From Words to Worlds: Beyond Current Evaluation Focus — MLLMs Fail at Higher-Order Theory of Mind

Anonymous ACL submission

Abstract

Theory of Mind (ToM), the ability to infer others' mental states from observable behavior, is fundamental to developing machines with human-level social intelligence. Despite the extensive evaluation dimensions for MLLMs, their ToM capabilities have not received sufficient attention. A comprehensive survey of ToM benchmarks reveals that while text-based ToM benchmarks are abundant, multimodal ones remain scarce. Moreover, existing multimodal ToM benchmarks fail to incorporate classic ToM task paradigms from cognitive science, higher-order ToM, or visual observations better aligned with the real world. Compared to the ambiguous expressions often present in text-based ToM benchmarks, visual information is intuitive, unambiguous, and better aligned with embodied applications. In this paper, we present From Words to Worlds (FWW), a novel pipeline that enables the generation of videos from multi-scene script templates, constructing a video-based ToM benchmark named FWW-V. We further propose a rubric-based evaluation framework, FWW-EVAL, to systematically assess 14 advanced MLLMs on FWW-V. Our results reveal that current models perform poorly on ToM tasks, with a striking disparity relative to human performance. Notably, for certain question types, the majority of MLLMs fail entirely, achieving 0% accuracy.

1 Introduction

As presented in mainstream Multimodal Large Language Models (MLLMs) technical reports such as Qwen3-VL (Team, 2025b), GLM-4.5V (Hong et al., 2025b), and InternVL3.5 (Wang et al., 2025b), current research focuses on evaluating and comparing model capabilities across a diverse range of tasks, such as STEM problem solving (Lu et al., 2024; Wang et al., 2024; Zhang et al., 2024), video understanding (Hu et al., 2025; Zhao et al., 2025), and GUI-based agents (Xie et al., 2024; Rawles et al., 2025). Video, in particular, reflects the dynamic nature of the real world, involving complex interactions between entities over time. Within the task dimension of video understanding, numerous benchmarks have recently emerged, such as long video understanding (Zhou et al., 2025; Geng et al., 2025; Wang et al., 2025a), fine-grained video motion understanding (Hong et al., 2025a), and road event scenario-specific video understanding (Parikh et al., 2025). Despite the extensive evaluation dimensions for MLLMs, their Theory of Mind (ToM) capabilities have still not received sufficient attention. As a core cognitive ability that enables inferring people's mental states based on their behavior, ToM in MLLMs is essential for achieving rich and meaningful human-machine interaction and represents a critical step toward artificial general intelligence.

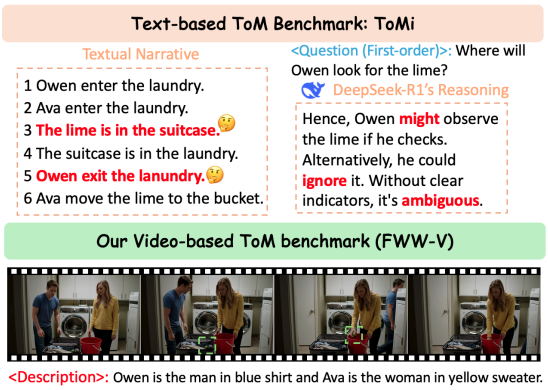

Figure 1: The ambiguities in textual narrative descriptions in ToMi and the comparison with our video-based ToM benchmark FWW-V, as well as the reasoning trajectory when evaluating DeepSeek-R1's ToM capabilities.

