MENTALWM-BENCH: EVALUATING MENTAL-STATE TRANSITIONS IN VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation models are increasingly used to build dynamic worlds with characters that act and interact. Coherent behavior in these worlds requires more than realistic appearance and motion: events must change characters' emotions and intentions, and these changes must shape subsequent behavior. Mental world modeling captures this connection between events, internal states, and actions. Yet visual fidelity, event occurrence, and a correct final expression alone cannot establish whether an event-driven mental-state transition unfolds in order and persists across interactions. We introduce MentalWM-Bench to evaluate these transitions through observable behavior. Three complementary tasks assess trigger execution, complete transitions, and two-turn continuation across individual, dyadic, and group scenarios. Evidence-based criteria examine events, initial and resulting states, emotion and intent changes, temporal order, and cross-turn continuity, alongside scene and physical fidelity. We evaluate seven video generators and compare automatic assessments with multi-rater human judgments. Correct outcomes can coexist with incomplete transitions, and successful first-turn states do not ensure subsequent preservation. Physical and mental-state scores remain weakly associated within generators. Emotion and intent updates score below endpoints, with lower performance in longer continuations and group interactions. MentalWM-Bench provides a framework for systematically evaluating mental-state dynamics in video, informing the development of next-generation world models with sustained, coherent character interactions.

1 INTRODUCTION

World models represent how environments evolve to support prediction and planning (Ha & Schmidhuber, 2018; Hafner et al., 2019). Video world models extend this ambition to visual environments that respond to actions (Yang et al., 2023; Bruce et al., 2024). When these worlds contain interacting characters, coherent dynamics involve not only physical motion but also how events change emotions and intentions, shaping subsequent behavior. Mental world modeling addresses this connection (Fei & Zhao, 2026; Baker et al., 2017; Rabinowitz et al., 2018; Zhou et al., 2026; Bai et al., 2026). For virtual characters and interactive storytelling, this makes responses reflect prior experience and gives subsequent interactions continuity. Both updating and preservation therefore matter for coherent interaction.

Existing benchmarks assess video quality, world generation, multi-turn interaction, and social responses(Huang et al., 2024; Duan et al., 2025; Ying et al., 2026; Peng et al., 2026), as well as temporal composition and object-state changes(Feng et al., 2024; Sun et al., 2025; Han et al., 2026). These perspectives are valuable, but scene stability, event execution, or terminal-expression correctness alone cannot establish an ordered, sustained mental-state transition. Our question is: *can a generator realize a specified event-driven change and carry its outcome into the next interaction?* This connects the visible process to the continuing state it should create.

Our design draws on emotion appraisal, which relates responses to an event's significance for an individual (Scherer, 2009; Moors et al., 2013), and event segmentation, which emphasizes updating representations as events unfold (Zacks et al., 2007). These perspectives motivate checking the event, response, and resulting state together. We operationalize measurable states through expressed

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.

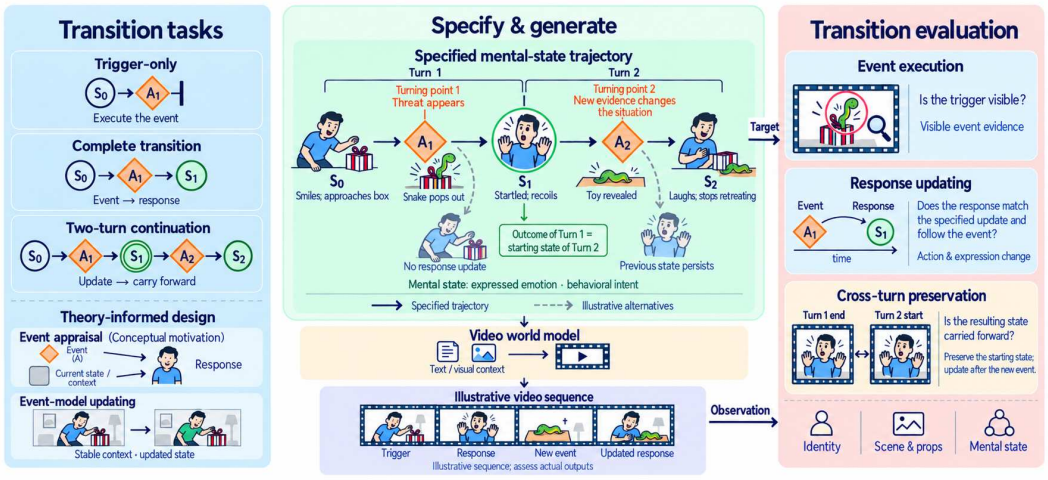


Figure 1: **Overview of MentalWM-Bench. Left:** Three tasks test trigger execution, complete transition, and continuation. **Center:** A birthday-box trajectory changes from anticipation to fear, then amusement when the snake is revealed as a toy. S_1 links both turns; dashed branches illustrate missed updates. **Right:** Generated evidence is checked for event execution, response updating, and continuity of identity, scene/props, and mental state.

emotion and goal-directed behavior, assessing whether they change after an event and persist across interactions.

MentalWM-Bench organizes this question into three complementary tasks: **trigger execution**, **complete transition**, and **two-turn continuation**. In $S_0 \xrightarrow{A_1} S_1 \xrightarrow{A_2} S_2$, S_1 is both an outcome and the next starting state. For example, a character opens a gift expectantly, recoils when a snake appears, then relaxes when it proves to be a toy (Figure 1): fear links the two turns. Supplying target responses in the latter tasks separates realizing a trajectory from inferring which response is appropriate. The benchmark covers 336 source scenarios and 471 prompts, with varied states and turning points across individual, dyadic, and group interactions; a five-turn pilot probes longer continuations.

Evidence-based criteria follow each trajectory: event execution, initial and resulting states, emotion/intent changes, timing, and cross-turn continuity, alongside scene and physical fidelity. This decomposition locates failures that an overall impression can conceal. We evaluate seven generators and compare automatic assessments with multi-rater human judgments. Conditional analyses ask whether process deficits remain after outcomes succeed, and whether preservation fails after a state is realized.

Our experiments show that **correct outcomes can coexist with incomplete transitions**, and **realized states are not always preserved** across interactions. **Physical fidelity does not summarize mental-state performance**: their weak association persists within generators. Emotion and intent updates score below endpoints, while longer continuations and group interactions present further challenges. Together, these findings motivate evaluating the full event–response process and its continuity, alongside visual and physical fidelity. For next-generation world models, they highlight the importance of maintaining character states across turns and updating them in response to new events. Our contributions are:

- A benchmark connecting event execution, mental-state updating, and preservation in interactive video worlds.
- Complementary tasks and observable criteria with diverse scenario coverage and multi-rater human–AI comparisons.
- Findings that distinguish plausible outcomes from coherent transitions and sustained states, identifying targets for future model development.

Table 1: Complementary benchmark coverage. MentalWM-Bench connects physical plausibility with emotion and intent, their event-linked updates, temporal realization, and preservation across generation turns. Together these checks follow a response from its initial conditions to subsequent interaction.

Benchmark	Phys.	Mental targets		Transition		Across turns	
		Emotion	Intent	Update	Timing	Multi.	Carry.
VBench	×	×	×	×	×	×	×
WorldScore	×	×	×	×	×	✓	×
TC-Bench	×	×	×	✓	✓	×	×
WBench	✓	×	×	×	×	✓	×
R^3-Bench	×	✓	✓	✓	×	×	×
SVBench	×	✓	✓	×	✓	×	×
Ours	✓	✓	✓	✓	✓	✓	✓

2 Related Work

Mental and social world models. Mental world modeling couples physical and mental states to predict agents' decisions (Fei & Zhao, 2026). Related work studies social perspectives, language-agent interaction, and collective beliefs (Zhou et al., 2026; 2024; Yu et al., 2026); MaineCoon extends social-interactive generation to streaming audio-visual worlds (Bai et al., 2026). Our benchmark tests whether generated behavior realizes and sustains specified changes, complementing models of how agents reason about others.

