When Video Verifier Scores Mislead: Capability Assessment and Verifier Selection

Anonymous authors
Paper under double-blind review

Abstract

We audit video verifiers, models that judge whether a candidate answer or trajectory agrees with a video, and show that their scores can mislead both conclusions about model ability and verifier selection. We vary two evaluation choices—negative type and judgment interface—while holding the tested models fixed. On controlled tasks, direct judgment yields win rates of 0.78–0.94 against visible-start negatives but only 0.46–0.58 against endpoint-matched negatives that require checking the internal trajectory. Yet on paired videos where the correct candidate reverses with the internal evidence, CoT judgment raises paired success from 0–6% to 75–88% for three of four models, showing that poor direct-judgment performance need not imply an absence of evidence-responsive discrimination. This CoT gain does not transfer to natural video: across 400 videos, CoT judgment yields no significant accuracy gain for any model and significantly reduces accuracy against event-denial negatives for three of four. Negative type also affects verifier selection: when content-matched negatives are the target, selecting on event-denial performance reduces held-out accuracy by 8.9 percentage points. These results separate success on the negative type tested, ability that can be elicited from the model, and suitability for the intended verification target.

1 Introduction

A video verifier judges whether a candidate answer or trajectory agrees with a video. Using such a model to evaluate or select outputs raises two questions: what does its measured performance establish about its abilities, and does the evaluation identify a useful verifier for the intended task? A strong video model is not automatically a reliable evaluator (Liu & Zhang, 2025). We study how the negative type and the judgment interface lead to different conclusions about the same frozen models (Figure 1).

Consider a verifier checking the trajectory of an object hidden under shuffled cups. A negative (incorrect) trajectory may begin under the wrong cup but otherwise follow the task's rules; the verifier can reject this *visible-start* negative from the initial reveal alone. A negative can instead begin and end under the correct cups and differ only in between. This *endpoint-matched* negative is also legal, since another sequence of allowed events would produce it, and rejecting it requires the internal evidence rather than an incorrect endpoint. A high score against the first type of negative says little about the second.

Poor performance on endpoint-matched negatives is open to two readings: the model cannot make the judgment, or the interface fails to elicit an ability the model has. We compare *direct judgment*, which reads the verdict from next-token answer scores, with *CoT judgment*, in which the model writes a rationale before its final answer (Section 2.2). Higher accuracy under CoT judgment alone does not show that the new decisions follow the visual evidence. We therefore pair videos that share their endpoints and both candidate trajectories but have opposite correct answers. Within a candidate order the text input is identical, so a verifier answers both videos correctly only if its choice changes with the internal evidence.

Two further questions concern natural video. Does CoT judgment also improve verification when the negatives are statements about real events? And when a verifier is chosen for its accuracy on one negative type, does the choice hold for another? We test four models on 400 TOC-Bench videos against two negative types, event-denial and content-matched, under both interfaces.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

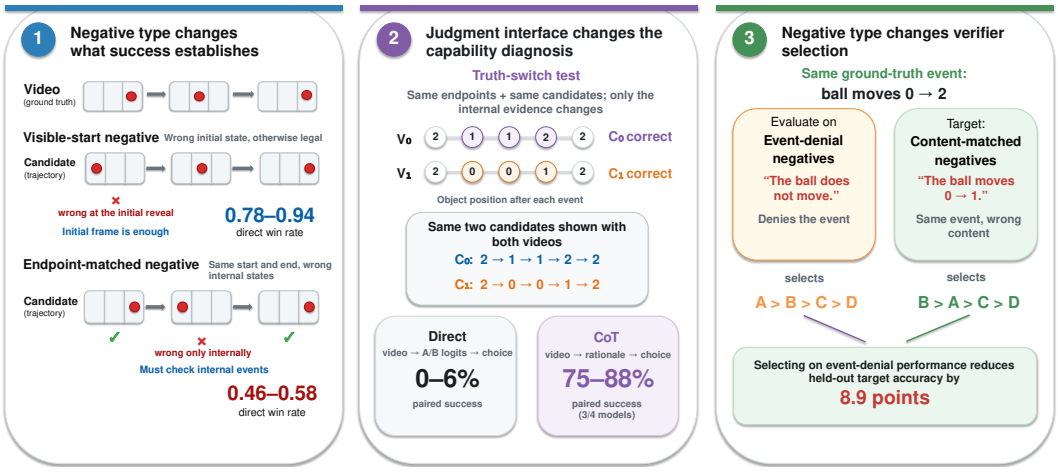

Figure 1: **What can a verifier score establish?** The same frozen verifiers support different conclusions under different evaluation designs. **Left:** The negative type determines what success demonstrates: visible-start negatives can be rejected from the initial frame, whereas endpoint-matched negatives require checking the internal events (win rates under direct judgment in the seven qualifying cells; Section 3). **Middle:** The judgment interface changes the capability diagnosis: on the truth-switch test, direct judgment fails where CoT judgment succeeds for three of four models (Section 4). **Right:** The negative type used for selection can change which verifier is chosen and reduce accuracy on the declared target (Section 5.5); the statements and rankings shown are schematic.

We make four contributions:

- **Endpoint-matched negatives.** We construct legal endpoint-matched negatives that separate success on visible-start negatives from verification of the internal trajectory: under direct judgment, win rates fall from 0.78–0.94 against visible-start negatives to 0.46–0.58, near chance (0.50), against endpoint-matched ones (Section 3).
- **Truth-switch test.** We introduce a truth-switch test showing that direct judgment can understate what a model can do: CoT judgment raises paired success from 0–6% to 75–88% for Qwen, GLM, and InternVL, while Molmo reaches only 14% (Section 4).
- **Natural-video transfer.** We show that the CoT gain does not transfer to natural video: CoT judgment brings no significant accuracy gain for any model, significantly lowers accuracy against event-denial negatives for three of four, and leaves model comparisons dependent on the negative type (Section 5).
- **Verifier selection.** We show that this dependence matters for selection: when content-matched negatives are the target, choosing among four models under both interfaces by event-denial accuracy costs 8.9 points of held-out accuracy (Section 5.5).

Evaluating a video verifier therefore requires separating three things that one score conflates: success on the negative type tested, ability that can be elicited from the model, and suitability for the intended verification target.

2 Audit Design

2.1 Negatives, Interfaces, and Evidence Controls

We vary two design choices, the negative type and the judgment interface, and hold the remaining inputs fixed within each experiment. Evidence controls (no video, and on natural video also shuffled frames) check whether a model's answers depend on the frames.

In the controlled tasks an executable program defines the faithful state trajectory. A trajectory is *legal* if some allowed event sequence produces it and *faithful* if the events in the presented video produce

Table 1: **The four studies.** Units are the resampling units for uncertainty.

Study	Units and models	Interfaces	Outcome
Construction (Section 3)	324 program units; six models	Direct, one candidate at a time	Win rate by negative type
Truth-switch (Section 4)	60 program pairs with two witness videos each; four models in two studies	Direct (answers and score-level test), CoT	Paired success: both videos correct within a candidate order
Natural-video (Section 5)	400 TOC-Bench videos; four models	Direct, CoT	Accuracy against event-denial and content-matched negatives
Verifier selection (Section 5.5)	Same 400 videos; eight configurations (model × interface)	Direct, CoT (existing responses)	Held-out content-matched accuracy of the configuration selected on each negative type

it. The tasks are *shell*, which tracks a hidden object among three cups, and *toggle*, which tracks three binary switches; a bounded-counter task serves as an extension. The construction study contains 324 program units, and the two truth-switch studies share 60 program pairs.

The natural-video study and the selection analysis use a fixed panel of 400 TOC-Bench videos (Chen et al., 2026): 120 event-order, 100 event-relative-state, 100 temporal-location, and 80 visibility-continuity cases. For each video, a reference proposition was drafted with an LLM from the video and question and verified by one of the authors against the full video (Section A.3). Each question refers to an *anchor event*. The two negative types pair the faithful proposition with either an *event-denial* negative, which states that the anchor event never occurs, or a *content-matched* negative, which keeps the anchor event and changes the queried detail by a fixed ontology-neighbor rule.

The truth-switch and natural-video studies evaluate Qwen3-VL-32B (Bai et al., 2025), GLM-4.5V (GLM-V Team, 2025), Molmo2-8B (Clark et al., 2026), and InternVL3.5-38B (Wang et al., 2025); the construction study adds Kimi-VL-A3B (Kimi Team et al., 2025) and Aria (Li et al., 2024). Table 1 summarizes the four studies, and Section A lists checkpoint revisions and inference settings.

2.2 Outcomes and inference

The construction study uses direct judgment only and scores each candidate on its own: a candidate's score is the next-token log-odds that it matches the evidence, and the *win rate* is the share of faithful–negative comparisons in which the faithful trajectory scores higher, with ties counted as one half. The other studies present both candidates together as options A and B. There, direct judgment selects the larger of the A and B next-token logits, and under CoT judgment a fixed parser extracts the final answer from the rationale. Ties, unparseable outputs, and truncations at the token cap are scored as incorrect. Paired success in the truth-switch studies requires correct answers on both videos of a pair (Section 4), and natural-video accuracy averages the two candidate orders within each video.

Uncertainty is computed by bootstrap at the program, program-pair, or video level, preserving each study's stratification. Simultaneous 95% intervals are reported within each declared family of estimates, and we call a contrast significant when its simultaneous interval excludes zero. Differences between percentages are in percentage points, written "points". Section F gives the families and procedures.

3 Visible-start success does not show internal-trajectory verification

Let a program produce the faithful trajectory $c^+ = (s_0, \ldots, s_H)$ through transitions $s_{t+1} = \tau(s_t, a_t)$. We construct a legal negative $c^- = (\tilde{s}_0, \ldots, \tilde{s}_H)$ satisfying

$$\tilde{s}_0 = s_0, \qquad \tilde{s}_H = s_H, \qquad \exists\, t \in \{1, \ldots, H-1\} :\ \tilde{s}_t \neq s_t. \tag{1}$$

Legality requires an allowed event sequence realizing every transition of c^-; the trajectory is nevertheless incorrect for the displayed program. We compare these *endpoint-matched* negatives with

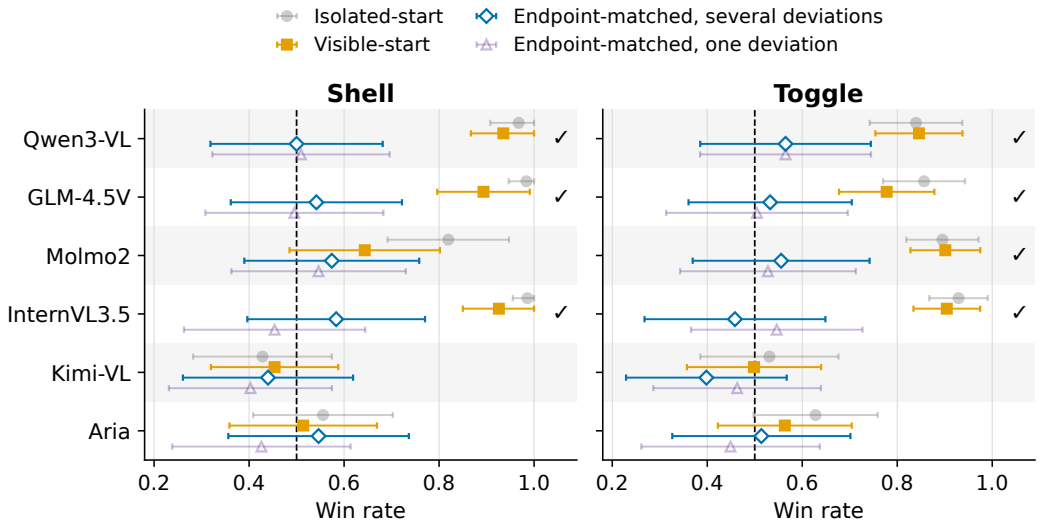

Figure 2: **Success on visible-start negatives does not carry over to endpoint-matched negatives.** Win rates of the six models on shell and toggle against all four negative types, with simultaneous 95% program-bootstrap intervals; the dashed line marks chance (0.50), and check marks identify the seven qualifying cells (Section B.2).

visible-start negatives, legal trajectories whose initial state is visibly incorrect. Two further negative types give context: an *isolated-start* negative changes only the initial state, and a *single-deviation* negative is endpoint-matched with one internal deviation (Section B.1).

The rendered input reveals the initial state, depicts each event without revealing the subsequent hidden state, and covers the final state. All candidate trajectories share one text schema. For a fixed video, endpoint matching removes the candidate endpoints as a basis for choosing between trajectories.

