

000  
001  HVCR: CAUSAL EVALUATION OF LARGE MULTI-  
002 MODAL MODELS IN HUMAN-LIKE VIDEO REASONING  
003

004 **Anonymous authors**  
005

006 Paper under double-blind review  
007

008 **ABSTRACT**  
009

010 The pursuit of human-like causal reasoning in large multimodal models (LMMs)  
011 is a critical yet challenging frontier. Current video causal reasoning benchmarks  
012 often lack a systematic design that aligns with the nuances of human causal cog-  
013 nition. To address this gap, we introduce the HVCR benchmark for evaluating  
014 human-like video causal reasoning in LMMs, systematically designed across three  
015 levels. At the **definitional** level, since causal relations in videos are inherently  
016 *specific* and naturally align with the field of actual causality, we adopt definitions  
017 from this domain that robustly handle complex scenarios like preemption, which  
018 can not be modeled by the simple “but-for” test. At the **goal-oriented** level, our  
019 aim is to simulate human causal judgments rather than fitting formal definitions  
020 or frameworks. Therefore, we establish our gold standard using human “con-  
021 sensus” from rigorous human experiments in cognitive science, leveraging seven  
022 well-studied causal scenarios as reliable references. At the **representational** level,  
023 we employ explicit *causal graphs* and a variant of *twin networks* to enable auto-  
024 matic generation of causal questions. The HVCR benchmark contains 300 videos  
025 (240 synthetic and 60 realistic) and 4,967 causal questions. These questions span  
026 three causal rungs (discovery, intervention, and counterfactual) and eight types,  
027 focusing on key aspects of human-like causal reasoning such as causal attribution  
028 and responsibility. Human evaluation shows that average observers achieve nearly  
029 80% accuracy on our synthetic videos, confirming their clarity. However, current  
030 LMMs underperform on both synthetic and real-world videos, revealing a signif-  
031 icant gap in their human-like causal reasoning capabilities. To our knowledge,  
032 HVCR is the first video causal reasoning benchmark to systematically integrate  
033 these three design levels, jointly consider synthetic and real-world settings, and  
034 focus exclusively on pure causal reasoning of LMMs.  
035

036 **1 INTRODUCTION**  
037

038 *Where causation is concerned, a grain of wise subjectivity tells us more about the*  
039 *real world than any amount of objectivity.*  
040

041 — Pearl & Mackenzie (2018)  
042

043 The pursuit of human-like causal reasoning is a critical frontier for AI (Chi et al., 2024; Joshi et al.,  
044 2024; Hong et al., 2024). However, current progress in video causal reasoning lacks a systematic  
045 design that simulates human causal judgments from three levels.  
046

047 **Definitional Level.** Causal relationships in videos are inherently *specific* rather than *general*. For  
048 example, in a video where Alice spills water onto a rice cooker, causing it to short-circuit and catch  
049 fire, one can infer the *specific* causal relation “Alice’s action caused the rice cooker to catch fire.” In  
050 contrast, the *general* claim “spilling water on an appliance causes short-circuit” cannot be justified  
051 from a single event. This shows that video causality naturally aligns with *specific* actual causality  
052 (Halpern, 2016), unlike textual scenarios where both general and specific causal relations coexist.  
053 Therefore, video causality should be modeled with techniques from actual causality. Nevertheless,  
054 much existing research do not specify definitions or still relies on the “but-for” test from legal theory,  
055 where  $X$  is a cause of  $Y$  if, but for  $X$ ,  $Y$  would not have occurred (Yi et al., 2020; Foss et al., 2025).  
056 While simple and general, this test often fails to capture human intuition. A classic case is late  
057

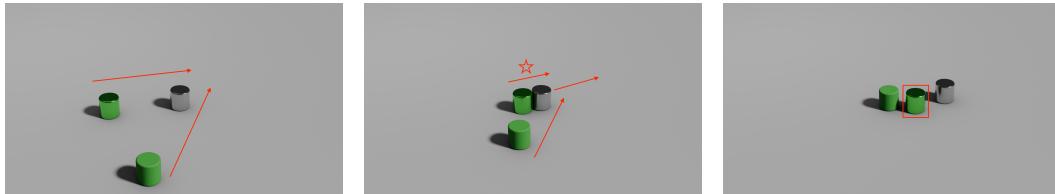


Figure 1: An example of Late Preemption.

preemption (Lewis, 2000): *Suzy and Billy both throw rocks at a bottle. Suzy’s rock arrives first and shatters it. Since both throws are perfectly accurate, Billy’s would have shattered the bottle had it not been preempted by Suzy’s.* This case can be readily mapped into a video, as shown in Figure 1. Here, the but-for test incorrectly concludes that Suzy’s throw is not a cause. This underscores the need for more precise definitions from actual causality, such as the Halpern-Pearl definition (Halpern, 2016), when modeling causal relations in videos.

**Goal-Oriented Level.** Beyond definitional foundations, the goal of causal reasoning in LMMs should be *simulating human causal cognition* rather than strictly *fitting a formal definition or framework*. Although theories such as Force Dynamics (Talmy, 1988; Wolff, 2007), Mental Models (Khemlani et al., 2014), Causal Models (Goldvarg & Johnson-Laird, 2001; Sloman & Sloman, 2009), and Counterfactual Simulation (Halpern & Pearl, 2005; Gerstenberg et al., 2021) aim to capture human causal cognition, their predictions often diverge from actual human judgments, especially in complex physical events (Mao et al., 2022). Large-scale crowdsourced annotation is one approach (Foss et al., 2025), but carefully controlled experiments, such as randomized controlled trials (RCTs) in psychology, philosophy, and cognitive science, provide more reliable references (Cartwright, 2010). These results serve as a “consensus” of human perception, reducing reliance on massive annotation while grounding models in empirically validated human reasoning.

**Representational Level.** Explicit causal representations are essential for video-based causal reasoning benchmarks. Most existing work omits explicit representations of causal relationships (Yi et al., 2020; Foss et al., 2025), while others do (Chen et al., 2024b; 2025; Yi et al., 2020).

In this work, we propose the HVCR benchmark for evaluating human-like video causal reasoning in LMMs, systematically designed across three levels. At the **definitional level**, we adopt definitions of actual causality as our foundation, which not only guides the selection of causal scenarios but also serves as a formal reference for comparison with human causal judgments. At the **goal-oriented level**, our aim is to *simulate human causal cognition*. To this end, we select seven well-studied causal scenarios that provide well-defined causal structures and human “consensus” gold standards (Kueffner, 2021). At the **representational level**, we employ *causal graphs* to explicitly encode general causal relationships, and a variant of *twin networks* to represent factual and counterfactual worlds, which are then used to automatically generate causal questions and answers.

