AAAI 2026
January 20 – 27, 2026
Singapore

CausalStep: A Benchmark for Explicit Stepwise Causal Reasoning in Videos

Xuchen Li[1,2,3*] Xuzhao Li[4*] Shiyu Hu[4] Kaiqi Huang[1,2†] Wentao Zhang[3,5†]

[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Zhongguancun Academy
[4]Nanyang Technological University
[5]Peking University
s-lxc24@bjzgca.edu.cn, kqhuang@ia.ac.cn, wentao.zhang@pku.edu.cn

Dr. Shiyu Hu

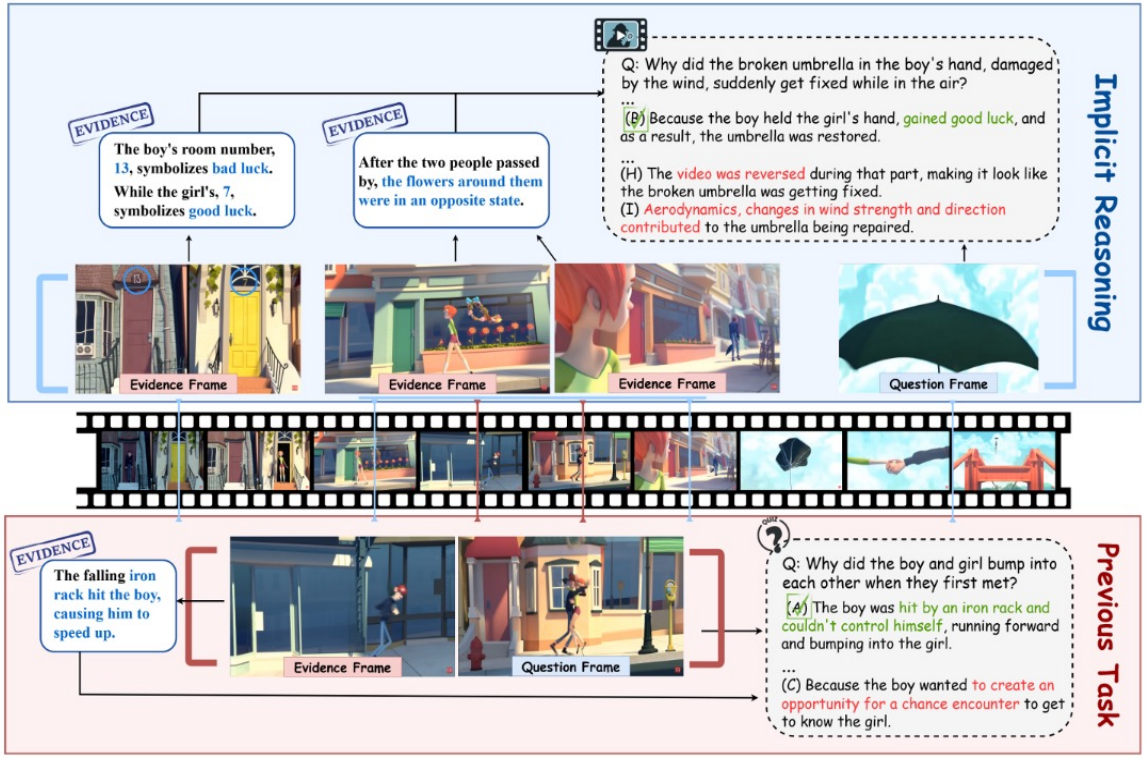
- Research Fellow in Nanyang Technological University (NTU)
- https://huuuuusy.github.io/
- shiyu.hu@ntu.edu.sg

Scan to download this slides

WeChat for the first author

Motivation

Most video reasoning benchmarks focus on **perception or shallow understanding**, requiring only the **identification of relevant frames or context**.



Implicit Reason

For Perception

Limitation 1: By providing the entire video as input, these benchmarks allow models to **exploit global information** or **shortcut strategies**, thereby failing to assess true causal and stepwise reasoning.

