Physics-Aware Video Generation: A Survey of First-Principles and Common-Sense Approaches

Mickael Pires[0009−0007−5651−6814] and Rodrigo de Bem[0000−0002−4860−0115]

Center of Computational Sciences, Federal University of Rio Grande,
Rio Grande, Brazil
{mickael.pires,rodrigobem}@furg.br

Abstract. Recent text- and image-conditioned video generators can produce visually convincing sequences, but their ability to preserve physically plausible dynamics remains limited. This survey focuses on simulator-free approaches to physics-aware video generation, i.e., methods that improve or assess physical plausibility without relying on explicit external physics engines. We organize the literature into two main families. The first consists of first-principles approaches, where physical structure is introduced through formal mechanics, analytic constraints, physical rewards, or benchmark scenarios grounded in laws such as Newtonian dynamics, energy conservation, and Lagrangian formulations. The second consists of common-sense approaches, where physical plausibility is encouraged through data-driven learning, language-based reasoning, self-supervised world modeling, or human-aligned judgments of expected physical behavior. Based on this taxonomy, we discuss the trade-offs between the precision and interpretability of first-principles methods and the scalability of common-sense methods in open-domain settings. We also highlight recurring limitations, including the difficulty of modeling discontinuous events such as collisions, the gap between perceptual plausibility and physical correctness, and the lack of robust out-of-distribution evaluation. We conclude by outlining research directions toward hybrid methods that combine explicit physical structure with scalable learned representations.

Keywords: Simulator-Free · First-Principles · Physical Common-Sense · Physics-Aware · Video Generation

1 Introduction

Recent generative video models can produce visually convincing sequences from text prompts, as in text-to-video systems such as Sora [16, 13] and CogVideoX [26], or from image conditioning, as in image-to-video models such as Stable Video Diffusion [4]. However, visual realism does not imply physical correctness. Generated videos may appear plausible at first glance while still violating basic constraints of motion, contact, causality, conservation, or object permanence. This gap is particularly problematic for applications in which video generation

is expected to support simulation, prediction, robotics, autonomous systems, or scientific reasoning rather than only visual synthesis.

Recent benchmarks have made this limitation more explicit. VideoPhy [2], PhyGenBench [14], and VideoPhy-2 [3] evaluate whether text-to-video models follow physical common sense under prompted real-world interactions. Other benchmarks focus on image-conditioned generation and physical reasoning. Physics-IQ [15] tests whether generative video models capture physical principles across controlled scenarios, while Morpheus [29] evaluates generated videos against real physical experiments and conservation-law-based criteria. Together, these works show that current models can achieve high perceptual quality while still failing on physically meaningful behavior. They also reveal a fragmented evaluation landscape, where each benchmark introduces different scenarios, assumptions, and metrics.

This survey focuses on simulator-free approaches to physics-aware video generation. By simulator-free, we refer to methods that improve, guide, or evaluate physical plausibility without relying on an explicit external physics engine. This scope is narrower than general physics-aware generation, but it is important: many recent video models aim to acquire physical structure from data, language, learned representations, rewards, or analytic constraints rather than from full simulation pipelines.

Existing surveys already provide useful broader perspectives. Liu et al. [12] organize physics-aware generation according to whether physical knowledge is incorporated through explicit simulation or implicit learning, while Lin et al. [11] analyze physics cognition in video generation from a cognitive-science perspective. In contrast, this paper examines simulator-free video generation through a different axis: whether physical awareness is induced from first-principles structure or from learned common-sense representations. First-principles approaches introduce explicit physical assumptions, constraints, rewards, or analytic formulations. Common-sense approaches instead rely on data-driven learning, language supervision, self-supervised world modeling, or human-aligned judgments of plausible physical behavior.

The remainder of the paper is organized as follows. Section 2 introduces the proposed taxonomy. Section 2.1 reviews simulator-free first-principles approaches, including methods and benchmarks that encode or evaluate formal physical structure. Section 2.2 reviews common-sense approaches, including methods and benchmarks that rely on learned physical priors or semantic plausibility. Finally, Section 3 discusses the main trade-offs, limitations, and open research directions suggested by this taxonomy.

2 Taxonomy

This section organizes simulator-free physics-aware video generation into two families: first-principles and common-sense approaches. The distinction is methodological rather than merely thematic. It is detailed in Figure 1 and provides the structural framework for the following literature review.

