Supplementary Material - 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

1. Model Details

We evaluate six state-of-the-art video generation models with distinct design philosophies. VideoCrafter2 is an open-source diffusion-based framework known for controllability and high-quality short clips. CogVideoX [6], a transformer-based model, emphasizes long-duration generation with improved temporal coherence. Wan2.1 advances photorealism and motion stability through refined denoising strategies. Video-MSG employs a controlled generation strategy getting high scores for T2VCompench [3] prompts. PhyT2V [5] is a model designed for physics video generation via CoT method. Table 1 represents the model details for each model.

Model	Duration (s)	FPS	Resolution
VideoCrafter2 [1]	5	8	512 x 320
CogVideoX-5b [6]	6	15	640 x 320
Wan2.1 [4]	6	15	832 x 480
Video-MSG [2]	6	28	720 x 480
PhyT2V [5]	6	8	720 x 480

Table 1. Details of duration, FPS, and resolution for each model are presented in the table.

2. Human Evaluation

A total of 500 videos were selected for human evaluation, covering outputs from VideoCrafter2 [1], CogVideoX [6], Wan2.1 [4], Video-MSG [2], and PhyT2V [5]. As shown in Figure 2, each video was evaluated by human judges who answered two specific questions designed to assess the video's content. The annotators followed a standardized set of instructions, shown in Figure 1, which ensured consistency and fairness across all assessments. The evaluation focused on how well the video adhered to the given prompt and whether it accurately conveyed the intended teaching point. These human judgments provide a benchmark for comparing automatic evaluation metrics against human perception.

3. Analysis of Score Mismatches Between PhyEduVideo and Human Evaluators

We analyzed cases where PhyEduVideo's scores for Semantic Adherence (SA) and Physics Commonsense (PC) did not align with human judgments, focusing on understanding the causes of mismatches (Figure 3) . Overall, the model performs well in straightforward scenarios, such as applying force to a shopping cart, where both human and model scores perfectly match. However, in more complex cases, PhyEduVideo tends to overestimate correctness, reflecting a limitation in capturing nuanced physics reasoning or semantic context. For example, in the rotating coil scenario, humans assigned low scores (SA = 0.34, PC = 0.34) due to partial recognition of the relation between current and rotation, while PhyEduVideo overestimated both (SA = 1.0, PC = 0.67). Similarly, in planetary orbit and charged particle in a magnetic field cases, the model assigned higher scores than humans, likely because it detected general motion or field presence but failed to capture detailed physics principles, such as orbital speed variation or circular trajectories. In meter bridge wire adjustment and projectile motion on a hill, PhyEduVideo again overestimated both SA and PC, misinterpreting visual cues as correct semantic and physics adherence, whereas humans recognized subtle discrepancies in the purpose or motion. In summary, PhyEduVideo generally aligns well with human judgments for clear and straightforward scenarios. In more complex situations requiring fine-grained reasoning, it sometimes assigns higher semantic and physics scores than humans, likely due to subtle physics nuances, partial contextual cues, or reliance on

visual detection of motion and objects.

References

[1] 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] 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. 1