Video and world-model benchmarks. VBench, WorldScore, and WBench evaluate video quality, world generation, and multi-turn interaction (Huang et al., 2024; Duan et al., 2025; Ying et al., 2026). Temporal and compositional benchmarks examine transitions, binding, and object state changes (Feng et al., 2024; Sun et al., 2025; Han et al., 2026). We connect these perspectives through event-linked emotion and intent changes, following their realization into subsequent generation turns.

Social reasoning and evaluation. Social QA benchmarks assess commonsense, video understanding, and conversational perspectives (Sap et al., 2019; Zadeh et al., 2019; Kim et al., 2023). SVBench withholds expected outcomes to test whether appropriate responses emerge (Peng et al., 2026); our complete and continuation tasks specify the response and assess its realization, timing, and preservation. R^3-Bench annotates mental–physical causal chains for video QA (Niu et al., 2026). Table 1 summarizes these differences. LLM-judge overestimation in social-agent evaluation (Wang et al., 2024) motivates explicit human–AI comparisons.

3 MentalWM-Bench

MentalWM-Bench connects what changes, why a response is required, and whether its outcome persists. We define the state–event trajectory, separate its requirements into three tasks, and populate them with varied states, turning points, and participant configurations.

3.1 Event-driven Mental-State Transitions

We evaluate whether a specified change in emotion or behavioral intent is visibly realized in a generated interaction. Emotion appraisal and event updating motivate considering the event and response together (Scherer, 2009; Moors et al., 2013; Zacks et al., 2007). A target trajectory is

$$S_0 \xrightarrow{A_1} S_1 \xrightarrow{A_2} S_2, \tag{1}$$

We represent the target state by emotion e_t and behavioral intent i_t:

$$S_t = (e_t, i_t), \qquad \Delta S_t = (e_t \to e_{t+1},\ i_t \to i_{t+1}). \tag{2}$$

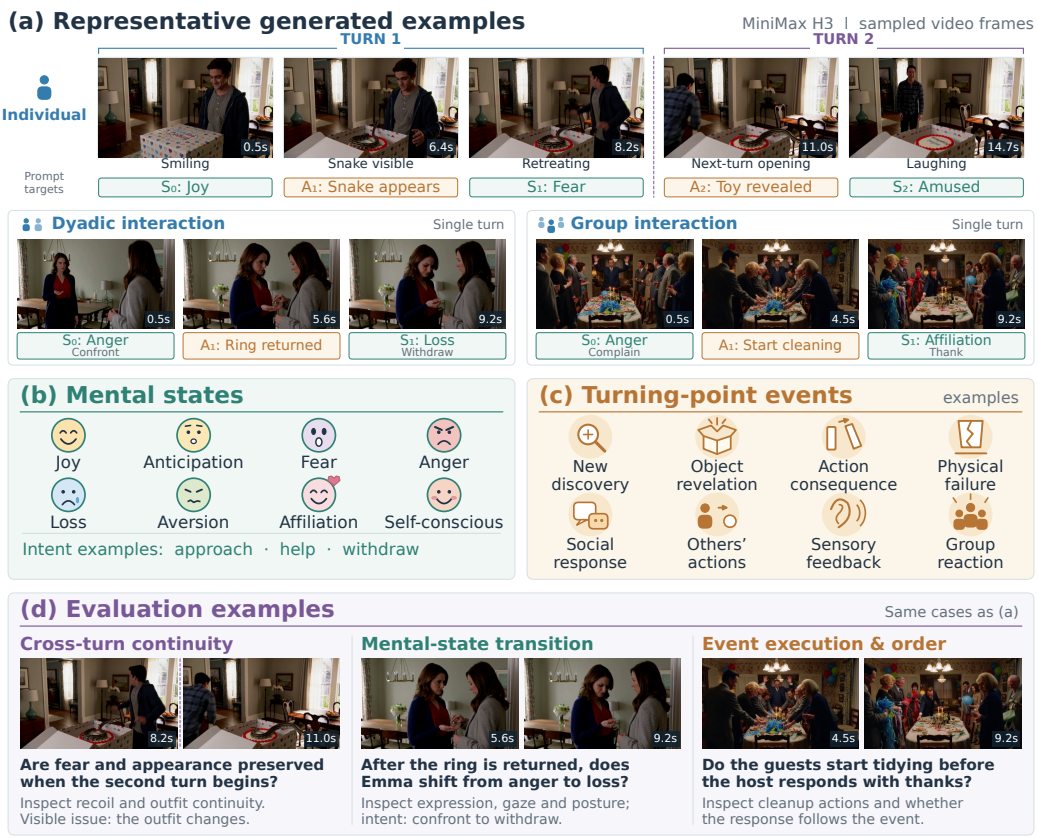


Figure 2: **From scenario coverage to observable evidence.** (a) MiniMax H3 frames span individual, dyadic, and group interactions; labels are prompt targets, not success annotations. (b,c) Varied states and turning points test responses beyond one expression or trigger. (d) Questions connect targets to visible evidence: outfit changes concern appearance; boundary behavior supports mental-state checks. Timestamps identify source frames; the dashed line marks the turn boundary.

Here A_t is an action or external event; ΔS_t denotes a categorical update, not a numerical difference. Expressions, posture, and goal-directed actions provide evidence for these targets. The birthday-box target links smiling, recoil, and laughter to successive events; each stage is evaluated separately.

Events admit multiple plausible responses. Human-reviewed targets give generators consistent requirements for trajectory realization and continuation. Laughter before the toy revelation matches an endpoint but violates timing.

3.2 Three Complementary Tasks

The tasks separate three questions that an overall video score can conflate. **Trigger-only** supplies S_0, A_1: checking execution without requesting a later response reveals whether the prerequisite event is realized. **Complete-transition** adds S_1: comparing both states and their timing distinguishes an event-linked update from merely displaying the right expression. **Two-turn continuation** adds A_2, S_2: the shared S_1 tests whether an achieved response becomes the context for another update. A five-turn pilot extends this question to longer interaction.

Figure 2 spans responses to objects, other people, and collective events, linking state/trigger categories to evidence questions. Outfit changes concern appearance continuity; mental-state preservation requires expressions, posture, and actions at the turn end and restart.

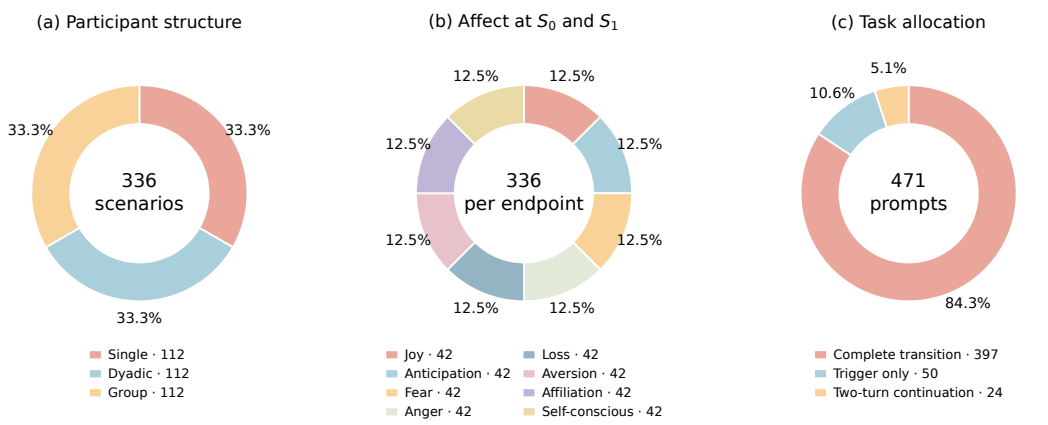


Figure 3: Dataset composition. (a) Participant structure of 336 unique scenarios. (b) Affect categories: 42 cases per category at each endpoint. (c) Allocation of 471 prompts. Percentages use the displayed denominators.

3.3 Dataset Construction and Coverage

Humans review every model-assisted or human-authored scenario before inclusion; preparation checks cover wording and the visibility of target states. Records specify participants, endpoint emotion/intent, the event, and marked S_0, A_1, S_1 spans, which scripts convert to continuous-shot prompts while retaining source IDs and versions. Behavioral rewrites express states as actions.