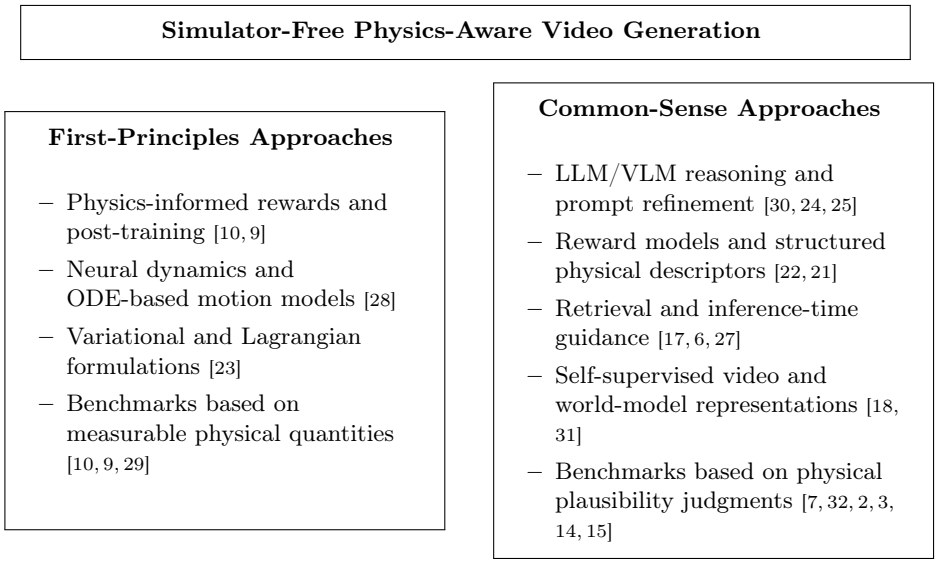

Fig. 1. Taxonomy of simulator-free physics-aware video generation. First-principles approaches introduce or evaluate explicit physical structure, whereas common-sense approaches rely on learned priors, language reasoning, retrieval, reward models, or evaluator judgments to improve physical plausibility.

2.1 First-Principles Approaches

First-principles approaches aim to make video generation physically grounded by introducing explicit structure from classical mechanics or by evaluating generated motion against physically meaningful constraints. In this survey, we include in this family methods and benchmarks that rely on Newtonian dynamics, conservation laws, differential equations, Lagrangian formulations, or physics-informed reward functions. The goal is not only to produce motion that looks plausible, but to constrain or assess whether the generated dynamics are compatible with known physical regularities.

This category is narrower than general physical common sense. A model that produces visually convincing falling objects, splashing fluids, or object interactions is not necessarily first-principles-aware. To be included here, a method must either encode physical structure into training, guidance, reward design, or latent dynamics or evaluate videos against explicit physical quantities such as trajectories, velocities, accelerations, energy, momentum, or governing equations. This distinction matters because first-principles methods are usually more interpretable than purely perceptual approaches, but they often remain limited to controlled scenarios, simplified dynamics, or restricted physical systems.

Existing works in this family can be grouped into two directions. The first uses post-training, reward modeling, or diagnostic evaluation to align generative video models with simple dynamical laws, such as free fall, projectile motion,

or ramp sliding [10, 9]. The second introduces stronger architectural or latent-space constraints, for example, by modeling state transitions through Neural Ordinary Differential Equations or variational formulations inspired by the least action principle [28, 23]. We also discuss scientific video generation methods that attempt to extract latent physical structure from observed physical processes [5]. Finally, we review benchmarks that quantify whether generated videos respect physical laws or are physically meaningful invariants.

Methods The simplest first-principles setting considered in recent work is object free fall. PISA [10] studies whether generative video models can be post-trained to model this basic Newtonian phenomenon in an image-to-video setting. The method fine-tunes a pretrained video generator on a small set of simulated falling-object videos and further improves the behavior through reward modeling. This is a useful starting point because free fall has a clear physical structure: object position, velocity, and acceleration can be measured against expected motion. However, the setting is also deliberately narrow. PISA reports generalization difficulties when the model is tested outside the training distribution, including unseen depths and heights. The work, therefore, shows both the promise and the fragility of post-training: simple physical behavior can be induced, but it does not automatically generalize to broader physical reasoning.

NewtonRewards [9] extends this idea from a single free-fall setting to a broader set of Newtonian motion primitives. Instead of relying only on visual realism, the framework converts generated videos into measurable proxies for physical variables. RAFT [19] is used to estimate motion, while V-JEPA2 [1] features are used as a proxy for object identity or mass-related consistency. These proxies support two reward functions: a Newtonian kinematic reward, which encourages constant-acceleration behavior across five motion primitives, and a mass conservation reward, which discourages degenerate solutions such as disappearing, freezing, or visually unstable objects. This makes the alignment objective more explicitly physical than generic preference-based video reward modeling. The limitation is that the physical signal is only as reliable as the proxy models used to estimate it. Moreover, the approach remains focused on simplified rigid-body scenarios and does not solve harder cases such as multi-object contact, collisions, deformation, or dissipative dynamics.