The resulting HVCR benchmark contains 300 videos and 4,967 causal questions. It comprises 240 synthetic videos generated from six scenarios, primarily involving object motion and interactions on an infinite horizontal tabletop, and 60 real-world videos covering sports events and traffic accidents. The questions span three causal rungs (Discovery, Intervention, and Counterfactual) and eight types, covering key aspects of human-like causal reasoning such as causal attribution and responsibility assignment. Human evaluation demonstrates that average observers can accurately understand the causal dynamics in synthetic videos, achieving nearly 80% accuracy. In contrast, current LMMs underperform on both synthetic and real-world videos. To our knowledge, HVCR is the first video causal reasoning benchmark systematically designed across three levels, jointly considers both synthetic and real-world settings, and focuses exclusively on evaluating pure causal reasoning abilities in LMMs.

## 2 PRELIMINARIES

### 2.1 DEFINITIONAL LEVEL: ACTUAL CAUSALITY

Causality research in computer science can be divided into two subfields (Kiciman et al., 2024). *Type causality* concerns general causal relationships between variables, e.g., “smoking causes lung

108 cancer,” while *actual causality* (or token causality) identifies the causes of specific events and as-  
 109 signs responsibility, e.g., “David’s 30 years of smoking caused his lung cancer” (Halpern, 2016;  
 110 Hausman, 2005). The aim of actual causality is to **axiomatize the intuitive form of causal rea-**  
 111 **soning employed by humans** (Kueffner, 2021), supporting applications such as legal reasoning,  
 112 machine failure debugging, root cause analysis, and explainable AI (Lagnado & Gerstenberg, 2017;  
 113 Andreas et al., 2023; Lu et al., 2023; Dubslaff et al., 2022; Sharma et al., 2022; Blake et al., 2025;  
 114 Chockler et al., 2024).

## 116 2.2 REPRESENTATIONAL LEVEL: CAUSAL GRAPHICAL MODELS

118 **Causal graphs** are directed acyclic graphs  
 119 (DAGs) that depict causal relationships  
 120 between variables (Figure 2a). In this  
 121 work, we use causal graphs to repre-  
 122 sent general causal relationships between  
 123 events. This is an implicit encoding of dif-  
 124 ferent worlds.

125 **Twin networks** can be viewed as an ex-  
 126 tension to causal graphs, where two inter-  
 127 linked networks are employed to represent  
 128 the factual and counterfactual worlds in  
 129 one network (Balke & Pearl, 1994) (Figure  
 130 2b). These two networks are identical in structure and share the same exogenous variables, as  
 131 exogenous variables are not modified by interventions on observed variables. In this work, we use  
 132 twin networks to represent the “states” of the factual and counterfactual worlds. That is, nodes rep-  
 133 resent assignments in both worlds, while edges have two states depicting whether the direct causal  
 134 influence actually takes effect.

## 136 2.3 THE LADDER OF CAUSATION

138 The Ladder of Causation (Pearl & Mackenzie, 2018) illustrates three rungs of causal reasoning: As-  
 139 sociation, Intervention, and Counterfactual. Since association does not involve causal inference, we  
 140 instead regard Discovery as Rung 1 (Chen et al., 2024a). **Discovery** (Rung 1) involves identifying  
 141 cause-effect pairs from observational data without prior knowledge, e.g., “*Is there a causal rela-*  
 142 *tionship between review frequency and exam scores?*” **Intervention** (Rung 2) involves exploring  
 143 the effects of manipulating variables using the *do*-operator, e.g., “*What if I review everyday, will my*  
 144 *exam scores improve?*” **Counterfactual** (Rung 3) considers hypothetical alternatives, e.g. “*What if*  
 145 *I have attended a party instead of reviewing, would my exam scores be good?*”

## 147 3 DATASET CONSTRUCTION

150 HVCR comprises 300 videos, of which 240 are synthetic and 60 are realistic. The dataset contains a  
 151 total of 4967 causality-focused QA pairs. The construction pipeline for HVCR is shown in Figure 3.

### 153 3.1 CAUSAL SCENARIO SELECTION

155 The field of actual causality (and causal judgment) often identify seven real-world causal scenarios  
 156 (Kueffner, 2021). These scenarios are notable because human causal judgments often conflict with  
 157 formal definitions and models of actual causality. This discrepancy has driven the development of  
 158 new definitions and more accurate causal models. Therefore, focusing on these specific scenarios  
 159 and addressing these conflicts naturally aligns with human intuition. This paper focuses on six of  
 160 these scenarios, summarized below. All six scenarios are included in the synthetic data, while the  
 161 realistic data only considers SW, LP, and BP, as they are more common in real-world situations. The  
 six scenarios are typically represented by the neuron diagrams shown in Figure 3.

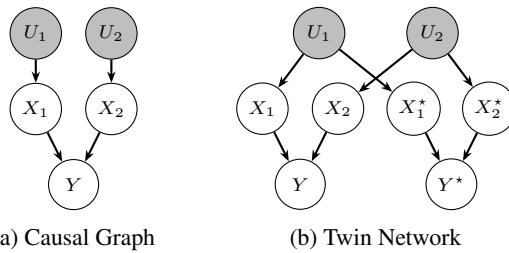


Figure 2: Example of causal graphical models.

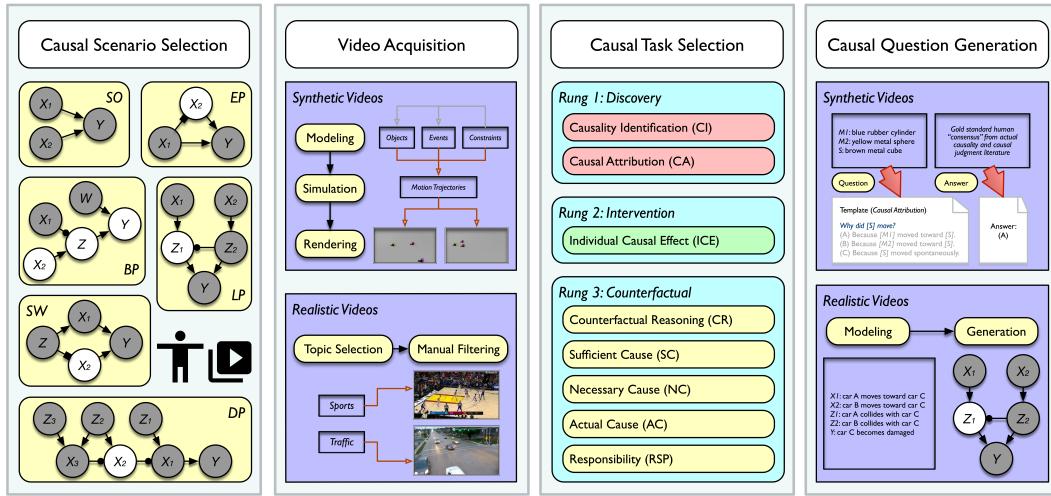


Figure 3: The Construction Pipeline of HVCR.