[3] 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

[4] 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

[5] 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

[6] 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

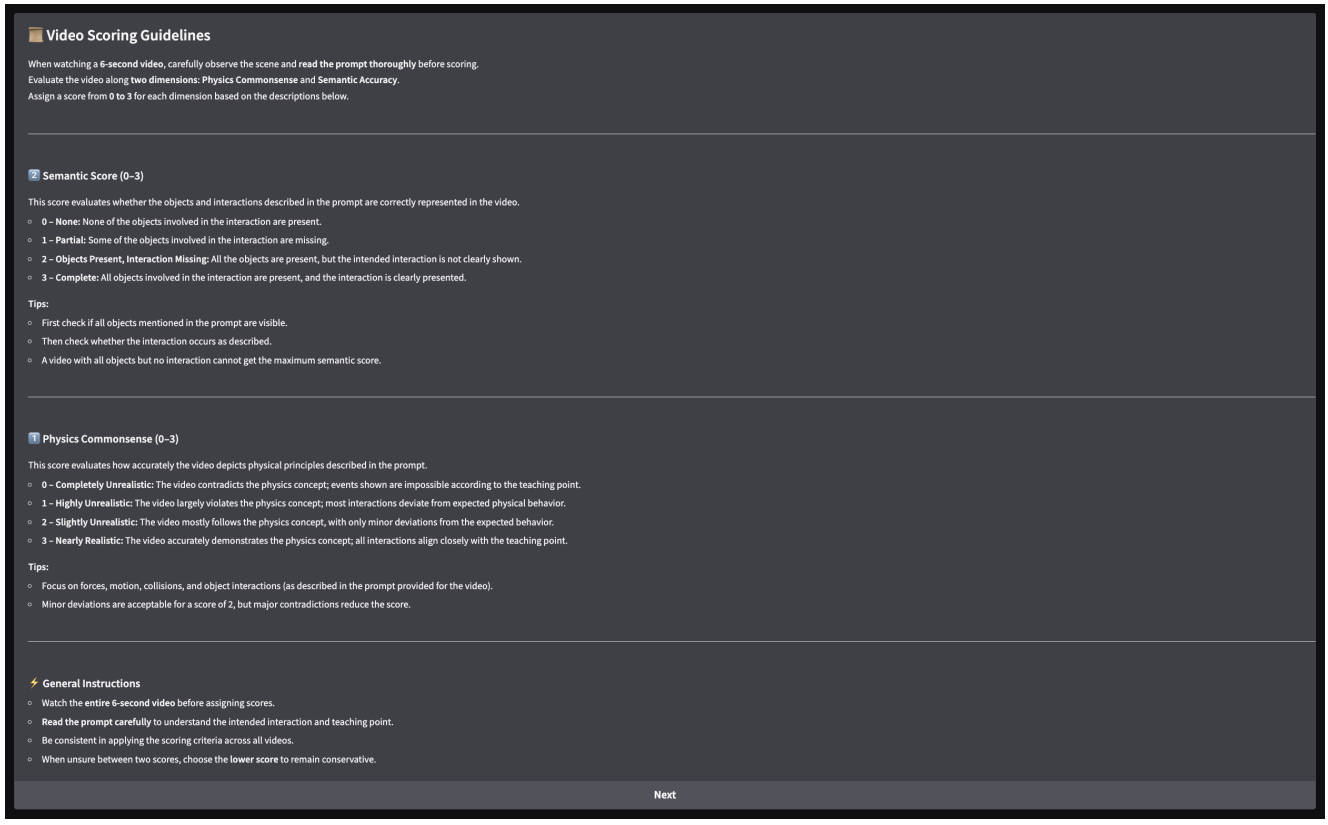


Figure 1. Guidelines and rules given to human annotators to ensure consistent and reliable evaluation.