The 336 source scenarios contain 112 individual, 112 dyadic, and 112 group cases, with 42 cases per affect category at each endpoint. Adding 61 individual-scene rewrites gives 397 complete-transition prompts; these reuse source scenarios. With 50 trigger-only and 24 two-turn prompts, the release contains 471 prompts (Figure 3). Targets specify requested content rather than certify generation success.

4 Evaluation Metrics

Gemini evaluators compare videos with target spans using visible evidence. A WBench-informed protocol checks execution and scene/physical fidelity; mental-state checks assess the resulting response dynamics. Multiple raters assess all three tasks, including changes and continuity, on overlapping videos. Section 6 reports correspondence; Appendix B gives the rules.

4.1 Event Execution and State Change

Drawing on WBench's adherence and physical criteria (Ying et al., 2026), **Interaction** checks whether events and subject actions are executed; **Setting** checks scene and subject adherence; **Physical** checks causal plausibility and human motion. Each averages its component scores; their equal-weight mean is the three-dimension **proxy**.

Mental-state transitions translate the appraisal and event-updating motivation into eight visible checks (Scherer, 2009; Moors et al., 2013; Zacks et al., 2007): initial/resulting emotion and intent establish the states; emotion/intent change tests updating; order and absence of premature response connect it to the event. This is *Social* in Table 2 and *Social-8* in released records. Applicable checks receive 0/1/2 for missing or contradictory, partial, or clear evidence. Logical constraints tie change scores to state and event evidence. Normalized scores are averaged; inapplicable items are excluded (Appendix D.2). Changes require visible post-event differences, order requires successive initial/event/result evidence, and premature final responses fail even when the endpoint matches.

The complete-transition Overall averages Interaction, Setting, Physical, and Social. Trigger-only uses the three-dimension proxy. These are averages of evidence-based criteria, not rates of fully successful trajectories.

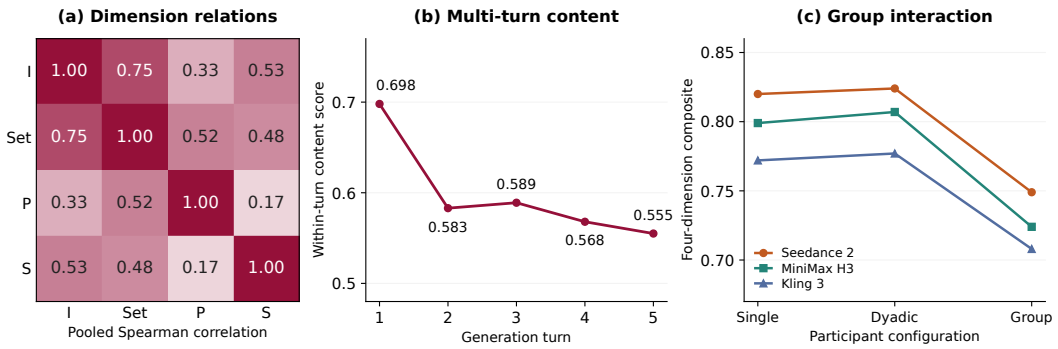


Figure 4: Complementary diagnostics. (a) Physical–mental scores correlate least across 2,657 videos (I/Set/P/S: Interaction/Setting/Physical/mental state). (b) Across 37 completed five-turn chains, model-equal content means drop most at the first continuation and stay below the initial level. (c) Group scores are lower for the displayed models; all seven appear in Appendix A. Panels use distinct protocols.

4.2 Continuation

Thirteen checks cover Turn 1 (initial state, event, result), the boundary (mental state, identity, scene/props), Turn 2 (event visibility, match, result), and the sequence (order, replay, physics, fidelity). Boundary state requires specified S_1 at the observed end and restart: a low score can reflect failed realization or failed preservation. We therefore report phase means, the mean of all 13 normalized checks, and conditional preservation results. The five-turn pilot has a separate five-check content protocol (Appendix B).

5 Results and Key Findings

We first compare models within each task, then ask what aggregate scores can miss: incomplete processes, failed preservation, and differences across physical fidelity, state changes, and participant configurations. Human comparisons in Section 6 provide a separate check on these automatic assessments.

5.1 Evaluation Setup

Seven generators produce 2,657 complete-transition and 350 trigger-only videos; six produce 144 two-turn sequences. Table 2A compares complete transitions on 336 shared base prompts per model. B isolates trigger execution on 50 shared cases; C examines state carryover and subsequent updating on 24 shared cases. Each task has its own Overall score. The 61 rewrites for five models are reported separately (Appendix A); diagnostic figures retain all 2,657 videos unless noted.

5.2 Core Findings

We follow execution, transition, and preservation in order. Even without a requested final response, trigger-only Setting exceeds Interaction for all seven models: Interaction/Setting/Physical means are 0.675/0.770/0.850 (Table 2B). This separates a convincing setting from realizing the event needed to drive a response.

Outcome realization does not guarantee process completeness. Automatic and human analyses both distinguish outcomes from complete processes. Among 168 videos with strict reassessment, 70 pass automatic event-completion and resulting-action checks; 8 (11.4%) fail full trajectory, including two order failures. Human endpoint means resemble those of the 62 passing both rules, while initial-state and timing scores are lower. Thus an accepted event and final action can coexist with incomplete initial-state or temporal evidence.

Table 2: Automatic results (0–1); human comparison: Section 6. A: complete transition on 336 shared prompts; B: trigger execution; C: continuation. A/B show dimension means, C phase means. Models rank by Overall; bold marks column best. Model means weight generators equally; full items: Appendix E.

A. Complete transition: 336 shared base prompts

Model	Interaction	Setting	Physical	Social	Overall
Seedance 2	**0.781**	0.861	0.876	**0.681**	**0.800**
MiniMax H3	0.761	**0.864**	**0.899**	0.577	0.775
Kling 3	0.769	0.833	0.806	0.588	0.749
Vidu Q2	0.670	0.806	0.834	0.501	0.703
Hailuo 2.3	0.643	0.792	0.805	0.539	0.695
Wan 2.2	0.478	0.649	0.707	0.320	0.538
Hunyuan 1.5	0.386	0.453	0.468	0.213	0.380
Model mean	0.641	0.751	0.771	0.489	0.663

B. Trigger-only execution

Model	Interaction	Setting	Physical	Overall
MiniMax H3	0.756	**0.837**	**0.940**	**0.844**
Seedance 2	**0.774**	0.836	0.877	0.829
Hailuo 2.3	0.736	0.835	0.877	0.816
Vidu Q2	0.732	0.819	0.837	0.796
Kling 3	0.688	0.807	0.887	0.794
Hunyuan 1.5	0.500	0.636	0.813	0.650
Wan 2.2	0.540	0.623	0.720	0.628
Model mean	0.675	0.770	0.850	0.765

C. Two-turn continuation

Model	Turn 1	Boundary	Turn 2	Global	Overall
Hailuo 2.3	0.882	**0.826**	**0.750**	**0.688**	**0.779**
MiniMax H3	**0.896**	0.820	0.653	0.615	0.736
Seedance 2	0.889	0.750	0.667	0.620	0.723
Kling 3	0.875	0.743	0.479	0.505	0.639
Wan 2.2	0.118	0.278	0.229	0.255	0.223
Hunyuan 1.5	0.076	0.264	0.188	0.203	0.184
Model mean	0.623	0.613	0.494	0.481	0.547

Using its own checklist and eligibility rule, the multi-rater analysis finds process deficits in 10/83 outcome-passing slots (12.0%); four have shortcomings reported by every rater. Only one flagged video overlaps the automatic analysis. These analyses support process checks; matched comparisons assess human–AI correspondence (Section 6, Appendix D.3). The modeling target is the ordered path to an outcome: a plausible final response should emerge from the preceding event, giving the interaction a coherent history rather than only an acceptable ending.

For a virtual character, an expression has meaning in relation to what preceded it. A fearful pose before a threat and the same pose after that threat may look similar in isolation, yet imply different interactions. Checking the initial state and event–response timing makes this distinction explicit. This suggests evaluating local visual plausibility together with the sequence of evidence that connects an event to a changed response.