**Symmetric Overdetermination** (SO): Multiple processes, all producing the same outcome, terminate at the same time.

**Switch** (SW): An event triggers one of two processes, both of which have the same outcome, making the event immaterial for the final result.

**Late Preemption** (LP): Two causal processes run in parallel; both would produce the same outcome, but one terminates before the other, rendering the latter irrelevant.

**Early Preemption** (EP): Two causal processes could produce the same outcome, but one terminates before the other even starts.

**Double Preemption** (DP): A process that would have prevented another process is itself prevented by a different process.

**Bogus Preemption** (BP): An action is taken to interrupt a process that was never active.

## 3.2 VIDEO ACQUISITION

### 3.2.1 PRINCIPLES

We followed these principles for generating synthetic videos and collecting realistic ones. **First, “one-take” exhibition.** A single-take video should be enough to convey all the prior knowledge needed for the causal scenarios and be easily understood by the average human observers. **Second, quality control.** We focused on creating or collecting high-quality videos that adhere to the first principle. **Third, diversity design.** We aimed for multi-dimensional diversity, including variations in data realism, entities, events, topics, causal structures, and perturbed settings, etc..

### 3.2.2 SYNTHETIC DATA GENERATION

**Objects & Events.** Objects in HVCR follow similar compositional attributes as in CLEVRER (Yi et al., 2020), including three shapes (cube, sphere, and cylinder), two materials (metal and rubber), and eight colors (gray, red, blue, green, brown, cyan, purple, and yellow). In each video, identical objects are prohibited, where each combination of the three attributes uniquely identifies a single object. We consider both unary and binary events. For **unary events**, we focus on the *motion status* of an object, which shows whether the object is moving or stationary at a time point. For **binary events**, we examine: 1) *collisions* and 2) *relative movement directions*. A moving object may approach a stationary one or another moving object, indicating potential collisions.

**Modeling.** First, we define unique **events and constraints** for each scenario and control initial object attributes (e.g., location, orientation, velocity, and angular velocity) to realize these events in videos. The constraints primarily involve 1) the number and sequence of collisions and 2) the “one-take” exhibition of knowledge from both factual and counterfactual worlds. For example, in LP, we consider the moving directions of M1 and M2 relative to S, the motion status of S, and the

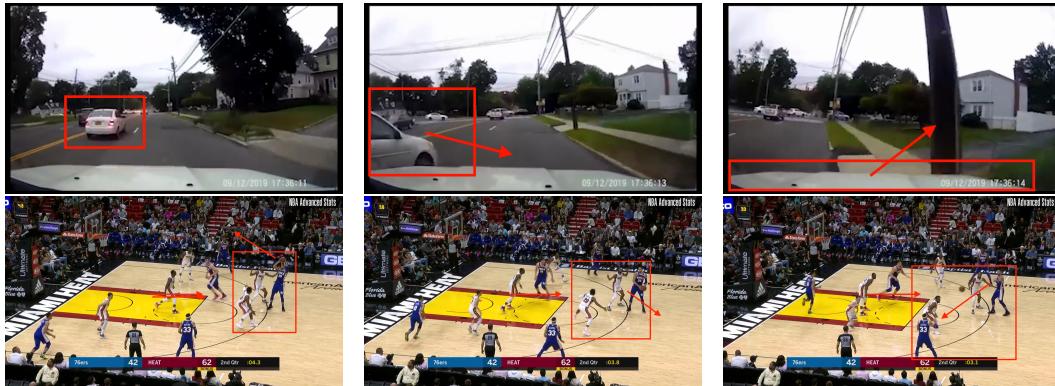


Figure 4: Top: SW The white car in front changes lanes suddenly, leaving the driver no time to stop. Whether staying on course or turning right, the vehicle is damaged. The driver ultimately turns right and crashes into a utility pole. Bottom: BP 76ers No. 25 (in blue) fakes a shot, causing HEAT No. 5 (in white) to jump, then passes to 76ers No. 33. HEAT No. 5 interrupts an inactive process.

collision between M1 and S. The number of collisions is exactly one, and M2 must be factually and counterfactually able to reach S’s initial position to ensure sufficiency. Variations in object attributes increase diversity, while constraints help uphold the “one-take” principle.

**Simulation.** Second, we use the Bullet physics engine (Coumans, 2010) to simulate the objects’ motion trajectories. Object attributes are implemented flexibly, with values randomized from a specified range. They are also adjusted to enable early modeling of counterfactual knowledge. For constraints, we designed *automatic filters* to remove simulations where the number and sequence of collisions did not align with our modeling.

**Rendering.** Third, we used Blender (Blender Online Community, 2016) to render the motion trajectories into realistic videos. Due to the challenges of “one-take” modeling of counterfactual knowledge and potential issues with the camera angle (such as objects being out of view or unclear collisions), we manually selected videos that adhered to the “one-take” principle.

### 3.2.3 REALISTIC DATA COLLECTION

Realistic videos are collected from existing video benchmarks that reflect at least one of our causal scenarios. Here, we focus on SW, LP, and DP, as they are common in real-world interactions.

**Topics & Sources of Videos.** We focused on two primary topics: *multi-person sports* and *traffic accidents*. Sports videos, which are excellent for capturing multi-person interactions fitting our scenarios, are collected from MultiSports (Li et al., 2021) and NSVA (Wu et al., 2022). These included Olympic and World Cup volleyball, as well as NBA basketball. Videos of traffic accidents, involving collisions between vehicles and people, were sourced from V-TIDB (Nishiyama et al., 2023), TAD (Xu et al., 2024), and TU-DAT (Pradeep Kumar & Kant, 2025). These videos included both ego-centric and camera-view perspectives. In total, we collected 20 videos for each scenario.

**Entities & Events.** The topics involve *multi-entity interactions*, including human-human (e.g., in volleyball and basketball games), human-object (e.g., vehicle-pedestrian collisions), and object-object (e.g., vehicle-vehicle collisions). Events vary by topic: for volleyball, they include *fake attacks* and *block attempts*; for basketball, they include *fake shots* and *securing jump balls*; and for traffic accidents, they involve events like *turning directions* and *collisions*. As a result, realistic videos contain a richer variety of both entities and events.