Numerous text-based ToM benchmarks have been proposed to evaluate the ToM capabilities of LLMs, such as ToMi (Le et al., 2019), HiToM (Wu et al., 2023), and OpenToM (Xu et al., 2024), which are constructed based on the classic "Sally-Anne" test (Baron-Cohen et al., 1985) from cognitive sci-

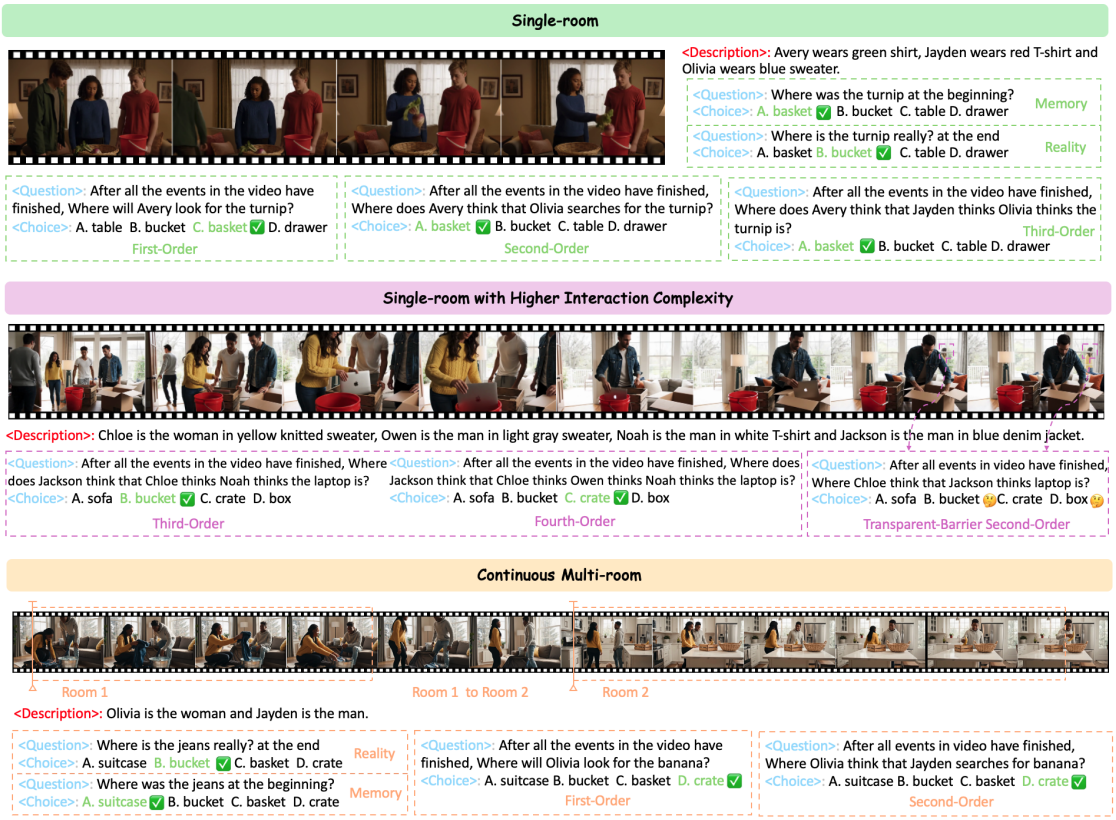

Figure 2: The three video categories in FWW-V, namely single-room, single-room with higher interaction complexity, and continuous multi-room, along with their corresponding QA pairs.

ence or templated variants of it. However, compared to the intuitive nature of visual information, textual narrative descriptions of social events are prone to ambiguity. For instance, in the third line of the textual narrative shown in Figure 1, only the fact that "the lime is in the suitcase" is described, without specifying whether the characters are aware of the lime's location. This introduces ambiguity in answering questions and affects the evaluation of LLMs' ToM capabilities, as demonstrated by the reasoning trajectory of DeepSeek-R1 (Guo et al., 2025) presented in Figure 1. In contrast to the abundance of text-based ToM benchmarks, multimodal ToM benchmarks remain scarce. Moreover, existing multimodal ToM benchmarks construct video data from simulated environments (MMToM-QA (Jin et al., 2024), MuMA-ToM (Shi et al., 2025)), Minecraft game worlds (SoMi-ToM (Fan et al., 2025)), or extract video clips from Ego4D (Grauman et al., 2025) with limited camera perspectives (EgoToM (Li et al., 2025)), failing to incorporate classic ToM task paradigms from cognitive science, higher-order ToM, or visual observations better aligned with the real world.

