VideoPhy2: Challenging Action-Centric Physical Commonsense Evaluation of Video Generation

Hritik Bansal*, **Clark Peng***, Yonatan Bitton*,
Roman Goldenberg, Aditya Grover, Kai-Wei Chang

ICLR 2026

UCLA

Google Research

Video Generative Models

CogVideoX

Wan

Meta Movie Gen

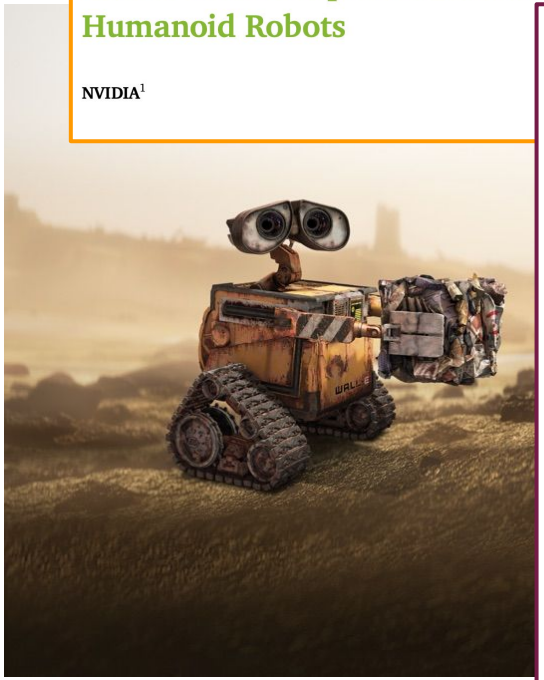
Physical World Simulators

NVIDIA

2025-3-28

GR00T N1: An Open Foundation Model for Generalist Humanoid Robots

NVIDIA[1]

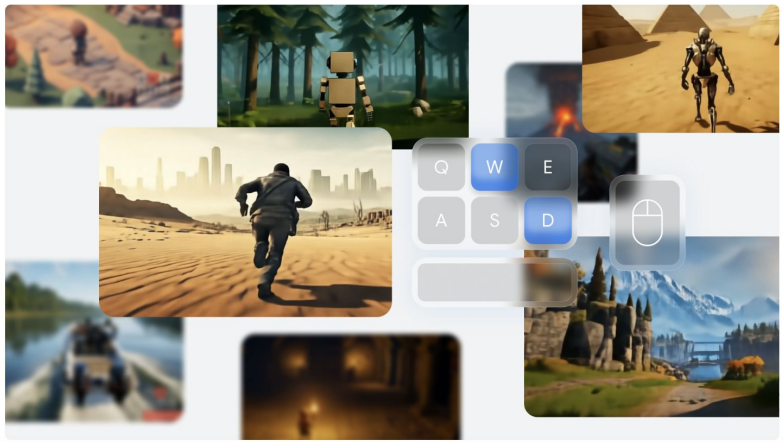


Genie 2: A large-scale foundation world model

4 DECEMBER 2024

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, Tim Rocktäschel

Share

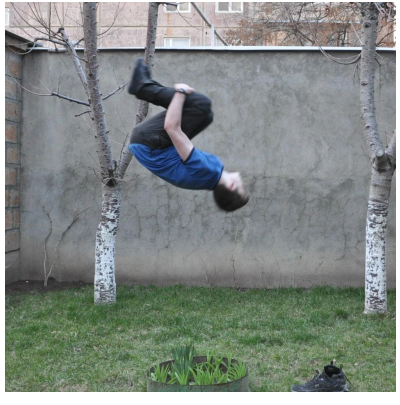


Learning Universal Policies via Text-Guided Video Generation

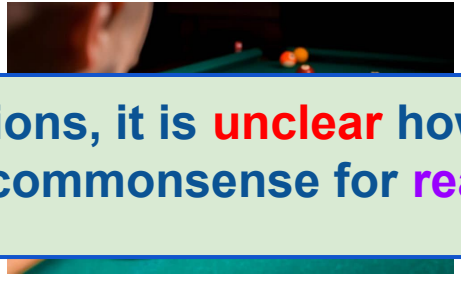
§, Bo Dai ‡¶, Hanjun Dai ‡, Ofir Nachum ‡,
Dale Schuurmans ‡‖, Pieter Abbeel §
erkeley§ Georgia Tech¶ University of Alberta‖
ersal-policy.github.io/

Humans Rely on Physical Commonsense In Everyday Activities

As humans, we develop an intuitive understanding of the object interactions through our experience with the real-world, without any formal education in physics.