Our main outcome is the *endpoint-matching drop*: the win rate against visible-start negatives minus the win rate against endpoint-matched negatives. A cell (one model on one task) *qualifies* when it meets three criteria fixed before the drop was inspected (Section B.2): a text-record test, in which the model receives the true state sequence as text; a visible-start criterion, which requires a high win rate that depends on the video; and a drop criterion, which requires a drop that is larger with video than without. Seven cells qualify: Qwen, GLM, and InternVL on both shell and toggle, and Molmo on toggle. In these cells, win rates range from 0.78 to 0.94 against visible-start negatives and from 0.46 to 0.58 against endpoint-matched ones; Qwen's shell win rate, for example, falls from 0.935 to 0.500. Figure 2 displays all twelve shell/toggle cells, including those that do not qualify.

No endpoint-matched shell/toggle interval excludes chance, although the intervals are too broad to establish equivalence to chance. Kimi fails the text-record test, Aria fails the visible-start criterion, and no counter cell qualifies.

Rejecting a legal trajectory with a visibly wrong start does not demonstrate detection of a legal deviation confined to its internal path. Whether the models can make that distinction under another interface is the question of the next section.

4 CoT JUDGMENT REVEALS EVIDENCE RESPONSIVENESS

4.1 A FIXED-CANDIDATE TRUTH-SWITCH TEST

A stronger interface could raise scores without making the decisions *evidence-responsive*, that is, without making the choice follow the video when only the internal evidence changes. We test this with two executable witness videos for each endpoint-matched pair of candidates. Video V_0 realizes trajectory C_0 and video V_1 realizes trajectory C_1; the programs have equal initial and final states but

Table 2: **Truth-switch paired success under direct and CoT judgment.** Paired success (Equation (2)) with simultaneous 95% bootstrap intervals. The upper rows are the Qwen–GLM study and the lower rows the Molmo–InternVL study, each analyzed as its own six-estimate family. The score-level column tests the logits of the direct pass: the share of pair-orders in which the A−B logit margin moves in the correct direction between the two videos (chance 50%; intervals in Section C.4).

Model	Direct (%)	Score-level (%)	CoT (%)	CoT − direct (pp)
Qwen3-VL-32B	5.83 [0.00, 15.00]	73.33	75.00 [65.83, 84.17]	+69.17 [60.00, 78.33]
GLM-4.5V	0.00 [0.00, 9.17]	60.83	88.33 [79.17, 97.50]	+88.33 [79.17, 97.50]
Molmo2-8B	3.33 [0.00, 10.83]	55.00	14.17 [6.67, 21.67]	+10.83 [3.33, 18.33]
InternVL3.5-38B	4.17 [0.00, 11.67]	59.58	88.33 [80.83, 95.83]	+84.17 [76.67, 91.67]

different internal trajectories. Both videos use the construction study's renderer, nuisance seed, and covered final state, and their decoded initial and final frames match exactly (Figure 4 in Section C.1).

For each order $o \in \{1, 2\}$ of the two candidates, each video is evaluated in a fresh conversation. The text prompt and input token IDs are identical across the two videos within an order and interface. Let $\hat{C}(V_j; o)$ be the candidate selected for video V_j. We define

$$B_{p,o} = \mathbf{1}\{\hat{C}(V_0; o) = C_0 \ \wedge \ \hat{C}(V_1; o) = C_1\}, \qquad \widehat{B} = \frac{1}{60} \sum_{p=1}^{60} \frac{B_{p,1} + B_{p,2}}{2}. \tag{2}$$

The *paired success* $\widehat{B}$ is the rate of answering both videos correctly, averaged over the two orders of each program pair p. A fixed preference for an answer position or a candidate cannot give the two opposite answers within an order; the rule follows Vinoground's text score (Zhang et al., 2024).

The 60 pairs were fixed in a pilot, 15 in each shell/toggle by four/six-event cell, without selection on model correctness. Two studies share them. Under CoT judgment, the Qwen–GLM study runs Qwen with a reasoning prompt (1,024-token cap) and GLM with native thinking (4,096), and the Molmo–InternVL study runs Molmo with a reasoning prompt (2,048) and InternVL with its documented thinking-mode system prompt (4,096). Each study contains 960 responses (two models, two interfaces, two videos, and two orders per pair) and is analyzed as its own six-estimate family (Section C).

4.2 CoT JUDGMENTS FOLLOW THE CHANGED EVIDENCE

CoT judgment raises paired success from 5.83% to 75.00% for Qwen and from 0.00% to 88.33% for GLM (Table 2); both gains are significant in the six-estimate family. In the Molmo–InternVL study, InternVL rises from 4.17% to 88.33% (+84.17 points [76.67, 91.67]), whereas Molmo rises only from 3.33% to 14.17% (+10.83 [3.33, 18.33]), and 35 of its 240 CoT responses decline to choose ("Both A and B", "Neither"). Under CoT judgment, three of the four models are evidence-responsive: the same candidate pair receives opposite, correct decisions when the internal video evidence changes, so their near-chance direct win rates in Section 3 understate what they can do. The size of the CoT gain differs by model.

Single-video accuracy hides most of the direct failure. GLM answers A on all 240 direct requests and InternVL answers B on 199, giving 50.00% and 51.25% single-video accuracy but 0.00% and 4.17% paired success; Qwen's direct single-video accuracy is 51.67%, against paired success of 5.83%. Under CoT judgment, single-video accuracy reaches 86.67% for Qwen, 93.75% for GLM and InternVL, and 43.33% for Molmo.

Does the direct failure lie in the scores or in reading an answer from them? A score-level test on the same direct pass compares the A-minus-B logit margins of the two videos within an order; a fixed preference for a position or a candidate scores 50%. Qwen's margins move in the correct direction in 73.33% of pair-orders (pointwise 95% interval [65.42, 81.25]), so its scores respond to the switched evidence. GLM and InternVL reach only 60.83% and 59.58%, with median margin movements of 0.28 and 0.13 logits, and Molmo's 55.00% [46.67, 63.33] includes chance. Removing the position bias by averaging each video's margin over the two candidate orders gives single-video accuracy of 60.00% for Qwen and 49.58–52.08% for the others, and de-biased paired success of 16.67%

for Qwen, against 86.67% single-video accuracy and 75.00% paired success under CoT judgment (Section C.4); the signal is too small to act on. Qwen's direct scores thus carry a signal that a strong position bias swamps, GLM's and InternVL's carry almost none, and for all three CoT judgment is what produces usable discrimination.

5 Natural video: no significant CoT gain, persistent negative-type dependence

5.1 Negative type and interface on natural video

We next ask whether CoT judgment improves natural-video verification and whether it removes negative-type dependence. The natural-video study crosses four models, the two negative types, the two interfaces, three evidence conditions (ordered, shuffled, and no video), and both candidate orders; each comparison uses the same candidate text, media (64, 48, 80, and 48 frames for Qwen, GLM, Molmo, and InternVL), and order. The design contains 38,400 responses.

The two negative types require different checks: an event-denial negative is rejected by confirming that the anchor event occurs, a content-matched negative only by checking the queried detail. For the question "At the moment the cell phone first shows up, what is the state of the jars?" the faithful statement is "At the event 'the cell phone first shows up', the entity 'the jars' is partly hidden.", the event-denial negative is "The anchor event 'the cell phone first shows up' does not occur anywhere in the video.", and the content-matched negative is "At the event 'the cell phone first shows up', the entity 'the jars' is fully visible." Which type is harder varies by model (Figure 3a).

5.2 CoT judgment does not improve natural-video accuracy

In the 12-estimate interface family, no CoT-minus-direct accuracy change is a significant gain, and four are significant declines: Qwen against event-denial negatives (-17.63 points $[-24.63, -10.63]$), GLM against event-denial (-32.13 $[-39.13, -25.13]$) and content-matched negatives (-7.38 $[-14.38, -0.38]$), and InternVL against event-denial negatives (-8.00 $[-15.00, -1.00]$). The two positive point estimates, Molmo against event-denial ($+6.38$ $[-0.63, 13.38]$) and InternVL against content-matched negatives ($+5.25$ $[-1.75, 12.25]$), are not significant. CoT judgment also changes each model's denial-minus-content-matched accuracy difference; all four changes are significant, and they differ in sign (Section D.2).

GLM's declines mix completion failure with judgment, since 194 of 800 CoT ordered-denial and 196 of 800 ordered-content-matched responses are invalid or truncated. Qwen's CoT ordered-denial cell has nine invalid and one truncated response out of 800, and crediting all ten as correct would raise its accuracy only to 45.0%, against 61.38% under direct judgment.

A rationale analysis of Qwen's ordered-video event-denial responses locates the decline (Section D.3). At video level it consists of 54 losses against 12 gains, concentrated in anchor events that presuppose a prior absence: on the 83 videos whose anchor event reads "comes back into view", accuracy falls from 73.5% to 33.1%. Of the 440 wrong CoT answers, 396 conclude that the anchor event never happens, and on the same frames CoT judgment against the content-matched negative is fully correct on 36 of the 54 loss videos (Table 3). A frame-level check of the 83 "comes back into view" videos against the 64 frames the model saw finds the reappearance visible in 47; on those, CoT accuracy is 41.5% against 85.1% direct, so the decline is not an artifact of frame sampling. On the 31 videos whose sampled frames show no reappearance, both interfaces are scored against a reference the model could not verify (CoT 25.8%, direct 62.9%; Section D.3).

Under CoT judgment the event-denial negative thus acts as a "none of the above" option: a rationale that disputes a detail, reads the event phrase strictly, or fails to locate the moment ends by choosing the one statement that does not assert the event.

5.3 Large model gaps need not survive a change of negative type

Let A_m^{den} and A_m^{cm} denote model m's ordered-video CoT accuracy against event-denial and content-matched negatives. For models a and b, the between-model gaps are $\Delta_{ab}^{\mathrm{den}} = A_a^{\mathrm{den}} - A_b^{\mathrm{den}}$ and

Table 3: **Same video, same frames, two candidate pairs.** Qwen's CoT rationales under the event-denial and content-matched pairs (faithful-first order; verbatim, omissions marked): against the event-denial negative it concludes that the anchor event never happens; against the content-matched negative it locates the event. The video is chosen by a fixed rule (Section D.3).

Question: At the moment the wooden cabinet comes back into view, what is the state of the blue pen? *A:* faithful statement in both pairs; *B:* The anchor event 'the wooden cabinet comes back into view' does not occur anywhere in the video. (left) or At the event 'the wooden cabinet comes back into view', the entity 'the blue pen' is partly hidden. (right).

Event-denial pair	**Content-matched pair**
The wooden cabinet is visible throughout the entire video sequence, from Frame 0 to Frame 63. Therefore, the event "the wooden cabinet comes back into view" does not occur as a distinct event—it is continuously present. The phrase "comes back into view" implies it was previously out of view, which is not the case here. Thus, the anchor event does not occur. Answer: B	The event "the wooden cabinet comes back into view" occurs at the beginning of the video, specifically in Frame 0, where the cabinet is clearly visible in the background. [...] There is no object or hand covering it, and it remains in the same position throughout the initial frames. In Frame 0, the blue pen is clearly visible on the table, and the cabinet is in full view. The pen remains visible in subsequent frames (e.g., Frame 1, 2, etc.) as the hand interacts with it, but at the moment the cabinet comes into view (Frame 0), the pen is fully visible. [...] Answer: A

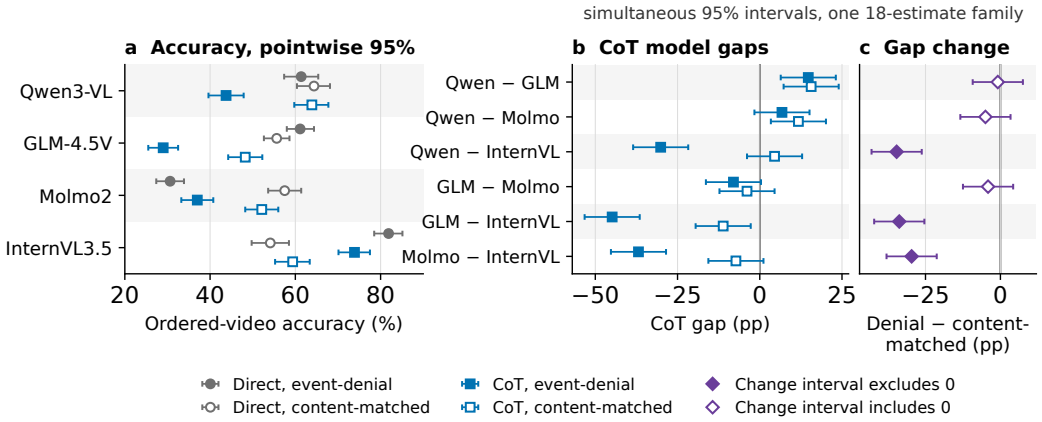

Figure 3: **Negative-type dependence persists under CoT judgment.** (a) Ordered-video accuracy for four models under both negative types and both interfaces (pointwise intervals). (b,c) Between-model CoT gaps against each negative type and their six denial-minus-content-matched changes, from one 18-estimate simultaneous family; three changes are significant.

