arXiv:2608.02150v1 [cs.CV] 3 Aug 2026

PhyCheck: Fine-Grained Evidence-Grounded Dataset for Physical Law Understanding in Video-LLMs

Zhongjie Ba[1,2], **Shengwang Xu**[1,2], **Peng Cheng**[1,2]*, **Jinyang Zou**[1,2], **Ting Yu**[3], **Zhibo Wang**[1,2], **Zhan Qin**[1,2]

[1] State Key Lab. of Blockchain and Data Security, Zhejiang University, Hangzhou, China
[2] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China
[3]Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
zhongjieba@zju.edu.cn, xushen9wan9@gmail.com, {peng_cheng, zou.jinyang}@zju.edu.cn,
ting.yu@mbzuai.ac.ae, {zhibowang, qinzhan}@zju.edu.cn

Source Code: https://github.com/masunozomi/PhyCheck

Abstract

Embodied intelligence and world models require video understanding systems to go beyond recognizing objects and actions and develop an understanding of physical regularities. However, despite their strong performance on general video understanding tasks, current video-language models still struggle to reliably determine whether an observed event conforms to specific physical laws. Existing benchmarks primarily assess the physical quality of generated videos, providing limited support for systematically evaluating and improving the physical-law understanding of Video Large Language Models (Video-LLMs). To address this gap, we introduce **PhyCheck**, a video question answering dataset organized at two complementary levels of granularity. The coarse-grained subset asks models to determine whether the phenomenon shown in a video conforms to or violates physical laws, while the fine-grained subset further examines whether models can capture physical details responsible for the violation or compliance. We use these subsets as structured supervision to improve physical understanding. In addition, the dataset contains a diagnostic subset with external causal context that reveal hidden factors affecting physical plausibility, assessing whether models can recalibrate their judgments accordingly. Experiments with Fine-tune Qwen2.5-VL show that training with the proposed data substantially improves the understanding of physical-consistency, while evaluations in the diagnostic subset reveal that current models still have difficulty incorporating additional causal conditions into their decisions. These findings highlight the gap between recognizing surface-level inconsistencies and understanding underlying physical mechanisms, and provide a foundation for evaluating and improving physical understanding in Video-LLMs.

1. Introduction

Building World Models is treated as a path towards achieving artificial general intelligence (AGI) LeCun [2022]. World Models are AI systems that understand and reason about the physical working mechanisms of the real world Ding et al. [2025]. Video modality has naturally become the primary testbed for exploring World Models, since this is the unique modality encompassing spatiotemporal dynamics reflecting physical principles, catalyzing the rapid proliferation of video understanding models, particularly the recent emergence of highly capable Video Large Language Models (Video-LLMs).

Existing Video-LLMs have exhibited strong commonsense reasoning and planning capabilities Chow et al. [2025]; Yue et al. [2024]; Lu et al. [2024]; Kim et al. [2024]; Niu et al. [2024]; Zhen et al. [2024], but they have been found lacking the understanding of the physical world Chow et al. [2025].

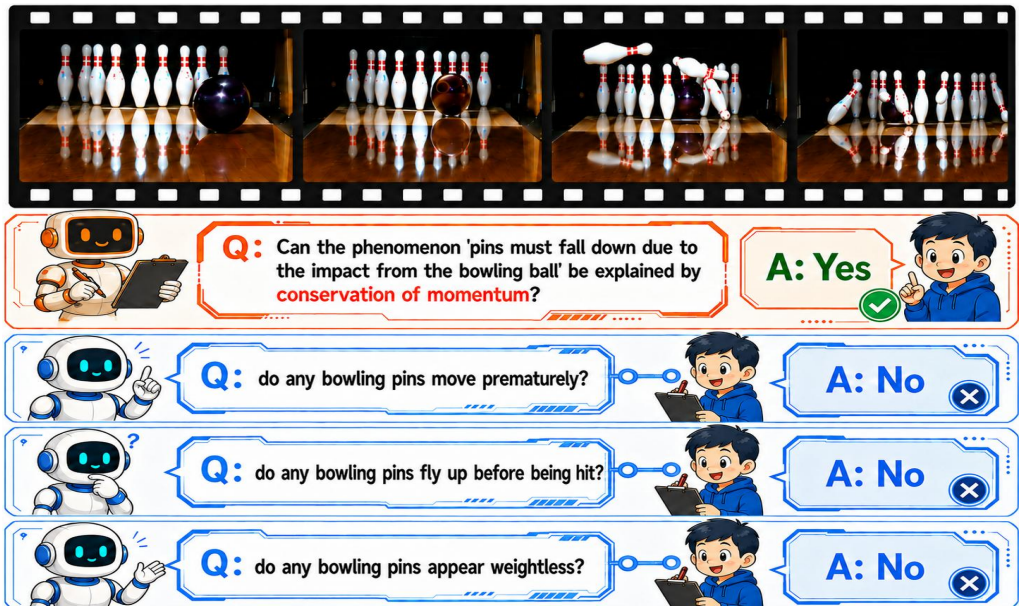

Figure 1: Example Annotations from PhyCheck

It is widely argued that current Video-LLMs fundamentally act as a sophisticated pattern recognizers driven by spurious statistical correlations, rather than mastering the causal relationships of underlying physical laws Motamed et al. [2026]–These models learn shortcuts without understanding.