To this end, we introduce From Words to Worlds (FWW), a carefully designed and novel pipeline that enables the generation of videos from structured multi-scene script templates, constructing a video-based ToM benchmark named FWW-V. Specifically, we first construct multi-scene script templates based on the narratives from ToMi, a classic text-based ToM benchmark containing narratives with true belief and false belief. Furthermore, we construct additional script templates with a larger number of scenes featuring higher interaction complexity, as well as continuous multi-room. We leverage the powerful Google Veo 3 (Google DeepMind, 2025) series models to generate videos based on these structured multi-scene script templates. In Figure 2, we present diverse videos generated from our constructed scripts, along with their corresponding questions. The top row shows videos generated from scripts based on ToMi narratives, while the middle row depicts videos generated from scripts containing a larger number of scenes with higher interaction complexity. Notably, as illustrated in the 7th and 8th frames, although the female character has left the room, she can still observe the male character's actions through a transparent glass window. The bottom row shows

videos generated from scripts involving continuous multi-room settings, where the female and male characters first interact in the living room and subsequently move to the kitchen, illustrating more naturalistic, everyday social scenarios that better reflect real-world dynamics.

To prevent models from arriving at correct answers through lucky guessing rather than establishing proper analysis processes, and to mitigate the influence of visual spurious correlations (Shapira et al., 2024; Ullman, 2023), commonsense spurious correlations, and priors from false-belief structure, we propose FWW-EVAL, a rubric-based (Gunjal et al., 2025) evaluation framework. FWW-EVAL employs multi-dimensional criteria spanning the perception level, cognition level, and conclusion level to comprehensively evaluate the correctness of models' analysis processes. We then apply FWW-EVAL to evaluate the analysis processes of multiple models by leveraging both humans and GPT-5 as judges. We find GPT-5 can serve as a proxy to human judgments on FWW-EVAL.

Using FWW-V and FWW-EVAL, we evaluate 14 state-of-the-art MLLMs, including GPT-5 (OpenAI, 2025b), Gemini-2.5-Pro (Team, 2025a), and Claude-Opus-4-1 (Anthropic, 2025). The evaluation results reveal: (i) Even powerful models such as GPT-5 fail to achieve satisfactory performance on higher-order false belief ToM tasks, with most MLLMs either completely failing or achieving only marginal performance on such tasks. Across all evaluated models, we observe a **striking disparity relative to human performance**. (ii) Unlike existing mainstream MLLM benchmarks, where open-source models demonstrate performance comparable to their closed-source counterparts, a substantial performance gap exists between open-source and closed-source models on ToM tasks. (iii) As video length increases from single-room to continuous multi-room scenarios, model accuracy on questions of equivalent difficulty exhibits an overall declining trend. (iv) Transparent-Barrier second-order tasks (perceptually accessible yet physically separated scenarios) proved challenging for all models, with none of the models' reasoning processes acknowledging that visual accessibility induces changes in cognitive states. Our contributions include: (1) We establish a **generative bridge from text-based ToM (words) to video-based ToM (worlds)**. (2) We construct FWW-V, which aligns with classic ToM task paradigms from cognitive science, while incorporating settings with **higher interaction complexity and continuous multi-room**. (3) We design FWW-EVAL, a **rubric-based evaluation** framework that eliminates confounding factors, including prior knowledge, spurious correlations, and lucky guessing, enabling accurate evaluation of MLLMs' ToM capabilities. (4) Our experimental results reveal the limited performance of state-of-the-art MLLMs, highlighting substantial room for improvement in ToM capabilities and opening promising directions for future research.

2 From Words to Worlds (FWW)

In this section, we present the pipeline for constructing FWW-V, a video-based ToM benchmark, its overall statistics, as well as the rubric-based evaluation framework FWW-EVAL, which is used to evaluate the performance of MLLMs on FWW-V.