Physical Commonsense



Despite physical motivations, it is unclear how well do generated videos follow physical commonsense for real-world activities!

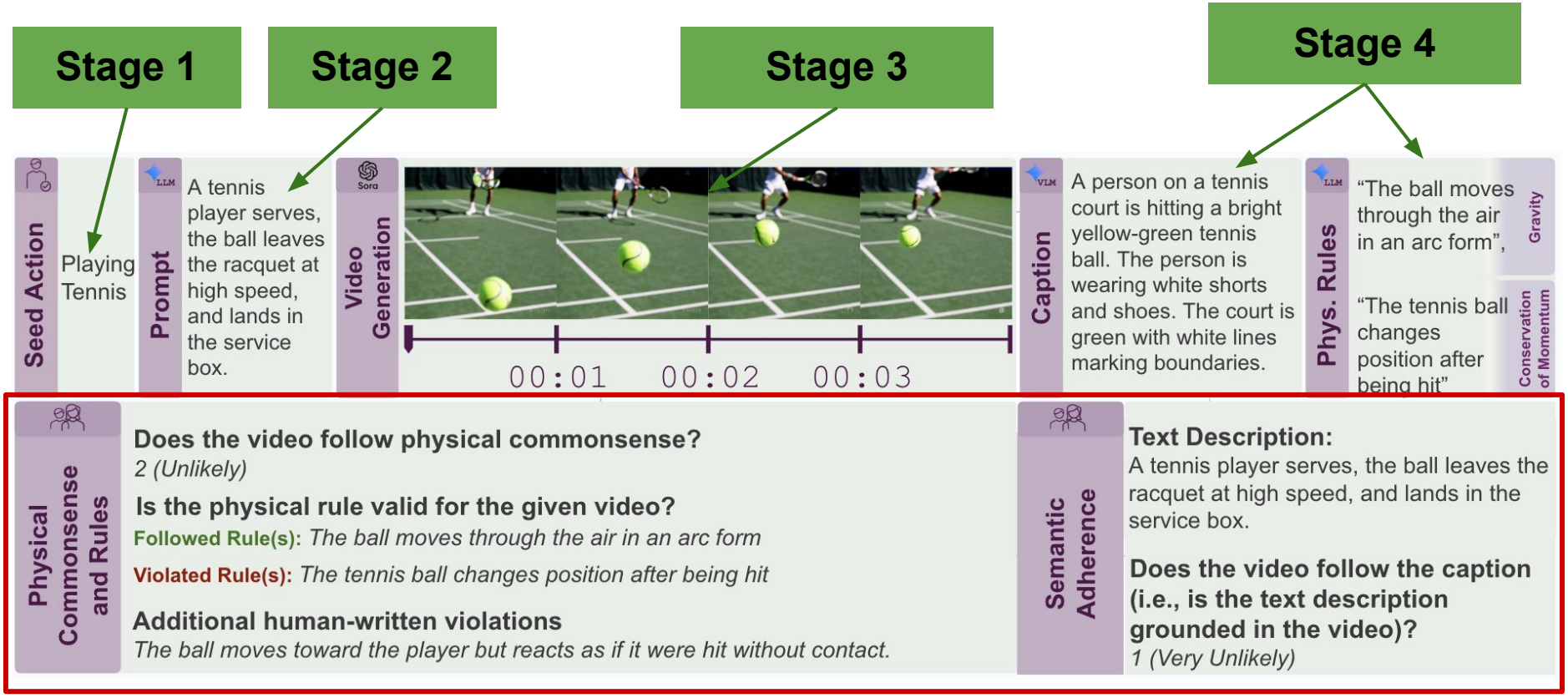
VideoPhy-2: Physical Commonsense for Real-World Actions

Categories
(e.g, sports, object interactions)