**Typical Examples.** In **SW**, a common example is a traffic accident where a vehicle in the left lane suddenly cuts in front of the following vehicle. The driver of the latter cannot slow down in time. Whether they try to brake or swerve right, the vehicle will collide with either the lane-changing car or the guardrail/another vehicle in the right lane. An example is shown in Figure 4. A typical **LP** situation in basketball involves two teammates, one behind the other, both going for a rebound. Regardless of who gets the ball, their team gains possession. In traffic, a variant of **LP** occurs when

270 an oncoming vehicle crashes into the front side of a car, and the driver cannot stop in time, resulting  
 271 in a collision. This variant, where the two causal processes are not independent, is called dependent  
 272 preemption (Moore, 2012). The **BP** scenario is particularly interesting and often involves feints  
 273 in sports. For instance, in basketball, a player might fake a shot to trick a defender into jumping  
 274 too early. In volleyball, a player might fake an attack jump, causing an opponent to misjudge and  
 275 commit to a defensive block prematurely. An example is shown in Figure 4.

### 278 3.3 CAUSAL TASK SELECTION

281 **Rung 1: Discovery.** We include Causality Identification (CI, e.g., “*Does the car turning left af- 282  
 283 fect whether it becomes damaged?*”) and Causal Attribution (CA, e.g., “*Why did the car become  
 damaged?*”) in this rung.

284 **Rung 2: Intervention.** We include Individual Causal Effect (ICE, e.g., “*If we force the car to stay 285  
 286 straight, will the car in front of it cause it to become damaged?*”) in this rung.

287 **Rung 3: Counterfactual.** We include Counterfactual Reasoning (CR, e.g., “*If the car had not 288  
 289 turned left, would it still have become damaged?*”), Sufficient Cause (SC, e.g., “*Was the fact that 290  
 291 the car turned left sufficient for it to become damaged?*”), Necessary Cause (NC, e.g., “*Was the fact 292  
 293 that the car turned left necessary for it to become damaged?*”), Actual Cause (AC, e.g., “*What is the 294  
 295 actual cause of the car becoming damaged?*”), and Responsibility (RSP, e.g., “*Was the fact that the 296  
 297 car turned left responsible for it becoming damaged?*”) in this rung.

298 We distinguish tasks reflecting natural human reasoning from those representing more abstract or  
 299 formal processes. “Human” tasks involve intuitive, token-level reasoning to explain a single oc-  
 300 currence. CA and RSP are “Human” as they align with human intuitions (Halpern & Pearl, 2005;  
 301 Halpern, 2016). CR, SC, and NC are also “Human”, foundational to actual causality and causal judg-  
 302 ment (Kiciman et al., 2024). ICE is “Human” since it concerns intervening on a single unit (Pearl  
 303 et al., 2016), mirroring the human cognitive process of predicting consequences through mental sim-  
 304 ulation (Sloman & Sloman, 2009). In contrast, CI is “Non-Human” as it learns a general causal rule,  
 305 not interpreting a single event. Similarly, AC is “Non-Human” since it outputs formal causal models  
 306 that aim to simulate human judgment but may diverge systematically (Lagnado & Channon, 2008).

### 307 3.4 CAUSAL QUESTION CONSTRUCTION

308 **Causal Structures.** In HVCR, we provide  
 309 causal graphs and twin networks for each video.  
 310 Synthetic videos have two settings: 1) Ba-  
 311 sic, which strictly follow the causal structures  
 312 defined by the scenario, with events uniquely  
 313 defined in the modeling stage; 2) Perturbed,  
 314 which add 1-2 isolated nodes (e.g., 1 stationary  
 315 object, 2 stationary objects, 1 moving object, or  
 316 2 moving objects). These objects do not inter-  
 317 act with the main scenario objects, and the per-  
 318 turbations increase the options for CA and AC.  
 319 For realistic videos, we manually define the en-  
 320 tities and events. The causal structures are sim-  
 321 ilar to the original scenarios but more diverse,  
 322 especially for LP (e.g., dependent preemption).  
 323 Entities and events also contribute to this diversity.

324 **Question Generation.** Based on the causal structure, we generate causal questions using a  
 325 template-based approach. For videos where the entities, events, and causal structures match the  
 326 original scenario, we use the corresponding answers directly. For videos with any changes, we  
 327 manually verify the consistency of the answers.

328 Table 1: Human evaluation results.

329 Question Type	330 # Videos	331 # QAs	332 Acc.
<i>333 Human</i>			
Causal Attribution	6	6	76.67%
Individual Causal Effect	6	14	75.71%
Counterfactual Reasoning	6	13	76.92%
Sufficient Cause	6	15	76.00%
Necessary Cause	6	15	78.67%
Responsibility	6	15	88.00%
<b>Overall</b>	36	78	<b>78.97%</b>
<i>334 Non-Human</i>			
Causality Identification	6	16	61.25%
Actual Cause	6	6	56.67%
<b>Overall</b>	12	22	60.00%
<b>Total</b>	48	100	60.00%

Rung/Type	# QAs	Avg Q/A Len
R1		
CI (YN)	804	21.1/1.0
CA (MC)	300	13.1/8.5
R2		
ICE (YN)	700	30.4/1.0
R3		
CR (YN)	640	23.3/1.0
SC (YN)	741	24.5/1.0
NC (YN)	741	24.5/1.0
AC (MC)	300	13.1/8.5
RSP (YN)	741	23.6/1.0
<b>Overall</b>	<b>4967</b>	<b>22.9/2.0</b>

Figure 5: Statistics of HVCR. YN, Yes/No; MC, Multiple Choice.

## 4 BASELINE EVALUATION

### 4.1 HUMAN EVALUATION

For synthetic videos, we have conducted a human evaluation to show *whether these generated videos truly reflect our principles*. In this test, we select 8 videos for each scenario in the Basic setting, each corresponding to a question type. The resulting size of the test is 48 videos and 100 questions. Totally, 5 people on the Prolific platform participated in the test. The results are shown in Table 1. For human tasks, the average human score is 78.97%. This suggests that our synthetic “one-take” videos are basically sufficient to convey causal knowledge of both the factual and counterfactual worlds, in line with human intuition. Also, this provides a small-scale baseline for evaluation.

### 4.2 LARGE MULTIMODAL MODELS

We evaluate a range of LMMs on the HVCR benchmark, including: 1) **Open-source “smaller” models**: LLaVA-OneVision-Qwen2-7B (Li et al., 2025a), Qwen2.5-VL-7B/32B-Instruct (Bai et al., 2025a), InternVL2.5-8B (Chen et al., 2024c), and Perception-LM-8B (Cho et al., 2025). Each model was run on a single NVIDIA H200 (141GB) GPU on our research cluster. Models are run with 16 uniformly sampled frames. Temperatures are set to zero for reproducibility. 2) **Closed-source “larger” models**: GPT-4o (OpenAI, 2024) and Gemini-2.5-Flash (Comanici et al., 2025). Both models are evaluated using 10 frames, uniformly sampled. Again, temperatures are set to zero.