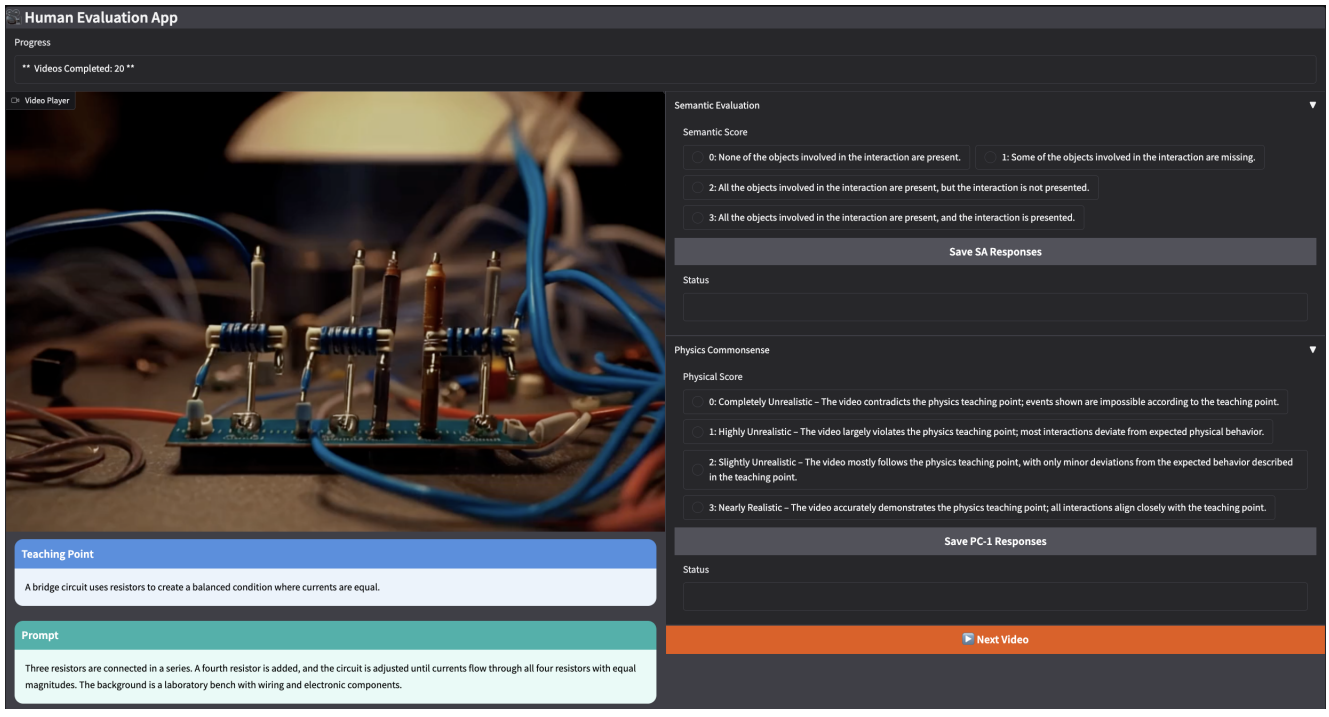


Figure 2. Questions provided for human evaluation and their respective scoring schemes are illustrated in the diagram above.