Realized states are not always preserved across turns. Identity continuity exceeds boundary mental-state scores for all six generators by 14.5–47.9 points (Table 13), motivating separate checks of appearance and state. Conditioning on a full-credit first-turn result still leaves sequences without full boundary credit. This separates failure to establish the target from failure to carry an established

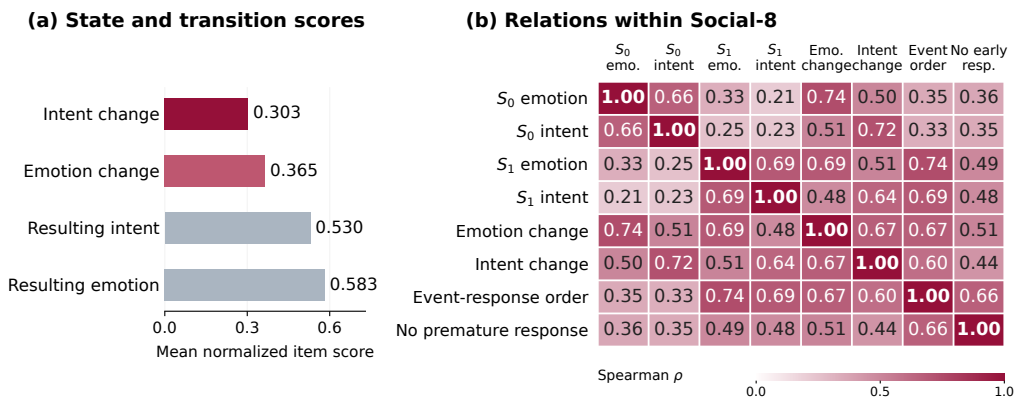


Figure 5: State updates and timing. (a) Change scores fall below resulting-state scores; applicable $N = 2,591/2,657/2,393/2,657$ top to bottom. (b) Eight-item Spearman correlations: pairwise $n = 2,174$–$2,657$; mean over 28 pairs is 0.517. Color scale: 0–1.

target across the generation boundary under the same judge. The full breakdown and the broader human boundary comparison are in Table 6.

Four of six generators score lower on second-turn content (Table 2C). Across 37 completed five-turn chains, content scores decline from 0.698 to 0.555 across successive targets, providing exploratory evidence of challenges in sustained content execution. Together, these results motivate treating the preceding outcome as an active condition for the next response. A character must retain the state established by prior interaction while remaining responsive to new events; preserving appearance alone is insufficient.

The five-turn profile in Figure 4b extends this concern to sustained interaction. The largest decrease occurs at the first continuation; later scores remain below the initial level. Because each turn requests different content and only completed chains are included, the curve describes this rollout setting rather than a controlled effect of length. Its practical implication is to assess continuation from the first handoff onward, rather than judge an interactive system only by its opening clip. A promising design direction is to make the current character state an explicit part of the conditioning context and update it with each new event.

Physical fidelity does not summarize mental-state performance. Physical–Social has the weakest pairwise correlation across 2,657 videos ($\rho = 0.17$; Figure 4a). Within-model correlations range from -0.166 to 0.078; removing each model's dimension means gives -0.050 (Table 3). The weak association thus persists beyond model differences: a high physical score does not establish successful mental-state realization. The two aspects therefore warrant separate evaluation: plausible motion and scenes do not by themselves show whether a character responds in a contextually consistent way.

Physical plausibility concerns whether motion and event consequences fit the visible world; mental-state realization concerns what the character does in response to that event. Their weak score association motivates checking character responses alongside physical improvements. This is an observation about the current evaluation system, which uses different Gemini judges for the physical and mental-state protocols, rather than evidence that the underlying abilities are independent.

Emotion and intent updates receive the lowest scores. Mental-state transitions score lowest for all seven generators: the shared-prompt mean is 0.489 versus 0.771 for Physical (Table 2A). Across all videos, intent/emotion change score 0.303/0.365, below resulting intent/emotion at 0.530/0.583 (Figure 5a). The dimension ordering persists before logical adjustment (Appendix D.2). Changes require both states and the event, so gaps locate unmet criteria, not calibrated capability differences. Separately, with full-credit states and event, 17/357 emotion and 6/255 intent cases retain incomplete process scores (Appendix D), motivating checks beyond endpoints.

The conditional analysis gives a more specific diagnosis than the marginal score gap. Most low change scores co-occur with an incomplete event or endpoint; a smaller subset remains after those

prerequisites succeed. Improving complete-transition generation therefore calls for both reliable realization of the requested event and states, and verification of the path between them. The remaining process deficits justify additional checks, without making them the dominant failure source. For development, this separation can help distinguish a missing response from a response that is present but appears at the wrong stage.

Order correlates with emotion/intent change at 0.67/0.60; initial/resulting states correlate at 0.33/0.23. No premature response correlates with order at 0.66 and state/change items at 0.35–0.51 (Figure 5b). The overlapping checks address different requirements: displaying an endpoint expression alone need not satisfy event-linked updating and timing.

Group interactions expose additional challenges. Group scenes score below individual scenes for all seven generators, whereas six score slightly higher on dyadic scenes (Figure 4c; Appendix A). The contrast makes group interaction a priority beyond single-character response generation. Such scenes require tracking who is affected by an event and coordinating distinct reactions without losing each character's ongoing state. These demands motivate role-specific state tracking and coordinated response generation as directions for future models. The present comparison identifies a scenario-level challenge rather than its cause: scenario sets differ, and five models' individual profiles include 61 rewrites.

Each response should remain attached to the appropriate role and prior state. A shared event may call for different responses from different participants. The group results motivate evaluating these requirements jointly, and developing representations that retain individual states while coordinating the interaction. These are design directions, not established failure mechanisms.

6 Human Agreement and AI Correspondence

Fifteen raters provided 455 task ratings across all three tasks, including mental-state and timing items, with generator identities hidden. Comparing normalized human means and automatic scores on the same video slots yields positive rank associations for the single-turn proxy ($\rho = 0.606$), two-turn Overall (0.844), and trigger-only score (0.529). Their respective within-tolerance agreements are 88.8%, 77.8%, and 74.2%, using an absolute normalized difference of at most 0.25.

The human reference also exhibits agreement across raters: task-level human–human tolerance agreement is 82.9%, 80.0%, and 88.9% for the three measures, respectively. Across pooled dimension means, 81.9% of rater pairs fall within the same tolerance, increasing to 83.2% after excluding the most prolific rater.

7 Limitations and Future Directions

We study specified, visibly expressed responses. The 24 two-turn cases and 37 completed five-turn chains cover differing targets and interfaces, limiting causal attribution. Broader cultural and scenario coverage, open-ended responses, longer interactions, and calibration of items with human–AI differences offer directions for extension.

8 Conclusion

MentalWM-Bench evaluates mental-state dynamics through event execution, response updating, and cross-turn preservation. Correct outcomes can conceal incomplete transitions; realized states may fail to persist; physical fidelity alone does not characterize these processes. Lower update scores and long-chain and group profiles identify further challenges. Human comparisons support shared trends while locating scoring differences. These findings motivate next-generation world models that maintain character states and update them with new events to sustain coherent interactions.

AI USE STATEMENT

Generative models produced the evaluated videos and Gemini-based evaluators produced the main automatic scores. An AI assistant supported literature lookup, methodological critique, statistical checking, translation, and manuscript revision. AI tools also assisted scenario drafting and conceptual illustrations; human reviewers checked every proposed scenario. These activities are separate from the human reference ratings. We take responsibility for the final content, including AI-assisted text and artifacts.

ETHICS STATEMENT

The paper uses generated-video examples and aggregated human judgments; individual raters are not identified. Emotion and intent labels describe expressed states in specified scenarios and are interpreted through visible evidence. These judgments depend on context and do not establish access to private mental states or agreement across cultures.

REPRODUCIBILITY STATEMENT

The code/data supplement provides prompts, archived evaluator code, scoring rules, per-video outputs, and offline result-reproduction scripts. Dataset and report links appear in Appendix C. Appendices B–C describe inputs, aggregation, human references, and retained evaluator records. Reconstructing reported scores requires the corresponding protocol and output versions; rerunning generation also depends on provider settings.