A more structural alternative is to place physical dynamics inside the generative pipeline rather than only scoring the final video. NewtonGen [28] argues that current video diffusion models tend to learn motion from visual surface patterns rather than from underlying dynamics. To address this, the authors introduce Neural Newtonian Dynamics, a module based on learnable second-order ordinary differential equations. The module is trained on a small set of clean synthetic videos and then predicts future physical states from initial conditions. These predicted states are rendered as optical flow and used to guide a motion-controlled video diffusion model. This design is stronger than post-hoc evaluation because it attempts to represent the evolution of a physical state before video synthesis. However, the continuous ODE formulation is naturally better suited to smooth

dynamics than to discontinuous events such as impacts and collisions. Its scalability to more complex scenes, multiple interacting objects, and open-domain text-to-video generation remains unresolved.

LaWM [23] introduces a related but more mechanics-driven formulation. Instead of modeling transitions with a generic latent dynamics module, it uses a latent variational integrator derived from a learned Lagrangian action functional. This connects video prediction to the least action principle, a central idea in classical mechanics. The motivation is clear: long-horizon prediction often suffers from accumulated error and physical drift, and a variational transition rule can help preserve more coherent dynamics over time. The method is relevant to this survey because it shows how first-principles structure can be introduced in latent space rather than imposed directly on pixels. Its weakness is also clear. The formulation is most natural for conservative or near-conservative systems, and it may struggle with dissipative effects, non-rigid deformation, friction, air resistance, or complex contacts. It also assumes that the learned latent coordinates capture physically meaningful degrees of freedom. If the encoder fails to recover those variables from visual observations, the Lagrangian transition mechanism cannot fix the representation.

Latent Knowledge-Guided Video Diffusion [5] represents a boundary case in this subsection. Unlike NewtonGen or LaWM, it does not explicitly encode a closed-form mechanics model into the generator. Instead, it aims to recover latent scientific structure from observed physical processes and use that structure to guide video diffusion. The method targets scientific phenomena such as fluid dynamics and typhoon evolution from a single initial frame. It extracts static knowledge through a masked autoencoder with a Vision Transformer backbone and dynamic knowledge through optical-flow-based motion estimation. These static and dynamic embeddings are then projected into the conditioning space of a pretrained video diffusion model and injected through parameter-efficient adaptation. This is useful for scientific video generation because textual prompts alone are often insufficient to describe the underlying evolution of a physical process. However, the method remains dependent on the information available in a single initial frame and on the reliability of pseudo-ground-truth optical flow. For complex systems with hidden variables or chaotic dynamics, the initial visual condition may not uniquely determine future evolution.

Overall, the methods in this subsection show that first-principles guidance can improve physical consistency, but mostly under restricted assumptions. Reward-based approaches are easier to attach to existing video generators, yet they depend heavily on measurement proxies and simulated training distributions. Latent dynamics approaches provide a more principled mechanism, but their assumptions often break under discontinuities, non-conservative systems, deformation, and dense multi-object interaction. This is the central trade-off of first-principles physics-aware generation: stronger physical structure usually comes at the cost of narrower scope.

Benchmarks Benchmarks for first-principles physics-aware video generation differ from general perceptual benchmarks because they evaluate physical quantities rather than only visual quality or semantic alignment. A useful benchmark in this family should make the relevant physical variables observable or measurable, define what counts as physically valid behavior, and distinguish visual plausibility from physical correctness. Recent work has moved in this direction by evaluating trajectories, velocities, accelerations, conservation laws, and governing equations.

Morpheus [29] is one of the clearest examples of this shift. It evaluates whether generated videos follow physical laws by grounding the assessment in real-world physical experiments. Rather than depending only on human judgments, multimodal language model scores, or pixel-level similarity, Morpheus extracts physical representations from video and evaluates them against conservation-law-based criteria. Its dataset contains real-world videos of Newtonian mechanics phenomena, including systems such as pendulums, projectile motion, and collisions. The evaluation pipeline tracks objects, estimates their trajectories, and uses physics-informed modeling to assess whether the generated dynamics are consistent with governing equations and physical invariants such as energy or momentum. This makes the benchmark more interpretable than purely perceptual evaluation. However, its scores remain proxies for physical validity. The evaluation depends on the reliability of object tracking, depth estimation, and fitted physical models, and it necessarily simplifies effects such as friction, air resistance, and other dissipative forces.

PisaBench [10] provides a narrower but more controlled diagnostic benchmark. It evaluates object free fall in image-to-video generation using both spatial and physics-specific metrics. The benchmark includes trajectory-based measures, such as position error, Chamfer Distance, and Intersection over Union, as well as velocity and acceleration errors. This design is useful because it makes the failure mode concrete: a video can look plausible while still producing the wrong falling trajectory or acceleration profile. The weakness is scope. PisaBench focuses on a single motion primitive and inherits ambiguity from the image-to-video setting, where object size, camera geometry, and initial height are not always recoverable from one frame. As a result, it is valuable as a diagnostic test for basic Newtonian behavior but not sufficient as a general benchmark for physical reasoning.

