WorldReasonBench

Human-Aligned Stress Testing of Video Generators as Future World-State Predictors

Can a realistic video show the right future?

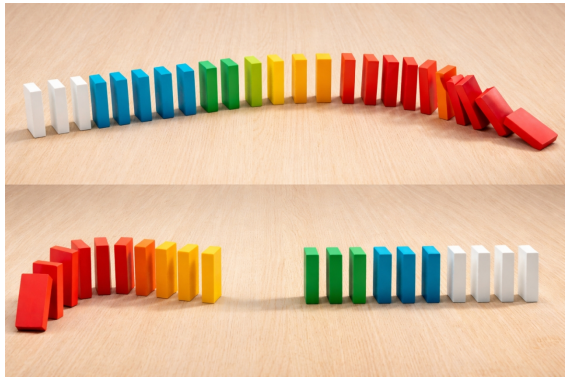
Keming Wu*, Yijing Cui*, Wenhan Xue, Qijie Wang, Xuan Luo, Zhiyuan Feng,
Zuhao Yang, Sudong Wang, Sicong Jiang, Haowei Zhu, Zihan Wang,
Ping Nie, Wenhu Chen, Bin Wang

Tsinghua University · NTU · Waterloo · HKUST (Guangzhou) · 2077 AI

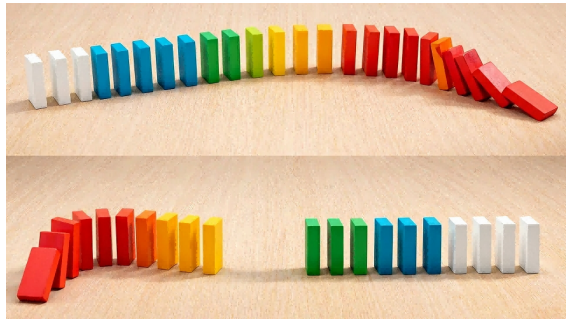
* Equal contribution

Initial image + instruction → evaluate the entire state transition

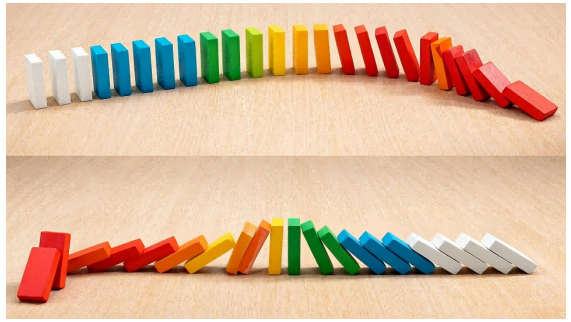
A correct endpoint can hide an incorrect process



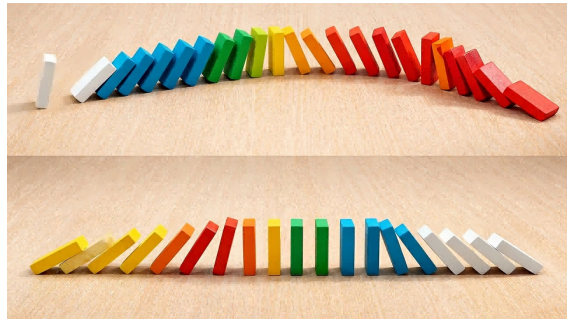
Initial state



Early frame



Middle frame



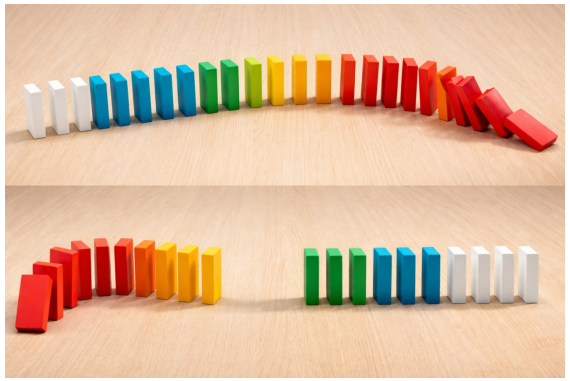
Late frame

Check contact, event order, and causal continuity.

Illustrative frames from the paper's qualitative comparison.

