PhyEduVideo: A Benchmark for Evaluating Text-to-Video Models for Physics Education

Megha Mariam K.M
IIIT Hyderabad, India
megha.km@research.iiit.ac.in

Aditya Arun
Adobe MDSR, India
adityaarun@adobe.com

Zakaria Laskar
IISER Thiruvananthapuram, India
zakaria.laskar@iisertvm.ac.in

C.V. Jawahar
IIIT Hyderabad, India
jawahar@iiit.ac.in

Abstract

Generative AI models, particularly Text-to-Video (T2V) systems, offer a promising avenue for transforming science education by automating the creation of engaging and intuitive visual explanations. In this work, we take a first step toward evaluating their potential in physics education by introducing a dedicated benchmark for explanatory video generation. The benchmark is designed to assess how well T2V models can convey core physics concepts through visual illustrations. Each physics concept in our benchmark is decomposed into granular teaching points, with each point accompanied by a carefully crafted prompt intended for visual explanation of the teaching point. T2V models are evaluated on their ability to generate accurate videos in response to these prompts. Our aim is to systematically explore the feasibility of using T2V models to generate high-quality, curriculum-aligned educational content—paving the way toward scalable, accessible, and personalized learning experiences powered by AI. Our evaluation reveals that current models produce visually coherent videos with smooth motion and minimal flickering, yet their conceptual accuracy is less reliable. Performance in areas such as mechanics, fluids, and optics is encouraging, but models struggle with electromagnetism and thermodynamics, where abstract interactions are harder to depict. These findings underscore the gap between visual quality and conceptual correctness in educational video generation. We hope this benchmark helps the community close that gap and move toward T2V systems that can deliver accurate, curriculum-aligned physics content at scale. The benchmark and accompanying codebase are publicly available at https://github.com/meghamariamkm/PhyEduVideo.

1. Introduction

Creating educational videos is a resource-intensive task that requires crafting clear explanations, designing effective visuals, and ensuring both accuracy and engagement. In subjects such as physics, videos are particularly powerful, as they can vividly illustrate abstract ideas—such as motion, force, or energy—that are otherwise difficult to convey through text alone.

In recent years, there has been growing interest in leveraging AI for educational content creation, ranging from generating textual explanations to building interactive tutors and, more recently, developing multimodal learning resources [4, 10, 32, 36]. Initiatives such as Khan Academy's integration with GPT-4 [16] and Socratic by Google [8] exemplify the promise of AI-powered tutoring, though they remain largely focused on text-based assistance rather than video generation. Similarly, research in intelligent tutoring systems (ITS) has advanced adaptive instruction and personalized feedback, but predominantly within textual or structured interaction formats.

Meanwhile, recent progress in text-to-video (T2V) models [3, 5, 6, 12, 20, 25, 28, 30, 31, 37] offers the potential to automatically generate rich visual explanations from natural language prompts. While these models can already produce aesthetically compelling videos, their educational utility—particularly in physics—remains underexplored [35, 38]. Harnessing them for instructional purposes could substantially reduce the effort required to produce high-quality learning resources, while also broadening access to scientifically accurate educational content.

To advance this vision, we introduce the first benchmark specifically designed to evaluate the capacity of T2V models to generate videos that explain physics concepts in pedagogically meaningful ways. Unlike existing benchmarks [1, 2, 13, 14, 21, 22, 24, 39], which emphasize gen-