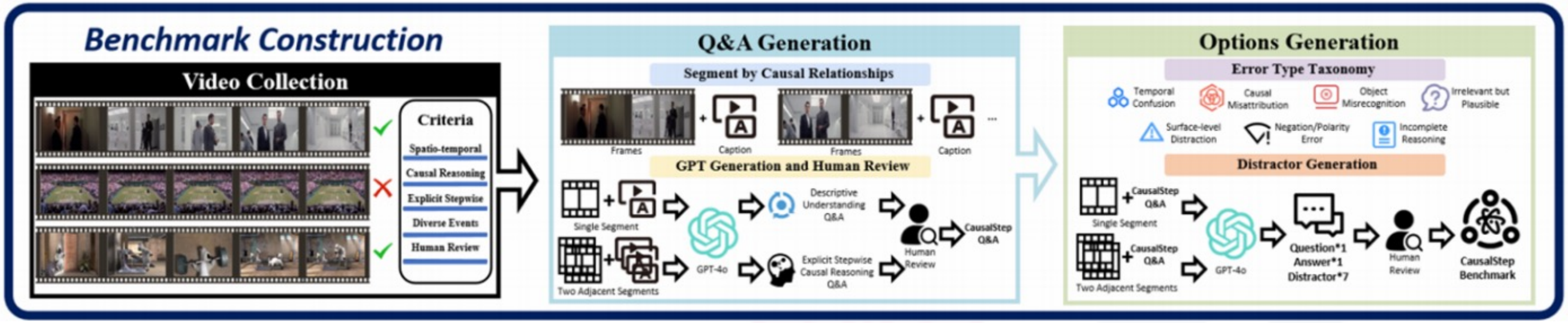
Limitation 2: The design of distractor options in multiple choice questions is often **unsystematic**, lacking systematic coverage of common reasoning errors.

Our Solution: CasualStep

A novel benchmark for explicit stepwise causal reasoning

Our Method: CausalStep

We introduce CausalStep,which **segments videos into causally linked units** and enforces a **strict stepwise QA protocol**, enabling rigorous evaluation of sequential, causally grounded reasoning in complex video narratives.

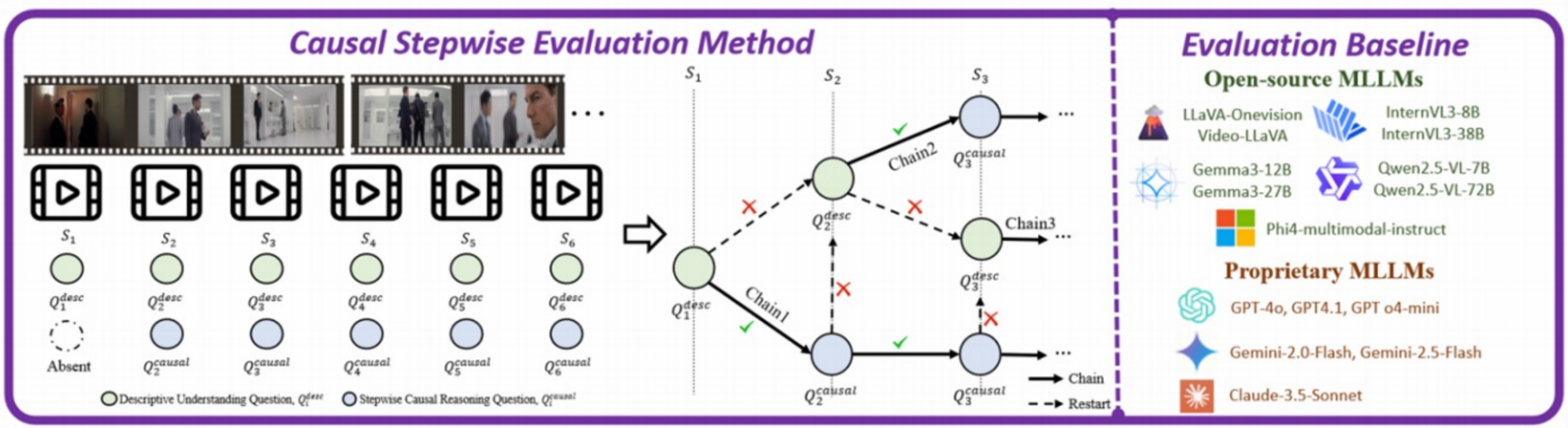


Data Contribution: Explicitly Embedding Causal Structure

- **Causal Segmentation:** Long videos are segmented into **causally linked event units**, rather than arbitrary clips.
- **Causal Question Design:** Questions are generated around **adjacent causal relations**, instead of simple perception or frame retrieval.
- **Taxonomy-based Distractors:** Distractor options are systematically designed based on common causal and temporal error types, improving **diagnostic power**.