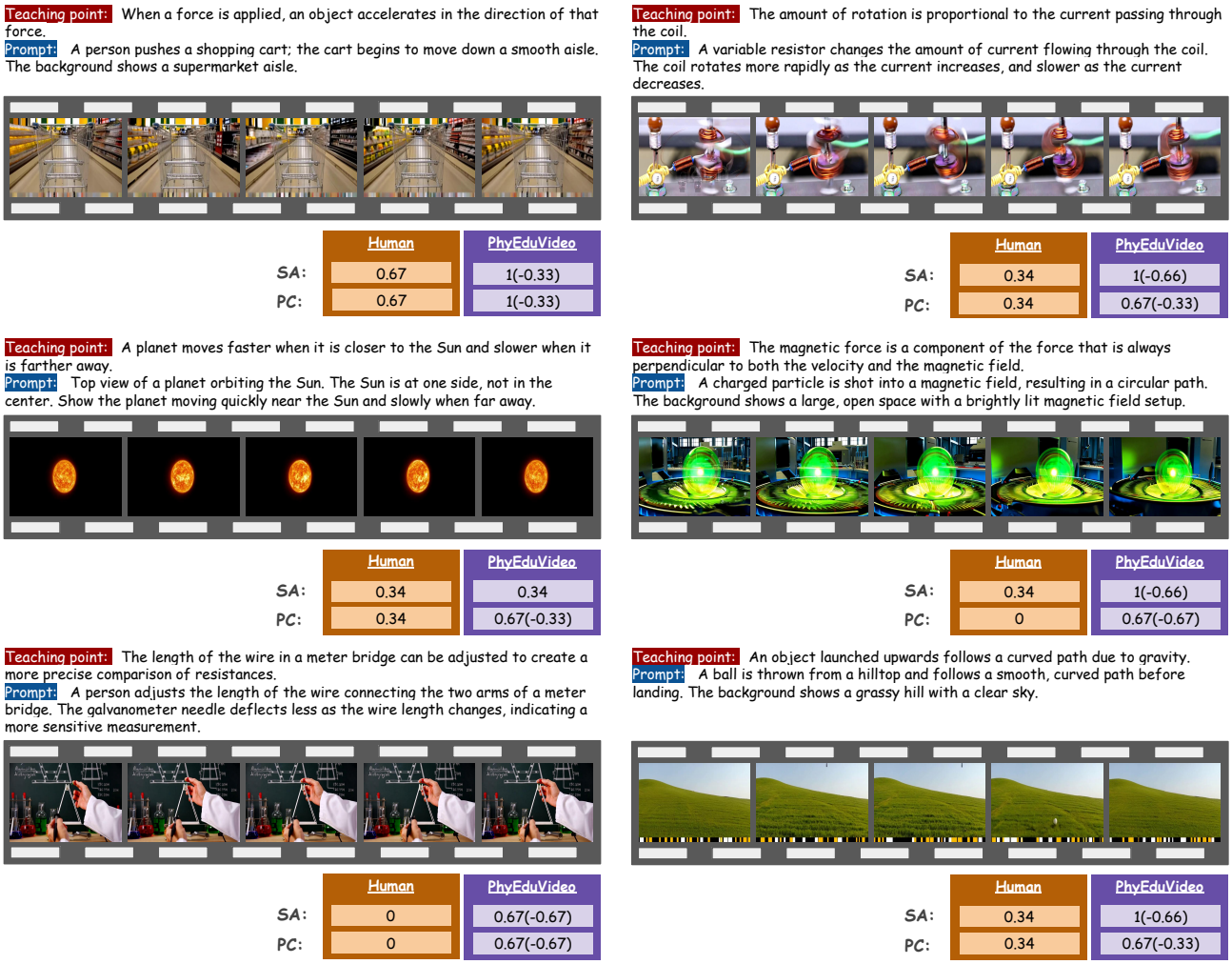


Figure 3. Comparison of SA (Semantic Adherence) and PC (Physics Commonsense) scores assigned by the Automatic Evaluator (PhyEduVideo) and humans.

Teaching point: If no external forces act on a system, the total linear momentum of the system remains the same before and after a collision.
Prompt: Two identical ice skaters glide toward each other on a frictionless ice rink. They collide gently and move together slowly after the collision. The background is a quiet, empty ice rink with no distractions.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: skater, ice rink
ACTION: A smaller skater collides with a larger stationary skater, and they slide together slowly across a frictionless ice rink.

KEY SEQ IDENTIFICATION

Q. After the collision, are both skaters moving together across the ice? Yes

Q. Is the two identical skaters gliding towards each other before the collision? Yes

ORDER VERIFICATION

Retr. prompt: At the moment the two skaters make contact

Description1: Two identical skaters are gliding towards each other.

Description2: both skaters slide together slowly after the collision.

OVERALL NATURALNESS EVALUATION

1) The skaters levitate, pass through each other without interaction, or explode in a flash of light after collision — completely ignoring momentum conservation.

2) The skaters bounce off each other and move away faster than before, or one suddenly speeds up while the other stops instantly, defying conservation laws.

3) The skaters' speeds or paths change slightly too early or too late relative to the moment of collision, or there's slight unnatural jittering, but overall momentum conservation is mostly maintained.

4) The skaters approach at equal speed, collide, and move together at the correct slower combined speed immediately after — matching the conservation of linear momentum almost perfectly.

Wan2.1
SA: 0.67
PC: 0.67

PhyT2V
SA: 0.67
PC: 0.67

Figure 4. Domain: Mechanics. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: Temperature describes the average kinetic energy of particles in a substance.
Prompt: Split the screen into two parts. On one side, show cold gas particles moving slowly and spaced far apart. On the other side, show hot gas particles moving rapidly and bouncing around quickly. Include a digital thermometer above each container showing low and high temperatures.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: gas particles, thermometer
ACTION: Cold gas particles move slowly; hot gas particles move rapidly.

KEY SEQ IDENTIFICATION

Q. Is the ice block completely melted after being in contact with the hot metal rod? Yes

Q. Is the metal rod no longer glowing after all the ice has melted? Yes

ORDER VERIFICATION

Retr. prompt: When the particles in both containers are visibly moving at different speeds

Description1: both containers show particles at rest or with minimal motion, and thermometers indicate similar low temperatures.

Description2: particles in the hot container move rapidly and collide frequently, while those in the cold container move slowly and less often, with thermometers showing a clear temperature difference.