This deficiency in physical understanding not only impedes the transition of these models to World Models capable of internalizing and simulating reality, but also limits their deployment in Embodied AI scenarios where safe and reliable interactions are critical Chow et al. [2025]; Wang et al. [2023]; Liu et al. [2024]; Guo et al. [2024]. For embodied AI to operate safely and reliably, and for current video understanding models to evolve into world models, mastering underlying physical laws (such as conservation, optical and mechanics laws) serves as an indispensable cornerstone for current video models.

Although recent studies have begun to recognize the importance of physical world understanding, the field still warrants deeper and more multifaceted exploration. Existing works primarily focus on evaluating current Video-LLMs' performance in terms of understanding physical world or interpreting physical commonsense. PhysGame proposed a benchmark to evaluate physical commonsense violations in gameplay videos Cao et al. [2024]. PhysBench pointed out that VLMs' ability to comprehend physical phenomena remains limited Chow et al. [2025], and they benchmark numerous VLMs to show their deficiencies in understanding the physical world, likely due to the absence of physical knowledge in training data. They further incorporated vision foundation models and a physics knowledge memory. There exist works studying other perspectives of video understanding (e.g., MVPBench evaluates complex interaction capability across multiple videos Bai et al. [2026]).

Although recent studies have begun to recognize the importance of physical world understanding, the field still warrants deeper and more multifaceted exploration. In this paper, we propose **PhyCheck**, a fine-grained, evidence-grounded video dataset designed for physical law understanding in Video Large Language Models (Video-LLMs). Unlike existing datasets that rely on coarse-grained overall descriptions or isolated judgments of physical common sense, PhyCheck systematically decomposes

complex physical laws into explicit visual evidence clues—such as identifying spontaneous object movement or unexpected physical deformation—backed by rigorous human annotations (see Figure 1 for an example). To ensure maximum reliability, we constructed a large-scale collection of approximately 50k human-verified VQA pairs. Detailed descriptions of our data construction, including video collection, annotation expansion, and rigorous human verification protocol, are provided in PhyCheck Dataset section.

In particular, PhyCheck bridges key gaps in current physical-understanding benchmarks through three distinctive design principles:

(1) Counter-Physical Anomaly Focus: Conventional datasets often rely on real-world videos depicting standard physical scenarios. The absence of counter-physical data restricts systematic evaluation and improvement of a model's ability to differentiate between physically consistent and inconsistent dynamics. When a dataset contains only physically plausible videos, models may attain strong performance by exploiting statistical biases and superficial correlations in the data, rather than learning the underlying physical principles that govern real-world interactions. As a result, apparent performance gains may reflect shortcut learning or memorization instead of genuine physical understanding. PhyCheck emphasizes explicit physical law violations. By strategically combining synthetic anomaly videos—built upon VideoPhy2 Bansal et al. [2025]—with curated real-world clips, our dataset provides a balanced set of positive and negative samples. This setup not only rigorously evaluates Video-LLMs on counter-physical edge cases, but also offers contrastive perspectives to effectively fine-tune and align model capabilities.

(2) Multi-Question Evidence-Grounded Decomposition: Instead of pairing each video with an isolated, high-level query (e.g., asking "*which car has a higher average speed?*" given a video of moving toy cars), PhyCheck equips every video sequence with a structured suite of multi-grained questions. We assume coarse-grained annotations are insufficient for capturing fine-grained physical phenomena and rule-specific inconsistencies. This lack of granularity constrains the development and assessment of models' physical reasoning capabilities. To address these limitations, we construct a hierarchical VQA-based annotation system. There is a major question querying the underlying physical principle related to the video content, followed by a suite of fine-grained questions probing explicit visual evidence. These question-and-answer pairs annotates both physically consistent and physically violating events at the level of specific physical principles—such as spontaneous motion, unnatural acceleration, or unexpected structural deformation- which forms an evidence-grounded reasoning chain before reaching a physical verdict. This multi-question evidence-grounded annotation structure enables more precise supervision and more rigorous evaluation of Video-LLMs' physical understanding capabilities.

(3) Context-Sensitive Physical Evaluation: During the construction of PhyCheck, we realized that physical validity is often context-dependent. For instance, a floating ball appears physically impossible in isolation, yet becomes entirely plausible when contextual cues—such as an active hairdryer underneath or a zero-gravity environment—are present. PhyCheck incorporates context-aware video pairs to uniquely assess whether Video-LLMs can dynamically adjust their physical logic based on environmental cues.

Based on PhyCheck, we systematically evaluate numerous Video-LLMs covering both open-source and commercial models across different sizes. Our experiments reveal that existing video-LLMs generally exhibit superficial physical understanding when encountering physically violating videos. Furthermore, we propose a reasoning-chain-based fine-tuning approach to enhance the physical comprehension capability of Qwen2.5-VL Bai et al. [2025].