→ Preventing shallow understanding based on relevant-frame identification.

Our Method: CausalStep



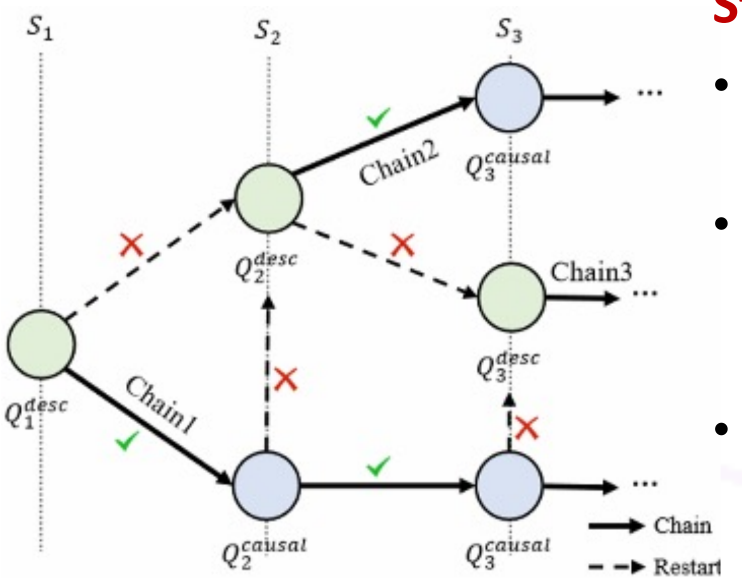
Evaluation Contribution: Enforcing Step-by-Step Causal Reasoning

- **Strict Stepwise QA Protocol:** At each step, the model can only access the **current causal segment**, with no future information.
- **Chain Dependency and Restart Mechanism:** Any incorrect step breaks the causal chain, making **stepwise reasoning mandatory**.
- **Process-level Evaluation Metrics:** We evaluate whether models can **consistently maintain causal chains**, beyond question-level accuracy.

→ Systematically eliminating shortcuts enabled by global context.

Data defines what causal structure is, while evaluation ensures that stepwise causal reasoning is the only viable strategy.

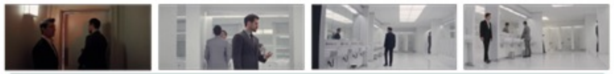
Details: Annotation Strategy



Stepwise Reasoning Chain Annotation

- The reasoning chain begins with the Q^{desc} for the segment S_1.
- If the current Q_i^{desc} is answered correctly, the chain will proceed to the Q^{causal} in the segment S_{i+1}. If any answer is incorrect, the reasoning chain is interrupted.
- At each step with a Q_i^{causal}, the model is provided with the current segment S_i and its direct preceding segment S_{i-1}, along with its previously correct answer.

Causal Segment 1



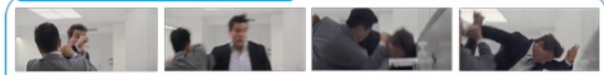
Q_1^{desc}: What are the two men primarily doing?
A: They just came out of the restroom and are preparing to leave the room.
B: One of the men is holding a mobile phone in his hand.
C: They are standing motionless in the middle of the hallway.
D: The restroom has many white sinks.
E: They are searching for a hidden secret entrance.
F: They first stand beside a door, then enter a bright restroom.
G: They are merely walking around in the restroom.
H: They are dining in a restaurant.

Causal Segment 2

Q_2^{desc}: What are the men primarily doing in the restroom?
A: One man is fixing his hair before leaving the restroom.
B: They gather in the restroom to conduct a secret transaction.
C: One man is talking to a female server.
D: The restroom has many white stalls and tiles.
E: One man walks past the restroom, while the other handles items on the counter.
F: Many men are wearing business suits and shirts.
G: No one in the restroom is handling items on the counter.
H: One man is merely passing by the restroom.