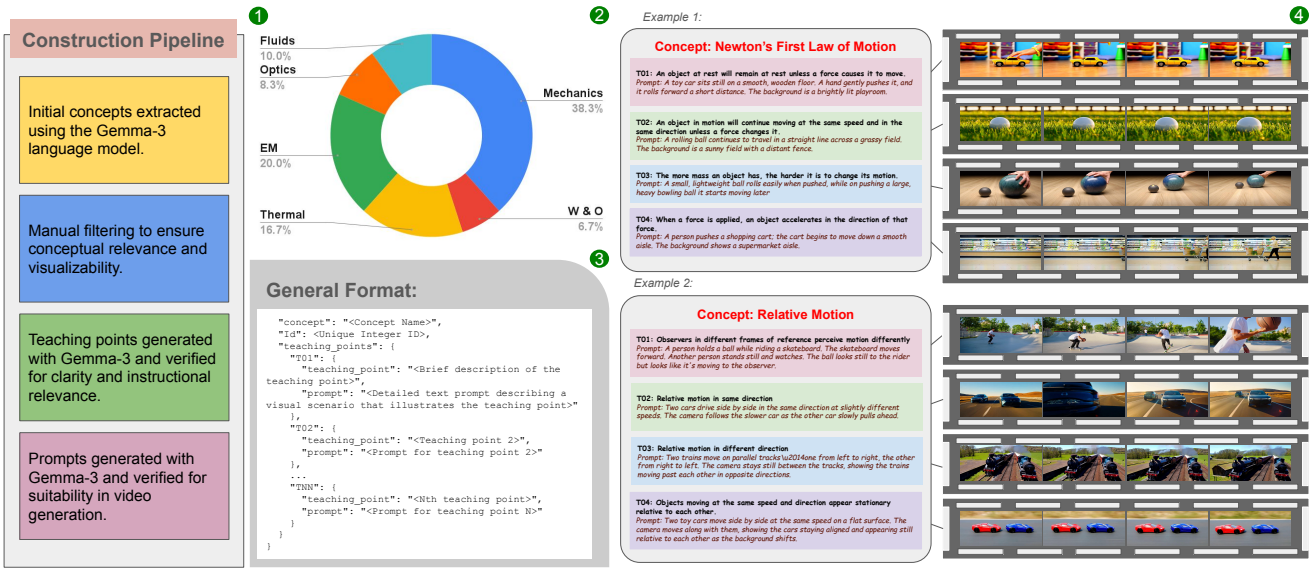


Figure 1. **Overview of the PhyEduVideo Benchmark.** 1 The construction pipeline, from concept extraction to prompt generation. 2 Concept distribution across five major physics domains: *Mechanics*, *Electromagnetism* (*EM*), *Optics*, *Thermodynamics* (Thermo), *Fluids*, and *Waves & Oscillations* (*W&O*). 3 Standardized representation of each concept, detailing key teaching points and corresponding video prompts. 4 Example concepts with teaching points, visual prompts, and representative generated video frames. As shown, current T2V models often fail to produce videos that are both semantically aligned and physically plausible—for example, in T04 (Relative Motion), the two toy cars were intended to move side by side at the same speed, but the generated video deviates from this.

eral video quality or physical plausibility, our benchmark prioritizes educational utility by grounding evaluation in well-defined physics concepts and their associated teaching points. Each concept is systematically decomposed into a set of teaching points that mirror how the concept would be introduced in instructional practice, ensuring both comprehensive coverage and pedagogical coherence. This structured design allows us to evaluate whether generated videos meaningfully support conceptual understanding rather than merely displaying visual plausibility. Figure 1 provides an overview of our benchmark. The PhyEduVideo benchmark consists of 205 prompts spanning 60 physics concepts, each decomposed into 1–5 teaching points that directly align with instructional goals. Breaking concepts down into teaching points ensures comprehensive coverage. The prompt associated with each teaching point has an average length of 16–45 words. Among the models we analyzed, Wan2.1 achieves the strongest overall performance, followed by PhyT2V. Domains such as Mechanics, Fluids, and Optics show relatively higher accuracy, whereas Electromagnetism and Thermodynamics remain more challenging, highlighting areas for future improvement. Our contributions are threefold:

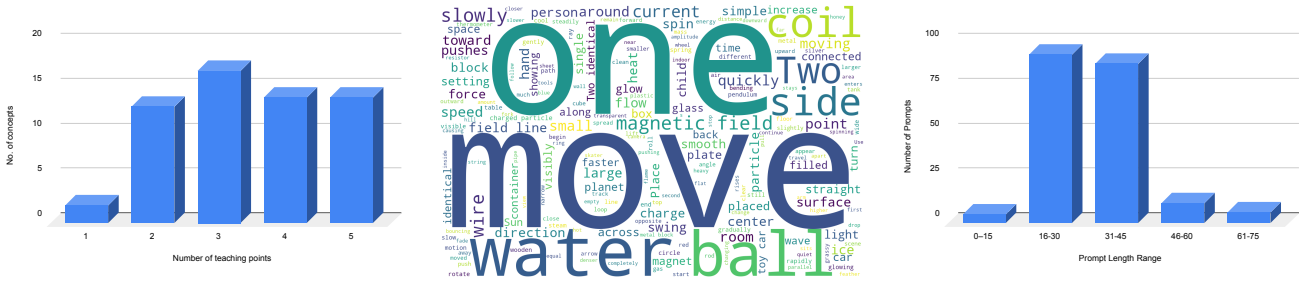
- We introduce PhyEduVideo, the first physics education benchmark designed to evaluate T2V generative models.
- We provide a structured framework that grounds evaluation in pedagogical units of analysis (teaching points), enabling fine-grained assessment of educational utility.
- We present empirical insights into the strengths and limitations of current T2V models in generating instructional videos, showing that while they produce visually coherent outputs, they often struggle with physics commonsense and semantic alignment.

2. Related Work

2.1. Text-to-Video Models

Text-to-video (T2V) generation has advanced rapidly, evolving from early GAN-based systems to diffusion and transformer architectures. Initial approaches such as MoCoGAN [27] and TGAN [7] introduced spatiotemporal discriminators but suffered from poor scalability, motion consistency, and text alignment. Diffusion models soon became dominant, with UNet-based architectures progressively denoising latent representations into coherent frames. Representative examples include ModelScope [29], VideoCrafter [5, 6], CogVideo [12], AnimateDiff [9], and Text2Video-Zero [15]. Large-scale efforts such as Imagen Video [11] and Make-A-Video [23] demonstrated high-resolution synthesis and spurred widespread adoption.