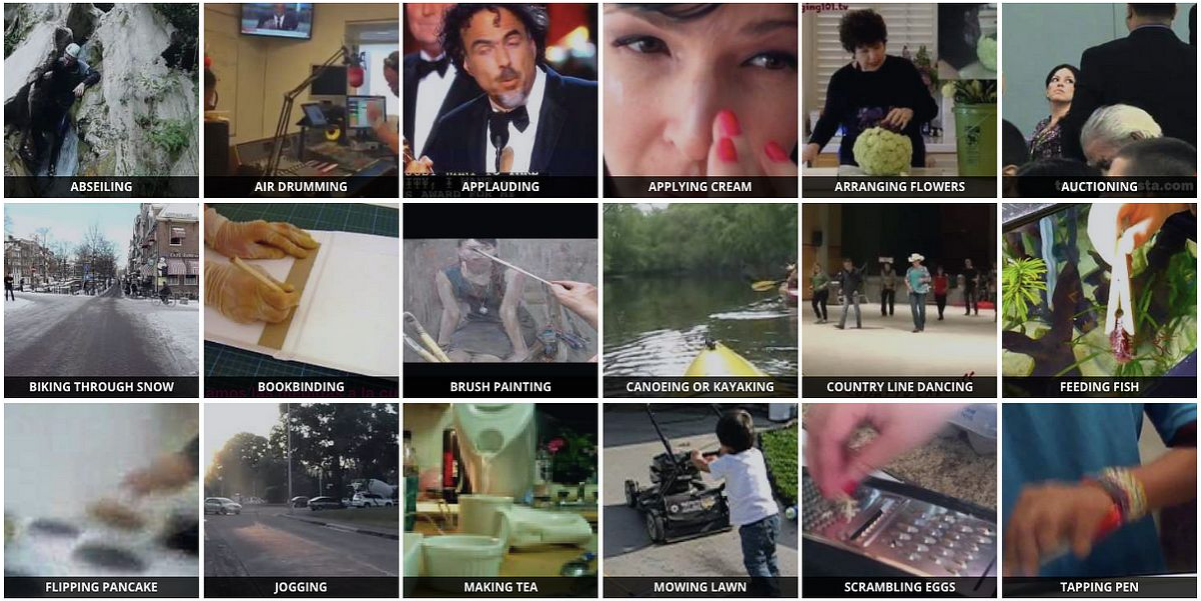
Diverse Physics Laws
(e.g, gravity, momentum)

Multiple Events
(e.g., moving and then stopping)

Pipeline

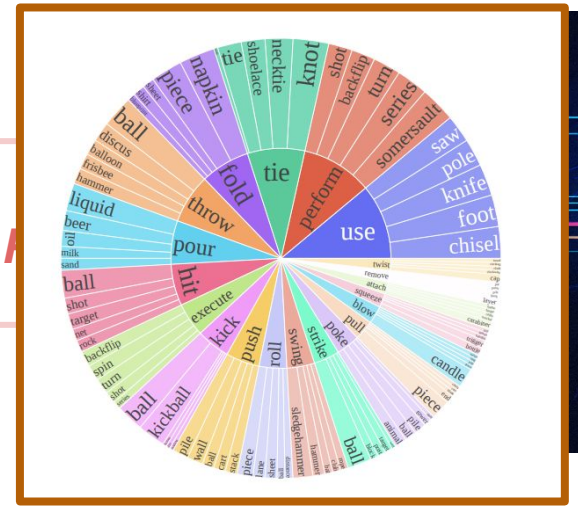

Stage 1
Stage 2
Stage 3
Stage 4
Seed Action
Playing Tennis
Prompt
A tennis player serves, the ball leaves the racquet at high speed, and lands in the service box.
Video Generation
00:01
00:02
00:03
Caption
A person on a tennis court is hitting a bright yellow-green tennis ball. The person is wearing white shorts and shoes. The court is green with white lines marking boundaries.
Phys. Rules
"The ball moves through the air in an arc form",
Gravity
"The tennis ball changes position after being hit"
Conservation of Momentum
Physical Commonsense and Rules
Does the video follow physical commonsense?
2 (Unlikely)
Is the physical rule valid for the given video?
Followed Rule(s): The ball moves through the air in an arc form
Violated Rule(s): The tennis ball changes position after being hit
Additional human-written violations
The ball moves toward the player but reacts as if it were hit without contact.
Semantic Adherence
Text Description:
A tennis player serves, the ball leaves the racquet at high speed, and lands in the service box.
Does the video follow the caption (i.e., is the text description grounded in the video)?
1 (Very Unlikely)
Stage 5: Human Eval