Q_2^{causal}: Why is the man to the left of the man standing in the middle positioned that way?
A: He is waiting for the right moment to act immediately after his companion completes the task.
B: He stands there to avoid being caught on the front face by the cameras in the room.
C: He is actually helping the man who seems to be unwell to keep his balance.
D: He is making a strategic deployment for a secret operation.
E: He is just adjusting his body's center of gravity to maintain an alert posture.
F: The color of the lining of his suit coordinates very well with the environment.
G: He stands there just to observe the reaction of the man in the middle.
H: He is not carrying out any secret deployment, but is observing the exit.

Causal Segment 3



Q_3^{desc}: What mainly happens among these men?
A: They are having a friendly physical training session.
B: The floor of the room is very smooth and reflective.
C: One man is helping another man do stretching exercises.
D: A fierce physical fight is taking place among these men.
E: These men are negotiating calmly and no conflicts have occurred.
F: They are apologizing to each other due to an accidental collision.
G: The movements of the characters in the picture are blurry, indicating rapid movement.
H: One of the men is falling down.

Q_3^{causal}: Why did the physical conflict in the restroom suddenly break out?
A: The man in the middle suddenly made a move, trying to snatch an item from one of them.
B: One of the men wearing a dark suit suddenly launched a hidden attack during the contact, triggering a fight.
C: They did not break out into a physical conflict; instead, they were performing a difficult collaborative act.
D: The mirror in the restroom vibrated due to the fight and made a noise.
E: There had been a long-standing conflict between them, and at this moment, the conflict finally erupted.
F: One of the men accidentally bumped into another man, leading to an unexpected friction.
G: Due to rapid movement and chaos, their actions look like they are fighting.
H: The conflict broke out because they failed to reach an agreement on the negotiation that had already taken place before.

Details: Annotation Strategy

Taxonomy-Based Distractor Generation

- For each question, we first define **several typical error types**. Distractor options are then systematically generated to cover these categories.
- **GPT-4o generates** plausible but incorrect alternatives that are contextually relevant and semantically similar to the correct answer.
- **Human annotators review** and edit these distractors, ensuring they are non-trivial, factually sound, and that each distractor fits its intended error type and maintains comparable plausibility.

Causal Segment 2

Q_2^{desc}: What are the men primarily doing in the restroom?

A: One man is fixing his hair before leaving the restroom. *Temporal Confusion*

B: They gather in the restroom to conduct a secret transaction. *Causal Misattribution*

C: One man is talking to a female server. *Object / Actor Misrecognition*

D: The restroom has many white stalls and tiles. *Irrelevant but Plausible*

E: One man walks past the restroom, while the other handles items on the counter. *Correct Answer*

F: Many men are wearing business suits and shirts. *Surface-level Distraction*

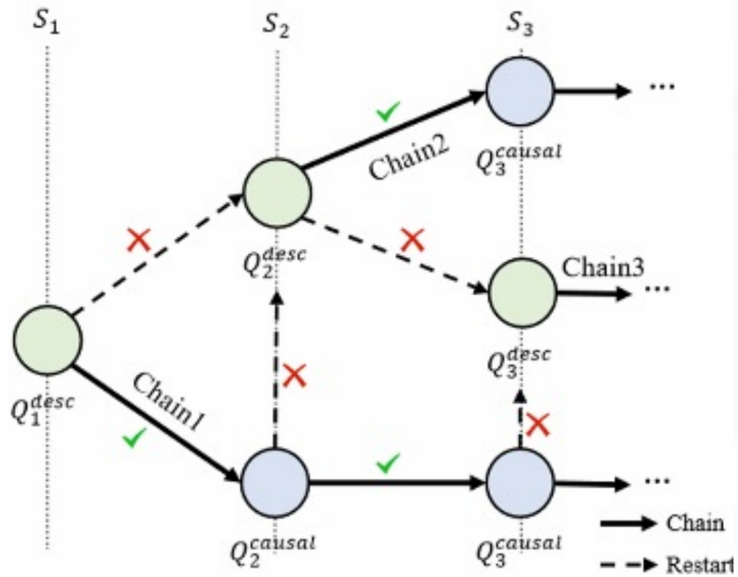
G: No one in the restroom is handling items on the counter. *Negation / Polarity Error*