Our main contributions are summarized as follows:

1. We introduce **PhyCheck**, a fine-grained, evidence-grounded video dataset specifically engineered for physical law understanding. To the best of our knowledge, **PhyCheck** is the first video QA dataset that explicitly probes both compliance with and violation of physical laws, forcing Video-LLMs to transcend superficial visual priors and evaluate true physical causality.
2. We design a multi-tier VQA annotation system that serves a dual purpose: benchmarking and model alignment. Each instance pairs a video clip with a high-level primary question evaluating physical law compliance, alongside a suite of fine-grained questions on explicit visual evidence. While the primary questions rigorously diagnose model capabilities, the fine-grained evidence chains serve as structured reasoning scaffolding that enables supervised fine-tuning for enhanced physical comprehension.
3. We construct a specialized context-sensitive subset to evaluate whether Video-LLMs can dynamically update their physical law judgments upon integrating environmental context. This suite goes beyond direct, static consistency recognition to assess higher-level contextual physics reasoning.
4. Through extensive evaluation across a diverse spectrum of open-source, proprietary, and specialized Video-LLMs, we reveal severe systemic limitations in current models' physical intuition. Specifically, while current Video-LLMs demonstrate strong performance on standard positive samples, their performance drops sharply on negative violation samples—highlighting an acute vulnerability to counter-physical anomalies.

2. Related Work

2.1. Video LLMs

Video LLMs extend vision-language models to video inputs by integrating temporal visual information with language reasoning. Recent studies have advanced video LLMs from multiple perspectives. Some works focus on enhancing video understanding capabilities: General-purpose video LLMs, Qwen and Gemini, provide broad multimodal instruction following and enhance complex reasoning over long and diverse video contexts. Other studies explore improved video modeling strategies, ranging from the compact spatiotemporal representations of LLaVA-OneVision2 An et al. [2026], which better preserve motion cues, to unified frameworks such as UniVid Yang et al. [2026] and UniVideo Wei et al. [2026], which integrate video understanding and generation within a shared architecture. Despite these advances, existing video LLMs mainly target semantic and temporal comprehension, whereas physical understanding requires models to go beyond observed events to infer hidden mechanisms and assess their consistency with physical laws.

2.2. Benchmarks for Evaluating Physical Understanding in Video-LLMs

To investigate whether Video LLMs can capture physical knowledge, several benchmarks have been proposed to evaluate physical reasoning capabilities from different perspectives. PhysBench Chow et al. [2025] provides a comprehensive evaluation of Vision-Language Models' understanding of the physical world, covering object properties, relationships, scene-level reasoning, and physics-based dynamics. However, it mainly focuses on general physical knowledge rather than identifying violations of physical laws in dynamic scenarios. PhysGame Cao et al. [2024] investigates whether video-language models can recognize physical commonsense violations in gameplay videos, such as abnormal motion and inconsistent interactions, but its evaluation is limited to game environments and lacks analysis of the underlying physical principles. MVP Krojer et al. [2025] proposes a shortcut-aware evaluation protocol based on minimal video pairs, aiming to distinguish genuine physical reasoning from reliance on superficial visual correlations. In short, existing benchmarks primarily provide coarse-grained evaluations of physical consistency, while fine-grained understanding of which physical laws are

Category	Representative laws
Mechanics Laws	Gravity, friction, inertia, and Newton's first/second/third law
Conservation Laws	Conservation of momentum, conservation of mass, conservation of angular momentum, and conservation of energy
Material Properties	Elasticity and hardness
Fluids and Interface Phenomena	Buoyancy, surface tension, and fluid dynamics
Optical Laws	Reflection
Thermal Laws	Melting and flame reactions

Table 1: Physical categories and representative laws.

violated and what visual evidence supports such judgments remains underexplored.

2.3. Video Generation Models and Physical Consistency Evaluation

Recent breakthroughs in generative architectures have driven the rapid evolution of video generation models (e.g., Sora Brooks et al. [2024], Cosmos NVIDIA et al. [2025], and Wan Wan et al. [2025]), enabling the synthesis of highly realistic and temporally coherent dynamic scenes.

As video synthesis advances toward simulating real-world dynamics, evaluating whether generated videos respect physical reality has become a critical research topic. Benchmarks such as VideoPhy Bansal et al. [2024] and VideoPhy2 Bansal et al. [2025] investigate whether video generation models can produce physically plausible events across diverse real-world scenarios. Concurrently, evaluation frameworks like VBench2 Zheng et al. [2025] and PhyGenBench Meng et al. [2024] systematically incorporate physics-related metrics to benchmark video generation quality.

While these efforts focus on the *synthesis side* by assessing whether generative models can render visually and physically plausible outcomes, our work targets a fundamentally distinct and complementary objective on the *comprehension side*. Rather than evaluating the perceptual realism of synthesized frames, **PhyCheck** investigates whether Video-LLMs can acquire an intrinsic understanding of the underlying physical mechanisms and explicitly reason about the physical laws governing both valid and anomalous visual outcomes.

3. PhyCheck Dataset

3.1. Overview

PhyCheck is a physics-laws-oriented video question answering dataset designed to evaluate and improve the ability of video understanding models to judge **whether observed phenomena comply with special physical laws**. It comprises **6,399** synthetic videos and contains **69,825** question–answer pairs, with an average of approximately **10.9** questions per video. Each question–answer pair is associated with one of six major categories of physical principles: (1) Mechanics Laws, (2) Conservation Laws, (3) Material Properties, (4) Fluids and Interface Phenomena, (5) Optical Laws, and (6) Thermal Laws. Notably, compared to the existing PhysGame and PhysBench dataset, all questions are presented in a binary yes/no format rather than as four-way multiple-choice questions, requiring models to directly determine whether a specific physical description is consistent with the observed video content.