NewtonBench-60K [9] broadens the evaluation beyond vertical free fall. It contains simulated videos covering five Newtonian motion primitives: free fall, horizontal throw, parabolic throw, upward ramp sliding, and downward ramp sliding. This makes it more diverse than single-task free-fall evaluation and allows models to be tested across different initial conditions, including out-of-distribution settings. The benchmark is useful because it connects evaluation to measurable kinematic regularities rather than relying only on visual preference. Still, its diversity should not be overstated. The scenarios remain simplified, with controlled rigid-body dynamics, fixed visual assumptions, and limited interaction complexity. It therefore tests whether models can learn or preserve

basic Newtonian motion, not whether they can handle the full range of physical phenomena present in open-world videos.

Taken together, these benchmarks expose an important limitation of current physics-aware video evaluation. First-principles benchmarks are more objective than general perceptual metrics, but they usually become objective by narrowing the problem. Free fall, projectile motion, ramp sliding, pendulums, and simple collisions are measurable and interpretable, but they do not cover the full difficulty of open-domain physical video generation. Future benchmarks need to preserve the interpretability of first-principles evaluation while expanding toward richer contact events, deformable objects, dissipative systems, multi-object interaction, and camera or scene complexity.

2.2 Common-sense Approaches

Physical common sense refers to the ability to judge whether object motion, contact, causality, state changes, and interactions are plausible under everyday physical expectations [2]. Unlike first-principles physics, this form of physical awareness is usually not expressed through explicit equations or conservation constraints. Instead, it is learned from visual experience, language, human feedback, retrieval, self-supervised video representations, or evaluator judgments. In video generation, common-sense approaches therefore aim to reduce physically implausible outputs by aligning generated videos with expected real-world behavior rather than by enforcing formal physical laws.

This family is broader and less rigid than the first-principles category. A method may be considered common-sense physics-aware when it improves physical plausibility through language reasoning, semantic conditioning, reward modeling, retrieval, world-model representations, or benchmark-driven feedback, even if it does not explicitly model acceleration, energy, momentum, or governing equations. This flexibility makes common-sense approaches more compatible with open-domain video generation, where scenes may contain deformable objects, fluids, occlusions, semantic actions, and diverse camera motions. The cost is weaker interpretability: these methods often improve plausibility without proving that the generator has learned physical laws rather than correlations associated with realistic-looking videos.

Existing works in this family can be grouped into four directions. The first uses language models or vision-language models to infer physical context, decompose prompts, predict coarse motion plans, or refine generation through reasoning [30, 24, 25]. The second uses reward models or structured physical descriptors to guide text-to-video generation toward more plausible dynamics [22, 21]. The third relies on geometry, retrieval, or inference-time steering to provide external motion cues without fully retraining the generator [20, 17, 6]. The fourth aligns video generators with self-supervised video representations or latent world models, attempting to transfer physical regularities from video understanding models into generative models [18, 27, 31].

Methods A first group of methods uses language-based reasoning to make physical context more explicit before or during generation. DiffPhy [30] uses an LLM to infer physical context from the input prompt and uses this information to guide video diffusion. PhyT2V [24] follows a related direction, using LLM-guided iterative self-refinement to improve physics-grounded text-to-video generation. VLIPP [25] extends this idea to image-to-video generation by using a vision-language model as a coarse motion planner that predicts rough trajectories or state changes before synthesis. These methods are attractive because they can be attached to existing generative pipelines without requiring a full physical simulator. However, their physical reasoning remains mediated by language or vision-language models. As a result, they may improve prompt interpretation and high-level causal structure while still failing on low-level temporal artifacts, trajectory consistency, contact dynamics, or long-horizon motion.

Reward-based methods attempt to move beyond prompt refinement by directly scoring generated videos according to physical plausibility. PhysCorr [22] introduces a physics-oriented reward formulation for text-to-video generation, targeting both object-level stability and interaction-level coherence. This type of method is important because it creates an optimization signal that is closer to the generated video than a rewritten prompt. Nevertheless, reward-based alignment inherits the weaknesses of the reward model. If the reward model detects only superficial cues, the generator can improve the score without truly improving physical consistency. Moreover, optimizing for physical plausibility may create trade-offs with semantic alignment, visual fidelity, or diversity when the reward is not well calibrated.

Another direction introduces structured physical information into the generative process. WISA [21] decomposes physical principles into textual descriptions, qualitative categories, and quantitative properties, and injects these attributes through a Mixture-of-Physical-Experts Attention mechanism. The associated WISA-32K dataset contains 32,000 videos covering 17 physical laws across dynamics, thermodynamics, and optics. This is stronger than generic text conditioning because it gives the model a structured interface to physical concepts. However, it still does not enforce physical laws in the first-principles sense. The model is guided by semantic and categorical descriptions of physical phenomena, so its behavior depends on the coverage and quality of the dataset and on whether the chosen physical categories transfer to open-domain prompts.