OVERALL NATURALNESS EVALUATION

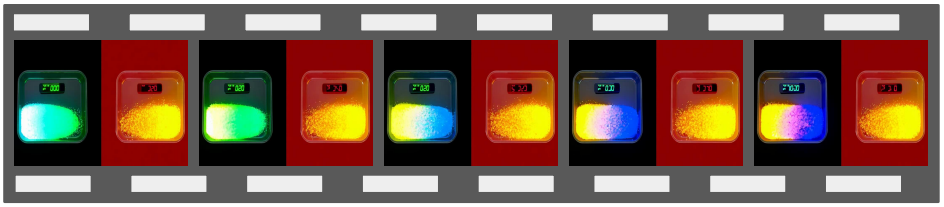
1) The animation shows gas particles on the cold side moving backward in time, changing shape, or merging and splitting at random. Thermometers flicker with nonsensical symbols. Particles teleport or transform into non-physical objects like animals or geometric shapes. The visual sequence is magical and completely ignores physical laws.

2) The cold gas particles are moving faster than the hot gas particles, or both sides have particles moving at the same speed regardless of thermometer reading. Particles may stop abruptly or pass through container walls without bouncing. The thermometer readings do not correlate with the observed particle speeds, clearly breaking the connection to kinetic energy.

3) The vast majority of particle motion is correct, but there are minor issues: perhaps a few collisions look awkward or a couple of particles move slightly faster or slower than they should for their side. The thermometer may have a slight delay in updating when the particle speeds change, but these are minor deviations that do not seriously undermine the teaching point.

4) The animation accurately shows cold gas particles moving slowly and spaced apart, and hot gas particles moving rapidly and bouncing energetically. Thermometers above each container display low and high temperatures that match the observed motion. All visual details closely align with the expected physical behavior and teaching point, with no noticeable errors.

Wan2.1
SA: 1
PC: 0.67



PhyT2V
SA: 1
PC: 0.67

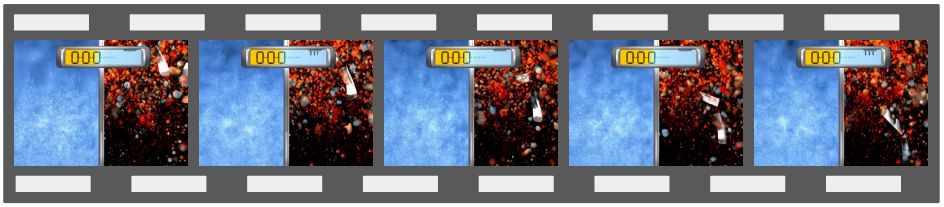


Figure 5. Domain: Thermodynamics. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: Temperature describes the average kinetic energy of particles in a substance.
Prompt: Split the screen into two parts. On one side, show cold gas particles moving slowly and spaced far apart. On the other side, show hot gas particles moving rapidly and bouncing around quickly. Include a digital thermometer above each container showing low and high temperatures.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: hair dryer, ping pong ball
ACTION: A ping pong ball floats in an upward stream of air and falls when the air stops.

KEY SEQ IDENTIFICATION

Q. Is the ping pong ball floating stably above the upward air stream from the hair dryer? Yes

Q. Does the ping pong ball start to fall as soon as the air stream stops? Yes

ORDER VERIFICATION

Retr. prompt: When the hair dryer turns off and the ball starts falling.

Description1: the ball floats steadily above the hair dryer in the fast-moving air stream.

Description2: the hair dryer is off and the ball falls straight down due to gravity.

OVERALL NATURALNESS EVALUATION

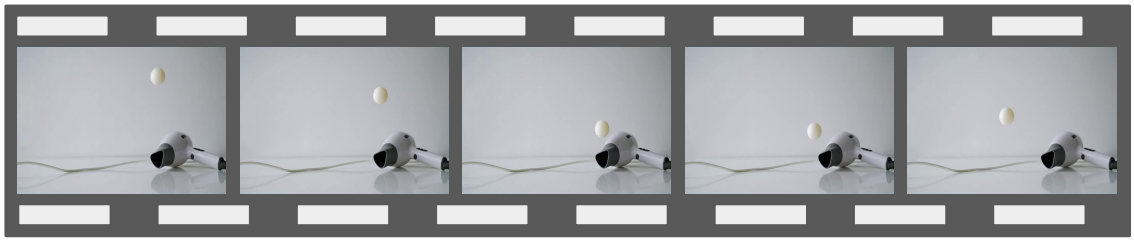
1) The ping pong ball levitates above the hair dryer, but it glows, spins in place with no air movement, and occasionally floats side to side or hovers even after the hair dryer is off. The ball might even rise higher when the dryer turns off or move in impossible ways, completely ignoring gravity and airflow.