$\Delta^{\mathrm{cm}}_{ab} = A^{\mathrm{cm}}_{a} - A^{\mathrm{cm}}_{b}$. We test the negative-type change $\Delta^{\mathrm{den}}_{ab} - \Delta^{\mathrm{cm}}_{ab}$ together with both gaps for all six model pairs, an 18-estimate simultaneous family.

Three of the six negative-type changes are significant, all involving InternVL (Figure 3b,c). Against event-denial negatives InternVL scores 73.88% and Qwen 43.75%; against content-matched negatives they score 59.38% and 63.88%. The Qwen-minus-InternVL gap is thus -30.13 points $[-38.50, -21.75]$ against event-denial negatives and $+4.50$ $[-3.88, 12.88]$ against content-matched ones, a change of -34.63 $[-43.00, -26.25]$. InternVL's large lead disappears, and Qwen's small lead is not significant. No model pair has significant gaps of opposite sign under the two negative types. CoT judgment does not make model comparisons independent of the negative type tested.

5.4 Evidence controls

Under CoT judgment, ordered video significantly improves over no video in five of the eight model-by-type comparisons of the 16-estimate evidence family, whereas all eight ordered-minus-shuffled intervals include zero. Three of the four models therefore use the frames (both of GLM's ordered-

Table 4: **Selecting on the wrong negative type reduces accuracy on the declared target.** Both policies are evaluated on held-out ordered-video content-matched comparisons. Intervals are approximate simultaneous 95% whole-video bootstrap intervals over the three selection costs, with selection repeated in each draw. The CoT-only cost is not significant.

Eligible set	Denial-selected (%)	Content-matched-selected (%)	Selection cost (pp)
All eight	54.13	63.01	$+8.88$ [2.46, 15.30]
Four direct	54.13	64.38	$+10.25$ [3.83, 16.67]
Four CoT	59.38	63.88	$+4.50$ [-1.92, 10.92]

minus-no-video intervals include zero), and this family detects no effect of frame order. Text-only preferences do not explain the main comparison either: subtracting each model's no-video accuracy leaves a Qwen–InternVL negative-type change of -25.13 points $[-35.88, -14.37]$ in a separate diagnostic family. Section D reports all cells, parser outcomes, and evidence contrasts.

5.5 Negative type changes which verifier is selected

Changed model gaps matter in practice only relative to a target. We therefore evaluate verifier selection with a declared target, distinguishing faithful propositions from content-matched negatives on the same 400 videos, using the existing responses. The analysis was specified after the cell accuracies in Section D were known; its rules, eligible sets, and partitions were fixed before any selection outcome was computed.

A *configuration* is one model under one interface. Within an eligible set of configurations, the *denial-selected* policy picks the configuration with the highest event-denial accuracy on the selection videos, and the *content-matched-selected* policy picks the one with the highest content-matched accuracy (Equation (4) in Section E); both are evaluated on content-matched negatives from held-out videos. The *selection cost* is the held-out accuracy of the content-matched-selected configuration minus that of the denial-selected one. We use twenty stratified five-fold partitions, with 320 selection and 80 evaluation videos per fold, and three eligible sets: all eight configurations, the four direct ones, and the four CoT ones. Whole-video bootstrap draws repeat the selection in each draw, and the three selection costs form one simultaneous family (Table 4).

Among direct configurations, every fold selects InternVL on event-denial accuracy and Qwen on content-matched accuracy, so the direct-only cost equals the difference of their content-matched accuracies over all 400 videos, $(515 - 433)/800 = 10.25$ points; its interval $[3.83, 16.67]$ reflects video-sampling uncertainty.

Across all eight configurations, the denial-selected policy always chooses direct InternVL, while the content-matched-selected policy chooses direct Qwen in 70% of folds and CoT Qwen in the rest; the selection cost is 8.88 points $[2.46, 15.30]$, ranging from 7.75 to 10.25 points across the twenty partitions. Among CoT configurations the same switch from InternVL to Qwen occurs, but the 4.50-point cost is not significant. On a separately fixed 200/200 split the direct-only cost is 9.25 points $[1.75, 16.75]$, while the all-eight (4.75) and CoT-only (2.00) costs are not significant (Section E.3).

Selecting on event-denial accuracy does what it is asked: on held-out event-denial comparisons, the denial-selected configuration scores 81.88%, versus 55.82% for the content-matched-selected configuration (all eight). The cost arises when a choice made on one negative type is transferred to another.

6 Related work

State tracking and temporal diagnostics. Temporal-perception benchmarks, atemporal probes, and evidence-perturbation audits show how appearance or unordered content can sustain video-language performance (Liu et al., 2024; Pătrăucean et al., 2023; Buch et al., 2022; Lei et al., 2023; Sevilla-Lara et al., 2021; Lee, 2026; Rahman, 2026; Chen et al., 2026; Meng et al., 2026). VET-Bench tracks visually identical objects through shuffles from motion alone, finds near-chance frozen models, and itself separates that setting from cup-game videos whose overlaid arrows disclose each swap (Liu & Lee, 2026); VSTAT finds that models solve textual transcriptions of state-tracking videos but fail on

the videos (Yu et al., 2026). Both evaluate solvers; our controlled frames disclose each event and we evaluate verifiers of supplied trajectories, so we neither address VET-Bench's motion-only problem nor contradict its findings.

Negative types and grounding controls. Contrast sets, Winoground, ARO, and SugarCrepe probe label-changing perturbations, order insensitivity, and artifact-bearing negatives (Gardner et al., 2020; Thrush et al., 2022; Yuksekgonul et al., 2023; Hsieh et al., 2023); VideoCon perturbs captions, including event order (Bansal et al., 2024); VELOCITI adds in-video negatives with absent-video and single-frame controls (Saravanan et al., 2025); Vinoground scores both members of paired video–caption sets (Zhang et al., 2024). We add endpoint-matched negatives, legal executable trajectories that are wrong only internally, and paired witness videos with exact truth conditions and fixed text inputs.

Judgment interfaces, judge reliability, and reward use. First-token probabilities can disagree with generated answers (Wang et al., 2024); judge studies and benchmarks document position, length, informativeness, and perceptual biases (Zheng et al., 2023; Chen et al., 2024b; Liu et al., 2023; Zhao et al., 2025; Lee et al., 2026; Zou et al., 2026; Park et al., 2026; Chen et al., 2024a; Li et al., 2025; He et al., 2024; Pi et al., 2025). Reranking, process supervision, and reward overoptimization (Cobbe et al., 2021; Lightman et al., 2024; Gao et al., 2023) motivate measuring what a verifier rewards. Structured and evidence-grounded verifiers (Shah et al., 2026; Liang et al., 2026; Wang et al., 2026a; Kim et al., 2026; Huang et al., 2026; Min et al., 2023; 2026; Wang et al., 2026b) are discussed in Section G.

7 Discussion and conclusion

Our results suggest three practices. First, a verifier should be evaluated on the negative type it will face: the endpoint-matching drop and the selection cost both show that performance on one negative type need not transfer to another. Second, accuracy under direct judgment should not be read as a ceiling on capability: Qwen's direct scores already respond to the evidence, and CoT judgment reveals discrimination in GLM and InternVL. Third, a CoT gain should be re-measured on the videos and negatives the verifier will face. The interface that raises paired success on the controlled tasks, where both candidates assert events, lowers natural-video accuracy for three of four models when one candidate denies the event.

These conclusions have several limitations. First, the controlled inputs are stylized keyframe sequences whose frames disclose each event, not continuous video, so the CoT gain we measure reflects responsiveness to disclosed event evidence; it does not show tracking of unannotated motion or of the full state sequence. Second, CoT judgment ran under one fixed prompt, parser, and token cap per model, chosen per study rather than compute-matched, so the results bound neither the best achievable CoT performance nor the effect of a common budget. The choice can matter: in a diagnostic run, GLM's CoT accuracy against endpoint-matched negatives on shell and toggle moved from 13% and 18% at a 1,024-token cap with the strict parser to 88% and 98% at 4,096 tokens with its native answer wrapper accepted. Third, the natural-video findings rest on one 400-video panel and its human-verified references. The selection result measures a declared content-matched target rather than an empirical distribution of deployment negatives; its CoT-only cost is not significant, and on a fixed 200/200 split only the direct-only cost is significant. Finally, we study frozen offline judgments of four mid-sized open models, not verifier training, adaptive attacks, or frontier models.

We introduced an executable audit of video verifiers built on legal endpoint-matched negatives, paired witness videos with exact truth conditions, and a natural-video study that varies negative type and judgment interface. Together, these experiments show that visible-start negatives can overstate what verifier success establishes, direct judgment can understate what a model can do, and evaluation on one negative type can select a less accurate verifier for another. Evaluating a video verifier therefore requires separating three things that one score conflates: success on the negative type tested, ability that can be elicited from the model, and suitability for the intended verification target.

Reproducibility Statement

Appendices A–G give the cohorts, checkpoint revisions, negative constructions, prompts, parsers, and statistical procedures. An anonymous artifact with the task generators, prompts, parsers, per-response outputs, and analysis scripts accompanies the submission; the full release, including the truth-switch and selection packages, will be published with the camera-ready version. Model checkpoints and TOC-Bench media are identified but not redistributed.

AI use statement

LLM tools were used as follows. Reference propositions for the natural-video panel were drafted by an LLM from each video and question and then verified by one of the authors against the full video for every item (Section A.3). The rationale codes in Section D.3 were assigned by an LLM assistant under a codebook fixed before coding, with a manually read validation sample; for the frame-level check in the same appendix, an LLM assistant proposed the per-video judgments from the sampled frames and one of the authors reviewed every item. LLM assistants also helped write analysis and experiment code, run and monitor experiments, and edit the manuscript. The authors reviewed these outputs and are responsible for all content.

Ethics statement

The natural-video study uses TOC-Bench videos under their documented licenses; we do not redistribute media. Human involvement was limited to the authors' verification of reference propositions and review of the frame-level check, and no personal data were collected.

References

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. VideoCon: Robust video-language alignment via contrast captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13927–13937, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Bansal_VideoCon_Robust_Video-Language_Alignment_via_Contrast_Captions_CVPR_2024_paper.html.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546):1434–1445, 2024. doi: 10.1080/01621459.2023.2197686. URL https://arxiv.org/abs/2104.00673.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2917–2927, 2022.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595. PMLR, 2024a. URL https://proceedings.mlr.press/v235/chen24h.html.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? A study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

Junzhe Chen, Siyuan Meng, Yuxi Chen, Man Zhao, Wenyao Gui, and Xiaojie Guo. TOC-Bench: A temporal object consistency benchmark for video large language models. *arXiv preprint arXiv:2605.09904*, 2026. doi: 10.48550/arXiv.2605.09904. URL https://arxiv.org/abs/2605.09904.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026. doi: 10.48550/arXiv.2601.10611. URL https://arxiv.org/abs/2601.10611.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip H. S. Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20224–20234, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Ding_MOSE_A_New_Dataset_for_Video_Object_Segmentation_in_Complex_ICCV_2023_paper.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://arxiv.org/abs/2004.07780.

GLM-V Team. GLM-4.1V-Thinking and GLM-4.5V: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. doi: 10.48550/arXiv.2507.01006. URL https://arxiv.org/abs/2507.01006.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.127. URL https://aclanthology.org/2024.emnlp-main.127/.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, 2023. doi: 10.52202/075280-1355. URL https://papers.neurips.cc/paper_files/paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html.

Zixuan Huang, Yang Zhou, Kaixuan Wang, Guli Zhang, et al. Deferred exposure of future trajectories for verifiable reasoning in autonomous driving VLMs. *arXiv preprint arXiv:2608.01755*, 2026. doi: 10.48550/arXiv.2608.01755. URL https://arxiv.org/abs/2608.01755.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Hyomin Kim, Junghye Kim, Joanie Hayoun Chung, Yoonjin Oh, et al. Plan-and-verify video reward reasoning with spatio-temporal scene graph grounding. *arXiv preprint arXiv:2606.11838*, 2026. doi: 10.48550/arXiv.2606.11838. URL https://arxiv.org/abs/2606.11838.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025. doi: 10.48550/arXiv.2504.07491. URL https://arxiv.org/abs/2504.07491.

Jae Joong Lee. Accuracy without grounding: Diagnosing visual dependency dissociation in video LLM benchmarks. *arXiv preprint arXiv:2607.13305*, 2026. URL https://arxiv.org/abs/2607.13305.