PhysVideo [20] addresses a different limitation: many physical failures in generated videos arise because motion occurs in three-dimensional space, while standard video models observe only view-dependent two-dimensional projections. The method uses cross-view geometry guidance, first generating physics-aware foreground dynamics from multiple orthogonal viewpoints and then using these signals to guide final video synthesis with background context. This makes the method more spatially grounded than text-only approaches and better suited to cases where motion consistency depends on three-dimensional structure. Its limitation is that it introduces additional assumptions and complexity: the generation process depends on foreground modeling, cross-view consistency, and

explicit physical attributes, which may restrict usability in highly dynamic open-world scenes.

Training-free and retrieval-based approaches avoid expensive model retraining by steering generation at inference time. RagMe [17] retrieves external video demonstrations and uses them as motion references to improve motion realism in generated videos. This is a practical strategy because real videos provide examples of plausible dynamics that may be difficult to encode manually. Reason-then-guide [6] takes a complementary route: instead of retrieving positive examples, it reasons about physically implausible alternatives and guides generation away from them. These methods are useful because they can be applied to pretrained generators, but their success depends on external signals. Retrieval-based methods require relevant motion examples, while counterfactual guidance depends on the quality of the reasoning process used to identify implausible behavior.

A final group of methods attempts to transfer physical regularities from self-supervised video representations or latent world models into video generation. Phantom [18] jointly models visual content and latent physical dynamics so that generation is conditioned not only on appearance but also on inferred physical state. WMReward [27] uses a latent world model to provide inference-time physical alignment signals. VideoREPA [31] distills physics-related representation structure from video foundation models into text-to-video generators through token relation alignment. These methods are promising because they target the internal representation gap between video understanding models and video generators. Their weakness is that the transferred physical knowledge is only as reliable as the teacher representation or latent world model. If the foundation model encodes correlations rather than robust physical structure, the generator may inherit the same limitations.

Overall, common-sense methods are more scalable than first-principles approaches because they do not require a closed-form description of the physical system. They can operate on natural prompts, diverse scenes, and open-domain video models. However, this flexibility is also their main weakness. Most methods improve physical plausibility indirectly, through language, learned rewards, retrieved examples, or representation alignment. They therefore make generated videos more convincing, but they do not necessarily establish physical correctness.

Benchmarks Benchmarks for common-sense physics-aware video generation evaluate whether generated videos look and behave plausibly under human or model-based judgments. They differ from first-principles benchmarks because they usually do not require direct measurement of physical quantities such as acceleration, force, energy, or momentum. Instead, they assess semantic adherence, temporal consistency, physical plausibility, causal ordering, and whether generated events match expected real-world behavior.

VBench [7] is a broad perceptual benchmark for video generation. It decomposes video quality into 16 dimensions, including subject consistency, background consistency, temporal flickering, motion smoothness, dynamic degree,

aesthetic quality, imaging quality, object class, multiple objects, human action, color, spatial relationship, scene composition, appearance style, temporal style, and overall consistency. VBench is valuable because it provides a fine-grained evaluation suite aligned with human perception. However, it was not designed specifically to test physical understanding. A model can score well on VBench while still violating contact dynamics, object permanence, causal ordering, or material behavior.

VBench 2.0 [32] addresses this limitation by extending the benchmark toward intrinsic faithfulness. It introduces five additional dimensions: human fidelity, controllability, creativity, physics, and common sense. The physics dimension considers phenomena such as state changes, geometric consistency, deformation, thermodynamic change, material change, and other physical transformations. The common-sense dimension evaluates whether generated actions produce plausible consequences. This makes VBench 2.0 more relevant to physics-aware generation than the original VBench. Still, it remains a broad benchmark rather than a physics-only evaluation suite, so its physical dimensions must be interpreted as part of a larger perceptual and semantic assessment.

VideoPhy [2] focuses more directly on physical common sense in text-to-video generation. It evaluates both semantic adherence and whether generated videos follow physical commonsense for real-world activities involving object and material interactions. This distinction is important because a video may match the prompt semantically while still failing physically. VideoPhy also introduced an automated evaluator to reduce the cost of large-scale human evaluation. However, its scoring remains relatively coarse: it is useful for detecting obvious physical failures, but it does not provide detailed measurements of why a failure occurs or which physical rule is violated.

VideoPhy 2 [3] improves this direction by making the benchmark more action-centric and by evaluating physical rule grounding in addition to semantic adherence and physical common sense. Rather than only asking whether a video appears physically plausible, it examines whether the generated action follows the physical rule implied by the prompt. This makes the benchmark more diagnostic than the original VideoPhy. Even so, VideoPhy 2 remains a commonsense benchmark: it evaluates physical plausibility and rule adherence through annotated actions and evaluator judgments rather than through formal physical measurements.