However, convolutional UNets struggle with long-range temporal dependencies, motivating the shift to Diffusion Transformers (DiTs), which use self-attention to model global spatial-temporal relationships. Models such as Sora [20], CogVideoX [37], Hunyuan [25], Wan2.1 [28], Pika [31], Lumiere [3], and Kling [30] exemplify this trend,



(a) (b) (c)

Figure 2. **Overview of benchmark statistics for the PhyEduVideo dataset.** (a) Distribution of teaching points across physics concepts, (b) Word cloud of frequent prompt terms, (c) Distribution of prompt lengths.

setting a new paradigm for T2V. This progression from GANs to UNet diffusion, and now to transformer-driven architectures, reflects a broader shift in generative modeling toward scalability, semantic fidelity, and controllability.

Complementary research explores multimodal pretraining for scalable video understanding, retrieval-augmented generation for stronger text–video alignment, and physics-aware conditioning for controllable dynamics. Notably, PhyT2V [35] combines LLM guidance with simulation priors, achieving 2.3× stronger physical compliance and 35% average gains on PhyGenBench [21]. These developments signal the maturation of T2V into a discipline uniting vision, language, and physical reasoning. Physics-aware models, in particular, show promise for education by offering intuitive, visual explanations of abstract concepts. By explicitly simulating physical interactions and constraints, they open new opportunities for delivering pedagogically grounded resources at scale. In this study, we systematically evaluate their strengths and limitations for instructional use.

2.2. Evaluation Benchmarks for Text-to-Video Models

While T2V models have made rapid progress in fidelity, stability, and semantic alignment, their evaluation has relied mostly on general-purpose metrics. VBench [13] introduced a hierarchical framework with dimensions such as prompt adherence, spatial coherence, and temporal consistency, later expanded in VBench++ [14] and VBench 2.0 [39] to include commonsense reasoning, physics realism, and aesthetics. Other benchmarks focus on compositional generalization (T2VCompBench), motion dynamics (DEVIL [19]), or controllability. Together, these efforts have established a solid foundation for large-scale and systematic T2V evaluation.

Physics-specific benchmarks test adherence to physical principles. VideoPhy [1] introduced Semantic Adherence (SA) and Physical Commonsense (PC) metrics, extended in VideoPhy2 [2] with a Physical Rules (PR) dimension. Physics-IQ [22] emphasized intuitive physical reasoning with real-world videos, while PhyGenBench [21] broadened coverage across mechanics, thermodynamics, and optics using simulation probes and LLM-based evaluators. These benchmarks represent an important step toward measuring physical realism, yet they are not explicitly designed for teaching contexts.

Despite these advances, existing benchmarks emphasize plausibility over pedagogy: they test if videos look realistic but not whether they *teach*. To address this gap, we propose the first benchmark tailored to physics education. Each concept is decomposed into fine-grained teaching points, enabling systematic evaluation of whether generated videos convey core ideas clearly and coherently. This reframing shifts evaluation from surface-level realism to instructional utility, offering a complementary perspective to prior benchmarks and advancing T2V research toward impactful educational applications.

3. PhyEduVideo

The PhyEduVideo benchmark is developed to systematically evaluate the capabilities of Text-to-Video (T2V) models in accurately visualizing foundational physics concepts for educational purposes. It encompasses a total of 60 core concepts drawn from seven major domains of classical physics: *Mechanics* (38.33%), *Waves & Oscillations* (6.67%), *Thermodynamics* (16.67%), *Electricity and Magnetism (Electromagnetism)* (20.00%), *Fluids* (10%), and *Optics* (8.33%) Figure 1 (2). *Mechanics* is the most represented domain, reflecting its foundational role in introductory physics education. A standardized format is provided in Figure 1 (3).

To construct the benchmark, we followed a structured multi-stage pipeline, visualized in Figure 1 (1):

1. **Concept Identification:** We began by using the Gemma-3 language model [26] to extract an initial set of classical physics concepts from standard K-12 and undergraduate physics curricula. This automated step ensured coverage across a wide conceptual space.
2. **Manual Filtering:** The extracted list was then manually reviewed by physics experts to retain only those concepts that are both pedagogically essential and visually realiz-

able. Abstract, redundant, or highly mathematical topics—such as Lagrangian mechanics, tensor calculus, or complex integrals—were excluded in favor of those that lend themselves to intuitive, observable phenomena like Newton's laws, simple harmonic motion, or conservation of energy.