2.1 FWW-V Construction

As illustrated in Figure 3, the construction of FWW-V primarily involves four steps:

1. **Forming the Multi-Scene Script Template Pool.** We first construct multi-scene script templates grounded in the narratives from ToMi, a text-based ToM benchmark encompassing true-belief and false-belief scenarios. Furthermore, we manually design supplementary script templates featuring a larger number of scenes with higher interaction complexity, as well as continuous multi-room settings.

2. **Template Selection and Instantiation.** We select a multi-scene script template from the pool and instantiate it by populating the placeholders `[Character]`, `[Location]`, `[Object]`, and `[Container]`. To this end, we pre-construct candidate sets for each placeholder type and impose commonsense constraints to ensure physical plausibility. For example, a large object cannot be placed into a small container. After these placeholders are filled in, the corresponding questions and answer options are generated accordingly.

3. **Prompt Design and Video Generation.** For different templates, we carefully design tailored prompts to guide the video generation process (detailed prompts can be found in the Appendix F). We then leverage the powerful Google Veo 3 series models to generate videos

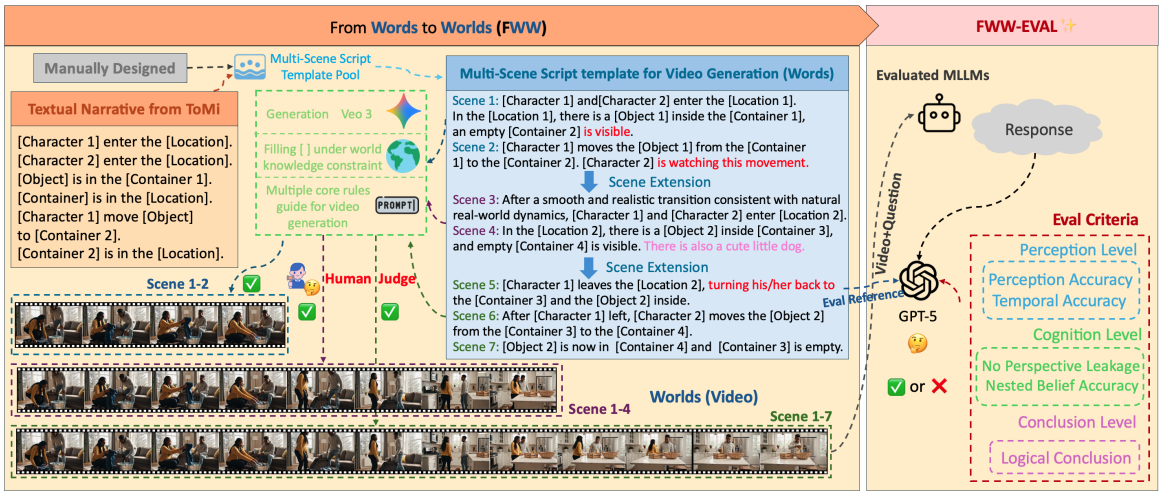

Figure 3: An overview of our FWW pipeline that generates videos from multi-scene script templates and the FWW-EVAL framework for evaluating MLLMs' responses.

based on these tailored prompts and instantiated scripts.

4. **Manual Quality Assurance and Curation.** We meticulously review and filter the generated videos to ensure a curated set that is unambiguous, logically aligned with the corresponding multi-scene script, and visually faithful and natural. Specifically, two authors independently review each video and reach consensus on its inclusion, ensuring the high quality of the final curated set.

2.2 FWW-V Statistics

Overall, we curate a total of 85 high-quality videos, along with their corresponding 704 questions. The duration, number of scenes, specific question types, and quantities for each video category (single-room, single-room with higher interaction complexity, and continuous multi-room) are illustrated in Figure 4. Higher interaction complexity features more frequent object movements, multiple characters leaving the room successively (key action for inducing false beliefs), and transparent-barrier scenarios (perceptually accessible yet physically separated). Each video scene corresponds to multiple social events described in the textual narratives, thereby making every video rich and information-dense. We describe the Data Release and Licensing information in Appendix A. **All types of videos are provided in our supplementary materials.**

2.3 FWW-EVAL

To ensure accurate evaluation of MLLMs' ToM capabilities, we design a rubric-based evaluation framework, FWW-EVAL, which synergistically examines MLLMs' question-analysis process across three levels: perception, cognition, and conclusion. For each question q_i, we provide the corresponding multi-scene script S_i as the reference for evaluation. Formally, the evaluation function is defined as $\mathcal{E}(q_i, r_i, S_i) \rightarrow \{c_1^P, c_2^P, c_1^C, c_2^C, c^L\}$, where r_i denotes the model's response and each $c \in \{0, 1\}$ indicates whether the corresponding criterion is satisfied.