### 4.3 RESULTS & DISCUSSION

#### 4.3.1 RESULTS ON SYNTHETIC VIDEOS

**Overall Results.** As shown in Table 2, Qwen2.5-VL-32B stands out with the best overall performance, achieving 58.33% accuracy, far surpassing large, closed-source models such as GPT-4o and Gemini-2.5-Flash. Among open-source models, InternVL2.5 leads, closely followed by Perception-LM, both outperforming closed-source alternatives. In contrast, models like LLaVA-OneVision and Qwen2.5-VL-7B show poorer results, especially the latter. Upon reviewing the failure modes of Qwen2.5-VL-7B, we found that it tends to select both answers for Yes/No questions when uncertain. The scaling effect is evident, with the 32B version showing a notable performance improvement. Closed-source models all fall below 50% accuracy, suggesting their open-source counterparts promising for video causal understanding.

**Human vs. Non-Human.** Of the 7 models tested, more than half (4 models) performed worse on Human tasks than on Non-Human tasks, indicating that the models still struggle with human-like causal reasoning. Furthermore, the high performance of these models on CI and AC tasks (especially Gemini-2.5-Flash, InternVL2.5, and Perception-LM, with AC accuracy above 80%) contrasts sharply with their lower performance on tasks like CA and RSP, suggesting that the models are better at matching formal frameworks than at reflecting human causal cognition.

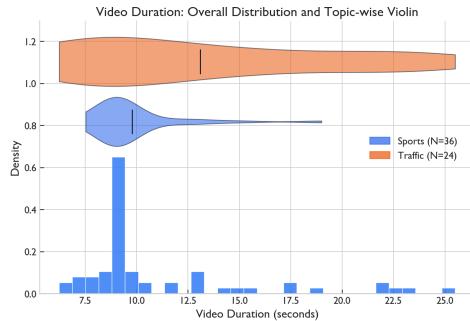


Figure 6: Duration distribution of realistic videos in HVCR.

378  
379  
Table 2: Performance of different models on synthetic videos in HVCR.  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431

			Closed			Open			
			GPT-4o	Gemini-2.5-Flash	InternVL2.5 (8)	LLaVA-OneVision (7)	Perception-LM (8)	Qwen2.5-VL (7)	Qwen2.5-VL (32)
# Videos # QAs									
<i>Accuracy per scenario per setting</i>									
SO	Basic	20	280	42.50%	55.36%	54.64%	21.07%	55.71%	10.00%
	Perturbed	20	280	38.93%	47.86%	55.36%	21.43%	53.57%	7.14%
SW	Basic	20	360	49.17%	51.39%	46.67%	39.72%	43.61%	26.11%
	Perturbed	20	360	46.67%	49.17%	48.33%	36.94%	51.78%	11.67%
LP	Basic	20	300	42.00%	44.67%	60.67%	23.33%	62.67%	15.00%
	Perturbed	20	300	35.33%	44.00%	53.33%	20.33%	55.33%	11.33%
EP	Basic	20	280	39.29%	41.79%	58.57%	26.79%	63.57%	19.64%
	Perturbed	20	280	28.93%	39.64%	51.43%	25.00%	62.86%	18.57%
DP	Basic	20	400	69.75%	62.75%	70.50%	36.75%	52.00%	44.75%
	Perturbed	20	400	64.50%	65.25%	70.00%	34.00%	53.75%	40.25%
BP	Basic	20	380	39.74%	41.58%	46.58%	44.74%	38.42%	38.16%
	Perturbed	20	380	35.26%	43.68%	51.84%	45.79%	39.21%	36.05%
<i>Accuracy per rung per type</i>									
R1	CI	240	640	63.44%	44.22%	56.09%	13.44%	66.25%	0.78%
	CA	240	240	49.17%	45.83%	47.92%	24.58%	51.25%	25.42%
R2	ICE	240	560	27.68%	43.93%	32.68%	35.71%	36.61%	31.07%
R3	CR	240	520	28.46%	41.15%	62.69%	15.96%	36.35%	23.85%
	SC	240	600	39.33%	41.83%	83.33%	28.50%	81.17%	16.00%
	NC	240	600	38.17%	50.50%	31.83%	67.83%	31.83%	48.33%
	AC	240	240	69.17%	85.42%	80.83%	49.17%	89.58%	31.25%
	RSP	240	600	60.00%	61.50%	61.33%	29.00%	56.00%	27.83%
<i>Accuracy of human vs. non-human tasks</i>									
Human	-	-	39.94%↓	47.85%↓	53.94%↓	35.06%↑	49.07%↓	29.23%↑	62.18%↑
Non-Human	-	-	65.00%	55.45%	62.84%	23.18%	72.61%	9.09%	44.66%
<i>Aggregated accuracy</i>									
Basic		120	2000	48.10%	50.00%	56.30%	33.20%	54.20%	27.30%
Perturbed		120	2000	42.80%	49.05%	55.50%	31.70%	54.30%	22.30%
<b>Overall</b>		240	4000	45.45%	49.53%	55.90%	32.45%	54.25%	24.80%

**Analysis by Scenario and Setting.** Models vary in their understanding of different scenarios. Apart from Qwen2.5-VL-32B, which achieves over 50% accuracy in all scenarios except DP, other models tend to perform well in certain scenarios but struggle in others. For instance, InternVL2.5 excels in understanding DP, but performs poorly in SW.

**Analysis by Rung and Type.** From the radar chart shown in Figure 7, it is evident that all models perform best and most consistently in the counterfactual rung, while there is a significant variation in performance across models in the discovery and intervention rungs. Additionally, some failure modes emerge. For instance, Qwen2.5-7B and LLaVA-OneVision primarily struggle with the discovery rung, while GPT-4o performs poorly in the intervention rung.

#### 4.3.2 RESULTS ON REALISTIC VIDEOS

**Overall Results.** Perception-LM is the only model to surpass 50% accuracy, achieving 58.22%, followed by Gemini-2.5-Flash and Qwen2.5-VL-32B. For synthetic data, the top-performing model, InternVL2.5, now ranks at the bottom, with Qwen2.5-VL-32B also falling below 50%. This highlights the gap between synthetic and real-

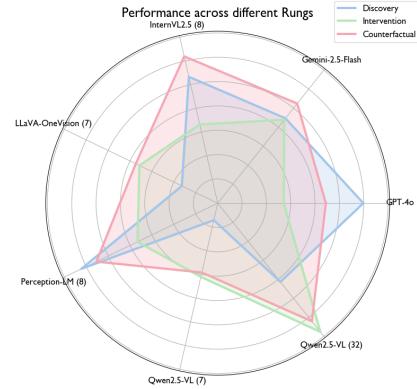


Figure 7: The performance of LMMs on synthetic videos by causal rung.