REFERENCES

Lichen Bai, Tianhao Zhang, Shitong Shao, Dingwei Tan, Qiyu Zhong, Zhengpeng Xie, Haopeng Li, Qinghao Huang, Dandan Shen, Tengjiao Ji, Wei Wang, Peicheng Wu, Yuxuan Zhao, Xiangyu Zhu, Welly Luo, Shurui Yang, and Zeke Xie. MaineCoon: Pursuing A Real-Time Audio-Visual Social World Model. arXiv:2606.17800, 2026. URL https://arxiv.org/abs/2606.17800.

Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4), 2017. doi: 10.1038/s41562-017-0064. URL https://doi.org/10.1038/s41562-017-0064.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative Interactive Environments. arXiv:2402.15391, 2024. URL https://arxiv.org/abs/2402.15391.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A Unified Evaluation Benchmark for World Generation. In *ICCV*, 2025. URL https://arxiv.org/abs/2504.00983.

Hao Fei and Yiran Zhao. Mental world modeling. *arXiv preprint arXiv:2607.27201*, 2026. URL https://arxiv.org/abs/2607.27201.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. TC-Bench: Benchmarking Temporal Compositionality in Text-to-Video and Image-to-Video Generation. arXiv:2406.08656, 2024. URL https://arxiv.org/abs/2406.08656.

David Ha and Jürgen Schmidhuber. World Models. arXiv:1803.10122, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. In *Proceedings of the*

36th International Conference on Machine Learning, volume 97, pp. 2555–2565, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Xianjing Han, Bin Zhu, Shiqi Hu, Franklin Mingzhe Li, Patrick Carrington, Roger Zimmermann, and Jingjing Chen. OSCBench: Benchmarking Object State Change in Text-to-Video Generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30867–30884, 2026. doi: 10.18653/v1/2026.acl-long.1425. URL https://aclanthology.org/2026.acl-long.1425/.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *CVPR*, 2024. URL https://arxiv.org/abs/2311.17982.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, 2023. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Agnes Moors, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, 5(2):119–124, 2013. doi: 10.1177/1754073912468165. URL https://doi.org/10.1177/1754073912468165.

Lixing Niu, Jiapeng Li, Xingping Yu, Xinyi Dong, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. Read the room: Video social reasoning with mental-physical causal chains. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TJilJnZjpw.

Wenshuo Peng, Gongxuan Wang, Tianmeng Yang, Chuanhao Li, Xiaojie Xu, Hui He, and Kaipeng Zhang. SVBench: Evaluation of Video Generation Models on Social Reasoning. In *CVPR*, 2026. URL https://arxiv.org/abs/2512.21507.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, and Matthew Botvinick. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4218–4227, 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Klaus R. Scherer. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23(7):1307–1351, 2009. doi: 10.1080/02699930902928969. URL https://doi.org/10.1080/02699930902928969.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. In *CVPR*, 2025. URL https://arxiv.org/abs/2407.14505.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. SOTOPIA-π: Interactive Learning of Socially Intelligent Language Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12912–12940, 2024. doi: 10.18653/v1/2024.acl-long.698. URL https://aclanthology.org/2024.acl-long.698/.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning Interactive Real-World Simulators. arXiv:2310.06114, 2023. URL https://arxiv.org/abs/2310.06114.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A Comprehensive Multi-turn Benchmark for Interactive Video World Model Evaluation. arXiv:2605.25874, 2026. URL https://arxiv.org/abs/2605.25874.

Haofei Yu, Yining Zhao, Guanyu Lin, and Jiaxuan You. Building Social World Models with Large Language Models. arXiv:2606.11482, 2026. URL https://arxiv.org/abs/2606.11482.

Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007. doi: 10.1037/0033-2909.133.2.273. URL https://doi.org/10.1037/0033-2909.133.2.273.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *CVPR*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zadeh_Social-IQ_A_Question_Answering_Benchmark_for_Artificial_Social_Intelligence_CVPR_2019_paper.html.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, and Maarten Sap. Social World Models. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2509.00559.

APPENDIX CONTENTS

The appendix follows the evidence behind the main claims: A details coverage and performance profiles; B specifies task and scoring protocols; C reports human correspondence; D separates outcome, process, and preservation failures; E provides the full item scores.

A Full Single-turn Results 13
B Task and Scoring Details 14
C Human References and Reproducibility 15
D Supplementary Diagnostics 15
E Item-level Results by Task 20

A FULL SINGLE-TURN RESULTS

Figure 6 gives all eight mental-state criteria. Table 3 reports participant profiles and within-generator Physical–Social correlations. The latter are computed separately within each generator's complete-transition videos. Centering both scores within generator and then recomputing Spearman gives -0.050 overall, and $-0.129/-0.044/0.020$ within individual/dyadic/group sets. Centering removes model means, not all scenario differences; ranks are computed after centering.

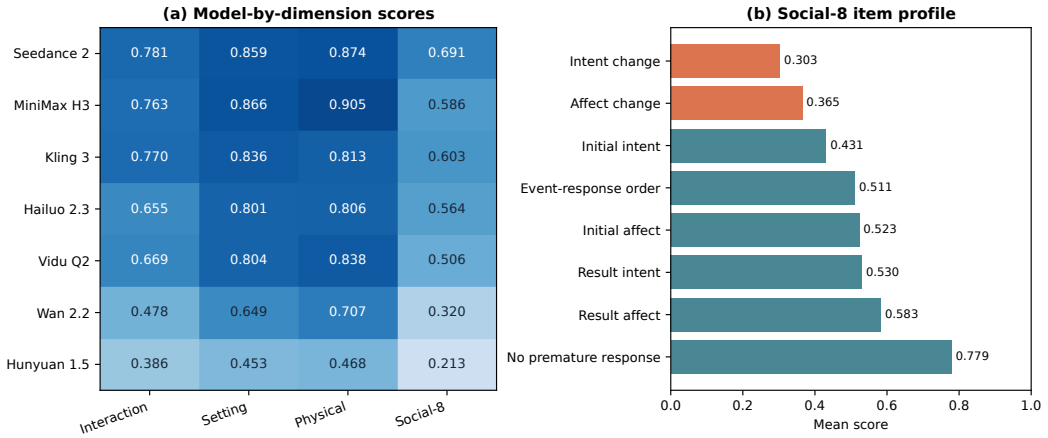


Figure 6: Full dimension and Social-8 profiles for the 2,657 complete-transition videos. Item means use applicable records; model coverage is unequal.

Table 3: Participant profiles and within-model correlations. Single/Dyadic/Group are four-dimension composites (0–1) on different scenario sets; n and $\rho_{P,S}$ describe each model's Physical–Social Spearman correlation.

Generator	Single	Dyadic	Group	n	$\rho_{P,S}$
Seedance 2	0.820	0.824	0.749	397	-0.022
MiniMax H3	0.799	0.807	0.724	397	-0.050
Kling 3	0.772	0.777	0.708	397	-0.035
Hailuo 2.3	0.709	0.714	0.695	397	-0.166
Vidu Q2	0.711	0.733	0.665	397	-0.013
Wan 2.2	0.538	0.552	0.525	336	-0.034
Hunyuan 1.5	0.397	0.380	0.363	336	0.078

Table 4: Coverage sensitivity. Shared: the same 336 base prompts per model; full: all available prompts (397 for five models, 336 for Wan/Hunyuan); rewrites: 61 additional individual-scene prompts where available. Scores are four-dimension means; intervals resample source cases 10,000 times with models paired. The rewrite subset is not a matched estimate of wording effects.

Generator	Shared (95% CI)	Full	Rewrites
Seedance 2	0.800 [0.783, 0.816]	0.801	0.810
MiniMax H3	0.775 [0.760, 0.790]	0.780	0.807
Kling 3	0.749 [0.731, 0.766]	0.755	0.791
Vidu Q2	0.703 [0.685, 0.720]	0.704	0.712
Hailuo 2.3	0.695 [0.678, 0.712]	0.707	0.771
Wan 2.2	0.538 [0.520, 0.557]	0.538	–
Hunyuan 1.5	0.380 [0.365, 0.395]	0.380	–

B Task and Scoring Details

B.1 Generation Protocols

Trigger-only. Seven generators share 50 cases (20 single, 15 dyadic, 15 group). Generation and evaluation receive S_0, A_1 only, excluding source S_1; the nominal five-second prompt ends after the event. **Complete transition.** Prompts supply S_0, A_1, S_1, with the response after the event. Five generators cover 397 prompts; Wan and Hunyuan cover 336.