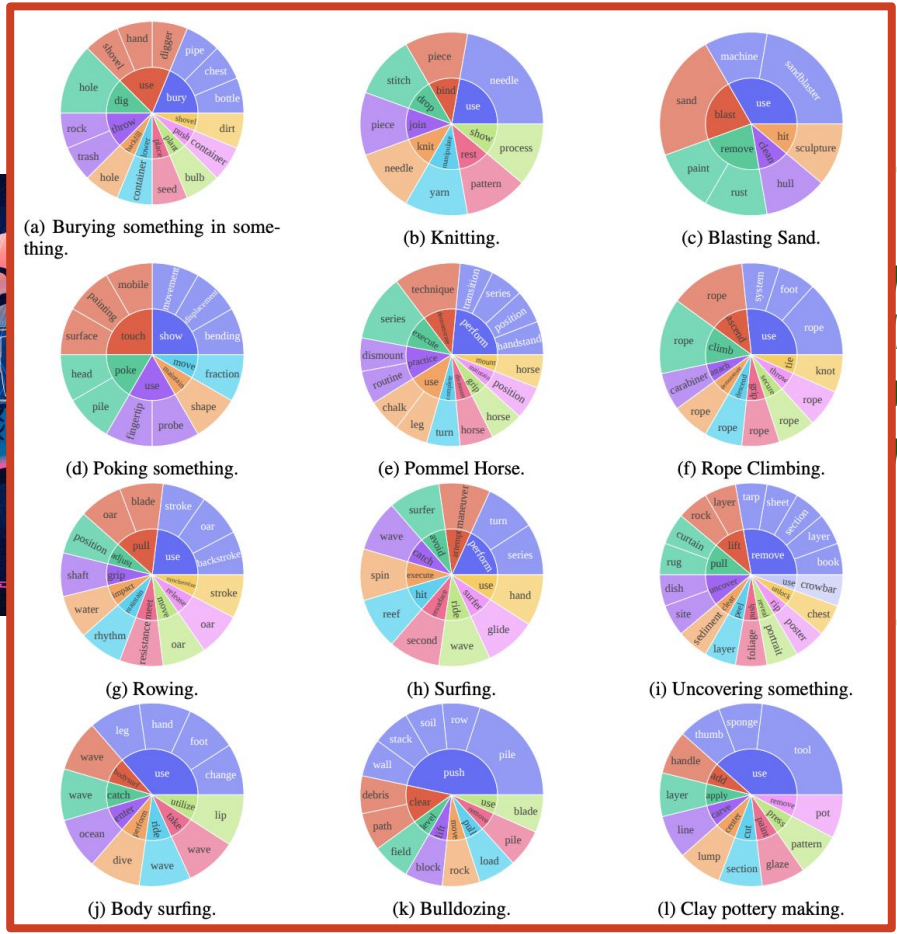
Stage 1: Seed Actions Are Taken from Video Datasets and Subsequently Filtered By Humans



200 actions are selected from 600 actions for physical commonsense evaluation!

Stage 2: LLM Creates A Prompt from Seed Action





Stage 3: Video Generation Conditioned on Prompt

Prompt:
A tennis player serves, the ball leaves the racquet at high speed, and lands in the service box.

Sora

Stage 4: Powerful Video-Language Model Captions And Lists Candidate Physical Rules

Gemini

Video Caption:
A person on a tennis court is hitting a bright yellow-green tennis ball.

Candidate Physical Rules:
The ball moves through the air in an arc form (Gravity).

The tennis ball changes position after being hit (Conversation of Momentum).