3. **Decomposition into Teaching Points:** Each validated concept was further broken down into multiple *teaching points*—fine-grained, pedagogically distinct subconcepts that capture specific physical behaviors or relationships. As shown in Figure 1 4 for example, the concept of "Newton's First Law" is divided into four teaching points: objects at rest, constant motion, inertia, and force-induced acceleration. This decomposition allows T2V models to be tested on precise subcomponents of conceptual understanding, rather than broad themes.
4. **Prompt Generation and Refinement:** For each teaching point, candidate prompts were first generated automatically using Gemma-3 and then refined by humans. These prompts provide short, clear descriptions for generating videos. Examples of the final prompts and their corresponding videos are shown in Figure 1 4.

Benchmark Statistics: The final PhyEduVideo benchmark comprises 205 prompts derived from the 60 physics concepts, each decomposed into between one and five teaching points (Figure 2(a)). Each prompt is written as a self-contained, visually descriptive scenario that maps directly to a teaching goal. The average prompt length falls in the 16–45 word range, with longer prompts offering additional context for more complex situations, as seen in Figure 2(c). Figure 2(b) shows the prompt vocabulary, which spans a wide range of physical entities (e.g., "ball," "coil," "current") and actions (e.g., "move," "push," "show"), reflecting both linguistic diversity and conceptual coverage. Collectively, these characteristics enable PhyEduVideo to serve as a rigorous and pedagogically grounded testbed for evaluating the scientific accuracy, temporal coherence, and visual fidelity of physics-focused T2V models. In comparison, PhyGenBench offers 160 prompts across 27 physical laws, T2VPhysBench provides 84 prompts spanning twelve laws, and VideoPhy focuses on interaction-driven scenarios—highlighting PhyEduVideo's broader, education-oriented design grounded in structured teaching points.

3.1. Metrics

To evaluate T2V models for physics education using the PhyEduVideo benchmark, we propose a structured framework assessing video generation quality, prompt adherence, and physics-specific fidelity. The evaluation is conducted across four dimensions: *Semantic Alignment (SA)*, *Physical Commonsense (PC)*, *Motion Smoothness (MS)*, and *Temporal Flickering (TF)*.

- **Semantic Alignment (SA):** [1, 2, 21] This metric measures how well a generated video matches the main idea of the input prompt. It checks if the core scenario, key actions, and important visual elements described in the text appear correctly and coherently in the video. For example, for the prompt "A rolling ball continues in a straight line," a semantically aligned video should show the ball moving steadily along a straight path. Semantic Alignment is scored from 0 to 3 using InternVL3.5 [33], which evaluates two components: object score (0 = none, 1 = some, 2 = all key objects present) and action score (0 = main action not depicted, 1 = main action depicted). A higher score means the video correctly represents both the described objects and actions.
- **Physics Commonsense (PC):** [1, 2, 21] This metric evaluates whether the generated video correctly follows the intended teaching point. For example, when ice is placed in water, it should melt gradually, $0°C$ until the ice is fully melted, and the water level should rise steadily as the ice turns into liquid. Following PhyGenBench [1], this metric is structured into three finer-grained evaluation stages:
  1. *Key Physical Phenomena Detection:* This sub-metric evaluates whether the video successfully captures the essential physical behavior described in the prompt. For example, if the prompt involves projectile motion, the video should display a curved parabolic trajectory, rather than an unrealistic linear path.
  2. *Physics Order Verification:* This stage assesses the temporal coherence of physical events within the video. It verifies whether the sequence of actions follows a logically and physically correct order. For instance, in a pendulum motion, the object must first be released before it begins to swing. To perform this evaluation automatically, we employ LLaVA-Interleave [17].
  3. *Overall Naturalness Evaluation:* This component assesses the naturalness of a video by examining whether objects and their movements appear physically plausible. To guide this evaluation, we define four GPT-generated descriptions for a given prompt, representing different levels of naturalness: Fantastical descriptions involve highly imaginative or impossible scenarios. Clearly unrealistic descriptions depict objects behaving in ways that blatantly violate fundamental physical principles, for example, a ball sinking through a solid table or two objects occupying the same space simultaneously. Slightly unrealistic descriptions generally follow physical principles but include minor inconsistencies or exaggerated effects, such as overly bouncy objects or frictionless slides. Realistic descriptions describe objects moving and interacting fully in accordance with real-world physics. InternVideo2 [34] is then employed to compare the

Metric	*EM*		*Mech*		*Fluids*		*Thermal*		*Optics*		*W&O*		Avg	
	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$\rho\uparrow$	$\tau\uparrow$
VideoPhy-SA	0.24	0.19	0.31	0.24	0.49	0.40	0.28	0.21	0.45	0.36	0.47	0.37	0.44	0.34
VideoPhy-PC	-0.01	-0.01	0.11	0.09	-0.11	-0.09	-0.05	-0.04	0.17	0.14	0.07	0.06	0.01	0.01
PhyEduVideo-SA	0.46	0.41	0.48	0.45	0.59	0.51	0.66	0.60	0.45	0.41	0.42	0.39	0.51	0.46
PhyEduVideo-PC	0.30	0.27	0.56	0.52	0.35	0.33	0.59	0.55	0.30	0.27	0.57	0.54	0.39	0.36

Table 1. Domain-wise correlations between human and model scores using Spearman's ρ and Kendall's τ. Models (VideoPhy, PhyEduVideo) are split into SA = Semantic Alignment and PC = Physics Commonsense. Domains are abbreviated as follows: EM: Electromagnetism, Mech: Mechanics, Thermal: Thermodynamics, W&O: Waves and Oscillations, and Avg: Average across all domains.