Two-turn continuation. The first turn's endpoint conditions $S_1 \to A_2 \to S_2$, requesting the same character, outfit, scene and props without replaying A_1. Six generators share 24 cases, selected from 61 source cases by Seedance continuation-interface availability. Durations and interfaces are not fully matched. **Five-turn pilot.** A separate five-check protocol covers 37 completed chains: eight each for H3, Hailuo, Kling and Wan, and five for Hunyuan. Unsuccessful requests and a separate eight-case Wan batch are excluded.

B.2 Checklists and Aggregation

Table 5 summarizes the evidence checked by the evaluators. In the strict rubric, 0/1/2 mean absent/contradictory, direct but incomplete, and clear target-specific evidence. Scores are divided by two and applicable items averaged. AI intent items may be inapplicable; missing required evidence or uncertainty alone scores zero.

Table 5: Evaluator evidence requirements: eight mental-state checks and thirteen two-turn items organized by phase. Human references assess corresponding concepts.

Criterion / phase	Checks	Evidence required
Social-8: complete transition		
Initial / final emotion	2	Facial and bodily/action evidence before / after the event
Initial / final intent	2	Goal-directed action, not inferred from emotion, gaze or props
Emotion / intent change	2	A visible change after the event; an unchanged or early response fails
Event–response order	1	Initial state, event and result separately visible in this order
No premature response	1	An identifiable initial baseline, without early final-response cues
Two-turn continuation		
Turn 1	3	Opening state, trigger event, resulting state
Boundary	3	Social-state match, identity, scene/prop continuity
Turn 2	3	Event visibility, event match, resulting state
Sequence	4	Temporal order, no replay, physical coherence, overall fidelity

The strict human and AI forms match all eight mental-state criteria. WBench-derived scores are compared at dimension level because native AI scales differ from human 0/1/2 ratings. Trigger-only compares an eight-item human checklist with the three-dimension AI proxy.

Task composites follow Section 4; they average evidence-based criteria rather than count fully successful trajectories.

C Human References and Reproducibility

Fifteen raters provided 455 task ratings across all three tasks, with generator identities hidden. Comparing normalized human means with automatic scores yields agreement of 88.8%, 77.8%, and 74.2% for the single-turn proxy, two-turn Overall, and trigger-only scores, respectively, with Spearman correlations of 0.606, 0.844, and 0.529. Agreement requires an absolute normalized score difference of at most 0.25. The inter-rater report covers 243 multi-rated video slots: 58, 118, and 67 have two, three, and four raters, respectively. Across 50 items and 19,568 within-slot rater pairs, exact agreement is 64.5%, agreement within one category on the original 0/1/2 scale is 93.7%, and median ordinal Krippendorff's α is 0.253. Dimension means show 81.9% agreement within ± 0.25 over 2,593 pairs; excluding the most prolific rater gives 67.2% exact item agreement and 83.2% dimension agreement. For emotion change, intent change, and boundary continuity, exact agreement is 64.0%, 65.7%, and 63.3%, with ordinal α of 0.297, 0.299, and 0.368, respectively.

D Supplementary Diagnostics

These analyses distinguish three questions: whether a scored outcome is realized, whether its preceding process is complete, and whether it is preserved. Conditional subsets make the evidence for each claim explicit; protocol comparisons assess sensitivity to how those claims are measured.

D.1 Preservation Conditional on the First Outcome

We select sequences with `turn1_result`=2 and report the existing `boundary_state_match` judgments. Table 6 accounts for all 144 two-turn records: 85 meet the condition. The total pools eligible sequences rather than weighting models equally. The 95% interval resamples the 24 source cases 10,000 times, preserving models within each case. As a sensitivity check, requiring a first-turn score of at least 1 yields 58 full-credit boundaries among 106 sequences (54.7%). These are conditional analyses of the original automatic scores, not new judgments.

Table 6: Boundary state after a full-credit first-turn result. Columns 0/1/2 count absent/contradictory, partial, and full-credit boundary evidence. Hunyuan has no eligible sequences; Wan's single case does not support a stable rate.

Generator	Eligible	Boundary 0	Boundary 1	Boundary 2	Full credit
Hailuo 2.3	21	5	2	14	66.7%
MiniMax H3	20	5	2	13	65.0%
Seedance 2	21	6	5	10	47.6%
Kling 3	22	5	2	15	68.2%
Wan 2.2	1	1	0	0	0/1
Hunyuan 1.5	0	0	0	0	–
Total	85	22	11	52	61.2%

D.2 Sensitivity to Consistency Adjustment

The strict protocol retains both `raw_model_criteria` and adjusted `criteria`. We recompute the eight-item mean from each on the same 2,657 records with identical applicability masks. Adjustment changes 786 emotion-change, 681 intent-change, and 278 order scores; the other five mental-state items are unchanged. Table 7 separates these rule effects from the judge's initial judgments. Lower adjusted change scores are not independent evidence of a distinct capability gap.

Table 7: Sensitivity of checklist scores to consistency adjustment. Means use all applicable records; the conditional counts use the same 336 base prompts per model as the main ranking.

Statistic	Before	After
Emotion change mean	0.547	0.365
Intent change mean	0.464	0.303
Event–response order mean	0.569	0.511
Models with mental-state score lowest	7/7	7/7
Emotion-conditioned process deficits	17/357	17/357
Intent-conditioned process deficits	6/255	6/255

The unadjusted mental-state means are 0.753/0.646/0.667/0.620/0.563/0.352/0.230 for Seedance/H3/Kling/Hailuo/Vidu/Wan/Hunyuan, respectively. Each remains below the same model's Interaction, Setting, and Physical means. Thus the dimension profile and residual conditional cases persist without post-processing, but the magnitude and ordering of individual-item gaps depend on the protocol. Neither version supplies an independent human reference.

D.3 Outcome Checks versus Complete-Process Checks

We use 168 distinct single-turn videos covered by the supplementary strict human reassessment. An outcome pass requires both existing Boolean checks `event_completes` and `resulting_action_shown` to pass. The process comparison additionally checks `event_integrated_in_order` and `full_action_trajectory`. Of 168 videos, 98 fail the outcome rule, 8 pass it but fail a process check, and 62 pass both rules. All eight fail full trajectory; two also fail ordering. Thus 11.4% of outcome-passing videos (4.8% of the full sample) have additional process shortcomings, which include missing initial actions as well as premature responses. These are nested rules assembled from existing automatic judgments, not a separately run final-frame model or external baseline.

Table 8: Human reassessment of outcome-passing videos (0–2 scale). Groups are defined by automatic process checks; the human scores are from one supplementary strict reassessment per video. Differences are descriptive.

Human criterion	Process fail ($n = 8$)	Both pass ($n = 62$)	Difference
Final emotion	1.25	1.35	-0.10
Final intent	1.50	1.39	$+0.11$
Initial emotion	0.75	1.55	-0.80
Initial intent	1.00	1.53	-0.53
Emotion change	1.00	1.29	-0.29
Intent change	1.25	1.45	-0.20
Causal order	0.62	1.39	-0.76
No premature response	0.75	1.44	-0.69

Table 8 shows similar mean endpoint scores but lower initial-state and timing scores in the process-fail group. The separate strict automatic protocol also reports lower order scores (1.25 versus 1.74) and more premature-response flags (38% versus 3%, rounded). These observations support complementary coverage, not statistical independence between evaluators or an estimate of detection accuracy.

For `s0a1s1_072_kling3`, the automatic records identify a smile before the fruit is peeled; the human reassessment assigns zero to order and no-premature-response, despite final emotion/intent scores of 1/2. For `s0a1s1_refined_029_vidu-q2`, the snake and retreat occur in order, but the requested initial fist-raising and hammer-seeking actions are absent. This second case illustrates incomplete realization rather than reversed order.