1. **Perception Level $\mathcal{L}_P$.** This level comprises two criteria: *Perception Accuracy* (c_1^P) evaluates whether the model correctly perceives the objects and containers in the video, while *Temporal Accuracy* (c_2^P) evaluates whether it accurately understands the temporal ordering of scenes.

2. **Cognition Level $\mathcal{L}_C$.** This level has two criteria: *No Perspective Leakage* (c_1^C) evaluates whether the model can distinguish between true world states and characters' belief states, while *Nested Belief Accuracy* (c_2^C) evaluates whether it performs genuine inference

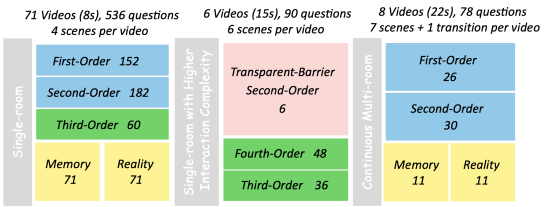

Figure 4: Statistical Overview of our FWW-V benchmark. Rectangular blocks show question types with their corresponding counts.

Benchmark	Modality	Data Description	Size	Highest Order	Answer Type	Evaluator
Text-only Benchmarks						
ToMi	T	Sally-Anne Test	100 narratives	2	Multiple Choice	Accuracy
HiToM	T	Sally-Anne Test	600 narratives	4	Multiple Choice	Accuracy
ToMBench	T	Multi-type mental state inference in diverse real-world scenarios	2,860 test samples	2	Multiple Choice	Accuracy
ToMATO	T	Mental state inference during conversations between agents	753 conversations	2	Multiple Choice	Accuracy
SimpleToM	T	Mental state inference, action prediction, and judgement	1,147 narratives	1	Multiple Choice	Accuracy
Multimodal Benchmarks						
MMToM-QA	T & V	Household activities **simulator**	134 videos	1	Multiple Choice	Accuracy
MuMA-ToM	T & V	Household activities **simulator**	225 examples	2	Multiple Choice	Accuracy
SoMi-ToM	T & I & V	**Minecraft** game	35 videos / 363 images	1	Multiple Choice	Accuracy
FWW-V (Ours)	T & V	**Bridge Text (Words) to Video (Worlds)**	85 videos / 704 questions	**4**	Multiple Choice	**Rubric-based**

Table 1: Comparison of ToM Benchmarks. Highest Order refers to the maximum depth of recursive mental state inference (e.g., 2nd-order: "A thinks that B thinks..."). T, I, and V denote Text, Image, and Video modalities, respectively. Due to space constraints, we have not listed all benchmarks. According to our survey, the ratio of existing text-only benchmarks to multimodal benchmarks is approximately $(4 \sim 5) : 1$.

when processing complex interpersonal inference chains such as "Where does A think B thinks...?"

3. **Conclusion Level $\mathcal{L}_L$.** This level comprises the *Logical Conclusion* criterion (c^L), which evaluates whether the final conclusion in the model's analysis process is the unavoidable logical result of the preceding steps.

As we employ the multi-scene script S_i corresponding to each problem as the evaluation reference, perfect correspondence between video frames and script descriptions cannot be guaranteed during the script-to-video generation process, potentially introducing misjudgments at the perception level. Notably, we observe that the conclusion level can effectively capture instances where significant perceptual errors impact question answering. Furthermore, when the cognition level criteria fail to detect priors of false-belief structure, the conclusion level provides complementary detection. A model's analysis is classified as erroneous if it violates criteria at either the cognition or conclusion level, with perception level criteria serving an auxiliary role. This multi-level synergistic coverage achieves human expert-level evaluation accuracy (see Section 3.3). Supplementary analysis, cases for multi-level synergistic coverage, and the specific prompt content used in the judge model's analysis process are in Appendix E.2 to E.4.