H: One man is merely passing by the restroom. *Incomplete Reasonin*

Details: Evaluation Mechanism

CausalStep Evaluation Framework

- **Five key metrics:**
  - Chain Success Rate (CSR)
  - Average Maximum Chain Length (AMCL)
  - Maximum Chain Length (MCL)
  - Restart Frequency (RF)
  - Weighted Score (WS)
- **Two supplementary indicators:**
  - Descriptive Understanding Accuracy (DUA)
  - Isolated Causal Reasoning Accuracy (ICRA)



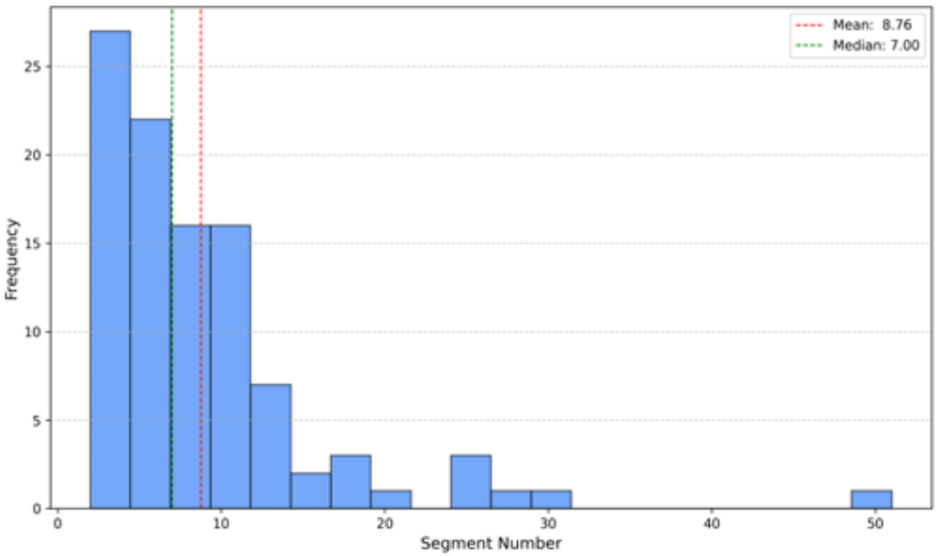
Algorithm 1 CausalStep Evaluation Framework

Input:
Segments $[S_1, S_2, \ldots, S_N]$;
Descriptive QA list $[Q_1^{desc}, Q_2^{desc}, \ldots, Q_N^{desc}]$;
Reasoning QA list $[Q_2^{causal}, \ldots, Q_N^{causal}]$;
Model $\mathtt{M}$
Output: Total score for the video

$score \leftarrow 0$;
$chain\_length \leftarrow 0$;
$i \leftarrow 1$;
$current\_question\_type \leftarrow$ 'desc';
while $i \leq N$ **do**
  if $current\_question\_type ==$ '*desc*' **then**
    $desc\_ans \leftarrow \mathtt{M.Answer}(Q_i^{desc}, S_i)$
    if $is\_correct(desc\_ans)$ **then**
      $chain\_length \leftarrow chain\_length + 1$;
      $score \leftarrow score + 1$;
      $i \leftarrow i + 1$;
      $current\_question\_type \leftarrow$ 'causal';
    else
      $chain\_length \leftarrow 0$; // Restart
      $i \leftarrow i + 1$;
      $current\_question\_type \leftarrow$ 'desc';
  if $current\_question\_type ==$ '*causal*' **then**
    if $i > N$ **then**
      break;
    $causal\_ans \leftarrow \mathtt{M.Answer}(Q_i^{causal}, A_{i-1}, [S_{i-1}, S_i])$
    if $is\_correct(causal\_ans)$ **then**
      $chain\_length \leftarrow chain\_length + 1$;
      $score \leftarrow score + chain\_length$;
      $i \leftarrow i + 1$;
      $current\_question\_type \leftarrow$ 'causal';
    else
      $chain\_length \leftarrow 0$; // Restart
      $i \leftarrow i + 1$;
      $current\_question\_type \leftarrow$ 'desc';
return: $score$;