Sua Lee, Sanghee Park, and Jinbae Im. MM-JudgeBias: A benchmark for evaluating compositional biases in MLLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25336–25373, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1162. URL https://aclanthology.org/2026.acl-long.1162/.

Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487–507, 2023. doi: 10.18653/v1/2023.acl-long.29.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. doi: 10.48550/arXiv.2410.05993. URL https://arxiv.org/abs/2410.05993.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. VL-RewardBench: A challenging benchmark for vision-language generative reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24657–24668, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_VL-RewardBench_A_Challenging_Benchmark_for_Vision-Language_Generative_Reward_Models_CVPR_2025_paper.html.

Shawn Liang, Sahil Shah, Chengwei Zhou, S P Sharan, Harsh Goel, Arnab Sanyal, Sandeep Chinchali, and Gourav Datta. RT-NeuS: Towards real-time neuro-symbolic video understanding via adaptive temporal verification. *arXiv preprint arXiv:2602.23553*, 2026. doi: 10.48550/arXiv.2602.23553. URL https://arxiv.org/abs/2602.23553v2.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Ming Liu and Wensheng Zhang. Is your video language model a reliable judge? In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m8yby1JfbU.

Tiedong Liu and Wee Sun Lee. Can vision-language models solve the shell game? *arXiv preprint arXiv:2603.08436*, 2026. URL https://arxiv.org/abs/2603.08436.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. VideoZeroBench: Probing the limits of video MLLMs with spatio-temporal evidence verification. *arXiv preprint arXiv:2604.01569*, 2026. doi: 10.48550/arXiv.2604.01569. URL https://arxiv.org/abs/2604.01569.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving vision-language models with perception-centric process reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33099–33109, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Min_Improving_Vision-language_Models_with_Perception-centric_Process_Reward_Models_CVPR_2026_paper.html.

Seojeong Park, Jiho Choi, Junyong Kang, Seonho Lee, Jaeyo Shin, and Hyunjung Shim. Mitigating perceptual judgment bias in multimodal LLM-as-a-judge via perceptual perturbation and reward modeling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.02578.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, 2023. URL https://arxiv.org/abs/2305.13786.

Renjie Pi, Haoping Bai, Qibin Chen, Xiaoming Simon Wang, Jiulong Shan, Xiaojiang Liu, and Meng Cao. MR. judge: Multimodal reasoner as a judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20181–20205, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1021. URL https://aclanthology.org/2025.emnlp-main.1021/.

Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip H. S. Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022. doi: 10.1007/s11263-022-01629-1. URL https://arxiv.org/abs/2102.01558.

Farrukh Rahman. What does a temporal benchmark score measure? Decomposing channel use in video VLM evaluation. *arXiv preprint arXiv:2607.12304*, 2026. URL https://arxiv.org/abs/2607.12304.

Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. VELOCITI: Benchmarking video-language compositional reasoning with strict entailment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18914–18924, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Saravanan_VELOCITI_Benchmarking_Video-Language_Compositional_Reasoning_with_Strict_Entailment_CVPR_2025_paper.html.

Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 535–544, 2021.

Sahil Shah, S P Sharan, Harsh Goel, Minkyu Choi, Mustafa Munir, Manvik Pasula, Radu Marculescu, and Sandeep Chinchali. NeuS-QA: Grounding long-form video understanding in temporal logic and neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8805–8813, 2026. doi: 10.1609/aaai.v40i11.37834. URL https://ojs.aaai.org/index.php/AAAI/article/view/37834.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision – ECCV 2016*, pp. 510–526, 2016. doi: 10.1007/978-3-319-46448-0_31. URL https://arxiv.org/abs/1604.01753.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022. URL https://arxiv.org/abs/2204.03162.

Junxin Wang, Dai Guan, Weijie Qiu, Zhihang Li, Yongbo Gai, Zhengyi Yang, Mengyu Zhou, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Grounding the score: Explicit visual premise verification for reliable vision-language process reward models. *arXiv preprint arXiv:2603.16253*, 2026a. URL https://arxiv.org/abs/2603.16253.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My Answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis–verification multi-agent framework for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33784–33793, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Think_Then_Verify_A_Hypothesis-Verification_Multi-Agent_Framework_for_Long_Video_CVPR_2026_paper.html.

Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, June Suk Choi, Ellis Brown, Oscar Michel, Boyang Zheng, Jinwoo Shin, and Saining Xie. Benchmarking visual state tracking in multimodal video understanding. *arXiv preprint arXiv:2606.03920*, 2026. doi: 10.48550/arXiv.2606.03920. URL https://arxiv.org/abs/2606.03920.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International*

Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing LMMs over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024. doi: 10.48550/arXiv.2410.02763. URL https://arxiv.org/abs/2410.02763.

Yulai Zhao, Haolin Liu, Dian Yu, Sunyuan Kung, Meijia Chen, Haitao Mi, and Dong Yu. One token to fool LLM-as-a-judge. *arXiv preprint arXiv:2507.08794*, 2025. URL https://arxiv.org/abs/2507.08794.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, 2023. doi: 10.52202/075280-2020. URL https://papers.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Xiaohan Zou, Roshan Sridhar, Mohammadtaher Safarzadeh, and Dan Roth. When vision-language models judge without seeing: Exposing informativeness bias. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15417–15448, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.703. URL https://aclanthology.org/2026.acl-long.703/.

A Cohorts, checkpoints, and reference annotation

A.1 Experiments and shared units

Table 1 summarizes the four studies. The resampling unit is the program unit in the construction study, the program pair in the truth-switch studies (60 pairs, each with two witness videos), and the video in the natural-video study and the selection analysis (the same 400 videos).

No truth-switch pair was filtered after the pilot, and witness programs are disjoint across pairs. Additional pairs and videos served technical pilot runs only and are excluded from every estimate.

A.2 Checkpoints and inference settings

All models were loaded from the fixed Hugging Face revisions in Table 5; the model reports are Bai et al. (2025), GLM-V Team (2025), Clark et al. (2026), Wang et al. (2025), Kimi Team et al. (2025), and Li et al. (2024). Inference used bfloat16 weights, greedy decoding, flash-attention-2 for Qwen, GLM, Kimi-VL, and Aria, SDPA attention for Molmo2 and InternVL, and Transformers 4.57.6 (4.50.3 for Kimi). The Molmo–InternVL truth-switch study ran each model on A100 40 GB GPUs (one for Molmo2, four for InternVL); other runs used H100 80 GB, A100 80 GB, or A100 40 GB GPUs.

Table 5: **Evaluated checkpoints.** Hugging Face repositories and revision identifiers of the six models.

Model	Repository	Revision
Qwen3-VL-32B	`Qwen/Qwen3-VL-32B-Instruct`	`0cfaf48183f594c314753d30a4c4974bc75f3ccb`
GLM-4.5V	`zai-org/GLM-4.5V`	`ed47433b37111465ec527affaaddceff371bca04`
Molmo2-8B	`allenai/Molmo2-8B`	`e28fa28597e5ec5e0cca2201dd8ab33d48bc4a1b`
InternVL3.5-38B	`OpenGVLab/InternVL3_5-38B-HF`	`7c830fc25e874f9a861087e3c70587f2b9b555f3`
Kimi-VL-A3B	`moonshotai/Kimi-VL-A3B-Instruct`	`398eede0903cd983a2bfa0cc634e9ac1d843f375`
Aria	`rhymes-ai/Aria`	`53a09c77895d7025efa6239ea71cf1a8991064e8`

Table 6: **Token caps for CoT judgment.** All decoding is greedy.

Study	Qwen cap	GLM cap	Molmo cap	InternVL cap
Truth-switch	1,024	4,096	2,048	4,096
Natural-video	8,192	4,096	2,048	2,048

Qwen and Molmo receive an explicit reasoning instruction for CoT judgment; GLM uses native thinking for CoT judgment and disables it for direct judgment; InternVL uses the thinking-mode system prompt documented by its authors. Parsers are described with each study (Sections C.2 and D.1).

A.3 Natural-video panel and reference annotation

The 400-video panel was drawn from TOC-Bench before any evaluated model was run: videos used in earlier development were excluded, the remainder were ordered by question and video identifier and shuffled within stratum with seed 20260817, and the first valid videos filled quotas of 120/100/100/80 for event order, event-relative state, temporal location, and visibility continuity. Videos are globally unique.

For each video, an LLM drafted a reference proposition from the video and the question, without access to benchmark answers or model outputs. One of the authors then checked every reference

against the full video and accepted it only when the target was evaluable and unambiguous, all required events were observed, and the categorical or numerical fields were valid; event-order onsets within 0.05 s were excluded as ties. The natural-video study uses these references unchanged. Shuffled evidence uses the identical decoded frame multiset with a per-video seed derived from 20260817. TOC-Bench draws on Charades (Sigurdsson et al., 2016), Perception Test (Pătrăucean et al., 2023), OVIS (Qi et al., 2022), and MOSE (Ding et al., 2023) under their documented licenses.

B Controlled negatives and the construction study

B.1 Transition legality and rendering

A controlled program fixes an initial state and an ordered list of events, and execution yields the faithful trajectory. Shell tracks the ball's slot in $\{0, 1, 2\}$, and an event swaps slots $(0, 1)$, $(1, 2)$, or $(0, 2)$; toggle tracks three binary switches, and an event flips one of them; counter tracks a level in $\{0, \ldots, 7\}$, and an event adds -2, -1, $+1$, or $+2$, with programs that leave this range excluded. A trajectory is legal when some allowed event sequence produces it; legality does not require that this be the event sequence shown in the input. Because any two shell slots are one swap apart, every well-formed shell state sequence is legal, so legality constrains toggle and counter candidates more than shell candidates. Endpoint-matched negatives also satisfy Equation (1).

Every program of length four or six is enumerated. A unit (P, Q, R) consists of three executable programs with the same initial and final states: Q differs from the faithful program P at exactly one intermediate state, and R differs at two or more intermediate states while using P's multiset of events. For shell and toggle, R is the unused program with the most differing intermediate states; for counter, it is the first program meeting these conditions in a seeded order. Seeded selection with an equal quota in every (initial state, final state, deviation position) stratum (shell and toggle) or seeded greedy packing (counter), followed by a hash-ordered pick balanced over the position of Q's deviation, gives 54 units per task and length, 324 in total, and no program appears in more than one unit. The four negative types are built per unit. The endpoint-matched negatives are the trajectories of R (several internal deviations, the main type) and Q (one localized deviation, the single-deviation type). The visible-start negative replays P's events from a changed initial state at the same component distance as Q's deviation and is kept only when every step remains legal. The isolated-start negative changes only the initial state and keeps P's later states; because every event is one-to-one, the first displayed event cannot produce its first transition. The endpoint-matching drop compares the visible-start negative with the several-deviation endpoint-matched negative; the single-deviation and isolated-start negatives give context.

The renderer (`latent-event-board-v1`) produces 320×256 RGB keyframes of a titled board: three panels centered at 0.24, 0.50, and 0.76 of the width for shell and toggle, and a tick-marked 0–7 number line in the counter's initial frame. The initial frame reveals the state. Every event frame carries an "EVENT k OF n" marker and indicates the event without drawing any state: a double-headed arrow between the swapped slots and a "SWAP i <-> j" label for shell, a pointer above the toggled switch and a "TOGGLE k" label with all switch states masked for toggle, and the signed step beside a masked level for counter. The final frame shows a mask labelled "STATE HIDDEN" and draws no state. A unit-level seed shifts the drawing by at most 3 px horizontally and 2 px vertically and scales the palette brightness by at most $\pm 3.5\%$, identically in every frame of P, Q, and R. Compared conditions reuse the corresponding program and seed, and the truth-switch witness videos are the renders of P and R.

B.2 Scores, prompts, and qualification criteria

For independently scored candidates, the win rate of Section 2.2 is

$$W = \mathbb{E}\big[\mathbf{1}\{s(c^+) > s(c^-)\} + \tfrac{1}{2}\mathbf{1}\{s(c^+) = s(c^-)\}\big]\,. \tag{3}$$

The score s of a candidate is the next-token log-odds of the answer letter that means it matches the evidence, averaged over both assignments of A and B to the two meanings. The pairwise template presents the two trajectories together and counterbalances their order.