Figure 3. Comparison of SA (Semantic Adherence) and PC (Physics Commonsense) scores assigned by the VideoPhy, Automatic Evaluator (PhyEduVideo) and humans. Detailed videos are available on the GitHub page.

generated video against these categories, assigning the most appropriate category to the video.

- **Motion Smoothness:** [13] This refers to the continuity and coherence of object motion and background in the video. Videos should not exhibit jerky, inconsistent, or mechanically impossible motion patterns. The motion in the video should be smooth and follow the physics concept.
- **Temporal Flickering:** [13] This evaluates the stability of visual properties (like object color, size, or shape) across

frames. Abrupt flickers, changes in object identity, or disappearing elements can break temporal coherence and degrade the viewing experience. A consistently rendered object across the video receives a high flickering score.

Overall, these four criteria provide a structured and holistic framework for evaluating generated videos in the context of physics education. By addressing both conceptual and visual aspects, the benchmark supports rigorous and pedagogically meaningful assessment of model outputs. This enables more targeted progress in developing text-to-video models that are both scientifically accurate and educationally effective.

3.2. Human Evaluation

To assess the alignment of automatic metrics with human perception, we conducted a human evaluation study on 500 videos, involving annotators who had formally studied physics up to the 12th grade. The results, summarized in Tables 1, show that **PhyEduVideo** achieves much stronger correlations with human judgments than **VideoPhy** [1]. VideoPhy is a benchmark that tests whether text-to-video models follow basic physical commonsense, such as correct object interactions, material behaviors, and physical laws along with semantic adherence. For both SA and PC, the highest correlations are observed in the *Thermodynamics* category, while the lowest are found in *Electromagnetism* and *Optics*. Overall, PhyEduVideo achieves a Spearman correlation of 0.509 and a Kendall correlation of 0.462 for SA—considerably higher than the corresponding values for VideoPhy (gap = 0.071 and 0.122). For PC, PhyEduVideo reaches 0.392 (Spearman) and 0.363 (Kendall), again significantly outperforming VideoPhy (0.008 and 0.006), with absolute gains of 0.384 and 0.357, respectively. This consistent gap underscores the value of our benchmark in better capturing human judgment. Importantly, these higher correlation numbers also indicate that **PhyEduVideo** more faithfully aligns with pedagogically accurate teaching points, ensuring that evaluation outcomes reflect not just visual plausibility but instructional relevance. Figure 3 presents qualitative examples where human scores are shown alongside predictions from VideoPhy and PhyEduVideo, further demonstrating how our benchmark provides more faithful and interpretable assessments.

4. Experiments

4.1. Evaluated Models

We evaluate five state-of-the-art text-to-video (T2V) generation models on our benchmark: CogVideoX [37], Wan2.1 [28], VideoCrafter2 [6], Video-MSG [18], and PhyT2V [21]. CogVideoX-5B, with demonstrated success on physics-focused evaluations, serves as a baseline for physics-grounded video generation due to its consistent

	Model	SA ↑	PC ↑	MS ↑	TF ↑
Mechanics	VideoCrafter2	0.75	0.52	0.94	0.92
	CogVideoX	0.85	0.57	0.98	0.97
	Wan2.1	0.86	0.66	0.99	0.98
	Video-MSG	0.75	0.53	0.99	0.99
	PhyT2V	0.80	0.59	0.98	0.97
W&O	VideoCrafter2	0.72	0.49	0.92	0.90
	CogVideoX	0.79	0.59	0.98	0.98
	Wan2.1	0.87	0.59	0.99	0.98
	Video-MSG	0.69	0.46	0.99	0.99
	PhyT2V	0.72	0.59	0.98	0.97
Fluids	VideoCrafter2	0.58	0.48	0.89	0.87
	CogVideoX	0.71	0.58	0.98	0.97
	Wan2.1	0.90	0.63	0.99	0.98
	Video-MSG	0.67	0.58	0.99	0.99
	PhyT2V	0.85	0.63	0.98	0.97
Thermal	VideoCrafter2	0.51	0.38	0.89	0.86
	CogVideoX	0.75	0.52	0.98	0.98
	Wan2.1	0.93	0.52	0.99	0.98
	Video-MSG	0.71	0.39	0.99	0.99
	PhyT2V	0.75	0.49	0.98	0.97
EM	VideoCrafter2	0.54	0.50	0.89	0.88
	CogVideoX	0.73	0.65	0.98	0.98
	Wan2.1	0.65	0.57	0.99	0.98
	Video-MSG	0.60	0.48	0.99	0.99
	PhyT2V	0.75	0.62	0.98	0.98
Optics	VideoCrafter2	0.64	0.62	0.88	0.83
	CogVideoX	0.69	0.60	0.99	0.98
	Wan2.1	0.78	0.64	0.99	0.98
	Video-MSG	0.69	0.64	0.99	0.99
	PhyT2V	0.80	0.71	0.99	0.98
Average	VideoCrafter2	0.62	0.50	0.90	0.88
	CogVideoX	0.75	0.59	0.98	0.98
	Wan2.1	0.83	0.60	0.99	0.98
	Video-MSG	0.68	0.52	0.99	0.99
	PhyT2V	0.78	0.60	0.98	0.97