Scope: short-horizon, initial-image-conditioned transition correctness

436 cases across four reasoning dimensions



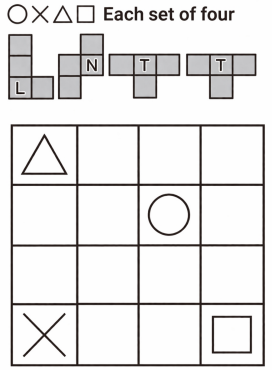
World knowledge

Physical transitions



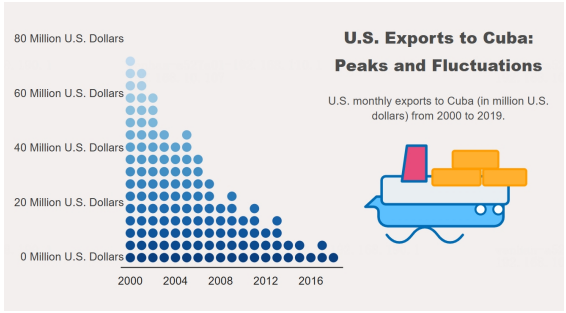
Human-centric

Context-conditioned actions



Logic reasoning

Rules and state changes



Information-based

Text and data preservation

22 subcategories · Structured ground-truth QA

Implicit and hinted prompts test sensitivity to instruction guidance

Measure outcomes, processes, and visual quality

Process-aware verification

Video-grounded answers → reference checking

State	What is present?
Temporal process	What happened, and when?
Fine-grained fidelity	Were details preserved?
Causal mechanism	Was the transition plausible?

Holistic quality

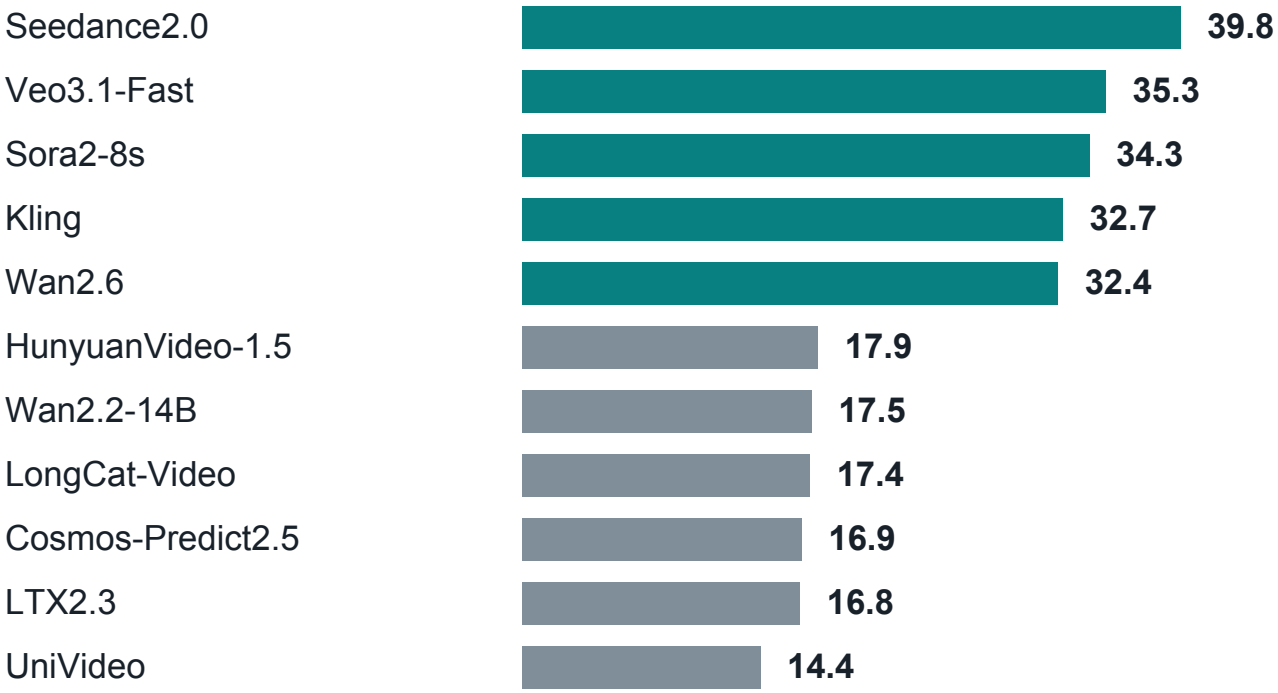
Reasoning quality
Temporal consistency
Visual aesthetics

Weighted score: 0.4 / 0.3 / 0.3

Process score = QA accuracy^0.8 × dynamic-phase accuracy^0.2

Tested generators remain far from saturation

Process-aware score (0–100)



32.4–39.8

Closed-source configurations

14.4–17.9

Open-source configurations

Logic reasoning is hardest on average; dynamic failures remain common

Process-aware rankings align with human preferences

5,969 preference pairs · 1,432 videos · 15 annotators

Spearman correlation with human Elo (reported model-ranking panel)



Model-level rank agreement does not guarantee reliable near-tie decisions

Controls clarify what the scores depend on

Video grounding

Removing video lowers binary QA accuracy by 20.75 pp on the tested 119-video subset.

Evaluator swaps

Generator-ranking correlations: 0.857–0.973 across the tested evaluator families.

Prompt length control

Pooled gain: +17.34 pp with neutral padding; +5.88 pp residual after matching word count.

Hint gains reflect prompt formulation as well as transition information

TAKEAWAY

Evaluate the right transition, not only realistic frames

WorldReasonBench

Structured cases + process-level diagnostics

WorldRewardBench

Expert preferences for automatic judge calibration



Data and evaluation code

unix-ai-lab.github.io/WorldReasonBench/

Future work: long horizons, systematic 3D tests, and reward-guided training