The pairwise template is, with braced fields filled from the unit:

```
You are comparing two proposed state traces against
evidence.
{evidence}
Task: {task}
Each candidate has exactly three fields: the initial state,
every intermediate post-event state in chronological order,
and the final state. The final answer is deliberately
omitted.
Candidate A:
{candidate_a}
Candidate B:
{candidate_b}
Choose the one candidate whose complete state trace exactly
matches the evidence. Reply with exactly one token: A or
B.
```

With video, the evidence line reads "You are given N frames in chronological order, labelled Frame 0 through Frame $N-1$." The shell task text is "A ball starts at one of three slots 0=left, 1=middle, 2=right. Each shown event swaps exactly the two named slots. Track the ball slot after every event." The toggle task text is "Three switches are ordered left, middle, right. Each shown event toggles exactly one named switch. A state is a three-bit string in left-to-right order, where 1=ON and 0=OFF." Candidate trajectories are written with the lines `Initial:`, `Intermediate:` (post-event states joined by `->`), and `Final:`. The absolute template opens "You are auditing one proposed state trace against evidence," shares the evidence, task, and field paragraphs, presents one candidate trajectory, and counterbalances A/B between "every stated state exactly matches the evidence" and "at least one stated state is wrong, missing, or extra."

Three criteria were fixed before the endpoint-matching drop was inspected. In the *text-record test*, the model receives the exact state sequence as text instead of video and must reach at least 0.90 accuracy, with a pointwise 95% lower bound above 0.75, against every negative type. The *visible-start criterion* requires a win rate against visible-start negatives of at least 0.75 with video (simultaneous lower bound above 0.50) and a *video effect*, the win rate with video minus the win rate without, of at least 0.15 (simultaneous lower bound above zero). The *drop criterion* requires an endpoint-matching drop of at least 0.15 with a simultaneous lower bound above zero, and a video effect on the drop (the drop with video minus the drop without video) whose simultaneous lower bound is positive. A cell *qualifies* when it meets all three criteria.

The construction study uses the studentized 864-estimate family of Section F.1; renderer and text-record diagnostics use separately declared pointwise procedures.

B.3 Complete results

Table 7 reports all six models, three tasks, and four negative types, including cells that do not qualify. On shell, GLM falls from 0.894 against visible-start negatives to 0.542 against endpoint-matched negatives and InternVL from 0.926 to 0.583; on toggle, Molmo falls from 0.901 to 0.556. Kimi fails the text-record test except on counter; Aria passes the test, but none of its cells qualifies. No counter cell qualifies either: Qwen's and InternVL's counter drops are significant, but no counter drop has a significantly positive video effect, and the counter renderer is sensitive to rendering details. The video effect on the drop is positive for Qwen, GLM, Molmo, and InternVL, so a text-only preference between trajectories does not by itself explain their drops; these diagnostics are included in the artifact.

Table 7: **Win rates of all six models against all four negative types.** Cells report the win rate (Equation (3)) with simultaneous 95% program-bootstrap intervals intersected with $[0, 1]$. The isolated-start negative changes only the initial state and keeps the faithful later states; the visible-start negative is a legal trajectory from a visibly wrong initial state; the two endpoint-matched negatives are legal trajectories with several internal deviations or one. Chance is 0.50.

				Endpoint-matched	
Model	Task	Isolated-start	Visible-start	Several deviations	One deviation
Qwen3-VL	Shell	0.97 [0.91, 1.00]	0.94 [0.87, 1.00]	0.50 [0.32, 0.68]	0.51 [0.32, 0.70]
	Toggle	0.84 [0.74, 0.94]	0.85 [0.75, 0.94]	0.56 [0.39, 0.74]	0.56 [0.38, 0.74]
	Counter	0.82 [0.70, 0.94]	0.83 [0.69, 0.96]	0.56 [0.38, 0.75]	0.46 [0.27, 0.65]
GLM-4.5V	Shell	0.98 [0.95, 1.00]	0.89 [0.80, 0.99]	0.54 [0.36, 0.72]	0.50 [0.31, 0.68]
	Toggle	0.86 [0.77, 0.94]	0.78 [0.68, 0.88]	0.53 [0.36, 0.70]	0.50 [0.31, 0.70]
	Counter	0.85 [0.74, 0.96]	0.75 [0.60, 0.90]	0.56 [0.38, 0.75]	0.52 [0.34, 0.71]
Molmo2	Shell	0.82 [0.69, 0.95]	0.64 [0.49, 0.80]	0.57 [0.39, 0.76]	0.55 [0.36, 0.73]
	Toggle	0.90 [0.82, 0.97]	0.90 [0.83, 0.97]	0.56 [0.37, 0.74]	0.53 [0.34, 0.71]
	Counter	0.85 [0.74, 0.95]	0.75 [0.60, 0.90]	0.54 [0.35, 0.72]	0.45 [0.26, 0.64]
InternVL3.5	Shell	0.99 [0.96, 1.00]	0.93 [0.85, 1.00]	0.58 [0.40, 0.77]	0.45 [0.26, 0.64]
	Toggle	0.93 [0.87, 0.99]	0.90 [0.83, 0.97]	0.46 [0.27, 0.65]	0.55 [0.37, 0.73]
	Counter	0.97 [0.92, 1.00]	0.95 [0.87, 1.00]	0.70 [0.53, 0.87]	0.64 [0.46, 0.82]
Kimi-VL	Shell	0.43 [0.28, 0.57]	0.45 [0.32, 0.59]	0.44 [0.26, 0.62]	0.40 [0.23, 0.57]
	Toggle	0.53 [0.39, 0.68]	0.50 [0.36, 0.64]	0.40 [0.23, 0.57]	0.46 [0.29, 0.64]
	Counter	0.47 [0.32, 0.62]	0.38 [0.21, 0.54]	0.41 [0.23, 0.58]	0.54 [0.36, 0.72]
Aria	Shell	0.56 [0.41, 0.70]	0.51 [0.36, 0.67]	0.55 [0.36, 0.74]	0.43 [0.24, 0.61]
	Toggle	0.63 [0.50, 0.76]	0.56 [0.42, 0.70]	0.51 [0.33, 0.70]	0.45 [0.26, 0.64]
	Counter	0.49 [0.33, 0.64]	0.46 [0.28, 0.63]	0.35 [0.17, 0.54]	0.46 [0.27, 0.65]

C FIXED-CANDIDATE TRUTH-SWITCH PROTOCOL AND DIAGNOSTICS

C.1 DESIGN AND INPUT CONTROLS

For each of the 60 candidate pairs, two witness programs realize the two different trajectories while preserving the initial and final states. Initial and final decoded pixels match within each pair. Candidate text is fixed across the videos. Text prompts and input token IDs match within each candidate order and interface, and each response is obtained in a fresh conversation. Figure 4 shows the first four-event shell pair in the fixed pair order. The Molmo–InternVL study reuses the same pairs, frames, prompts, and request matrix.

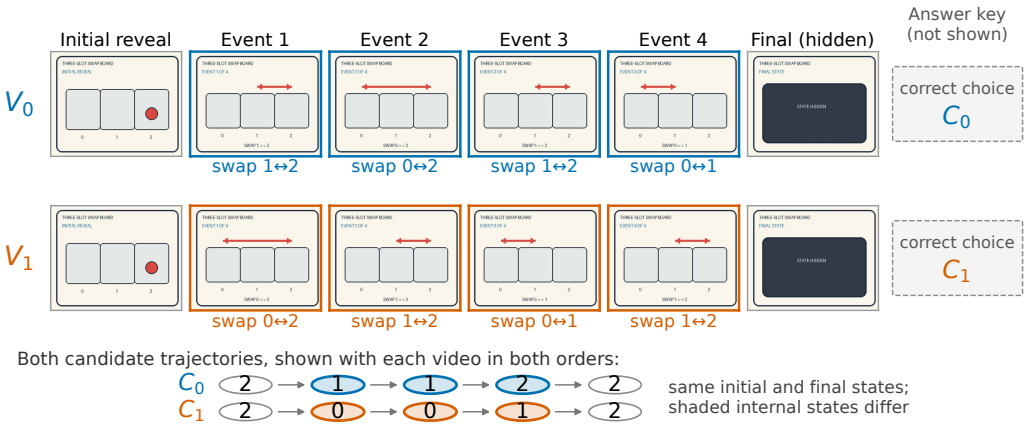

Figure 4: **A truth-switch pair.** The first four-event shell pair in the fixed pair order, chosen without reading any model output. Both witness videos share the initial reveal and the covered final frame; their events differ, so video V_0 realizes trajectory C_0 and video V_1 realizes C_1. Both candidate trajectories are shown with each video in both orders, so a verifier answers both videos correctly only if its choice follows the internal events.

Each four-event program supplies six frames, and each six-event program supplies eight: an initial frame, the event frames, and the covered final frame. In every pair, the two witness programs contain

Table 8: **Counterbalancing in the truth-switch studies.** The paired-success indicator requires both video answers to be correct within one candidate order. The two indicators are averaged within each program pair before estimation.

Witness video	Candidate order	Correct label
V_0	C_0, C_1	A
V_1	C_0, C_1	B
V_0	C_1, C_0	B
V_1	C_1, C_0	A

the same events in different orders (Section B.1). Because each event frame also prints its position, the frame multisets differ, and paired success cannot separate use of presentation order from use of the printed event labels and counters. The design is an intervention on internal-event evidence with exact truth conditions rather than a pure chronology intervention.

C.2 Prompts, parsing, and outcome definitions

Direct judgment uses the pairwise template of Section B.2 unchanged, with thinking disabled, and selects the unique larger of the A and B next-token logits; a tie is scored as incorrect. CoT judgment replaces the template's final sentence, "Reply with exactly one token: A or B.", with "Work through the visible initial state and each event in order, compare both complete traces, and explain the decisive mismatch concisely. End with exactly one final line in the form `Answer: A` or `Answer: B`." Qwen receives this prompt as its reasoning instruction; GLM also runs in native thinking mode. Decoding is greedy.

The parser reads the last nonempty line and accepts `Answer: X` for X in {A, B}; in the Qwen–GLM study it also accepts GLM's native final-answer wrapper `Answer: <|begin_of_box|>X<|end_of_box|>`, and the Molmo–InternVL study accepts only the requested line. Truncated CoT outputs, unaccepted answers, and tied direct maxima are scored as incorrect. Answer switch and candidate-identity order disagreement are computed over parsed, nontruncated responses.

In the Molmo–InternVL study, Molmo receives the same reasoning prompt and InternVL also receives the thinking-mode system prompt used in the natural-video study; direct judgment is unchanged.

Table 9: **Response-level diagnostics for the truth-switch studies.** Each model and interface has 240 responses (60 pairs $\times$ 2 videos $\times$ 2 orders); rates are percentages, and accuracy is single-video accuracy; neither-correct includes output failures; the invalid, truncated, tie, and format columns overlap; the last column counts A answers, B answers, and responses without an accepted answer. Upper block: Qwen–GLM study; lower block: Molmo–InternVL study.

Model	Interface	Correct/240	Accuracy	Paired	Neither	Switch	Order disagr.
Qwen3-VL-32B	Direct	124	51.67	5.83	2.50	6.67	82.50
Qwen3-VL-32B	CoT	208	86.67	75.00	1.67	75.00	0.00
GLM-4.5V	Direct	120	50.00	0.00	0.00	0.00	100.00
GLM-4.5V	CoT	225	93.75	88.33	0.83	88.33	0.00
Molmo2-8B	Direct	122	50.83	3.33	1.67	5.00	40.00
Molmo2-8B	CoT	104	43.33	14.17	27.50	24.17	42.50
InternVL3.5-38B	Direct	123	51.25	4.17	1.67	4.17	71.67
InternVL3.5-38B	CoT	225	93.75	88.33	0.83	88.33	0.00

Model	Interface	Invalid	Truncated	Ties	Format fail.	Format corr./240	A / B / none
Qwen3-VL-32B	Direct	4	0	4	4	124	219 / 17 / 4
Qwen3-VL-32B	CoT	32	31	0	32	208	97 / 111 / 32
GLM-4.5V	Direct	0	0	0	0	120	240 / 0 / 0
GLM-4.5V	CoT	15	15	0	110	130	115 / 110 / 15
Molmo2-8B	Direct	0	0	0	0	122	118 / 122 / 0
Molmo2-8B	CoT	35	0	0	35	104	126 / 79 / 35
InternVL3.5-38B	Direct	5	0	5	5	123	36 / 199 / 5
InternVL3.5-38B	CoT	15	14	0	15	225	114 / 111 / 15

Molmo's 35 CoT non-choices end with lines such as "Both A and B" (20), "Neither" (10), "A or B" (4), or "Cannot determine" (1), and its accepted CoT answers are near chance (104 of 205 correct); every accepted CoT answer of the other three models is correct.