Replication using original multi-rater judgments. The original export contains 142 rating slots for 133 distinct videos, with complete eight-item WBench-derived ratings from 2–4 raters (16 with

two, 62 with three, and 64 with four), yielding 586 within-slot rater pairs. Nine videos appear twice; we retain their task-specific slots. The export identifies one rater in every slot as the project's `admin` account. Excluding this account leaves 126 slots (118 distinct videos) with at least two raters. At the same 1.5/2 threshold, 86 pass outcome checks and 9 fail a process check (10.5%). Within these 86, human and AI process decisions both pass in 35 and both fail in 5; 42 human passes fail AI checks and 4 human failures pass AI. Thus excluding this account preserves the outcome–process distinction, without resolving item-level AI–human disagreement. Primary statistics below retain all available raters. Outcome passing requires mean ratings of at least 1.5/2 on both completion and resulting state; process passing additionally requires the same threshold on temporal integration and full trajectory. Of 83 outcome-passing slots, 10 fail a process check (12.0%). Table 9 reports all eight items. Four of these ten have at least one process item below full credit for every rater; the other six contain mixed judgments. Only one of the ten overlaps with the eight automatically flagged videos above. The analyses support the same distinction in different rated sets, not a matched sensitivity or precision estimate for the AI evaluator.

Table 9: Original multi-rater outcome–process analysis. Means use 0–2 ratings. Agreement uses all 142 slots and 586 rater pairs per item; exact agreement and ordinal α describe individual ratings, not reliability of the mean. Outcome-pass groups are defined by human scores.

Human item	Process fail (10)	Both pass (73)	Exact (%)	α
Completion	1.65	1.85	62.6	0.13
Resulting state	1.78	1.83	58.5	0.29
Temporal integration	1.36	1.86	65.7	0.21
Full trajectory	1.46	1.84	58.0	0.21
Event start	1.53	1.81	63.0	0.27
Initial state	1.59	1.75	63.3	0.27
Event match	1.15	1.55	49.8	0.32
Reaction to event	1.58	1.84	61.4	0.18

With the stricter rule requiring every rater to give full credit, 23 slots pass both outcome items, of which 8 fail a process item (34.8%). Equal endpoint means under this rule follow from selection and do not establish statistical equivalence. Likewise, lower process means in the primary comparison partly follow from the group definition; event matching is also lower. Temporal integration and full trajectory have within-one-category agreement of 93.5% and 94.5%, respectively, but both have ordinal $\alpha = 0.21$. These figures establish multi-rater support for outcome–process distinctions while retaining individual disagreement. This human checklist is separate from the strict eight-item mental-state reassessment above.

Matched item-level AI–human comparison. We pair all 142 slots with the original automatic check outputs for their 133 distinct videos. The eight fields correspond to event start, match, completion and temporal integration, and opening action, reaction, resulting action and full trajectory. Human means are divided by two; automatic checks are binary. Table 10 reports rank correlation, mean absolute error (MAE), and agreement after thresholding human means at 0.75. Intervals use 10,000 bootstrap resamples of distinct videos, retaining repeated slots together; they do not account for dependence across generators sharing a source scenario.

Among the 83 human outcome-passing slots, human and automatic process checks both fail in 7 cases and both pass in 37. The remaining cases comprise 3 human failures passed by AI and 36 human passes failed by AI. Thus the two references share some diagnosed process deficits but differ substantially in pass thresholds or judgments. These matched results support partial correspondence, not interchangeable process labels or a fully validated error detector. They also explain why composite agreement should be supplemented with item-level comparisons. Of the 36 human-pass/AI-fail slots, 34 also fail an automatic outcome check, including 25 failing event completion. Most disagreements therefore extend beyond process judgments. With human thresholds of 1.0, 1.5, and 2.0 on the original 0–2 scale, respectively, outcome-eligible sets contain 123, 83, and 23 slots; their human-pass/AI-fail process counts are 68, 36, and 0. The eligible sets change with the threshold, so this sensitivity analysis does not establish improved accuracy at the strict threshold: four of its eight human process failures still pass AI checks. We retain 1.5 as the primary rule.

Table 10: Matched original-checklist AI–human comparison (142 slots, 133 videos). Human references average 2–4 raters. Decision agreement uses human mean ≥ 0.75 and automatic `yes`; MAE uses normalized scores. These are the original event/action checks, not the supplementary strict mental-state checklist.

Item	ρ	MAE	Agreement (%)	95% interval
Completion	0.441	0.408	64.8	[56.6, 72.9]
Resulting state	0.447	0.280	70.4	[62.5, 77.9]
Temporal integration	0.448	0.426	63.4	[55.1, 71.6]
Full trajectory	0.454	0.447	60.6	[52.1, 68.8]
Event start	0.367	0.219	76.1	[68.7, 82.9]
Initial state	0.424	0.245	76.1	[68.7, 82.9]
Event match	0.508	0.358	70.4	[62.7, 77.6]
Reaction to event	0.327	0.272	75.4	[67.8, 82.6]

D.4 Process Checks Conditional on Endpoints and Event

The common-prompt analysis uses 2,352 videos (336 prompts per model). Eligibility requires full credit (2/2) for the applicable initial and resulting emotion (or intent), trigger visibility, and exact trigger match. We then inspect change, event–response order, and premature response separately; inapplicable intent-change checks are excluded. Among 357 emotion-eligible videos, 13 lack full change credit, 4 lack full order credit, and 10 lack full premature-response credit; their union is 17 videos (4.8%). Among 255 intent-eligible videos, 238 have an applicable intent change: 2 lack full change credit, 1 lacks full order credit, and 5 lack full premature-response credit; the union is 6 videos (2.4%). These subsets can overlap. Case-cluster bootstrap intervals (10,000 draws from 336 source cases, retaining all models within a case) are 1.9–8.2% for the emotion residual rate and 0.4–5.4% for intent.

Using all available prompts gives 17/434 (3.9%) emotion cases and 6/292 (2.1%) intent cases. The small residuals qualify the marginal score gaps: most low change scores co-occur with incomplete endpoint or event realization. Residual cases establish criterion disagreement under the same judge, not human-confirmed false negatives of a separate baseline. Per-video membership and aggregation rules are retained with the analysis script.

Five-turn extension. The five equal-generator content means are 0.698, 0.583, 0.589, 0.568, and 0.555 across 37 completed chains. This descriptive extension retains its own protocol and is secondary to the conditional two-turn analysis.

D.5 TWO-TURN PROFILES

Figure 7 reports the full continuation profile and case-level uncertainty. The turns have different targets; comparisons are descriptive.

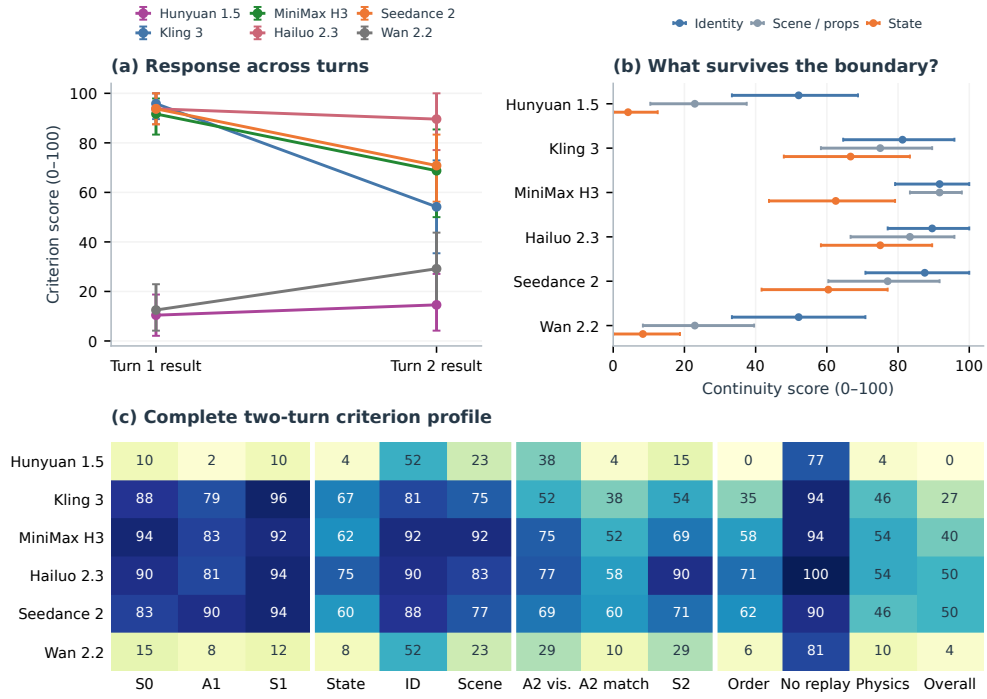


Figure 7: Two-turn results on 24 shared cases per generator. Top: response and boundary scores. Bottom: all thirteen criteria (0–100). Error bars are 95% intervals from 5,000 case-bootstrap resamples preserving paired turns. ID denotes identity; State denotes boundary mental-state match.