Table 2. Comparison of five video generation models across six physics domains, along with their overall averages. Metrics include Semantic Adherence (SA), Physics Commonsense (PC), Motion Smoothness (MS), and Temporal Flickering (TF). Best scores are highlighted in cyan, and second-best in light cyan. Domains are abbreviated as follows: EM: Electromagnetism, Thermal: Thermodynamics and W&O: Waves and Oscillations

high scores in physics-following benchmarks. Wan2.1 is a strong, general-purpose T2V model that achieves high scores across a wide range of benchmarks, providing insight into the generalization capabilities of current systems. VideoCrafter2 is known for generating high-resolution, visually coherent videos, making it useful for assessing visual quality and detail. In addition, we include models with more specialized architectures. Video-MSG employs a training-free, structured guidance pipeline that closely follows input prompts. Its generation proceeds in three

stages: (1) Background Planning, where a multimodal large language model (MLLM, specifically GPT-4o) produces detailed background descriptions, rendered via a text-to-image (T2I) model and animated with an image-to-video (I2V) model; (2) Foreground Object Layout and Trajectory Planning, where object positions and motions are inferred with MLLM guidance; and (3) Video Generation, where the planned layout is denoised to produce the final video. This compositional approach has shown strong performance on T2V-CompBench [24] and is evaluated here for its ability to produce visually coherent, pedagogically meaningful physics content. PhyT2V [35], in contrast, is specifically engineered for physics-aware generation: it integrates large language models with physics simulation priors to iteratively refine video content, ensuring adherence to physical laws while maintaining semantic and temporal coherence. Its performance on PhyGenBench [21] demonstrates notable gains in physical plausibility and instructional clarity, making it uniquely suited for evaluating T2V models in educational contexts. Comprehensive implementation details, including model configurations and evaluation protocols, are provided in the supplementary material.

4.2. Quantitative Evaluation

Quantitative evaluations, summarized in Tables 2, highlight clear trends in both perceptual quality and correctness-based performance of text-to-video (T2V) models. All evaluated models achieve consistently high scores in Motion Smoothness (MS) and Temporal Flickering (TF), with values typically above 0.85, demonstrating that current systems are capable of generating visually coherent and temporally stable videos. However, this strength contrasts sharply with the lower scores observed in correctness-oriented metrics such as Semantic Adherence (SA) and Physics Commonsense (PC), which are critical for ensuring educational and conceptually accurate content. Among the models, Wan2.1 [28] stands out as the overall best performer, achieving the highest SA and PC scores across most domains, followed closely by PhyT2V [35], which maintains competitive reasoning ability while delivering visually stable results. In comparison, VideoCrafter2 [6] ranks lowest in both SA and PC despite its strong performance on temporal flickering and motion smoothness. Video-MSG [18] similarly excels in video quality metrics but does not achieve a significant boost in physics commonsense, suggesting that compositional control alone is insufficient for capturing complex physics concepts.

A category-level analysis reveals notable differences across physics domains. *Mechanics* emerges as a relatively solvable domain, with models achieving SA scores above 0.75 and PC scores exceeding 0.50, reflecting that visually grounded concepts like motion and collisions are easier to represent. *Fluids* and *Optics* stand out as the best-performing domains overall (across all 4 metrics), reaching the highest SA (up to 0.90) and PC (up to 0.71), indicating that distinctive visual dynamics such as flow patterns or light interactions are more learnable by current models. By contrast, *Thermodynamics* and *Electromagnetism* show the weakest correctness performance: in Thermodynamics, VideoCrafter2 drops to an SA of 0.51 and a PC of 0.38, while in Electromagnetism, most models record PC values below 0.50. *Waves & Oscillations* show moderate performance, better than Thermodynamics and Electromagnetism but trailing behind Fluids and Optics. These results reveal a consistent reasoning–perception gap: while models reliably generate smooth and visually appealing content, their semantic adherence and physics commonsense remain limited. Wan2.1 and PhyT2V perform comparatively better, showing greater stability, coherence, and conceptual alignment, making them more suitable for physics-focused educational content. A key challenge arises in domains such as *Electromagnetism*, where core concepts involve charges, magnetic fields, and electric fields—phenomena that are not directly visible. For teaching, however, it is crucial to make such invisible entities perceivable in order to build understanding. This is precisely where our benchmark stands out from existing physics-based benchmarks: rather than only checking whether generated videos obey physical laws, we emphasize the educational dimension, requiring models to represent abstract and invisible concepts in a way that aids learning.