PhyGenBench [14] evaluates physical commonsense correctness through structured physical scenarios. It contains 160 prompts across 27 physical laws and four domains, and is paired with PhyGenEval, a hierarchical evaluation framework based on language and vision-language models. The benchmark is useful because it decomposes evaluation into physical phenomena and causal verification, rather than treating physical plausibility as a single undifferentiated score. Its main limitation is that it still depends on model-based evaluators to interpret generated video content. As with other VLM-based metrics, evaluation quality depends on whether the evaluator can reliably detect temporal events, object interactions, and causal sequences.

Physics-IQ [15] provides a stronger test of whether generative video models capture physical principles rather than only producing visually realistic content. It evaluates generated videos across five domains: solid mechanics, fluid dynamics, optics, magnetism, and thermodynamics. The dataset contains 396 real-world videos spanning 66 scenarios, with repeated recordings and multiple viewpoints used to estimate expected physical variability. This design is important because it distinguishes physically meaningful deviations from natural variation in real physical systems. Physics-IQ also separates physical comprehension from visual authenticity, showing that perceptual realism does not necessarily imply physical understanding. Compared with VideoPhy or PhyGenBench, Physics-IQ is closer to first-principles evaluation, but we consider it still belongs in this subsection because it is often used to assess whether open-domain generative video models exhibit learned physical understanding rather than explicitly enforced mechanics. Indeed, despite introducing the concept of *physical variance* to capture differences between repeated takes, Physics-IQ evaluates empirical statistical consistency instead of validating underlying mathematical equations or explicit physical laws.

Taken together, these benchmarks show that the field has moved beyond generic visual quality metrics, but the evaluation problem remains unresolved. General benchmarks such as VBench are useful for perceptual quality but weak for physical reasoning. Specialized benchmarks such as VideoPhy, VideoPhy 2, and PhyGenBench better target physical common sense, but they rely heavily on evaluator judgments and discrete scenario design. Physics-IQ introduces a more rigorous comparison against real physical variability, but it still cannot cover the full diversity of open-world physical interactions. The central challenge is therefore to design evaluations that are broad enough for open-domain video generation while still precise enough to distinguish physical plausibility from visual realism. Finally, in this section, Table 1 summarizes the main benchmarks examined in the two groups and the methods that best fit each.

3 Discussion and Conclusion

The proposed taxonomy exposes a central trade-off in simulator-free physics-aware video generation, illustrated in Fig. 2. First-principles approaches offer clearer interpretability because they connect generation or evaluation to explicit physical quantities, such as trajectories, velocities, accelerations, conservation laws, or governing equations. This makes their failures easier to diagnose and their evaluation less dependent on subjective judgments. However, this strength comes from restricting the problem. Most current first-principles methods focus on simplified dynamics, such as free fall, projectile motion, ramp sliding, pendulums, or other controlled systems [10, 9, 29]. These settings are useful diagnostic tools, but they remain far from the complexity of open-domain video generation, where scenes may involve occlusion, deformable objects, fluids, friction, contact-rich interaction, camera motion, and ambiguous initial conditions.

Table 1. Coverage and qualitative method fit of physics-aware video generation benchmarks. **Scope** [C = Controlled, OD = Open-domain], **FP** (First-Principles) and **PCS** (Physical-Common-Sense) [F = focused, P = partial, N = not targeted]. The method order indicates benchmark fit, not leaderboard performance.

	Benchmark	Scope	FP	PCS	Best-matched methods	Limitation
First-principles	PisaBench [10]	C	F	N	Post-training [10] / Newtonian rewards [9]	Free fall only.
	NewtonBench-60K [9]	C	F	N	Newtonian rewards [9]	Simplified rigid bodies.
	Morpheus [29]	C	F	P	Physical consistency [29] / latent dynamics [28, 23]	Tracking/model fitting.
Physical-common-sense	PhyGenBench [14]	OD	P	F	Common-sense LLM/VLM reasoning [30, 24, 25] / structured descriptors [21]	VLM verification.
	VideoPhy [2]	OD	N	F	Common-sense alignment [18, 31] / retrieval/guidance [17, 6]	Coarse diagnosis.
	VideoPhy 2 [3]	OD	P	F	Rule-grounded reasoning [30, 24] / inference-time alignment [27]	Judgment-based.
	VBench 2.0 [32]	OD	P	F	Open-domain plausibility [22, 17] / world-model alignment [18, 27]	Broad suite.
	VBench [7]	OD	N	P	Motion realism [17, 20] / general video quality	Mostly perceptual.
	Physics-IQ [15]	C	P	F	Physical-principle comprehension [15] / representation/world-model alignment [18, 27]	Limited scenarios.