Details: Data Statistics

- **100 videos** (average duration 430.5 seconds, ranging from 149 to 994.4 seconds)
- **6 diverse categories** (Cartoons, Movies & TV Shows, Outdoor Sports, Regular Sports)
- An average of **8.76 causal segments** (ranging from 2 to 51 segments per video)
- A total of **1,852 multiple-choice QA pairs**, covering descriptive understanding questions and causal reasoning questions
- Each question **averages 8 options**, including 1 correct answer and 7 challenging distractors



Segments Distribution

Statistic	Value
#Videos	100
Video duration (mean)	430.5 s
Video duration (min / max)	149 s / 994.4 s
#QA pairs	1,852
QA type	Multiple-choice
Options per question	8
#Categories	6
Avg. segments per video	8.76
Segments per video (min / max)	2 / 51
Annotation	AI-assisted + Manual
Distractor design	Error-type taxonomy
Descriptive QA pairs	926
Reasoning QA pairs	926

Main Results

We provide the performance of a diverse set of open-source and proprietary models, alongside human baselines.

Model	CSR(%) ↑	AMCL ↑	MCL ↑	RF ↓	WS ↑	DUA(%) ↑	ICRA(%) ↑
Open-source models							
LLaVA-Onevision [18]	7	5.20	4	3.14	30.85	67.1	15.2
Video-LLaVA [25]	10	5.15	5	3.13	32.94	68.6	20.1
Phi4-multimodal-instruct [1]	13	5.33	4	3.01	33.78	70.1	21.4
Qwen2.5-VL-7B [45]	16	5.61	9	2.68	35.42	71.0	21.8
InternVL3-8B [52]	19	5.59	8	2.87	35.26	69.2	23.1
Gemma3-12b-it [16]	21	5.53	11	2.81	36.22	72.9	24.5
InternVL3-38B [52]	24	5.75	13	2.57	36.89	75.3	25.1
Qwen2.5-VL-72B [45]	26	5.89	17	2.47	37.69	76.1	25.2
Gemma3-27b-it [16]	29	5.94	20	2.42	37.64	77.7	26.3
Proprietary models							
Gemini-2.0-Flash [34]	31	6.04	21	2.45	39.60	79.4	27.1
Claude-3.5-Sonnet-20241022 [2]	35	5.87	23	2.37	38.58	80.9	28.5
GPT-4o-2024-11-20 [29]	39	5.94	23	2.17	38.88	80.0	29.7
Gemini-2.0-Flash-thinking [34]	41	6.15	25	2.15	40.65	81.1	30.2
GPT-4.1-2025-04-14 [31]	42	6.63	26	1.85	45.59	82.8	32.3
Gemini-2.5-Flash [10]	48	6.90	27	**1.68**	47.63	84.6	36.2
o4-mini-2025-04-16 [32]	**51**	**7.19**	**30**	1.69	**55.06**	**85.2**	**39.8**
Best Performance of Models	51	7.19	30	1.68	55.06	85.2	39.8
Human	79	8.03	46	0.74	62.39	92.0	76.8
Maximum	100	8.76	51	0	68.76	100.0	100.0

Analysis and Discussion

Experimental analysis: MLLMs' Strengths and Limitations in CausalStep

- A **substantial and persistent gap** between current **MLLMs and human-level performance** across all diagnostic metrics, underscoring the demanding nature of the CausalStep benchmark.

- Current models **struggle to perform accurate causal reasoning** when presented solely with an isolated segment pair, without the benefit of a preceding, correctly established reasoning chain.

- Even **the most advanced proprietary models remain considerably behind human-level performance**.

- We believe that CausalStep will serve as a vital tool to inspire and guide the community in **pushing the boundaries of video reasoning and advancing towards human-level causal intelligence** in complex, real-world scenarios.

Thanks for listening!

2026.01.25 in Singapore

Dr. Shiyu Hu

- Research Fellow in Nanyang Technological University (NTU)
- https://huuuuusy.github.io/
- shiyu.hu@ntu.edu.sg

Scan to download this slides

WeChat for the first author