3.2. Dataset Construction Framework

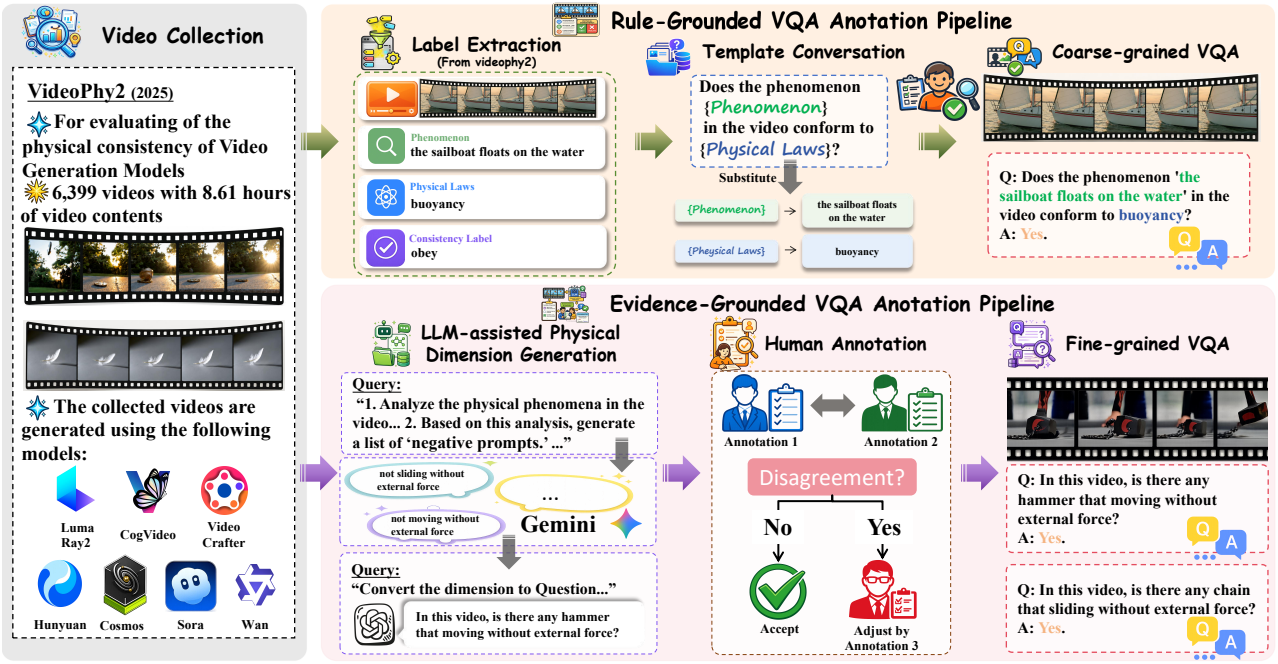

Figure 2: Overview of the PhyCheck dataset construction framework.

Data Source. We collect synthetic video samples from VideoPhy2, a large-scale benchmark for evaluating physical consistency in generated videos. Specifically, we reuse the videos from the training and test splits of VideoPhy2, resulting in a total of 6,399 videos with 8.61 hours of video content. These videos are generated by seven representative text-to-video generation models, including (1) Luma Ray2 LumaAI [2025], (2) CogVideo Yang et al. [2025], (3) Cosmos NVIDIA et al. [2025], (4) Hunyuan Kong et al. [2025], (5) VideoCrafter Chen et al. [2024], (6) Sora Brooks et al. [2024], and (7) Wan Wan et al. [2025].

While VideoPhy2 released these raw video files originally created for benchmarking video generation models, we substantially extend their utility for physical comprehension assessment and improvement. Instead of relying on their high-level generation-side video summary descriptions, we re-annotate and restructure these videos into a fine-grained, evidence-grounded VQA dataset through rigorous multi-tier prompt engineering and 556 working hours of intensive human verification.

Rule-Grounded Physical VQA Construction Pipeline. To construct VQA pairs that determine whether a video phenomenon conforms to or violates physical rules, we develop a rule-grounded physical VQA construction pipeline, as shown in Figure 2.We first leverage the physical rule annotations provided by VideoPhy2, where each synthetic video is associated with candidate physical rules and corresponding labels indicating whether the observed phenomenon is followed, violated, or undetermined. Since ambiguous cases provide unreliable supervision, we retain only the followed and violated samples and discard the undetermined ones.

Rather than directly transforming the original rule descriptions into questions, we adopt a template-based generation strategy to ensure consistent question formulation and balanced supervision. Specifically, we design two types of rule-aware templates: (1) consistency templates, which ask whether the

observed phenomenon conforms to a given physical rule, and (2) violation templates, which query whether the phenomenon violates the corresponding rule. The answers are deterministically assigned according to the original VideoPhy2 annotations. For consistency templates, followed cases are assigned "Yes" while violated cases are assigned "No". For violation templates, the answer mapping is reversed. This design balances question polarity and reduces potential biases caused by fixed answer patterns, while preserving the explicit correspondence between each question and its underlying physical law.

Each sample consists of a video, a rule-oriented question, and a binary answer, with the associated physical law retained as metadata for fine-grained analysis. When a video contains multiple valid physical rule annotations, we generate multiple VQA instances, each corresponding to a specific phenomenon–rule pair. Finally, all questions are manually reviewed to ensure that they are visually grounded, unambiguous, and answerable based on the provided video content.