Common-sense approaches address a different part of the problem. They scale more naturally to open-domain prompts because they rely on learned visual priors, language reasoning, retrieval, reward models, self-supervised representations, or human-aligned judgments. This makes them better suited to broad text-to-video and image-to-video generation pipelines. However, their notion of physical correctness is usually indirect. A video is treated as physically plausible when it matches expected visual patterns, satisfies an evaluator, or aligns with human judgments of everyday behavior [2, 3, 14]. This is useful, but it is not the same as obeying physical laws. A generated sequence may look plausible while still violating conservation, causality, object permanence, or consistent motion.

This distinction matters because perceptual plausibility can hide physical error. Human observers often tolerate or miss violations when the generated motion is visually smooth, semantically coherent, or short enough to avoid obvious contradictions. Conversely, physically correct behavior may look unintuitive when visual cues are incomplete, as in cases involving perspective, hidden forces, material properties, or air resistance. Benchmarks such as Physics-IQ and Morpheus are important because they make this gap explicit: visual realism does not necessarily imply physical understanding [15, 29]. For this reason, future evaluation should avoid treating physical plausibility as a single scalar score. It should separate semantic adherence, perceptual quality, temporal coherence, common-sense plausibility, and measurable physical consistency.

A second limitation concerns out-of-distribution generalization. Several recent studies suggest that current video generators can reproduce familiar physical patterns but struggle to extrapolate to unseen conditions, new combinations of

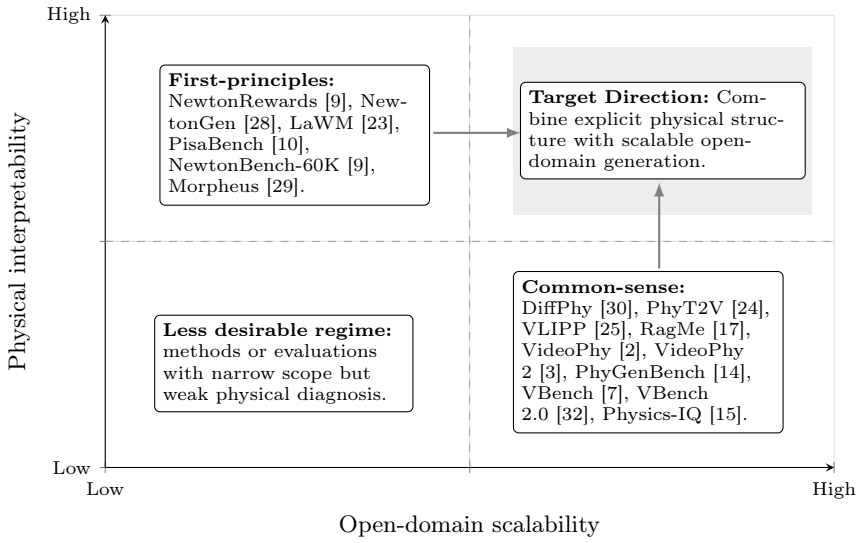

Fig. 2. Conceptual trade-off between physical interpretability and open-domain scalability in simulator-free physics-aware video generation. First-principles methods and controlled benchmarks provide stronger diagnostic value but usually operate in restricted settings, whereas common-sense methods and broad benchmarks scale better to open-domain generation but provide weaker evidence of physical correctness. The shaded region indicates the desired direction for future hybrid methods.

variables, or dynamics outside the training distribution [15, 10, 8]. This weakens strong claims that large video models have already learned general-purpose world models. The more defensible interpretation is narrower: current models can capture useful statistical regularities about motion and interaction, but they do not reliably abstract physical laws in a way that supports robust extrapolation.

The most promising direction is therefore not to choose between first-principles and common-sense approaches, but to combine them. Hybrid methods could use common-sense models for broad scene understanding, semantic grounding, and open-domain generation, while using first-principles constraints or physically meaningful latent variables to regularize motion, contact, conservation, and temporal evolution. Such methods would not need to simulate every detail of the world explicitly, but they should expose enough physical structure to make errors measurable and correctable. This requires better representations of state, force, mass, material properties, contact, and uncertainty, rather than relying only on pixels, prompts, or post-hoc evaluator scores.

Finally, the taxonomy also shows that the field needs stronger evaluation protocols. General benchmarks are useful for comparing video quality, but they are too coarse to establish physical understanding. Narrow first-principles benchmarks are more rigorous, but they risk becoming toy problems. A stronger evaluation suite should combine controlled physical diagnostics with open-domain

scenarios, report failure modes rather than only aggregate scores, and test extrapolation across unseen objects, materials, viewpoints, and initial conditions. Without this, progress in physics-aware video generation will remain difficult to interpret: models may become better at producing videos that look physically plausible without becoming better physical reasoners.

