Zucker, S.: Intphys: A benchmark for visual intuitive physics reasoning. arXiv preprint arXiv:1803.07616 (2018), <https://arxiv.org/abs/1803.07616>
32. of Science, T.W.: The wonder of science: Phenomena resource library. <https://thewonderofscience.com/phenomenal> (2025), open-access list of observable scientific phenomena aligned with NGSS standards.
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
34. Spearman, C.: The proof and measurement of association between two things. *The American Journal of Psychology* **15**(1), 72–101 (1904). <https://doi.org/10.2307/1412159>
35. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. arXiv preprint arXiv:2407.14505 (2024), <https://arxiv.org/pdf/2407.14505>
36. Team, W.A.: Wan 2.5 preview: Next-generation text-to-video & image-to-video model. <https://wan2-1.com/wan2-5> or <https://blog.fal.ai/wan-2-5-preview-is-now-available-on-fal/> (Sep 2025), preview release; supports 1080p, 10-second clips, synchronized audio, text and image input.
37. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision (ECCV). pp. 402–419 (2020)
38. Tong, J., Mou, Y., Li, H., Li, M., Yang, Y., Zhang, M., Chen, Q., Liang, T., Hu, X., Zheng, Y., Chen, X., Zhao, J., Huang, X., Qiu, X.: Thinking with video: Video generation as a promising multimodal reasoning paradigm (2025), <https://arxiv.org/abs/2511.04570>
39. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018), <https://arxiv.org/abs/1812.01717>
40. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners (2025), <https://arxiv.org/abs/2509.20328>
41. Xu, Y., Wu, Y., Yu, J., Yan, Z., Jiang, T., He, Y., Zhao, Q., Chen, K., Qiao, Y., Wang, L., Okumura, M., Wang, Y.: Expvid: A benchmark for experiment video understanding & reasoning. arXiv preprint arXiv:2510.11606 (2025)
42. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

43. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European Conference on Computer Vision (ECCV). vol. 13682, pp. 1–21. Springer (2022)
44. Zou, M., Zeng, Q., Miao, Y., Liu, S., Wang, Z., Liu, H., Zhou, W.: Physlab: A benchmark dataset for multi-granularity visual parsing of physics experiments. arXiv preprint arXiv:2506.06631 (2025)