432 Table 3: Performance of different models on realistic videos in HVCR.  
433

			Closed			Open						
			GPT-4o	Gemini-2.5-Flash	InternLL2.5 (8)	LLaVa-OneVision (7)	Perception-LM (8)	Qwen2.5-VL (7)	Qwen2.5-VL (32)			
# Videos # QAs												
<i>Accuracy per scenario per setting</i>												
SW	Basic	20	303	43.56%	51.49%	43.89%	41.58%	51.16%	40.59%	53.47%		
LP	Basic	20	284	45.77%	44.01%	33.80%	33.10%	64.79%	27.82%	45.07%		
BP	Basic	20	380	39.21%	52.89%	38.16%	40.26%	58.95%	40.79%	48.16%		
<i>Accuracy per rung per type</i>												
R1	CI	60	164	62.20%	62.80%	30.49%	15.85%	79.27%	23.78%	74.39%		
	CA	60	60	23.33%	50.00%	43.33%	41.67%	38.33%	45.00%	26.67%		
R2	CDE	60	140	53.57%	50.00%	45.00%	35.71%	55.71%	37.14%	55.71%		
R3	CR	60	120	64.17%	63.33%	31.67%	18.33%	62.50%	23.33%	60.00%		
	SC	60	141	19.15%	26.95%	22.70%	19.86%	70.92%	24.11%	29.08%		
	NC	60	141	31.21%	48.23%	47.52%	82.98%	36.88%	67.38%	37.59%		
	AC	60	60	55.00%	55.00%	68.33%	65.00%	75.00%	45.00%	60.00%		
	RSP	60	141	27.66%	45.39%	40.43%	46.81%	42.55%	39.01%	39.01%		
<i>Accuracy of human vs. non-human tasks</i>												
Human		60	743	37.15%↓	46.57%↓	38.09%↓	41.45%↑	52.22%↓	39.17%↑	42.40%↓		
Other		60	224	60.27%	60.71%	40.63%	29.02%	78.13%	29.46%	70.54%		
<i>Aggregated accuracy</i>												
Overall		60	967	42.50%	49.84%	38.68%	38.57%	58.22%	36.92%	48.91%		

456 world evaluations, suggesting that *perception* ability might be a crucial factor in real-world scenarios.

457 **Human vs. Non-Human.** The gap between human and non-human tasks becomes even more  
458 pronounced in real-world situations. The number of models with lower performance on human tasks  
459 increases from 4 to 5, and the disparity is particularly significant for GPT-4o and Qwen2.5-VL-32B.

## 463 5 RELATED WORK

464 The field of video causal reasoning has seen the introduction of a variety of benchmarks designed  
465 to assess model capabilities in understanding dynamic scenes. Datasets like CLEVRER (Yi et al.,  
466 2020), CRAFT Ates et al. (2022), and IntPhys (Rochet et al., 2018) use synthetically generated  
467 videos to control for visual complexity while focusing on physical interactions and causal events.  
468 CausalVQA (Foss et al., 2025) and NExT-QA (Xiao et al., 2021) shift the focus to real-world videos,  
469 with the former specifically designed to test physical causality in egocentric settings. To challenge  
470 models with more complex, long-form videos and reasoning tasks, benchmarks such as CausalStep  
471 (Li et al., 2025b), MECD (Chen et al., 2024b), and MECD+ (Chen et al., 2025) were introduced,  
472 with MECD focusing on discovering comprehensive event-level causal graphs. Other benchmarks,  
473 including MVP-Bench (Li et al., 2024), GRASP (Jassim et al., 2024), VCRBench (Sarkar & Etemad,  
474 2025), and Video-Holmes (Cheng et al., 2025), evaluate more fine-grained reasoning abilities like  
475 detecting shortcuts, grounding language to physical actions, or actively seeking clues in long videos.

## 477 6 CONCLUSION

480 HVCR introduces a novel and systematic approach to evaluating human-like causal reasoning in  
481 large multimodal models (LMMs), providing a critical tool for assessing the nuanced capabilities  
482 required for complex video causal cognition. By aligning with the principles of actual causality, in-  
483 corporating human consensus in goal-oriented evaluation, and leveraging causal graphical models as  
484 representations, HVCR offers a comprehensive framework that extends beyond current benchmarks.  
485 Despite promising results in human evaluation, the performance gap observed in LMMs highlights  
the challenges these models face in replicating human-like causal judgments.

486  
487

## 7 ETHICS STATEMENT

488  
489  
490  
491  
492  
493  
494

The authors of this paper have read and adhere to the ICLR Code of Ethics. Our research involves human evaluation, with labels collected through the Prolific platform. All participants were compensated for their time and provided informed consent. The study was conducted in a manner that respects the privacy and well-being of all human subjects, and no personally identifiable information was collected. We have taken care to ensure that the benchmark is created and released in a manner that avoids potential misuse or harm, aligning with our goal of responsibly advancing the field of large multimodal models.

495

496  
497

## 8 REPRODUCIBILITY STATEMENT

498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539

To ensure the reproducibility of our work during the reviewing process, we provide the data and code via the following link: <https://anonymous.4open.science/r/HVCR-A9B6/>.

540 REFERENCES  
541

542 Holger Andreas, Matthias Armgardt, and Mario Gunther. Counterfactuals for causal responsibility  
543 in legal contexts. *Artificial intelligence and law*, 31(1):115–132, 2023.

544 Tayfun Ates, M Ateşoğlu, Çağatay Yiğit, İlker Keser, Mert Kobas, Erkut Erdem, Aykut Erdem,  
545 Tilbe Goksun, and Deniz Yuret. Craft: A benchmark for causal reasoning about forces and inter-  
546 actions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2602–2627,  
547 2022.

548 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,  
549 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. [arXiv preprint arXiv:2502.13923](https://arxiv.org/abs/2502.13923),  
550 2025a.

552 Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. In *Forty-second International  
553 Conference on Machine Learning*, 2025b. URL <https://openreview.net/forum?id=MNSW6U5zUA>.

555 Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings  
556 of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI’94, pp. 230–237.  
557 AAAI Press, 1994.

559 Nathan Blake, David A Kelly, Akchunya Chanchal, Sarah Kapllani-Mucaj, Geraint Thomas,  
560 and Hana Chockler. Specrex: Explainable ai for raman spectroscopy. [arXiv preprint  
561 arXiv:2503.14567](https://arxiv.org/abs/2503.14567), 2025.

562 Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation,  
563 Blender Institute, Amsterdam, 2016. URL <http://www.blender.org>.

565 Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel  
566 Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environ-  
567 ments. [arXiv preprint arXiv:2506.09849](https://arxiv.org/abs/2506.09849), 2025.

568 Nancy Cartwright. What are randomised controlled trials good for? *Philosophical studies*, 147(1):  
569 59–70, 2010.