4.3. Qualitative Evaluation

Figure 4 qualitatively compares generations across six key physics education domains—*Mechanics*, *Waves & Oscillations*, *Fluids*, *Thermodynamics*, *Electromagnetism*, and *Optics*—revealing strengths and limitations in how current models visually communicate scientific concepts. Corresponding to each domain, an example teaching point, textual prompt to query the T2V and images from start, middle and end part of the generated output video are shown in the Figure 4. Detailed videos are in the Supplementary. Wan 2.1 [28] shows strong educational potential, generating coherent and semantically grounded sequences that align well with physical principles, such as realistic projectile motion in Mechanics and light refraction in Optics. CogVideoX [37] performs well in *Mechanics* and *Fluids*, where object interactions are simpler and more grounded in visual cues, though often hindered by structural inconsistencies. VideoCrafter2 [6] consistently delivers visually smooth outputs but lacks the semantic fidelity needed for instructional clarity, especially in abstract domains like *Electromagnetism* and *Waves*. Video-MSG [18] maintains temporal stability and shows potential in controlled categories like Mechanics and Thermodynamics, yet struggles with conveying deeper causal relationships and dynamic vari-

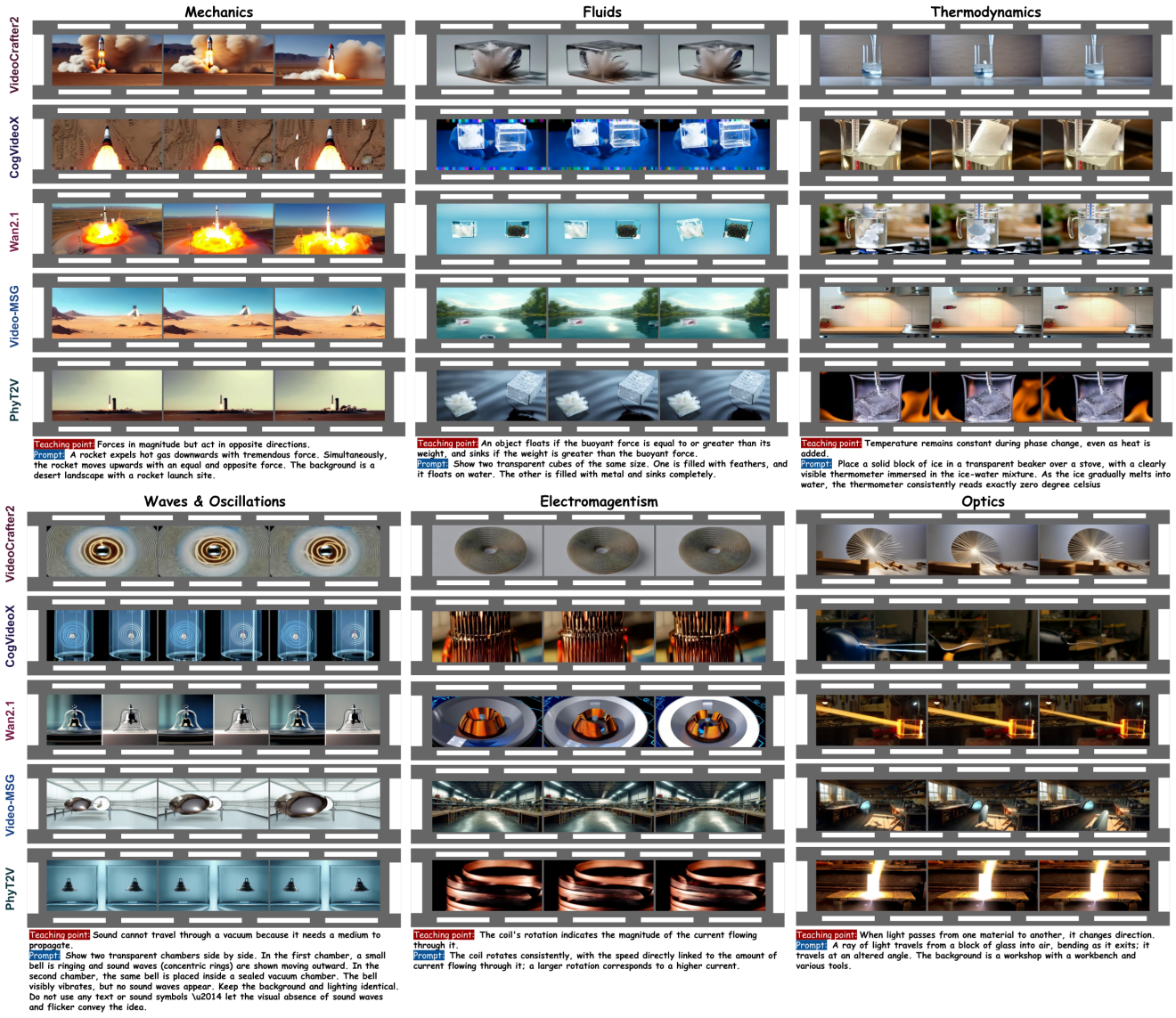


Figure 4. Qualitative comparisons of generated videos across six classical physics categories—*Mechanics*, *Waves & Oscillations*, *Fluids*, *Thermodynamics*, *Electromagnetism*, and *Optics*—for five T2V models: VideoCrafter2, CogVideoX, Wan2.1, Video-MSG, and PhyT2V. Detailed videos are available on the GitHub page.

ations essential for physics learning. PhyT2V [35], designed with physics-awareness in mind, achieves a strong balance between visual stability and conceptual fidelity, excelling particularly in scenarios that require accurate physical reasoning, such as current-induced effects in *Electromagnetism*. These observations underscore the gap between visual quality and conceptual fidelity in current models, emphasizing the need for physics-aware architectures to support meaningful and accurate science education content.

5. Conclusion

This work introduces a benchmark for evaluating text-to-video (T2V) generation in physics education. Unlike prior efforts that mainly test adherence to physical laws, our benchmark emphasizes educational relevance. Each physics concept is broken into granular teaching points, with prompts targeting their visual explanation. This enables evaluation of whether models generate videos that not only look realistic but also support teaching by making abstract or invisible entities—such as charges, fields, or wave interactions—visually understandable. Using this benchmark, we evaluate CogVideoX, Wan2.1, VideoCrafter2, Video-MSG, and PhyT2V. While models produce coherent motion with reasonable smoothness, they often struggle with semantic adherence and physics commonsense. Wan2.1 and PhyT2V perform comparatively better but still have room for improvement, highlighting the need for physics-aware, education-focused T2V systems.

Acknowledgments

We acknowledge the support of Google Cloud credits provided through a GCP Award as part of the Gemma Academic Program.

References

[1] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *ICLR*, 2025. 1, 3, 4, 6

[2] 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. *arXiv preprint arXiv:2503.06800*, 2025. 1, 3, 4

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia*, 2024. 1, 2