Stage 5: Human Evaluation Via Amazon Mechanical Turk

Semantic Adherence

Rate: 1-5

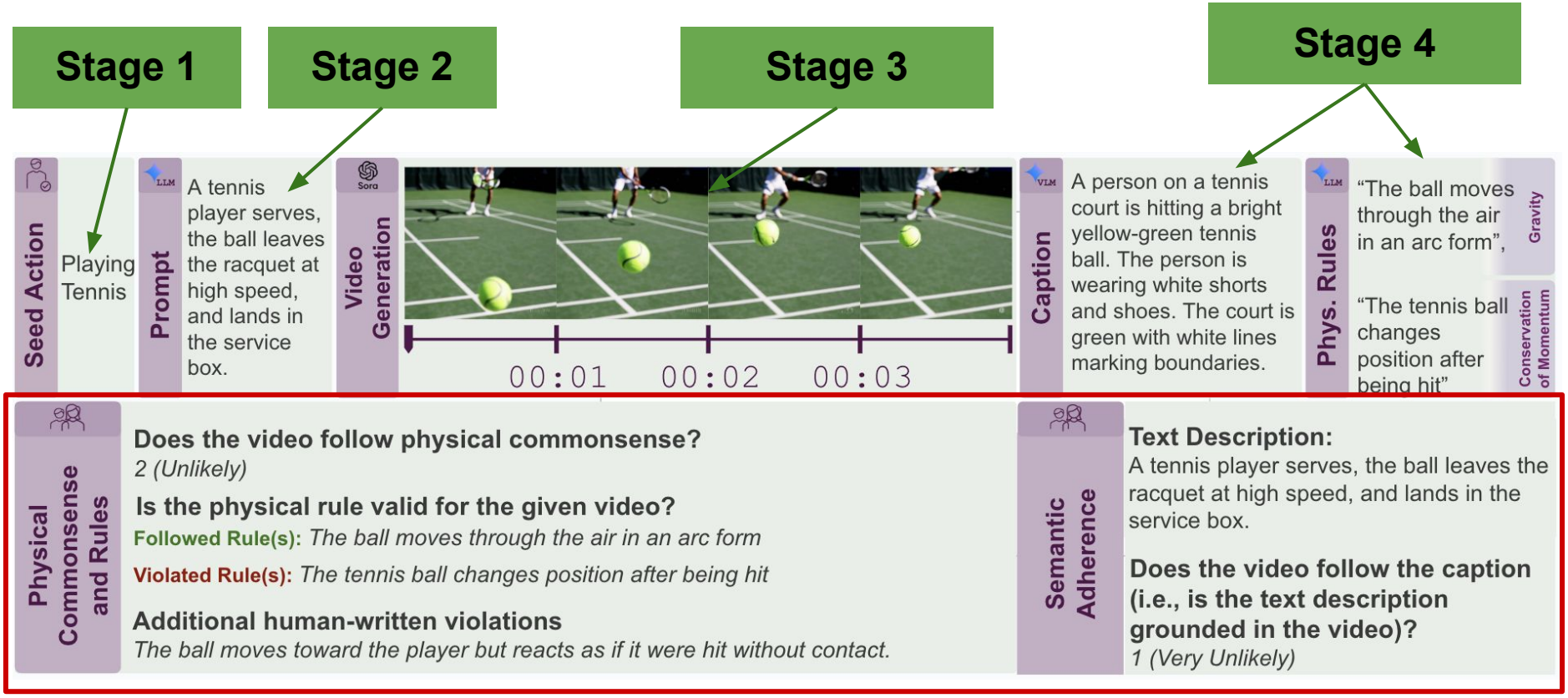
Physical Commonsense

Rate: 1-5

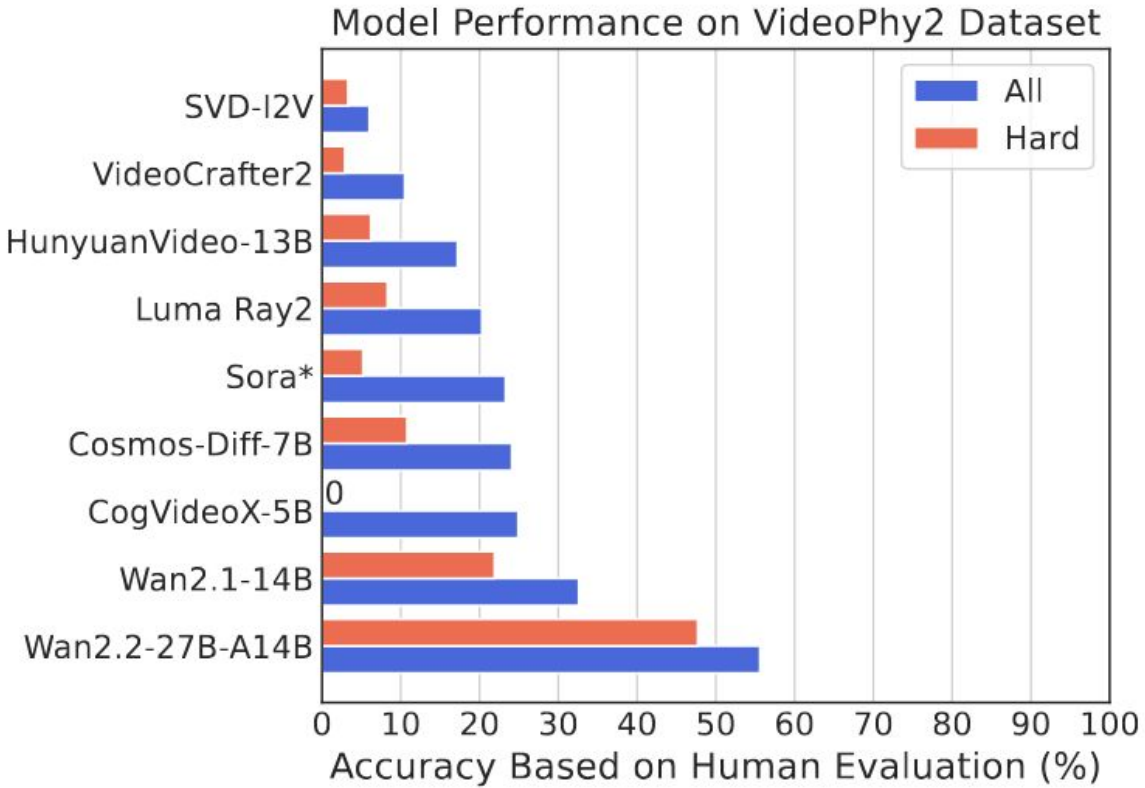
Physical Rule Classification

Rate: Yes-No

Pipeline


Stage 1
Stage 2
Stage 3
Stage 4
Seed Action
Playing Tennis
Prompt
A tennis player serves, the ball leaves the racquet at high speed, and lands in the service box.
Video Generation
00:01
00:02
00:03
Caption
A person on a tennis court is hitting a bright yellow-green tennis ball. The person is wearing white shorts and shoes. The court is green with white lines marking boundaries.
Phys. Rules
"The ball moves through the air in an arc form",
Gravity
"The tennis ball changes position after being hit"
Conservation of Momentum
Physical Commonsense and Rules