C.3 Task-specific results and uncertainty

Each study's primary family contains six estimates: CoT paired success, direct paired success, and their difference for each of its two models (Table 10). The bootstrap uses 10,000 paired draws of whole program pairs within the four task-by-length strata; the two studies are analyzed as separate families. The saved bounded-family companions remain useful when empirical outcomes are constant. For example, GLM's direct empirical paired-success rate is zero, but its simultaneous bootstrap interval is $[0.00, 9.17]\%$ and its bounded-family companion is $[0.00, 21.37]\%$. A zero count still admits a nonzero rate, and the bounded interval reports the range consistent with it. The Qwen, GLM, and InternVL improvements also exclude zero under the bounded companions; Molmo's does not ($[-31.91, 53.58]$), so its small gain rests on the bootstrap family alone.

Table 10: **All twelve primary truth-switch estimates, six per study.** Pointwise intervals are bootstrap percentiles; simultaneous intervals use the common radius described in Section F within each study; bounded-family intervals are Hoeffding intervals with a Bonferroni split over each study's six estimates, clipped to the parameter range. Absolute rates are percentages and differences are percentage points.

Model	Quantity	Estimate	Pointwise 95%	Simultaneous 95%	Bounded family 95%
Qwen3-VL-32B	Direct paired success (%)	5.83	[2.50, 10.00]	[0.00, 15.00]	[0.00, 27.20]
Qwen3-VL-32B	CoT paired success (%)	75.00	[67.50, 82.50]	[65.83, 84.17]	[53.63, 96.37]
Qwen3-VL-32B	CoT − direct (pp)	+69.17	[60.00, 78.33]	[60.00, 78.33]	[26.42, 100.00]
GLM-4.5V	Direct paired success (%)	0.00	[0.00, 0.00]	[0.00, 9.17]	[0.00, 21.37]
GLM-4.5V	CoT paired success (%)	88.33	[82.50, 93.33]	[79.17, 97.50]	[66.96, 100.00]
GLM-4.5V	CoT − direct (pp)	+88.33	[82.50, 93.33]	[79.17, 97.50]	[45.59, 100.00]
Molmo2-8B	Direct paired success (%)	3.33	[0.83, 6.67]	[0.00, 10.83]	[0.00, 24.70]
Molmo2-8B	CoT paired success (%)	14.17	[9.17, 20.00]	[6.67, 21.67]	[0.00, 35.54]
Molmo2-8B	CoT − direct (pp)	+10.83	[5.00, 16.67]	[3.33, 18.33]	[−31.91, 53.58]
InternVL3.5-38B	Direct paired success (%)	4.17	[0.83, 7.50]	[0.00, 11.67]	[0.00, 25.54]
InternVL3.5-38B	CoT paired success (%)	88.33	[81.67, 94.17]	[80.83, 95.83]	[66.96, 100.00]
InternVL3.5-38B	CoT − direct (pp)	+84.17	[76.67, 90.83]	[76.67, 91.67]	[41.42, 100.00]

Table 11: **Task and length summaries.** These are descriptive secondary summaries with pointwise 95% intervals; they do not reuse the primary simultaneous intervals. Each task aggregate contains 30 pairs and each task/length cell 15, for each of the four models. For constant observed outcomes, percentile intervals can have identical bounds; the saved bounded companions for all 72 estimates (36 per study) are retained in the supplementary table file.

Task	Events	Model	Pairs	Direct (%)	CoT (%)	CoT − direct (pp)
Shell	4 and 6	Qwen3-VL-32B	30	10.00 [3.33, 18.33]	61.67 [48.33, 75.00]	+51.67 [35.00, 68.33]
Shell	4 and 6	GLM-4.5V	30	0.00 [0.00, 0.00]	80.00 [70.00, 90.00]	+80.00 [70.00, 90.00]
Shell	4 and 6	Molmo2-8B	30	5.00 [0.00, 10.00]	18.33 [10.00, 26.67]	+13.33 [3.33, 23.33]
Shell	4 and 6	InternVL3.5-38B	30	3.33 [0.00, 8.33]	91.67 [83.33, 98.33]	+88.33 [78.33, 96.67]
Shell	4	Qwen3-VL-32B	15	10.00 [0.00, 20.00]	63.33 [43.33, 83.33]	+53.33 [26.67, 76.67]
Shell	4	GLM-4.5V	15	0.00 [0.00, 0.00]	83.33 [70.00, 93.33]	+83.33 [70.00, 93.33]
Shell	4	Molmo2-8B	15	3.33 [0.00, 10.00]	20.00 [6.67, 33.33]	+16.67 [0.00, 30.00]
Shell	4	InternVL3.5-38B	15	0.00 [0.00, 0.00]	93.33 [83.33, 100.00]	+93.33 [83.33, 100.00]
Shell	6	Qwen3-VL-32B	15	10.00 [0.00, 20.00]	60.00 [40.00, 80.00]	+50.00 [30.00, 70.00]
Shell	6	GLM-4.5V	15	0.00 [0.00, 0.00]	76.67 [60.00, 90.00]	+76.67 [60.00, 90.00]
Shell	6	Molmo2-8B	15	6.67 [0.00, 16.67]	16.67 [6.67, 30.00]	+10.00 [−3.33, 23.33]
Shell	6	InternVL3.5-38B	15	6.67 [0.00, 16.67]	90.00 [73.33, 100.00]	+83.33 [66.67, 96.67]
Toggle	4 and 6	Qwen3-VL-32B	30	1.67 [0.00, 5.00]	88.33 [80.00, 95.00]	+86.67 [78.33, 93.33]
Toggle	4 and 6	GLM-4.5V	30	0.00 [0.00, 0.00]	96.67 [91.67, 100.00]	+96.67 [91.67, 100.00]
Toggle	4 and 6	Molmo2-8B	30	1.67 [0.00, 5.00]	10.00 [3.33, 16.67]	+8.33 [3.33, 15.00]
Toggle	4 and 6	InternVL3.5-38B	30	5.00 [0.00, 10.00]	85.00 [73.33, 95.00]	+80.00 [70.00, 90.00]
Toggle	4	Qwen3-VL-32B	15	0.00 [0.00, 0.00]	86.67 [73.33, 96.67]	+86.67 [73.33, 96.67]
Toggle	4	GLM-4.5V	15	0.00 [0.00, 0.00]	100.00 [100.00, 100.00]	+100.00 [100.00, 100.00]
Toggle	4	Molmo2-8B	15	0.00 [0.00, 0.00]	3.33 [0.00, 10.00]	+3.33 [0.00, 10.00]
Toggle	4	InternVL3.5-38B	15	0.00 [0.00, 0.00]	86.67 [73.33, 96.67]	+86.67 [73.33, 96.67]
Toggle	6	Qwen3-VL-32B	15	3.33 [0.00, 10.00]	90.00 [80.00, 100.00]	+86.67 [76.67, 96.67]
Toggle	6	GLM-4.5V	15	0.00 [0.00, 0.00]	93.33 [83.33, 100.00]	+93.33 [83.33, 100.00]
Toggle	6	Molmo2-8B	15	3.33 [0.00, 10.00]	16.67 [6.67, 30.00]	+13.33 [3.33, 26.67]
Toggle	6	InternVL3.5-38B	15	10.00 [0.00, 20.00]	83.33 [63.33, 100.00]	+73.33 [53.33, 90.00]

C.4 Score-level test of direct judgment

For pair p and candidate order o, let $\mu(V)$ be the direct A-minus-B next-token logit margin for witness video V; one video has correct answer A and the other B within the order. The score-level outcome is $M_{p,o} = \mathbf{1}\{\mu(V_{\mathrm{A}}) > \mu(V_{\mathrm{B}})\}$, with ties counted one half, averaged over the two orders and the 60 pairs. Any fixed position or candidate preference gives 50%, and both-orders-correct gives 25%. Averaging each video's margin over the two candidate orders removes the position bias and yields a single-video score whose sign should favor the video's own trajectory; its accuracy over the 120 witness videos and the corresponding paired success are the operational counterparts of the within-pair test. Table 12 reports the four models, with pointwise 95% intervals from 10,000 stratified pair-bootstrap draws for the within-pair test; these estimates are descriptive companions to the answer-level families and are not part of them. Qwen's margins are large and mostly positive (91% of the 240 margins), so its argmax answers are A on 219 of 240 requests although the margins separate the two videos in 73.33% of pair-orders. GLM's margins are compressed (standard deviation 0.65 logits, all positive), InternVL's are coarsely quantized (15.8% of pair-orders tie exactly), and Molmo's are centered near zero with no directional signal.

Table 12: **Score-level test of direct judgment.** All entries are percentages: the within-pair margin test (pointwise 95% intervals), de-biased single-video accuracy and paired success from order-averaged margins, direct and CoT paired success from Table 2, and the share of positive margins (position bias).

Model	Within-pair	De-biased single-video	De-biased paired	Direct paired	CoT paired	Margins > 0
Qwen3-VL-32B	73.33 [65.42, 81.25]	60.00	16.67	5.83	75.00	91.25
GLM-4.5V	60.83 [54.58, 67.08]	49.58	5.00	0.00	88.33	100.00
Molmo2-8B	55.00 [46.67, 63.33]	51.25	3.33	3.33	14.17	49.17
InternVL3.5-38B	59.58 [51.25, 67.92]	52.08	6.67	4.17	88.33	15.00

D NATURAL-VIDEO STUDY DETAILS

D.1 DESIGN, PROMPTS, AND PARSING

The natural-video study uses the 400 videos and reference propositions of Section A.3 and the negatives of Section 2.1. Each cell (model, interface, evidence condition, negative type) contains 800 requests: 400 videos in two candidate orders. Direct and CoT conditions use the same task, candidate text, media, and order. The three evidence conditions are ordered frames, the fixed shuffled-frame control, and no video.

Nominal frame counts are 64 for Qwen, 48 for GLM, 80 for Molmo, and 48 for InternVL, followed by each model's native processor. Ordered and shuffled inputs use the same decoded frame multiset within a model. The prompt, with the question and candidate propositions filled per video, is:

```
You are given {N} frames from a video in the displayed
sequence.
Question: {question}
Choose the statement that more accurately describes the
target video. Use the supplied visual evidence when
available. Evaluate the statements themselves, including
whether their events occur.
A. {first proposition}
B. {second proposition}
{instruction}
```

Without video, the first line reads "No video frames are provided." Ordered and shuffled media receive identical wording, which neither reveals the control assignment nor implies that shuffled frames are chronological. The direct instruction is "Give your choice immediately. Reply with exactly one token: A or B." The CoT instruction is "Reason through the available evidence before deciding. End with a separate final line exactly in the form Answer: A or Answer: B."

CoT answers are read from the final `Answer: A` or `Answer: B` line, allowing minor formatting variants (markdown emphasis, a trailing period, an echoed proposition, and GLM's native answer wrapper); unparseable or truncated outputs and tied direct maxima are scored as incorrect.

D.2 STATISTICAL FAMILIES AND COMPLETE RESULTS

The analysis uses 10,000 shared paired bootstrap draws within the four video strata, seed 20260911. Four families are analyzed separately: the 18-estimate primary family of CoT model comparisons, the 12-estimate interface family, the 16-estimate evidence family, and the 18-estimate no-video-subtracted diagnostic family. The primary family contains both CoT gaps (event-denial and content-matched) for each of the six model pairs, together with their negative-type changes (Table 14). The interface family contains, for each model, the CoT-minus-direct change in event-denial accuracy, in content-matched accuracy, and in the denial-minus-content-matched difference (Table 13).

Table 13: **CoT minus direct judgment under ordered video.** Entries are percentage-point changes in event-denial accuracy, in content-matched accuracy, and in the denial-minus-content-matched accuracy difference, with simultaneous 95% intervals over this 12-estimate interface family. Qwen's CoT ordered-denial cell contains nine invalid and one truncated response out of 800; GLM's CoT ordered cells contain 194 (denial) and 196 (content-matched) invalid-or-truncated responses out of 800, so GLM's changes mix completion failure with judgment.

Model	Event-denial (pp)	Content-matched (pp)	Difference (pp)
Qwen3-VL-32B	-17.63 $[-24.63, -10.63]$	-0.50 $[-7.50, 6.50]$	-17.13 $[-24.13, -10.13]$
GLM-4.5V	-32.13 $[-39.13, -25.13]$	-7.38 $[-14.38, -0.38]$	-24.75 $[-31.75, -17.75]$
Molmo2-8B	$+6.38$ $[-0.63, 13.38]$	-5.38 $[-12.38, 1.63]$	$+11.75$ $[4.75, 18.75]$
InternVL3.5-38B	-8.00 $[-15.00, -1.00]$	$+5.25$ $[-1.75, 12.25]$	-13.25 $[-20.25, -6.25]$

The CoT evidence family shows significant ordered-minus-no-video effects for Qwen on both negative types, Molmo on both, and InternVL on denial (Table 15). The no-video-subtracted comparison

Table 14: **All 18 primary CoT comparison estimates.** Each gap is the first model minus the second in ordered-video CoT accuracy, and the change is the event-denial gap minus the content-matched gap, all in percentage points. Intervals are simultaneous 95% over this 18-estimate family.