Alternative judge protocol. A separate 21-video pilot (three cases, seven generators) first records video observations without targets, then scores those observations against targets without viewing the video. On unchanged videos, the mean proxy changes from 0.7296 to 0.7563 (case-averaged changes: 0.0167, 0.0135, 0.0500). This measures protocol sensitivity; no independent accuracy reference is used.

E ITEM-LEVEL RESULTS BY TASK

Tables 11–13 report all available videos, including rewrites, with 24 complete-transition, 16 trigger-only and 13 continuation checks, and items as rows and models as columns. Overall retains each task's composite; it does not average all displayed rows. Final-column means weight models equally, whereas Figure 5a averages applicable videos. Scores retain three-decimal precision.

Table 11: Complete-transition item scores (0–1). Models are ordered by Overall. Bold: best model per item. The final column weights models equally.

Item	Seedance 2	MiniMax H3	Kling 3	Hailuo 2.3	Vidu Q2	Wan 2.2	Hunyuan 1.5	Model mean
Interaction								
Participants visible	0.985	0.995	**0.997**	0.990	0.995	0.949	0.940	0.979
Event begins	0.904	**0.940**	0.917	0.781	0.793	0.568	0.479	0.769
Event matches	**0.617**	0.605	0.589	0.436	0.484	0.214	0.122	0.438
Event completes	**0.620**	0.607	0.587	0.441	0.486	0.211	0.125	0.440
Event in order	**0.599**	0.572	0.564	0.411	0.448	0.167	0.104	0.409
Subject visible	0.997	**1.000**	0.997	**1.000**	**1.000**	0.991	0.988	0.996
Opening action	0.811	0.783	**0.861**	0.763	0.736	0.565	0.488	0.715
Reaction shown	**0.929**	0.874	0.897	0.753	0.791	0.536	0.330	0.730
Resulting action	**0.806**	0.763	0.796	0.650	0.635	0.446	0.238	0.619
Full trajectory	**0.544**	0.489	0.494	0.327	0.320	0.128	0.045	0.335
Setting								
Scene maintenance	0.954	**0.982**	0.933	0.949	0.947	0.836	0.501	0.872
Required elements	0.935	**0.977**	0.932	0.947	0.965	0.792	0.708	0.894
Appearance maintenance	0.975	**0.987**	0.930	0.937	0.950	0.825	0.532	0.877
Requested action	**0.574**	0.516	0.549	0.373	0.355	0.143	0.071	0.369
Physical								
Causal physics	0.798	**0.856**	0.733	0.724	0.756	0.638	0.404	0.701
Human physics	0.950	**0.956**	0.893	0.887	0.920	0.776	0.533	0.845
Social								
S_0 emotion	**0.673**	0.540	0.554	0.577	0.529	0.390	0.353	0.516
S_0 intent	**0.543**	0.489	0.504	0.476	0.431	0.283	0.244	0.424
S_1 emotion	**0.798**	0.634	0.681	0.647	0.586	0.409	0.246	0.572
S_1 intent	**0.703**	0.656	0.665	0.582	0.531	0.340	0.162	0.520
Emotion shift	**0.586**	0.406	0.437	0.426	0.354	0.183	0.098	0.356
Intent shift	**0.457**	0.381	0.403	0.349	0.280	0.139	0.047	0.294
Causal order	**0.764**	0.637	0.644	0.563	0.495	0.247	0.131	0.497
No leakage	**0.953**	0.917	0.899	0.845	0.809	0.539	0.394	0.765
Overall	**0.801**	0.780	0.755	0.707	0.704	0.538	0.380	0.667

Table 12: Trigger-only item scores (0–1). Models are ordered by Overall. Bold: best model per item. The final column weights models equally.

Item	MiniMax H3	Seedance 2	Hailuo 2.3	Vidu Q2	Kling 3	Hunyuan 1.5	Wan 2.2	Model mean
Interaction								
Participants visible	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.980	0.960	0.991
A_1 begins	0.780	**0.880**	**0.880**	0.820	0.800	0.520	0.700	0.769
A_1 matches	**0.620**	0.600	0.560	0.600	0.460	0.220	0.200	0.466
A_1 completes	0.640	**0.660**	0.580	0.600	0.480	0.240	0.320	0.503
$S_0 \rightarrow A_1$ order	0.580	**0.600**	0.500	0.520	0.420	0.220	0.240	0.440
Subject visible	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.000
S_0 behavior	**0.880**	0.820	0.800	0.760	0.840	0.520	0.600	0.746
People visible	0.940	**1.000**	**1.000**	0.960	0.980	0.900	0.940	0.960
Interaction visible	0.620	**0.660**	0.600	0.620	0.540	0.240	0.280	0.509
Two-beat trajectory	0.500	**0.520**	0.440	0.440	0.360	0.160	0.160	0.369
Setting								
Scene maintenance	0.980	0.912	**0.992**	0.940	0.924	0.884	0.812	0.921
Required elements	0.900	0.960	**0.980**	0.940	**0.980**	0.680	0.800	0.891
Appearance maintenance	0.948	**0.952**	0.948	0.916	0.944	0.820	0.720	0.893
Requested action	**0.520**	**0.520**	0.420	0.480	0.380	0.160	0.160	0.377
Physical								
Causal physics	**0.920**	0.813	0.820	0.773	0.807	0.767	0.633	0.790
Human physics	0.960	0.940	0.933	0.900	**0.967**	0.860	0.807	0.910
Overall	**0.844**	0.829	0.816	0.796	0.794	0.650	0.628	0.765

Table 13: Two-turn continuation item scores (0–1). T1/T2 denote turns. Models are ordered by Overall. Bold: best model per item. The final column weights models equally.

Item	Hailuo 2.3	MiniMax H3	Seedance 2	Kling 3	Wan 2.2	Hunyuan 1.5	Model mean
Interaction							
T1 event	0.812	0.833	**0.896**	0.792	0.083	0.021	0.573
T2 event visible	**0.771**	0.750	0.688	0.521	0.292	0.375	0.566
T2 event match	0.583	0.521	**0.604**	0.375	0.104	0.042	0.372
Setting							
Identity continuity	0.896	**0.917**	0.875	0.812	0.521	0.521	0.757
Scene / props	0.833	**0.917**	0.771	0.750	0.229	0.229	0.622
Physical							
Coherence	**0.542**	**0.542**	0.458	0.458	0.104	0.042	0.358
Social							
T1 initial state	0.896	**0.938**	0.833	0.875	0.146	0.104	0.632
T1 result state	0.938	0.917	0.938	**0.958**	0.125	0.104	0.663
Boundary state	**0.750**	0.625	0.604	0.667	0.083	0.042	0.462
T2 result state	**0.896**	0.688	0.708	0.542	0.292	0.146	0.545
Sequence							
Temporal order	**0.708**	0.583	0.625	0.354	0.062	0.000	0.389
No replay	**1.000**	0.938	0.896	0.938	0.812	0.771	0.892
Overall fidelity	**0.500**	0.396	**0.500**	0.271	0.042	0.000	0.285
Overall	**0.779**	0.736	0.723	0.639	0.223	0.184	0.547