Does the video follow physical commonsense?
2 (Unlikely)
Is the physical rule valid for the given video?
Followed Rule(s): The ball moves through the air in an arc form
Violated Rule(s): The tennis ball changes position after being hit
Additional human-written violations
The ball moves toward the player but reacts as if it were hit without contact.
Semantic Adherence
Text Description:
A tennis player serves, the ball leaves the racquet at high speed, and lands in the service box.
Does the video follow the caption (i.e., is the text description grounded in the video)?
1 (Very Unlikely)
Stage 5: Human Eval

Results

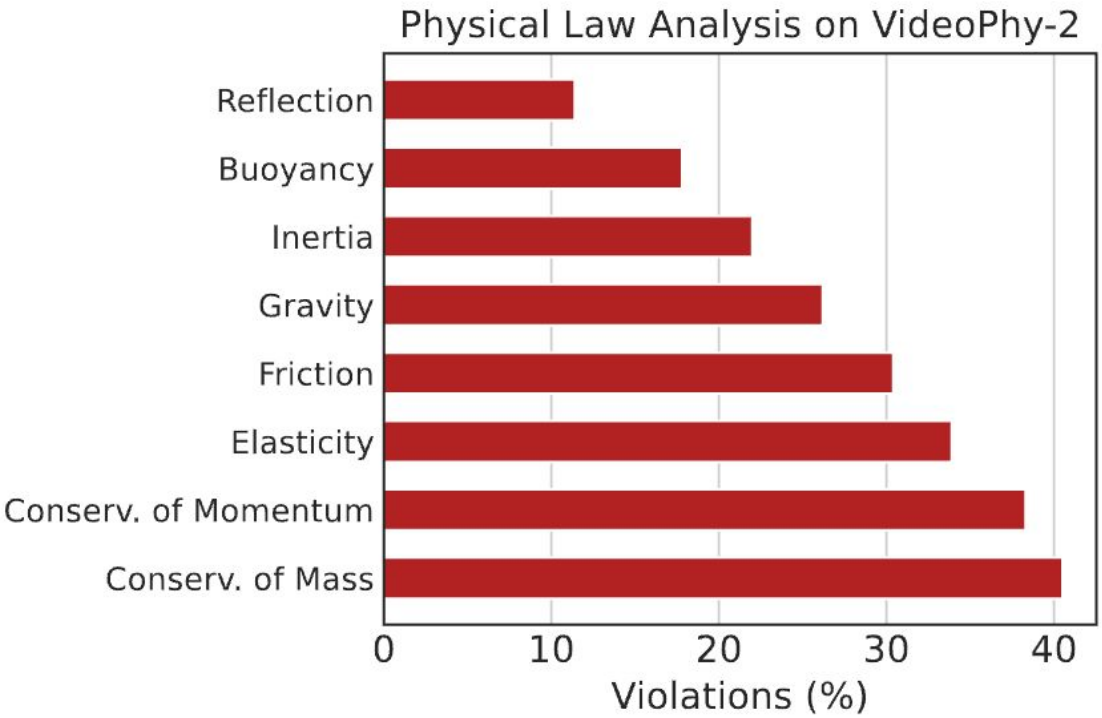


Joint score
1 if (SA>=4 and PC>=4)

Modern video generative models (closed and open) are far from simulating the real-world actions!