Model pair	Event-denial gap	Content-matched gap	Change
Qwen − GLM	+14.75 [6.38, 23.13]	+15.63 [7.25, 24.00]	−0.88 [−9.25, 7.50]
Qwen − Molmo	+6.75 [−1.63, 15.13]	+11.75 [3.38, 20.13]	−5.00 [−13.38, 3.38]
Qwen − InternVL	−30.13 [−38.50, −21.75]	+4.50 [−3.88, 12.88]	−34.63 [−43.00, −26.25]
GLM − Molmo	−8.00 [−16.38, 0.38]	−3.88 [−12.25, 4.50]	−4.13 [−12.50, 4.25]
GLM − InternVL	−44.88 [−53.25, −36.50]	−11.13 [−19.50, −2.75]	−33.75 [−42.13, −25.38]
Molmo − InternVL	−36.88 [−45.25, −28.50]	−7.25 [−15.63, 1.13]	−29.63 [−38.00, −21.25]

repeats the CoT between-model analysis after subtracting no-video accuracy within model and negative type (Table 16). It retains three significant negative-type changes.

Table 15: **CoT evidence contrasts.** Percentage-point differences between CoT ordered-video accuracy and CoT no-video or shuffled accuracy, with simultaneous 95% intervals over this 16-estimate family.

	Event-denial negatives		Content-matched negatives	
Model	Ordered − no video	Ordered − shuffled	Ordered − no video	Ordered − shuffled
Qwen3-VL-32B	+21.50 [14.50, 28.50]	+3.88 [−3.13, 10.88]	+11.38 [4.38, 18.38]	+2.50 [−4.50, 9.50]
GLM-4.5V	+5.50 [−1.50, 12.50]	−2.38 [−9.38, 4.63]	+0.50 [−6.50, 7.50]	+1.13 [−5.88, 8.13]
Molmo2-8B	+10.00 [3.00, 17.00]	+6.50 [−0.50, 13.50]	+7.75 [0.75, 14.75]	+2.25 [−4.75, 9.25]
InternVL3.5-38B	+42.13 [35.13, 49.13]	+1.38 [−5.63, 8.38]	+6.88 [−0.13, 13.88]	+1.63 [−5.38, 8.63]

Table 16: **No-video-subtracted CoT comparisons (diagnostic).** As in Table 14, after subtracting each model's no-video accuracy within negative type. Intervals are simultaneous 95% over this separate 18-estimate family.

Model pair	Event-denial gap	Content-matched gap	Change
Qwen − GLM	+16.00 [5.24, 26.76]	+10.88 [0.12, 21.63]	+5.13 [−5.63, 15.88]
Qwen − Molmo	+11.50 [0.74, 22.26]	+3.63 [−7.13, 14.38]	+7.88 [−2.88, 18.63]
Qwen − InternVL	−20.63 [−31.38, −9.87]	+4.50 [−6.26, 15.26]	−25.13 [−35.88, −14.37]
GLM − Molmo	−4.50 [−15.26, 6.26]	−7.25 [−18.01, 3.51]	+2.75 [−8.01, 13.51]
GLM − InternVL	−36.63 [−47.38, −25.87]	−6.38 [−17.13, 4.38]	−30.25 [−41.01, −19.49]
Molmo − InternVL	−32.13 [−42.88, −21.37]	+0.88 [−9.88, 11.63]	−33.00 [−43.76, −22.24]

Table 17: **All 48 natural-video accuracy cells.** Accuracy averages two candidate orders within each of 400 videos; brackets are pointwise 95% intervals.

Model	Interface	Evidence	Event-denial accuracy (%)	Content-matched accuracy (%)
Qwen3-VL-32B	Direct	Ordered	61.38 [57.38, 65.38]	64.38 [60.38, 68.13]
		Shuffled	56.38 [52.25, 60.38]	63.50 [59.75, 67.13]
		No video	2.88 [1.88, 4.00]	58.63 [55.00, 62.25]
	CoT	Ordered	43.75 [39.63, 47.88]	63.88 [59.75, 67.75]
		Shuffled	39.88 [35.63, 44.00]	61.38 [57.25, 65.50]
		No video	22.25 [19.38, 25.13]	52.50 [48.63, 56.50]
GLM-4.5V	Direct	Ordered	61.13 [58.00, 64.38]	55.63 [52.63, 58.63]
		Shuffled	61.13 [58.00, 64.38]	54.38 [51.63, 57.13]
		No video	43.13 [40.75, 45.63]	50.13 [48.13, 52.13]
	CoT	Ordered	29.00 [25.50, 32.50]	48.25 [44.25, 52.25]
		Shuffled	31.38 [28.00, 34.75]	47.13 [43.38, 51.00]
		No video	23.50 [21.25, 25.75]	47.75 [44.13, 51.38]
Molmo2-8B	Direct	Ordered	30.63 [27.38, 33.88]	57.50 [53.63, 61.38]
		Shuffled	28.88 [25.75, 32.00]	56.75 [52.88, 60.63]
		No video	40.63 [38.50, 42.88]	53.75 [49.75, 57.75]
	CoT	Ordered	37.00 [33.25, 40.75]	52.13 [48.25, 56.00]
		Shuffled	30.50 [26.88, 34.25]	49.88 [45.88, 54.00]
		No video	27.00 [24.13, 29.75]	44.38 [40.63, 48.13]
InternVL3.5-38B	Direct	Ordered	81.88 [78.50, 85.13]	54.13 [49.75, 58.50]
		Shuffled	84.63 [81.38, 87.63]	55.50 [51.63, 59.50]
		No video	66.88 [63.75, 70.00]	51.75 [48.75, 54.75]
	CoT	Ordered	73.88 [70.13, 77.50]	59.38 [55.25, 63.38]
		Shuffled	72.50 [68.88, 76.00]	57.75 [53.63, 61.88]
		No video	31.75 [28.88, 34.63]	52.50 [48.63, 56.38]

Table 18: Output diagnostics for all 48 natural-video cells. Each cell has 800 requests (400 videos in two candidate orders); Type is the negative type and Fmt the requested format. Invalid counts exclude truncations and include direct ties; format columns use the requested `Answer:` line only. Order columns count videos whose correctness or chosen label differs between the two candidate orders.

Model	Interface	Evidence	Pop.	Correct	Invalid	Trunc.	Ties	Fmt fail	Fmt corr.	Ord. corr.	Ord. ans.
Qwen	Direct	Ordered	Denial	491	12	0	12	12	491	69	76
Qwen	Direct	Ordered	Content	515	18	0	18	18	515	107	114
Qwen	Direct	Shuffled	Denial	451	10	0	10	10	451	81	86
Qwen	Direct	Shuffled	Content	508	13	0	13	13	508	144	146
Qwen	Direct	No video	Denial	23	3	0	3	3	23	23	26
Qwen	Direct	No video	Content	469	7	0	7	7	469	143	147
Qwen	CoT	Ordered	Denial	350	9	1	0	9	350	104	106
Qwen	CoT	Ordered	Content	511	1	3	0	1	511	101	102
Qwen	CoT	Shuffled	Denial	319	6	0	0	6	319	79	81
Qwen	CoT	Shuffled	Content	491	2	10	0	2	491	99	106
Qwen	CoT	No video	Denial	178	0	0	0	0	178	104	104
Qwen	CoT	No video	Content	420	1	0	0	1	420	122	122
GLM	Direct	Ordered	Denial	489	5	0	5	5	489	157	160
GLM	Direct	Ordered	Content	445	6	0	6	6	445	247	250
GLM	Direct	Shuffled	Denial	489	4	0	4	4	489	165	167
GLM	Direct	Shuffled	Content	435	5	0	5	5	435	263	266
GLM	Direct	No video	Denial	345	11	0	11	11	345	231	237
GLM	Direct	No video	Content	401	7	0	7	7	401	329	335
GLM	CoT	Ordered	Denial	232	184	10	0	727	28	92	168
GLM	CoT	Ordered	Content	386	179	17	0	704	45	112	158
GLM	CoT	Shuffled	Denial	251	172	10	0	727	26	91	148
GLM	CoT	Shuffled	Content	377	189	8	0	708	54	135	183
GLM	CoT	No video	Denial	188	7	0	0	323	143	70	70
GLM	CoT	No video	Content	382	15	0	0	200	298	156	168
Molmo	Direct	Ordered	Denial	245	0	0	0	0	245	87	87
Molmo	Direct	Ordered	Content	460	0	0	0	0	460	142	142
Molmo	Direct	Shuffled	Denial	231	0	0	0	0	231	87	87
Molmo	Direct	Shuffled	Content	454	0	0	0	0	454	138	138
Molmo	Direct	No video	Denial	325	0	0	0	0	325	73	73
Molmo	Direct	No video	Content	430	1	0	1	1	430	124	124
Molmo	CoT	Ordered	Denial	296	16	5	0	113	277	120	131
Molmo	CoT	Ordered	Content	417	43	3	0	81	390	139	154
Molmo	CoT	Shuffled	Denial	244	19	1	0	140	227	96	108
Molmo	CoT	Shuffled	Content	399	53	4	0	102	376	121	147
Molmo	CoT	No video	Denial	216	18	0	0	58	202	110	126
Molmo	CoT	No video	Content	355	75	0	0	175	306	155	177
InternVL	Direct	Ordered	Denial	655	11	0	11	11	655	51	54
InternVL	Direct	Ordered	Content	433	34	0	34	34	433	81	99
InternVL	Direct	Shuffled	Denial	677	9	0	9	9	677	43	47
InternVL	Direct	Shuffled	Content	444	44	0	44	44	444	112	133
InternVL	Direct	No video	Denial	535	40	0	40	40	535	117	124
InternVL	Direct	No video	Content	414	40	0	40	40	414	232	244
InternVL	CoT	Ordered	Denial	591	1	0	0	1	591	85	85
InternVL	CoT	Ordered	Content	475	2	0	0	2	475	107	107
InternVL	CoT	Shuffled	Denial	580	1	0	0	1	580	104	105
InternVL	CoT	Shuffled	Content	462	0	5	0	0	462	106	108
InternVL	CoT	No video	Denial	254	0	101	0	0	254	138	173
InternVL	CoT	No video	Content	420	1	19	0	1	420	122	127

Stratum-level accuracies for all 192 cells are provided in the supplementary CSV.

D.3 Why CoT judgment loses on event-denial negatives: a rationale analysis

This analysis asks why Qwen's ordered-video accuracy on event-denial negatives falls from 61.38% under direct judgment to 43.75% under CoT judgment while content-matched accuracy changes by -0.50 points. It uses the 800 CoT and 800 direct responses of the ordered-video event-denial cells, the same model's responses under the other evidence conditions and under the content-matched pair, and InternVL's denial cells as a comparator. The analyses, the rationale codebook, and the coding sample were fixed before any rationale was read.

Where the decline lives. At video level, after averaging the two candidate orders, 54 videos move from correct to wrong and 12 from wrong to correct (Table 19); 53 of the 54 losses receive the denial in both orders. Losses concentrate in the event-relative-state and visibility-continuity strata and, across strata, in anchor events that presuppose a prior absence (Table 20): on the 83 videos whose anchor event reads "comes back into view", accuracy falls from 73.5% to 33.1%, whereas "gets

partly hidden" and "leaves the frame" anchor events change by less than five points. Video duration is unrelated to the losses. Without video, both interfaces favor the denial (direct 96.8% of responses, CoT 77.8%); with ordered video the direct denial choice falls to 37.1% and the CoT choice only to 55.0%.

Table 19: **Video-level transitions from direct to CoT judgment, event-denial negatives, ordered video.** Each video's correctness is averaged over the two candidate orders; partial rows are videos with one correct order under either interface.

Stratum	Videos	Both right	Loss	Gain	Both wrong	Partial	Direct (%)	CoT (%)
Order	120	24	12	3	27	54	57.92	41.67
State	100	28	25	1	13	33	77.50	44.50
Location	100	17	4	7	36	36	39.00	38.50
Visibility	80	28	13	1	10	28	74.38	52.50
All	400	97	54	12	86	151	61.38	43.75

Table 20: **Accuracy by the phrasing of the anchor event.** Ordered video, both interfaces; the change is CoT minus direct in percentage points; loss and gain counts are videos.