Evidence-Grounded VQA Annotation Pipeline. Our fine-grained annotation pipeline aims to identify interpretable physical details that contribute to the violation or compliance of physical laws, thereby providing more informative supervision for improving VLMs' physical reasoning ability. Different from coarse-grained physical consistency labels that only indicate whether a phenomenon follows or violates a physical rule, fine-grained annotations focus on the specific visual and physical cues underlying the decision. The annotation process combines LLM-assisted physical reasoning with human verification, where large multimodal models generate candidate physical details, followed by manual annotation to ensure reliability.

Specifically, we first employ Gemini-3 to analyze each video and generate detailed descriptions of the observed physical phenomena. Given the video description and the associated physical law labels from VideoPhy2, Gemini-3 is further instructed to act as an expert in video generation and analyze potential physical inconsistencies. Based on this analysis, the model produces negative prompts that describe specific physical details which should be avoided when generating physically correct videos. The generated negative prompts which serve as interpretable hypotheses of possible violations are then provided to GPT-5 for refinement. GPT-5 transforms each physical inconsistency description into a clear and answerable question format, asking whether a specific physical detail exists in the video. This process converts implicit physical reasoning cues into structured fine-grained VQA annotations while maintaining their connection to the corresponding physical laws.

Finally, all generated questions and candidate physical details undergo human verification. Each video-question pair is independently reviewed by two annotators to determine whether the queried detail is supported by the video. Samples with agreement are directly labeled, while disagreements are resolved by a third annotator through majority voting. Annotators also check whether the question accurately captures the underlying physical phenomenon and provides meaningful evidence for physical consistency or violation. To evaluate the reliability of human verification, we measure the agreement between the two initial annotators. They achieve a raw agreement rate of 99.74% and a Cohen's κ coefficient of 0.9959, indicating almost perfect inter-annotator agreement. The resulting dataset contains video-level binary judgments together with fine-grained physical detail annotations, enabling both physical consistency evaluation and detailed analysis of Video-LLM behaviors.

3.3. Context-Sensitive Physical Evaluation (pilot study)

Beyond universal physical compliance, real-world physical validity can be context-dependent due to partial observations. To probe whether Video-LLMs can dynamically update their judgments given external causal context, we curate a small-scale pilot subset of 50 hand-crafted pairs (kept compact

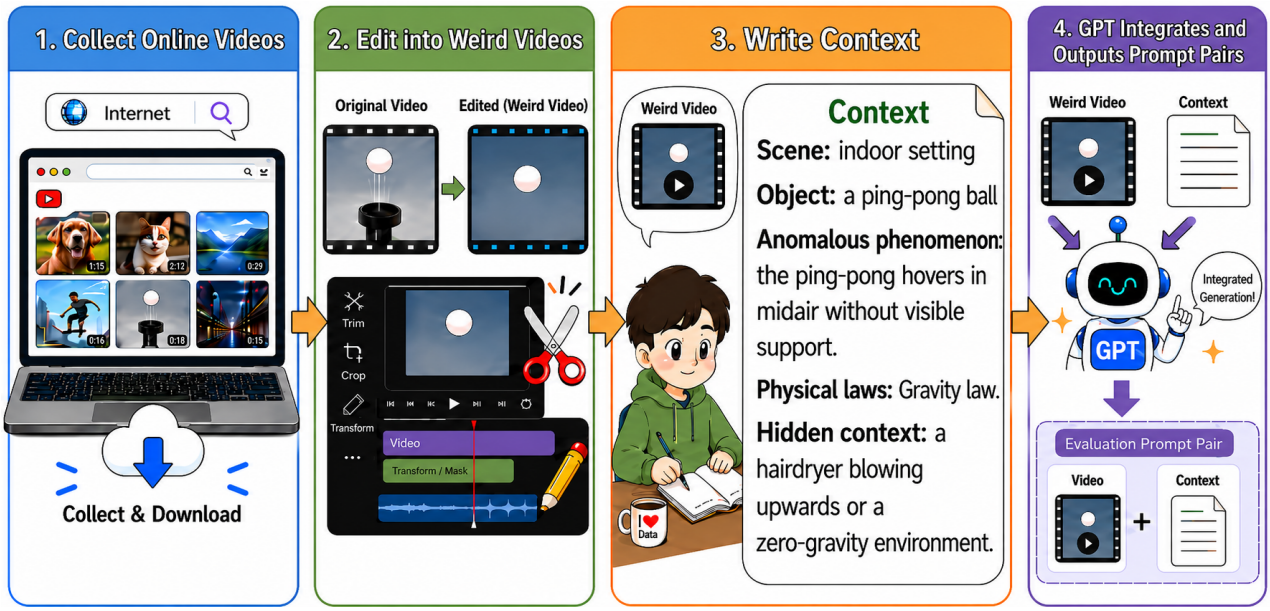

Figure 3: Overview of the context-sensitive dataset construction framework.

due to intensive manual editing, as primary efforts were dedicated to the main PhyCheck dataset). Specifically, we select real-world plausible videos and crop out the visible causal factors (e.g., removing a hairdryer to leave a floating ball), rendering the cropped clip seemingly implausible. We then supplement each clip with a manually written context describing the cropped part, namely the hidden cause (e.g., upward airflow). Models are evaluated on whether they can recognize the implausibility of the isolated visual effect alone and update their judgments once provided with external causal context (see Figure 3).