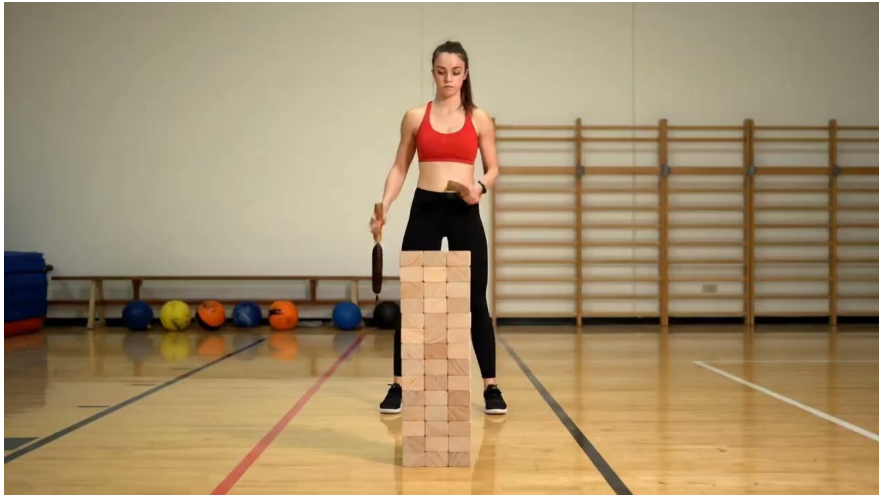
Wan2.1-14B achieves the highest joint performance.

Our Analysis Suggests Conservation Laws are the Most Violated



Examples (Veo3)

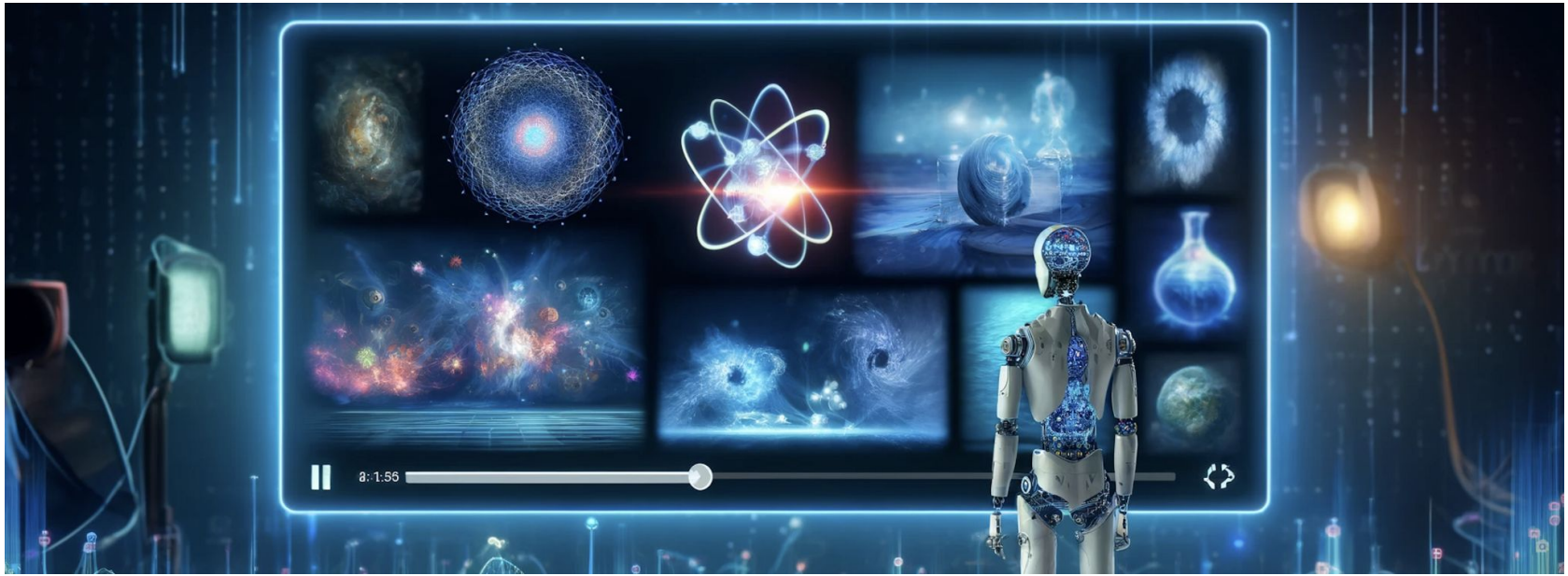
an archer draws a recurve bow, the string stretching taut, then releases the arrow, which hits a target's center.