2) The ball hovers, but its motion is inconsistent with airflow: it may drift far outside the airstream and still stay aloft, or it falls very slowly after the dryer is turned off, appearing to ignore gravity for several seconds. The ball might also bounce up and down repeatedly without any plausible reason.

3) The ball mostly stays in the air stream, levitating as expected, but there may be a slight lag between the dryer turning off and the ball beginning to fall, or the ball's motion is a bit jerky when it stabilizes in the air. The fall looks mostly natural but might be a bit too smooth or too abrupt.

4) The ping pong ball remains directly above the hair dryer, stably floating in the upward air stream; when the hair dryer is turned off, the ball immediately and naturally falls straight down under gravity. The timing and motion match real-world expectations of the Bernoulli effect and airflow.

Wan2.1
SA: 1
PC: 0.67



PhyT2V
SA: 1
PC: 0.67

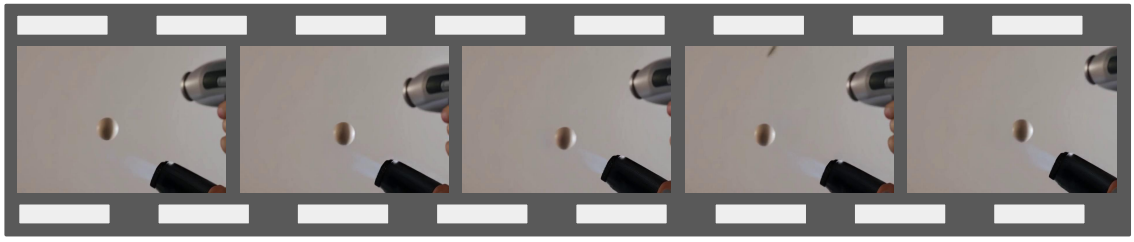


Figure 6. Domain: Fluids. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: When light passes from one material to another, it changes direction.
Prompt: A ray of light travels from a block of glass into air, bending as it exits; it travels at an altered angle. The background is a workshop with a workbench and various tools.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: light ray, glass block
ACTION: Light ray exits glass into air and bends away from normal.

KEY SEQ IDENTIFICATION

Q. Does the light ray bend at the boundary between glass and air? Yes

Q. Is the angle of the light ray in air different from its angle in glass? Yes

ORDER VERIFICATION

Retr. prompt: Ray going from glass to air - the point of ray enters air.

Description1: the ray of light moves through the glass block.

Description2: the ray of light exits the glass block into air, bending away from the normal; the angle of the ray changes, showing refraction, while the background and other objects remain the same.

OVERALL NATURALNESS EVALUATION

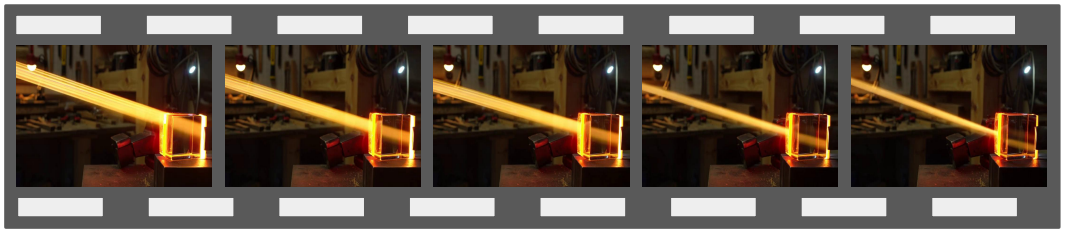
1) The light ray passes through the glass and into the air without changing direction at all, or bends in an impossible way (e.g., it loops, zigzags, or splits into multiple rays of different colors spontaneously). Alternatively, the ray transforms into a physical object or displays magical effects like sparking or levitating tools in the workshop.

2) The light ray noticeably ignores the interface between glass and air: it continues in a straight line, or bends in the wrong direction (toward the normal instead of away), or moves erratically for much of its path. The timing or sequence is inconsistent with normal behavior, such as the ray pausing mid-air or reflecting off surfaces that should be transparent.