4. Model Evaluation and Enhancement

4.1. Settings

Evaluated Models. Our evaluation covers several representative open-source and proprietary Video-LLMs, comprising the Qwen series, Gemini-3.5-Flash, LLaVA-OneVision2, UniVid, and UniVideo, spanning diverse video understanding paradigms. UniVid and UniVideo are unified video models that reuse existing video understanding backbones, based on BAGEL-7B-MoT Deng et al. [2025] and Qwen2.5-VL, respectively. Additionally, Skyra Li et al. [2026] is specifically designed for AI-generated video detection through grounded artifact reasoning. It identifies and localizes physics-related inconsistencies caused by violations of real-world physical principles, providing a specialized baseline for physical consistency understanding.

Evaluation Setup. **(1) Direct Evaluations.** We first directly evaluate existing Video-LLMs on the PhyCheck test dataset to investigate whether current models can verify physics-related statements grounded in video content. **(2) Supervised Fine-Tuning Utility.** To validate the training utility of PhyCheck, we fine-tune Qwen2.5-VL on the training split and compare its performance with existing baselines under the same evaluation protocol. *Exploratory Pilot Probe:* In addition, we conduct a lightweight pilot probe on our 50-pair context-sensitive subset. Using a three-setting evaluation protocol (video-only, context-assisted, and joint accuracy), we evaluate whether models can dynamically adjust

Model	Params	Overall	Conform Example			Voilate Example		
		Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	Precision ↑	Recall ↑	F1 ↑
Qwen3-Instruct	8B	62.82	67.41	84.68	75.06	40.41	20.24	26.97
Qwen3-Thinking	8B	60.28	66.64	79.91	72.67	36.03	22.05	27.36
Qwen3.5	9B	59.42	65.91	79.92	72.24	33.23	19.46	24.55
Qwen3.5	27B	58.81	65.71	78.80	71.66	32.48	19.87	24.66
Qwen3.6	27B	57.77	65.65	75.70	70.32	32.53	22.83	26.83
LlaVa-OneVision2	8B	**63.11**	66.68	**88.28**	**75.98**	38.12	14.07	20.55
UniVid	7B	59.79	66.18	80.05	72.46	34.32	20.31	25.51
UniVideo	7B	50.93	65.73	53.78	59.16	33.51	**45.38**	**38.55**
Gemini-3.5-Flash	-	62.24	67.05	84.29	74.69	38.65	19.28	25.73
Skyra-SFT	7B	61.57	**68.47**	77.59	72.74	**41.03**	30.38	34.91
Skyra-RL	7B	61.31	68.38	77.11	72.48	40.64	30.54	34.87

Table 2: Overall Physical Understanding Evaluation results of current Video-LLMs on PhyCheck.

Coarse-grained VQA	Fine-grained VQA	Overall	Conform Example			Violate Example		
		Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	Precision ↑	Recall ↑	F1 ↑
		50.93	65.73	53.78	59.16	33.51	45.38	38.55
	✓	51.43	65.49	56.00	60.38	33.15	42.51	37.26
✓		66.83	77.02	70.99	73.88	50.96	**58.73**	54.57
✓ ①	✓ ②	51.89	66.67	54.39	59.91	34.61	47.04	39.88
✓ ②	✓ ①	**81.00**	**80.68**	**93.68**	**86.69**	**82.05**	56.30	**66.78**

Table 3: Dataset Utility Validation of PhyCheck with different supervision settings.

their physical judgments upon integrating external causal context, rather than relying on rigid visual shortcuts.

Metrics. For the main evaluation on PhyCheck, We report overall accuracy, class-wise precision, recall, and F1 scores for both Conform and Violate examples. These metrics provide a comprehensive evaluation of models' ability to verify physical statements and reveal potential prediction biases toward either category.

Implementation Details. We adopt Qwen2.5-VL as the base model for its open-source availability and broad adoption in video understanding studies. It is fully fine-tuned on the PhyCheck training set for 12 epochs. All experiments are conducted using PyTorch on NVIDIA A100 GPUs.

4.2. Overall Physical Understanding of Video-LLMs

As shown in Table 2, existing models achieve moderate performance on PhyCheck, with most models obtaining accuracies around 60%, indicating that current Video-LLMs struggle with reliable physical-law judgment. Beyond overall accuracy, we analyze class-specific metrics for Conform and Violate examples. Most models exhibit higher recall on Conform examples than Violate examples, indicating a conservative