571 Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie  
572 Zhao, Yu Qiao, and Chaochao Lu. Causal evaluation of language models. [arXiv preprint  
573 arXiv:2405.00622](https://arxiv.org/abs/2405.00622), 2024a.

575 Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang  
576 Zhang, Yingxue Wang, Hui Lin, et al. Mecd: Unlocking multi-event causal discovery in video  
577 reasoning. *Advances in Neural Information Processing Systems*, 37:92554–92580, 2024b.

581 Tieyuan Chen, Huabin Liu, Yi Wang, Yihang Chen, Tianyao He, Chaofan Gan, Huanyu He, and  
582 Weiyao Lin. Mecd+: Unlocking event-level causal graph discovery for video reasoning. [arXiv  
583 preprint arXiv:2501.07227](https://arxiv.org/abs/2501.07227), 2025.

586 Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shen-  
587 glong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source  
588 multimodal models with model, data, and test-time scaling. [arXiv preprint arXiv:2412.05271](https://arxiv.org/abs/2412.05271),  
589 2024c.

592 Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can  
593 mllm think like holmes for complex video reasoning? [arXiv preprint arXiv:2505.21374](https://arxiv.org/abs/2505.21374), 2025.

598 Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and  
600 Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? *Advances  
601 in Neural Information Processing Systems*, 37:96640–96670, 2024.

605 Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan,  
606 Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-  
607 access data and models for detailed visual understanding. [arXiv preprint arXiv:2504.13180](https://arxiv.org/abs/2504.13180), 2025.

594 Hana Chockler, David A Kelly, Daniel Kroening, and Youcheng Sun. Causal explanations for image  
 595 classifiers. [arXiv preprint arXiv:2411.08875](https://arxiv.org/abs/2411.08875), 2024.

596

597 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit  
 598 Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the  
 599 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-  
 600 bilities. [arXiv preprint arXiv:2507.06261](https://arxiv.org/abs/2507.06261), 2025.

601 Erwin Coumans. Bullet Physics Engine. [Open Source Software: http://bulletphysics.org](http://bulletphysics.org), 2010.

602

603 Clemens Dubslaff, Kallistos Weis, Christel Baier, and Sven Apel. Causality in configurable software  
 604 systems. In [Proceedings of the 44th International Conference on Software Engineering](#), pp. 325–  
 605 337, 2022.

606

607 Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao.  
 608 Causalvqa: A physically grounded causal reasoning benchmark for video models. [arXiv preprint  
 609 arXiv:2506.09943](https://arxiv.org/abs/2506.09943), 2025.

610 Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. A counter-  
 611 factual simulation model of causal judgments for physical events. [Psychological review](#), 128(5):  
 612 936, 2021.

613

614 Eugenia Goldvarg and Philip N Johnson-Laird. Naive causality: A mental model theory of causal  
 615 meaning and reasoning. [Cognitive science](#), 25(4):565–610, 2001.

616 Joseph Y Halpern. [Actual causality](#). MIT Press, 2016.

617

618 Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i:  
 619 Causes. [The British journal for the philosophy of science](#), 2005.

620

621 Daniel Murray Hausman. Causal relata: Tokens, types, or variables? [Erkenntnis](#), 63(1):33–54,  
 622 2005.

623

624 Xudong Hong, Margarita Ryzhova, Daniel Biondi, and Vera Demberg. Do large language models  
 625 and humans have similar behaviours in causal inference with script knowledge? In [Proceedings of  
 626 the 13th Joint Conference on Lexical and Computational Semantics \(\\* SEM 2024\)](#), pp. 421–437,  
 627 2024.

628

629 Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni.  
 630 Grasp: A novel benchmark for evaluating language grounding and situated physics understanding  
 631 in multimodal language models. In [IJCAI](#), 2024.

632

633 Abhinav Joshi, Ashutosh Modi, et al. Cold: Causal reasoning in closed daily activities. [Advances  
 634 in Neural Information Processing Systems](#), 37:5145–5187, 2024.

635

636 Sangeet S Khemlani, Aron K Barbey, and Philip N Johnson-Laird. Causal reasoning with mental  
 637 models. [Frontiers in human neuroscience](#), 8:849, 2014.

638

639 Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language  
 640 models: Opening a new frontier for causality. [Transactions on machine learning research](#), 2024.

641

642 Konstantin Raphael Kueffner. A comprehensive survey of the actual causality literature. 2021.

643

644 David A Lagnado and Shelley Channon. Judgments of cause and blame: The effects of intentionality  
 645 and foreseeability. [Cognition](#), 108(3):754–770, 2008.

646

647 David A Lagnado and Tobias Gerstenberg. Causation in legal and moral reasoning. 2017.

648

649 Thao Minh Le, Vuong Le, Kien Do, Sunil Gupta, Svetha Venkatesh, and Truyen Tran. Finding the  
 650 trigger: Causal abductive reasoning on video events. [arXiv preprint arXiv:2501.09304](https://arxiv.org/abs/2501.09304), 2025.

651

652 David Lewis. Causation as influence. [The Journal of Philosophy](#), 97(4):182–197, 2000.

648 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan  
 649 Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task trans-  
 650 fer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL <https://openreview.net/forum?id=zKv8qULV6n>.

651

652 Guanzhen Li, Yuxi Xie, and Min-Yen Kan. Mvp-bench: Can large vision-language models conduct  
 653 multi-level visual perception like humans? In *Findings of the Association for Computational  
 654 Linguistics: EMNLP 2024*, pp. 13505–13527, 2024.

655

656 Xuchen Li, Xuzhao Li, Shiyu Hu, Kaiqi Huang, and Wentao Zhang. Causalstep: A benchmark for  
 657 explicit stepwise causal reasoning in videos. *arXiv preprint arXiv:2507.16878*, 2025b.

658

659 Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A  
 660 multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the  
 661 IEEE/CVF International Conference on Computer Vision*, pp. 13536–13545, 2021.

662

663 Pengyuan Lu, Ivan Ruchkin, Matthew Cleaveland, Oleg Sokolsky, and Insup Lee. Causal repair  
 664 of learning-enabled cyber-physical systems. In *2023 IEEE International Conference on Assured  
 665 Autonomy (ICAA)*, pp. 1–10. IEEE, 2023.

666

667 Jiayuan Mao, Xuelin Yang, Xikun Zhang, Noah Goodman, and Jiajun Wu. Clevrer-humans: De-  
 668 scribing physical and causal events the human way. *Advances in Neural Information Processing  
 669 Systems*, 35:7755–7768, 2022.

670

671 Dwayne Moore. Causal exclusion and dependent overdetermination. *Erkenntnis*, 76(3):319–335,  
 672 2012.

673

674 Shota Nishiyama, Takuma Saito, Ryo Nakamura, Go Ohtani, Hirokatsu Kataoka, and Kensho Hara.  
 675 Traffic incident database with multiple labels including various perspective environmental infor-  
 676 mation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*,  
 677 pp. 11377–11383. IEEE, 2023.