[4] Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences*, 2025. 1

[5] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 1, 2

[6] 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 *CVPR*, 2024. 1, 2, 6, 7

[7] Zihan Ding, Xiao-Yang Liu, Miao Yin, and Linghe Kong. Tgan: Deep tensor generative adversarial nets for large image generation. *arXiv preprint arXiv:1901.09953*, 2019. 2

[8] Google. Socratic by google help center. https://support.google.com/socratic/?hl=en. 1

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 2

[10] Ville Heilala, Roberto Araya, and Raija Hämäläinen. Beyond text-to-text: An overview of multimodal and generative artificial intelligence for education using topic modeling. In *SIGAPP*, 2025. 1

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[12] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 1, 2

[13] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 1, 3, 5

[14] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–18, 2025. 1, 3

[15] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *ICCV*, 2023. 2

[16] Khan Academy. Harnessing ai so that all students benefit: A nonprofit approach for equal access. https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/, 2024. 1

[17] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *ICLR*, 2025. 4

[18] Jialu Li, Shoubin Yu, Han Lin, Jaemin Cho, Jaehong Yoon, and Mohit Bansal. Training-free guidance in text-to-video generation via multimodal planning and structured noise initialization. *ArXiv2504.08641*, 2025. 6, 7

[19] Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, et al. Evaluation of text-to-video generation models: A dynamics perspective. *NeurIPS*, 2024. 3

[20] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 1, 2

[21] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 43781–43806. PMLR, 2025. 1, 3, 4, 6, 7

[22] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. 1, 3

[23] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 2

[24] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *CVPR*, 2025. 1, 7

[25] Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, et al. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*, 2024. 1, 2

[26] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 3

[27] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018. 2

[28] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2, 6, 7

[29] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[30] Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, et al. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*, 2025. 1, 2

[31] Leijie Wang, Nicholas Vincent, Julija Rukanskaitė, and Amy Xian Zhang. Pika: Empowering non-programmers to author executable governance policies in online communities. In *CHI*, 2024. 1, 2

[32] Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252:124167, 2024. 1

[33] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 4

[34] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024. 4

[35] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *CVPR*, 2025. 1, 3, 7, 8

[36] Antoun Yaacoub, Sansiri Tarnpradab, Phattara Khumprom, Zainab Assaghir, Lionel Prevost, and Jérôme Da-Rugna. Enhancing ai-driven education: Integrating cognitive frameworks, linguistic feedback analysis, and ethical considerations for improved content generation. *arXiv preprint arXiv:2505.00339*, 2025. 1

[37] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 1, 2, 6, 7

[38] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 1

[39] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 1, 3