Anchor-event phrasing	Videos	Direct (%)	CoT (%)	Change (pp)	Losses	Gains
"comes back into view"	83	73.49	33.13	−40.36	23	0
"gets partly hidden"	80	58.75	55.63	−3.12	5	5
"becomes non-visible"	72	71.53	50.69	−20.84	11	1
"enters or first shows up"	64	71.88	54.69	−17.19	6	2
"gets fully hidden"	51	38.24	20.59	−17.65	8	0
"leaves the frame"	25	34.00	30.00	−4.00	0	1
"disappears for good"	23	45.65	54.35	+8.70	1	3
other	2	75.00	50.00	−25.00	0	0

What the wrong rationales say. All 440 wrong CoT answers were coded under the four categories of Table 21. A seeded sample of 60 wrong and 30 correct responses was read in full; the remaining wrong responses were screened by cue lists into those containing conflict language, those containing an explicit absence assertion, and neither; every flagged or unclassified rationale was read in full (89 more), and a seeded 30-rationale excerpt check of the explicit-assertion stratum found 29 absence conclusions and one timing misreading, so the remaining 264 explicit-assertion rationales were assigned that code by rule. In total 176 rationales were read individually.

Of the 440 wrong CoT answers, 396 conclude that the anchor event never happens; 23 acknowledge the event, dispute a detail of the faithful statement, and choose the denial by elimination; 16 are long oscillating rationales (median 17,070 characters against 666 for the first category) whose final letter contradicts their own evaluation; and 5 misread the denial as a timing claim.

Among the 132 absence conclusions read individually, 31 first locate the event at a specific frame and then reason to its denial after disputing a detail; the others report that the entity never left view or was already present (27), that the event fails a strict threshold such as "fully hidden" (21), that it never happens (42), that no reappearance was found (7), or that the entity was not found (7). Mechanism disputes dominate the visibility-continuity losses: the rationale affirms the disappearance but rejects the reference's mechanism (occluded within the frame versus leaving it) and still picks the statement that denies the event.

The same frames, a different negative. Among the 54 loss videos, CoT judgment against the content-matched negative is fully correct on 36 and half-correct on 13. Eight loss videos were read under both pairs; in all eight the content-matched-pair rationale locates the event that the denial-pair rationale declared absent. Table 3 shows one such pair chosen by a fixed rule; in another, the denial-pair rationale states in both orders that "at no point in the video is the bottle fully hidden from view" while the content-matched-pair rationale states that "by frame 11, it is fully hidden behind the cloth" and that it "becomes visible again starting from Frame 14".

Table 21: **Codes for the 440 wrong CoT answers.** The rule-assigned rows come from the explicit-assertion stratum described in the text.

Code	Description	All 440	Read individually
C1	Concludes that the anchor event never happens	396	132
C2	Acknowledges the event, disputes a detail, denies by elimination	23	23
C4	Oscillating rationale whose final letter contradicts its evaluation	16	16
C3	Reads the denial as a claim about timing	5	5
	All wrong CoT answers	440	176

Reading. Besides the routes listed in Section 5.2, a presupposition check ("comes back" requires a prior absence) also ends at the denial. Direct judgment writes no rationale and so takes no such route, although it still chooses the denial in 37.1% of ordered-video responses. In a content-matched pair both statements assert the event, so neither offers this exit, and the controlled truth-switch pairs likewise offer two event-asserting trajectories. InternVL shows the same direction at smaller size (18 losses, 4 gains; denial choices rise from 134 to 208 of 800 with video).

Frame-level check. The references were verified against the full videos, whereas the model saw 64 uniformly sampled frames, so a short absence can fall between frames. For the 83 videos whose anchor event reads "comes back into view", the 64 frames presented to Qwen were rendered as labelled contact sheets and each video was classified, without reference to any model output or loss status, as A (the named object is out of view in at least one sampled frame and visible again later), B (visible in every sampled frame), C (leaves view without returning), or D (unclear). An LLM assistant proposed the classification with the absent and return frames noted, and one of the authors reviewed every item and accepted all 83. Result: A 47, B 25, C 6, D 5; of the 23 loss videos, 15 are A, 7 B, and 1 D. Qwen's ordered-denial accuracy on the 47 A-videos is 85.1% direct and 41.5% CoT; on the 31 B/C-videos it is 62.9% and 25.8%. The decline persists where the reappearance is visible in the model's input. On the B/C-videos the sampled frames do not show the reappearance the reference asserts, so neither interface can be scored against the frames there; these videos hold 7 of the 23 losses. The complete tables, per-response codes, quotations, and code are in the artifact's `data/m11-trace-audit` folder, and the frame-check manifest, responses, and analysis under `data/m11-frame-check`.

E Held-out verifier-selection analysis

E.1 Objective, policies, and partitions

The target is ordered-video faithful-versus-content-matched accuracy on the same 400 videos. Eligible sets contain all eight configurations, the four direct configurations, or the four CoT configurations. For an eligible set $\mathcal{K}$ and selection videos S,

$$\hat{k}_{\mathrm{den}}(S) \in \arg\max_{k \in \mathcal{K}} \hat{A}_k^{\mathrm{den}}(S), \qquad \hat{k}_{\mathrm{cm}}(S) \in \arg\max_{k \in \mathcal{K}} \hat{A}_k^{\mathrm{cm}}(S), \tag{4}$$

where $\hat{A}_k^{\mathrm{den}}$ and $\hat{A}_k^{\mathrm{cm}}$ are stratum-weighted event-denial and content-matched accuracies on S; tied maximizers are selected uniformly. Both selected configurations are evaluated by content-matched accuracy on the held-out videos. Twenty stratified five-fold partitions (seed 20260916) give evaluation folds of 24/20/20/16 videos across the four strata; every response of a video stays with that video.

E.2 Uncertainty and selection frequencies

Uncertainty uses 10,000 whole-video stratified bootstrap draws (seed 20260917); each duplicated video keeps its fold assignment, and selection on the selection folds is repeated under the bootstrap weights, so the intervals include selection variability. The three selection costs form one simultaneous family (Table 22). Fold-to-fold variation is not an independent-sample standard error (Bates et al., 2024), so the range and standard deviation over the twenty partitions are descriptive only. Table 23 lists the configurations each policy selects.

Table 22: **Repeated cross-validation summaries.** Selection costs are content-matched-selected minus denial-selected held-out content-matched accuracy, in percentage points. Range and standard deviation describe the twenty fixed partitions, which reuse the same 400 videos. The last two columns evaluate the same two policies on held-out event-denial comparisons.

				20 partitions		Denial accuracy (%)	
Eligible set	Cost	Pointwise 95%	Simult. 95%	Range	SD	Denial-sel.	Content-sel.
All eight	+8.88	[4.73, 14.13]	[2.46, 15.30]	7.75–10.25	0.72	81.88	55.82
Four direct	+10.25	[5.88, 14.25]	[3.83, 16.67]	10.25–10.25	0.00	81.88	61.38
Four CoT	+4.50	[−2.17, 8.88]	[−1.92, 10.92]	4.50–4.50	0.00	73.88	43.75

Table 23: **Configuration selection frequencies** over the 100 selections (20 partitions × 5 held-out folds) for each eligible set and criterion.

Eligible set	Selection criterion	Configurations chosen in 100 selections
All eight	Event-denial	Direct InternVL (100%)
	Content-matched	Direct Qwen (70%), CoT Qwen (30%)
Four direct	Event-denial	Direct InternVL (100%)
	Content-matched	Direct Qwen (100%)
Four CoT	Event-denial	CoT InternVL (100%)
	Content-matched	CoT Qwen (100%)

E.3 Separately fixed split

As a sensitivity check, a separately fixed split (seed 20260918) selects on 200 videos and evaluates on the other 200, with uncertainty from 10,000 bootstrap draws of the evaluation videos (seed 20261017) conditional on the selected configurations; the three costs form their own simultaneous family (Table 24). The direct-only cost remains significant; the all-eight and CoT-only costs are not significant.

Table 24: **Selection on a separately fixed 200/200 split.** Selected configurations (denial-selected / content-matched-selected; -dir direct, -cot CoT), their content-matched accuracy on the 200 evaluation videos, and the selection cost with approximate simultaneous, pointwise, and bounded 95% intervals.

Eligible set	Selected	Target (%)	Cost (pp)	Pointwise 95%	Bounded 95%
All eight	InternVL-dir / Qwen-cot	54.75 / 59.50	+4.75 [−2.75, 12.25]	[−1.75, 11.50]	[−17.13, 26.63]
Four direct	InternVL-dir / Qwen-dir	54.75 / 64.00	+9.25 [1.75, 16.75]	[3.50, 15.00]	[−12.63, 31.13]
Four CoT	InternVL-cot / Qwen-cot	57.50 / 59.50	+2.00 [−5.50, 9.50]	[−4.25, 8.25]	[−19.88, 23.88]

F Statistical procedures, validation, and artifact

F.1 Estimands and inference families

The paper reports four outcomes: win rate (construction study), paired success (truth-switch studies), accuracy (natural-video study), and held-out target accuracy (selection analysis). Uncertainty is computed by bootstrap over the independent unit of each study, and simultaneous 95% intervals are formed within declared families: the 864-estimate construction family; the two six-estimate truth-switch families; the 18-estimate primary, 12-estimate interface, 16-estimate evidence, and 18-estimate no-video-subtracted natural-video families; and the three-estimate selection family. No multiplicity correction is applied across families.

- The construction family is a studentized max-$|t|$ family over 864 estimates (50,000 semantic-program replicates, seed 20260731, critical value 4.05).
- The truth-switch, natural, and selection families use an unstudentized common radius: with bootstrap replicates $\widehat{\theta}_j^*$ of the J family estimates $\widehat{\theta}_j$, the radius is the 95th percentile over replicates of $\max_j |\widehat{\theta}_j^* - \widehat{\theta}_j|$, and every interval is $\widehat{\theta}_j$ plus or minus that radius, clipped to the parameter range, so a low-variance estimate receives the family radius; for example,

the direct GLM truth-switch interval is $[0.00, 9.17]\%$ although its percentile interval is degenerate.

- Pointwise intervals are 2.5/97.5 bootstrap percentiles. Where empirical bootstraps can degenerate under constant outcomes, Hoeffding bounds with a Bonferroni split over the family are reported as conservative companions.

Seeds are 20260915 and 20260919 for the truth-switch studies, 20260921 for the score-level test, 20260911 for the natural-video study, and 20260916 and 20260917 for the selection partitions and bootstrap.

F.2 Validation

All summary statistics were reproduced from the per-response outputs by independent code, and every study checks the completeness of its model/unit/condition/order keys, parser outcomes, and candidate-order aggregation. Infrastructure failures were retried with identical inputs and are excluded from all counts.

F.3 Artifact

The anonymous artifact contains the controlled-task generators and renderer, prompts and parsers, per-response outputs of the construction study and the natural-video study (including, for all 400 videos, the question, the reference, and both negative propositions), the rationale-analysis tables and codes, and the analysis scripts that reproduce the tables and figures. The truth-switch and selection packages will be added in the camera-ready release.

G Extended related work

Measurement validity and negatives. Shortcut learning (Geirhos et al., 2020) and construct validity (Cronbach & Meehl, 1955; Jacobs & Wallach, 2021) explain benchmark success that fails under distributions closer to the intended construct. Hard negatives expose order and relation insensitivity (Thrush et al., 2022; Yuksekgonul et al., 2023), generated negatives carry exploitable textual artifacts (Hsieh et al., 2023), and leaderboard rankings move with ordering, prompting, and scoring (Alzahrani et al., 2024). Contrast sets probe local decision boundaries with small label-changing perturbations (Gardner et al., 2020); our endpoint-matched negatives apply the same logic to executable state trajectories while preserving both endpoints, and our witness videos apply it to the video rather than the candidate.

Contrastive video-language evaluation. VELOCITI (Saravanan et al., 2025) separates absolute decisions from relative ranking and tests answer-position bias with absent-video and single-frame controls; its appendix reports reduced performance under chain-of-thought prompting for its two tested models. Vinoground (Zhang et al., 2024) pairs short natural videos with temporally counterfactual captions and reports a text score that requires the correct caption for both videos of a pair; our paired success follows this convention with executable references, shared endpoint pixels, and fixed text inputs within a candidate order.

Structured and evidence-grounded verification. Atomic and structured systems decompose judgments into facts, scenes, claims, clues, or visual premises (Min et al., 2023; Wang et al., 2026b; Min et al., 2026); NeuS-QA grounds question-derived propositions and applies temporal-logic model checking before answering (Shah et al., 2026); RT-NeuS reduces acquisition cost through adaptive sampling and cache reuse (Liang et al., 2026); EVPV checks candidate reasoning premises against visual constraints extracted once from the image and question (Wang et al., 2026a); SG-PVR trains video reward reasoning from atomic claims and scene graphs (Kim et al., 2026); DEFT defers trajectory options until after a scene-derived decision (Huang et al., 2026). These systems change the verification interface; our study measures how the interface and the negative type change what an evaluation establishes about a fixed verifier.