678

679 OpenAI. Gpt-4o, 2024. URL <https://openai.com/index/gpt-4o-system-card/>.

680

681 Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic  
 682 books, 2018.

683

684 Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*.  
 685 John Wiley & Sons, 2016.

686

687 Pavana Pradeep Kumar and Krishna Kant. Tu-dat: A computer vision dataset on road traffic anom-  
 688 alies. *Sensors*, 25(11):3259, 2025.

689

690 Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique  
 691 Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics  
 692 reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

693

694 Pritam Sarkar and Ali Etemad. Vcrbench: Exploring long-form causal reasoning capabilities of  
 695 large video language models. *arXiv preprint arXiv:2505.08455*, 2025.

696

697 Amit Sharma, Hua Li, and Jian Jiao. The counterfactual-shapley value: Attributing change in system  
 698 metrics. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

699

700 Steven Sloman and Steven A Sloman. *Causal models: How people think about the world and its  
 701 alternatives*. Oxford University Press, 2009.

702

703 Leonard Talmy. Force dynamics in language and cognition. *Cognitive science*, 12(1):49–100, 1988.

704

705 Phillip Wolff. Representing causation. *Journal of experimental psychology: General*, 136(1):82,  
 706 2007.

707

708 Dekun Wu, He Zhao, Xingce Bao, and Richard P Wildes. Sports video analysis on large-scale data.  
 709 In *European conference on computer vision*, pp. 19–36. Springer, 2022.

702 Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-  
703 answering to explaining temporal actions. In Proceedings of the IEEE/CVF conference on  
704 computer vision and pattern recognition, pp. 9777–9786, 2021.

705  
706 Yajun Xu, Huan Hu, Chuwen Huang, Yibing Nan, Yuyao Liu, Kai Wang, Zhaoxiang Liu, and Shiguo  
707 Lian. Tad: A large-scale benchmark for traffic accidents detection from video surveillance. IEEE  
708 Access, 2024.

709 Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B  
710 Tenenbaum. Clevrer: Collision events for video representation and reasoning. In International  
711 Conference on Learning Representations, 2020.

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 A THE USE OF LARGE LANGUAGE MODELS (LLMs)  
757758 In this study, LLMs are used solely for the purpose of polishing the writing and enhancing the clarity  
759 of the manuscript. They assist in refining sentence structure, improving grammar, and ensuring  
760 overall coherence. It did not contribute to the ideation or content generation of the research. The  
761 authors take full responsibility for the integrity of the research and the final content presented in this  
762 paper.764 B SUMMARY OF CAUSALITY-RELATED VIDEO REASONING BENCHMARKS  
765

766 767 Dataset	768 Realism		769 Labeling		770 Definition	771 Evaluation Goal	
	772 Synth	773 Real	774 Auto	775 Human		776 Human	777 Formal
778 IntPhys (Riochet et al., 2018)	✓	✗	✓	✗	779 VOE	✓	✗
780 IntPhys 2 (Bordes et al., 2025)	✓	✗	✓	✗	781 VOE	✓	✗
782 CLEVRER (Yi et al., 2020)	✓	✗	✓	✗	783 Heuristic	✗	✓
784 CLEVRER-Humans (Mao et al., 2022)	✓	✗	✗	✓	785 Human Judgment	✓	✗
786 CRAFT (Ates et al., 2022)	✓	✗	✓	✗	787 Force Dynamics	✓	✗
788 MECD (Chen et al., 2024b)	✗	✓	✓	✗	789 Granger Causality	✗	✓
790 MECD+ (Chen et al., 2025)	✗	✓	✗	✓	791 Granger Causality	✗	✓
792 CausalVQA (Foss et al., 2025)	✗	✓	✓	✓	793 Physical	✓	✗
794 CausalStep (Li et al., 2025b)	✗	✓	✗	✓	795 Stepwise Reasoning	✓	✗
796 VCRBench (Sarkar & Etemad, 2025)	✗	✓	✗	✓	797 Event Dependency	✓	✗
798 NExT-QA (Xiao et al., 2021)	✗	✓	✗	✓	799 Visible Cause	✓	✗
800 Finding the Trigger (Le et al., 2025)	✓	✓	✓	✗	801 Causal Abduction	✗	✓
802 Impossible Videos (Bai et al., 2025b)	✓	✗	✗	✓	803 Anti-Commonsense	✓	✗
804 GRASP (Jassim et al., 2024)	✓	✗	✓	✗	805 Intuitive Physics	✓	✗
806 Video-Holmes (Cheng et al., 2025)	✗	✓	✗	✓	807 Narrative Logic	✓	✗

785 Table 4: Comparison of HVCR with other causality-related video reasoning benchmarks.  
786

810 B.1 HUMAN EVALUATION  
811812 Labeling Instruction  
813814 **Goal**815 The purpose of this project is to investigate human intuition about causal relationships in  
816 videos. Your task is to carefully observe the motion and interactions of objects in each video  
817 and answer the accompanying questions based on your immediate perceptions.818 **Important:** Focus on your intuition rather than whether your answers are “correct” or “in-  
819 correct.” Respond based on how events appear to you.820 **Data**  
821822 1. *Video Type*: Synthetically generated videos that follow predefined causal structures.  
823 2. *Scene Description*: Objects move or remain stationary on a horizontally infinite tabletop.  
824 3. *Video Length*: 5 seconds (125 frames per video).  
825 4. *Object Properties*:826     • Each object has a unique combination of color, material, and shape.  
827     • No two objects within the same video share the exact same combination.828 5. *Question Types*:  
829     • Yes/No questions  
830     • Multiple-choice questions (more than one option may apply)831 **Guidelines**832 1. *Video Viewing Recommendations*: Watch each video multiple times for a comprehensive  
833 understanding. Some movements occur rapidly; a single viewing may not capture all  
834 relevant details.835 2. *Event Types to Observe*:836 (a) Unary Events (single-object events)  
837     • Motion Status: Determine whether an object is moving or stationary.  
838     • Motion Transitions: Identify when an object starts or stops moving.  
839     • Visibility Changes: Note when an object enters or exits the visible area.840 (b) Binary Events (interactions between two objects)  
841     • Collisions: Identify whether objects collide.  
842     • Movement Direction Relative to Another Object:  
843         – Typical scenario: a moving object approaches a stationary object.  
844         – Critical scenario: a moving object approaches another moving object, creating  
845             a potential collision.846 3. *Attention to Question Tense*: Pay attention to the tense used in each question (past,  
847 present, or continuous) and ensure your responses align with the chronological sequence  
848 of events observed in the video.849 We wish you an efficient and insightful labeling experience!  
850851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863