3) The ray exits the glass and bends at the interface, but the angle is slightly off (e.g., a small deviation from what Snell's Law would predict), or the bending animation seems abrupt or a little delayed. There might be a tiny visual glitch, like the ray edge blurring, but the overall sequence matches the teaching point.

4) The ray clearly changes direction as it exits the glass block into air, following the correct angle relative to the normal—bending away as expected. The transition is smooth and matches the physical principle, with no distracting artifacts or unrealistic motion. The background elements (workbench, tools) remain neutral and do not interfere with the physics depiction.

Wan2.1
SA: 1
PC: 0.67



PhyT2V
SA: 0.67
PC: 0.67

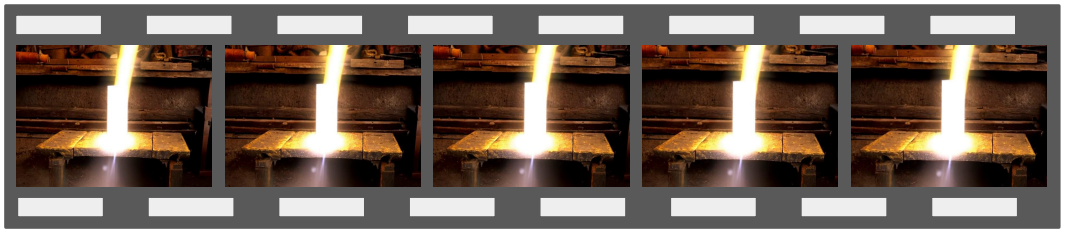


Figure 7. Domain: Optics. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: A capacitor stores electrical energy in an electric field created between its plates.
Prompt: Two metal plates are positioned close to each other. An arrow visually indicates the flow of electrons from one plate to the other, creating a visible electric field between the plates. A faint glow emanates from the region between the plates.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: metal plate, electrons, electric field
ACTION: Electrons flow between plates, generating electric field glow.

KEY SEQ IDENTIFICATION

Q. Is there a visible electric field (such as lines or a glow) shown between the two plates? Yes

Q. Is there an arrow showing electrons moving from one plate to the other?Yes

ORDER VERIFICATION

Retr. prompt: Middle Frame

Description1: From the first to the middle frame, the plates start out neutral, and then one plate becomes more negatively charged while the other becomes more positively charged. The electric field between the plates begins to form, and the faint glow starts to appear.

Description2: From the middle frame to the last frame, the electric field between the plates becomes stronger and the faint glow intensifies, indicating increased energy storage. The charge separation on the plates is now at its maximum, with the field fully established.

OVERALL NATURALNESS EVALUATION

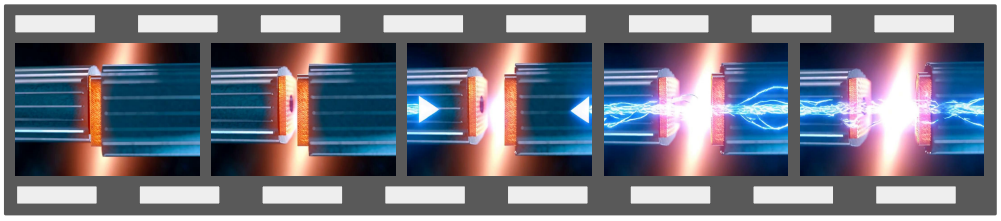
1) The plates float in midair emitting swirling, multicolored lightning bolts. Electrons visibly teleport between plates, and the plates levitate or morph shape. The 'electric field' manifests as an animated, pulsating wave that lifts objects or produces magical effects. The glow between plates pulses to the beat of music. None of these effects correspond to real physical behavior.

2) Electrons are shown moving in continuous loops between the plates even after the power source is removed, or the electric field causes the plates to attract or move towards each other dramatically. The glow becomes intensely bright, illuminating the whole scene. Arrows reverse direction randomly, and the plates spark or vibrate violently. These effects clearly contradict basic physical expectations for a capacitor.

3) The electron flow and field formation are correct, but there is a slight delay between electron motion and the appearance of the electric field. The faint glow between the plates may fade in or out a bit too slowly, or the electron arrow wiggles awkwardly. The sequence is almost correct, with only minor, brief timing or motion oddities.