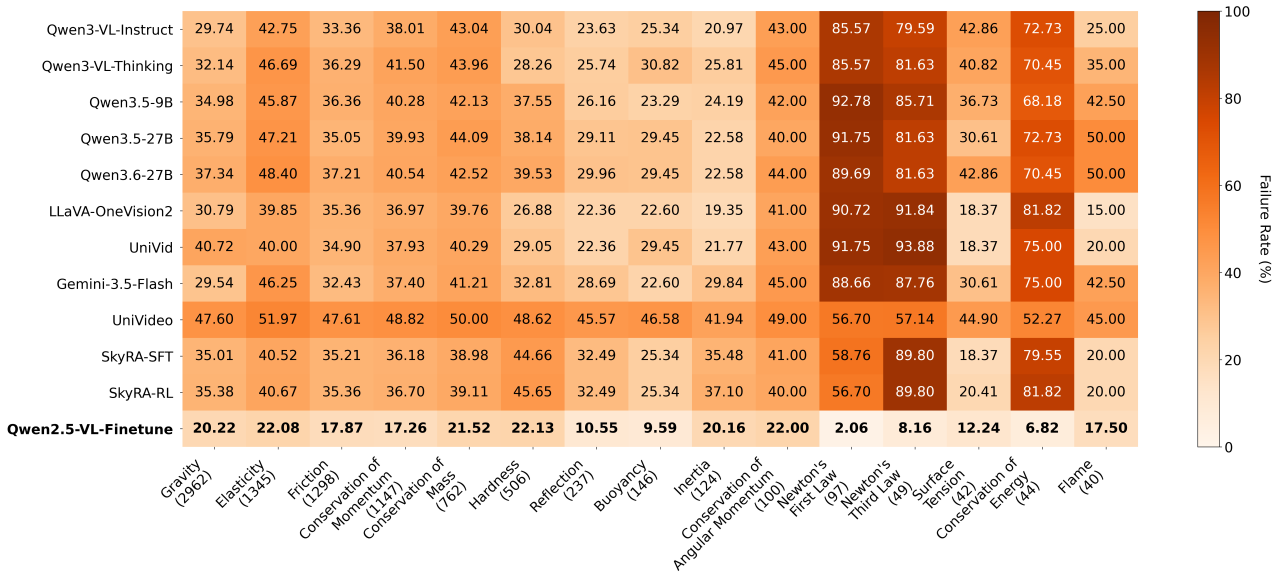

Figure 4: Failure Rates across Physical Law Categories (Top 15). Parentheses indicate sample counts.

prediction tendency when verifying physics-related statements. Skyra achieves competitive performance despite being primarily designed for AI-generated video detection through artifact reasoning. However, its performance does not consistently surpass general-purpose Video-LLMs, indicating that detecting physics-related artifacts is not equivalent to understanding underlying physical laws.

4.3. Dataset Utility Validation

To investigate the contribution of different supervision components in PhyCheck, we conduct a dataset utility analysis by fine-tuning Qwen2.5-VL with different subsets and training strategies, including coarse-grained VQA, fine-grained VQA, and their sequential combinations. As shown in Table 3, different supervision settings lead to distinct performance variations.

Training with fine-grained VQA alone provides only marginal improvement over the original model, with accuracy increasing from 50.93% to 51.43%. This suggests that fine-grained questions alone may not provide sufficient supervision for establishing fundamental physical consistency understanding. In contrast, training with coarse-grained VQA achieves a substantial improvement, increasing accuracy to 66.83%, highlighting the effectiveness of coarse-grained physical consistency supervision in providing foundational signals for physical understanding. Furthermore, we investigate the impact of combining different supervision components through sequential training. Training with coarse-grained VQA followed by fine-grained VQA results in limited improvement, whereas reversing the training order achieves the best performance of 81.00%. These results indicate that learning fine-grained physical consistency before coarse-grained physical reasoning provides a more effective learning trajectory, where fine-grained supervision establishes the learned physical concepts and coarse-grained supervision further enhances physical reasoning capability. Overall, these results validate the effectiveness of the hierarchical supervision design in PhyCheck, demonstrating that different annotation granularities provide complementary signals for improving Video-LLMs' physical law understanding.

Model	VideoCrafter 1984	Wan 1893	Cosmos 1889	CogVideo 1866	Ray2 1237	Hunyuan 403	Sora 180
Qwen3-Instruct	43.40	31.59	37.37	33.39	36.22	52.36	37.22
Qwen3-Thinking	45.26	36.29	39.07	35.42	39.37	51.36	42.22
Qwen3.5-9B	45.46	35.71	40.18	37.30	40.99	55.83	39.44
Qwen3.5-27B	46.77	35.97	40.50	38.00	42.12	54.84	37.78
Qwen3.6-27B	46.52	38.14	41.34	39.92	42.60	55.33	39.44
LlaVa-OneVision2	42.99	32.07	36.42	31.73	37.11	56.08	34.44
UniVid	45.51	36.61	39.44	36.07	39.69	53.10	45.56
UniVideo	48.84	49.02	49.07	49.20	48.75	51.36	47.78
Gemini-3.5-Flash	43.20	32.59	38.06	34.03	37.35	54.09	34.44
Skyra-SFT	41.99	35.97	39.92	33.28	39.13	45.66	41.67
Skyra-RL	42.04	36.24	40.02	33.76	39.45	46.15	42.78
Qwen2.5-Finetune	**23.74**	**16.22**	**18.95**	**17.42**	**19.32**	**15.14**	**19.44**

Table 4: Failure Rates by Generation Category. Subscripts denote evaluated video counts.

4.4. Category-wise Failure Analysis

Physical Law Categories: As shown in Figure 4, the prediction failure rates vary substantially across different physical phenomena, suggesting that current Video-LLMs exhibit uneven capabilities across physical concepts rather than a unified understanding of physical laws. Off-the-shelf Video-LLMs generally achieve lower failure rates on visually intuitive phenomena, such as Inertia, Reflection, and Buoyancy, where relevant physical cues are directly observable from object motions and interactions. In contrast, higher failure rates are observed in categories requiring more implicit physical reasoning, including Conservation laws and Newtonian mechanics. These categories require models to infer latent physical states and reason about abstract relationships that cannot be directly captured from visual appearances alone. After fine-tuning on PhyCheck, Qwen2.5-VL substantially reduces failure rates across all physical categories, with particularly notable improvements on Conservation laws and Newtonian mechanics.