a person uses nunchucks to break a stack of wooden blocks, the blocks scattering on impact.

VideoPhy-AutoEval: Automatic Evaluator For Scalable Evals!

A subset of human annotations finetune a video language model



VideoPhy2-AutoEval Outperforms Gemini-2.0-Flash!

Table 4: **Auto-rater evaluation results.** We present the pearson's correlation (×100) between the predicted scores and ground-truth scores (1-5) on the unseen prompts and unseen video models.

	Unseen prompts			Unseen video models		
	Avg.	**SA**	**PC**	**Avg.**	**SA**	**PC**
VideoCon-Physics [5]	28.5	32.0	25.0	26.5	27.0	26.0
VideoCon [3]	12.5	23.0	2.0	8.9	17.0	0.8
VideoLlava [34]	16.0	30.0	2.0	19.0	33.0	5.0
VideoScore [20]	13.5	17.0	10.0	9.0	5.0	13.0
Gemini-2.0-Flash-Exp	18.5	26.0	11.0	21.0	31.0	11.0
VIDEOPHY-2-AUTOEVAL	42.0	47.0	37.0	41.0	45.0	37.0
Rel. to Best (%)	+47.4	+46.9	+48.0	+49.0	+36.4	+61.5
Rel. to Gemini (%)	+127.0	+80.8	+236.4	+107.1	+45.2	+281.8

Table 6: **Auto-rater evaluation on physical rule classification.** We present the accuracy results for VIDEOPHY-2-AUTOEVAL and other video-language models on the rule classification tasks.

	Unseen prompts	**Unseen video models**
Random	34.5	31.2
VideoLlava [34]	38.1	38.7
Gemini-2.0-Flash-Exp	59.2	57.1
VIDEOPHY-2-AUTOEVAL	78.7	72.9
Rel. to Best (%)	+32.9	+27.7

Community Adoption of VideoPhy Series

Benchmark

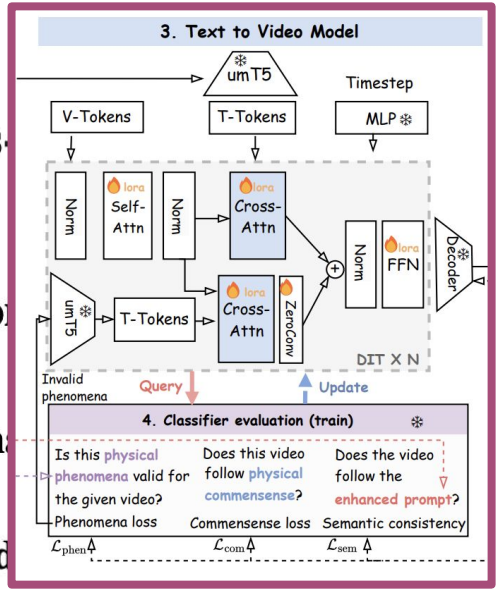
Reward Model

Think Before You Diffuse: LLMs-Guided Physics-Aware Video Generation

Articulated Kinematics Distillation from Video Diffusio

PhysMotion: Physics-Grounded Dynamics from a Single Im

WISA: World Simulator Assistant for Physics-Aware Text-to-Vid



TL;DR

- **VideoPhy2 is a high-quality dataset for physical commonsense evaluation of the open and closed video generative models.**

- **VideoPhy2-AutoEval is an automatic evaluator for assessing physical commonsense for the novel video models.**

Paper, code, and dataset are publicly available

VideoPhy 2

Challenging Action-Centric Physical Commonsense Evaluation of Video Generation

Hritik Bansal*[1], Clark Peng*[1], Yonatan Bitton*[2], Roman Goldenberg[2], Aditya Grover[1] Kai-Wei Chang[1],

(* Equal Contribution)

[1]University of California, Los Angeles

[2]Google Research