References

1. Assran, M., Bardes, A., Fan, D., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning (2025). https://doi.org/10.48550/arXiv.2506.09985
2. Bansal, H., Lin, Z., Xie, T., Zong, Z., et al.: Videophy: Evaluating physical commonsense for video generation. In: ICLR. pp. 102075–102121 (2025)
3. Bansal, H., Peng, C., Bitton, Y., et al.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation (2025). https://doi.org/10.48550/arXiv.2503.06800, accessed: 2026-06-17
4. Blattmann, A., Dockhorn, T., Kulal, S., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023). https://doi.org/10.48550/arXiv.2311.15127
5. Cao, Q., Li, X., Wang, D., et al.: Latent knowledge-guided video diffusion for scientific phenomena generation from a single initial frame. In: AAAI. pp. 2625–2633 (2026). https://doi.org/10.1609/aaai.v40i4.37250
6. Hao, Y., Chen, C., Mian, A.S., et al.: Enhancing physical plausibility in video generation by reasoning the implausibility (2025). https://doi.org/10.48550/arXiv.2509.24702
7. Huang, Z., He, Y., Yu, J., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR. pp. 21807–21818 (2024). https://doi.org/10.1109/cvpr52733.2024.02060
8. Kang, B., Yue, Y., Lu, R., et al.: How far is video generation from world model: A physical law perspective. In: ICML. pp. 28991–29017 (2025)
9. Le, M.Q., Zhu, Y., Kalogeiton, V., et al.: What about gravity in video generation? post-training newton's laws with verifiable rewards (2025). https://doi.org/10.48550/arXiv.2512.00425
10. Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., Xie, S.: Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In: ICML. pp. 35685–35709 (2025)
11. Lin, M., Wang, X., Wang, Y., Wang, S., Dai, F., Ding, P., et al.: Exploring the evolution of physics cognition in video generation: A survey (2025). https://doi.org/10.48550/arXiv.2503.21765
12. Liu, D., Zhang, J., Dinh, A.D., Park, E., Zhang, S., Mian, A., et al.: Generative physical ai in vision: A survey (2025). https://doi.org/10.48550/arXiv.2501.10928
13. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models (2024). https://doi.org/10.48550/arXiv.2402.17177
14. Meng, F., Liao, J., Tan, X., Lu, Q., Shao, W., Zhang, K., Cheng, Y., et al.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In: ICML. pp. 43781–43806 (2025)
15. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? In: WACV. pp. 948–958 (2026). https://doi.org/10.1109/wacv61042.2026.00099

16. OpenAI: Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/ (2024), accessed: 2026-05-31
17. Peruzzo, E., Xu, D., Xu, X., Shi, H., Sebe, N.: Ragme: Retrieval augmented video generation for enhanced motion realism. In: ACM MM. pp. 1081–1090 (2025). https://doi.org/10.1145/3731715.3733417
18. Shen, Y., Xiong, J., Yu, T., Lourentzou, I.: Phantom: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In: CVPR. pp. 11185–11194 (2026)
19. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020). https://doi.org/10.24963/ijcai.2021/662
20. Wang, C., Zhu, H., Tang, X., Luo, J., Jin, X., Chen, L., et al.: Physvideo: Physically plausible video generation with cross-view geometry guidance (2026). https://doi.org/10.48550/arXiv.2603.18639
21. Wang, J., Ma, A., Cao, K., Zheng, J., Feng, J., Zhang, Z., et al.: Wisa: World simulator assistant for physics-aware text-to-video generation. NeurIPS **38**, 5388–5416 (2026)
22. Wang, P., Wang, W., Li, Q.: Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection (2025). https://doi.org/10.48550/arXiv.2511.03997
23. Xiao, Q., Ghaffari, M.: Lawm: Least action world models for long-horizon physical consistency from visual observations (2026). https://doi.org/10.48550/arXiv.2605.08279
24. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In: CVPR. pp. 18826–18836 (2025). https://doi.org/10.1109/CVPR52734.2025.01754
25. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., et al.: Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In: ICCV. pp. 12360–12370 (2025). https://doi.org/10.1109/iccv51701.2025.01149
26. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR. pp. 83048–83077 (2025)
27. Yuan, J., Zhang, X., Friedrich, F., Beltran-Velez, N., Hall, M., Askari-Hemmat, R., et al.: Inference-time physics alignment of video generative models with latent world models (2026). https://doi.org/10.48550/arXiv.2601.10553
28. Yuan, Y., Wang, X., Wickremasinghe, T., et al.: Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics (2026). https://doi.org/10.48550/arXiv.2509.21309
29. Zhang, C., Cherniavskii, D., Zadaianchuk, A., et al.: Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments (2025). https://doi.org/10.48550/arXiv.2504.02918
30. Zhang, K., Xiao, C., Xu, J., et al.: Think before you diffuse: Infusing physical rules into video diffusion (2025). https://doi.org/10.48550/arXiv.2505.21653
31. Zhang, X., Liao, J., Zhang, S., et al.: Videorepa: Learning physics for video generation through relational alignment with foundation models. NeurIPS **38**, 122647–122676 (2026)
32. Zheng, D., Huang, Z., Liu, H., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness (2025). https://doi.org/10.48550/arXiv.2503.21755