Video Generation Sources: As shown in Table 4, Video-LLMs show relatively consistent failure patterns across different video generation sources, indicating that the generation model is not the primary factor affecting physical law understanding. These results suggest that current failures mainly originate from insufficient physical reasoning capabilities rather than source-specific visual characteristics.

4.5. Context-assisted physical Understanding.

As an initial exploratory probe, we evaluate the model behaviors on our 50-pair context-sensitive subset (refer to results in Appendix C3). Baselines achieve high accuracy under the video-only setting (e.g., 0.96 for Qwen3.5-27B), many struggle to update their predictions after context is introduced. This trend suggests a potential over-reliance on rigid visual shortcuts, where models tend to prematurely treat partial visual anomalies as definitive physical violations.

In contrast, our fine-tuned **Qwen2.5-Finetune** exhibits a distinct shift. Under the video-only setting, it shows a lower performance (0.18), which may reflect a form of epistemic caution. Having been trained on diverse physical rules, the model appears less inclined to strictly penalize visual observations. Remarkably, when context is provided, its performance rises to **0.98**.

We admit these preliminary results should be interpreted cautiously considering the limited scale of this subset, but they tentatively indicate that model alignment on PhyCheck can foster context-sensitive

reasoning, encouraging models to dynamically integrate hidden physical factors rather than adhering solely to surface visual cues.

5. Conclusion

In this work, we introduce **PhyCheck**, a physics-law-oriented video question answering dataset designed to evaluate and improve the physical law understanding ability of Video-LLMs. PhyCheck provides both fine-grained physical-law questions and a context-sensitive subset for evaluating physical judgments. Through comprehensive evaluation of representative open-source, proprietary, and specialized Video-LLMs, we reveal substantial limitations in current models' physical law understanding. Furthermore, fine-tuning experiments with Qwen2.5-VL demonstrate that PhyCheck provides effective supervision for improving physical understanding, highlighting its value as both an evaluation benchmark and a training resource for developing more physically grounded video understanding models.

References

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. Llava-onevision-2: Towards next-generation perceptual intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

Purui Bai, Tao Wu, Jiayang Sun, Xinyue Liu, Huaibo Huang, and Ran He. Mvpbench: A multi-video perception evaluation benchmark for multi-modal video understanding, 2026. URL https://arxiv.org/abs/2603.22756.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation, 2024. URL https://arxiv.org/abs/2406.03520.

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, 2025. URL https://arxiv.org/abs/2503.06800.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024. Accessed:2024-02-15.

Meng Cao, Haoran Tang, Haoze Zhao, Hangyu Guo, Jiaheng Liu, Ge Zhang, Ruyang Liu, Qiang Sun, Ian Reid, and Xiaodan Liang. Physgame: Uncovering physical commonsense violations in gameplay videos, 2024. URL https://arxiv.org/abs/2412.01800.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan.

Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7310–7320, Seattle, WA, USA, 2024. IEEE.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*, pages 97959–98108, Singapore, 2025. International Conference on Learning Representations (ICLR).

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Dingkun Guo, Yuqi Xiang, Shuqi Zhao, Xinghao Zhu, Masayoshi Tomizuka, Mingyu Ding, and Wei Zhan. Phygrasp: Generalizing robotic grasping with physics-informed large multimodal models, 2024. URL https://arxiv.org/abs/2402.16836.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs, 2025. URL https://arxiv.org/abs/2506.09987.

Yann LeCun. A path towards autonomous machine intelligence. https://openreview.net/forum?id=BZ5a1r-kVsf, 2022. Accessed:2022-06-27.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: Ai-generated video detection via grounded artifact reasoning, 2026. URL https://arxiv.org/abs/2512.15693.

Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting, 2024. URL https://arxiv.org/abs/2403.03174.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, pages

23439–23554, Vienna, Austria, 2024. OpenReview.net.

LumaAI. Luma ray2. https://lumalabs.ai/ray, 2025. Accessed:2025-07-08.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation, 2024. URL https://arxiv.org/abs/2410.05363.

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)*, pages 948–958, Tucson, Arizona, USA, 2026. IEEE.

Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning, 2024. URL https://arxiv.org/abs/2406.11815.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

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

Yi Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. NEWTON: Are large language models capable of physical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9743–9758, Singapore, 2023. Association for Computational Linguistics.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos, 2026. URL https://arxiv.org/abs/2510.08377.

Kejuan Yang, Yizhuo Zhang, Mingyuan Du, Yue Zhang, Dixin Zheng, Kaili Zhao, Yang Xiao, Hanzhong Liang, and Kenan Xiao. UNIVID: Unified Vision-Language Model for Video Moderation. In *Proceedings*

of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026), pages 467–479, San Diego, California, United States, 2026. Association for Computational Linguistics.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, pages 83048–83077, Singapore, 2025. OpenReview.net.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, Los Alamitos, CA, USA, 2024. IEEE Computer Society.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model, 2024. URL https://arxiv.org/abs/2403.09631.

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