4) Electrons are shown moving from one plate to the other in a brief, clear burst, with the electric field appearing steadily and symmetrically between the plates. The faint glow grows smoothly as the field builds, with all elements behaving as expected. The sequence accurately reflects the physical process of energy storage in a capacitor, with no noticeable deviations.

Wan2.1
SA: 1
PC: 0.67



PhyT2V
SA: 0.34
PC: 0.34

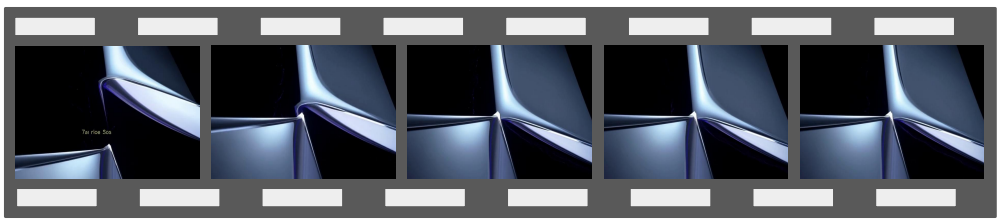


Figure 8. Domain: Electromagnetism. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.

Teaching point: Resonance occurs when a system is driven by an external force at its natural frequency, leading to large amplitude oscillations.
Prompt: Show a child sitting on a swing. Initially, the pushes are irregular, and the swing barely moves. Then, demonstrate the child being pushed at regular intervals matching the swing\u2019s natural back-and-forth motion. With each well-timed push, the swing\u2019s amplitude increases noticeably. Clearly highlight that the energy transfer is most efficient when the pushing frequency matches the swing\u2019s natural frequency.

SEMANTIC ADHERENCE

"The main interactions and objects involved in it."

OBJECTS: child, swing
ACTION: Regular pushes matching swing's frequency increase its amplitude efficiently.

KEY SEQ IDENTIFICATION

Q. Is the swing reaching a much higher amplitude when the pushes are given at regular intervals matching its natural frequency? Yes

Q. IDoes the swing remain at a low amplitude when the pushes are irregular? Yes

ORDER VERIFICATION

Retr. prompt: Show the frame where the swing first begins to noticeably increase its arc due to well-timed pushes (after the irregular pushes).

Description1: Between the first frame (where the swing barely moves with irregular pushes) and the retrieval frame (the first frame showing well-timed pushes), the swing\u2019s arc starts to noticeably increase, and the child swings higher than before.

Description2: Between the retrieval frame (first noticeable increase in arc) and the last frame (after several well-timed pushes), the swing\u2019s arc grows even larger, and the child reaches a much greater height, clearly showing the effect of resonance.

OVERALL NATURALNESS EVALUATION

1) The swing begins to levitate, spin, or move in impossible ways regardless of how or when pushes are applied. The child might fly off at random, or the swing reaches infinite amplitude instantly. There are magical effects such as glowing energy waves, or the swing responds to pushes even when no one is pushing, completely disregarding the laws of motion.

2) The swing's motion does not correspond at all to the timing or strength of pushes: for example, the swing slows down or stops entirely when pushed at its natural frequency, or gains maximum height from random, weak, or mistimed pushes. The amplitude might decrease or stay constant no matter how well-timed the pushes are, contradicting resonance. The sequence shows persistent impossible behaviors (e.g., the swing passes through the support structure, or pushes act with a visible delay of many seconds).

3) Most of the animation matches expected behavior, but there are small flaws: the swing might respond a bit too quickly or slowly to changes in push timing, or the amplitude increases are slightly exaggerated. There could be a brief moment where a mistimed push has a larger effect than expected, or the swing's motion looks a little awkward or jerky, but overall the resonance effect is clear and mostly accurate.

4) The swing only gains significant amplitude when pushed at regular intervals matching its natural frequency; irregular pushes have little effect as expected. The amplitude builds up gradually over several well-timed pushes, and the swing's motion is smooth and physically plausible. Energy transfer is clearly most efficient at resonance, and all details (timing, amplitude growth, damping if included) faithfully reflect the real physics of resonance in a playground swing.

Wan2.1
SA: 1
PC: 0.67

PhyT2V
SA: 1
PC: 0.67

Figure 9. Domain: Waves & Oscillations. Questions used for evaluation along with outputs from Wan2.1 and PhyT2V.