3 Experiments

3.1 Settings

We evaluate 14 state-of-the-art MLLMs from various model families, covering different model sizes and training methodologies. For open-source models, we include GLM-4.5V (Hong et al., 2025b), InternVL3.5-241B-A28B (Wang et al., 2025b), Qwen3-VL-235B-A22B-Instruct (Team, 2025b), and Qwen3-VL-235B-A22B-Thinking (Team, 2025b); For closed-source models, we include GPT-4.1-mini-20250414 (OpenAI, 2025a), Claude-Opus-4-1-20250805 (Anthropic, 2025), Gemini-2.5-Pro (Team, 2025a), and GPT-5 (OpenAI, 2025b). The complete list of evaluated models can be found in Table 2. All models are evaluated based on their official implementations or publicly available APIs, with all evaluations conducted in a zero-shot manner. More details about the evaluation are provided in the Appendix D.

To establish a human performance baseline, we recruit graduate-level participants with no specialized training in ToM research via academic group chats and online forums. Each participant completes the same set of 106 questions randomly sampled from FWW-V, spanning diverse question types. We report the average accuracy across participants for each question type as the human baseline.

3.2 Evaluation Results on FWW-V

The overall evaluation results for all evaluated MLLMs in the FWW-V benchmark are shown in Table 2, 3 and 4. All evaluated models fall significantly short of human performance. In Appendix C, we provide two additional experiments: (1) Does a **newer model version perform better** on our benchmark (GPT-5.2 vs. GPT-5)? (2) Is model performance **robust under a new video data variant**: scenarios where a character is present in the room but facing away (with no line of sight)?

Models	0-Order (Basic)		1st-Order (F-)		2nd-Order (S-)		3rd-Order (T-)	
	Mem	Rel	F-O	F-N	S-O	S-N	T-O	T-N
Open-source MLLMs								
Qwen2.5-VL-32B-Instruct	87.72	66.04	52.75	6.45	17.19	2.86	7.69	0.00
Qwen2.5-VL-72B-Instruct	87.32	85.92	73.64	9.76	38.02	17.65	5.56	0.00
Qwen3-VL-32B-Instruct	97.10	91.43	88.07	19.51	39.17	17.65	27.78	22.22
Qwen3-VL-32B-Thinking	97.18	97.18	85.45	12.20	28.63	27.45	5.56	0.00
GLM-4.1V-Thinking-Flash	96.83	79.37	53.54	13.89	24.17	28.89	6.25	0.00
GLM-4.5V	95.77	95.77	88.64	6.25	25.00	35.00	25.00	0.00
InternVL3.5-241B-A28B	88.10	83.33	80.30	8.33	28.75	13.33	33.34	16.67
Qwen3-VL-235B-A22B-Instruct	83.10	84.51	84.69	53.66	41.57	29.41	0.00	0.00
Qwen3-VL-235B-A22B-Thinking	98.59	91.55	90.09	12.20	36.39	37.25	38.89	44.44
Closed-source MLLMs								
GPT-4o-2024-11-20	73.24	61.97	55.86	51.22	28.85	19.61	5.56	0.00
GPT-4.1-mini-2025-04-14	90.14	66.20	50.00	23.81	31.73	19.61	22.22	0.00
Claude-Opus-4-1-20250805	94.37	83.10	78.38	58.54	59.87	54.90	44.44	33.33
Gemini-2.5-Pro	90.14	88.73	92.79	78.05	78.24	70.59	61.12	33.33
GPT-5	98.59	95.77	92.79	90.00	90.13	72.92	77.78	55.56
Human Performance	99.67		96.13				95.33	

Table 2: Evaluation results of MLLMs on the single-room video category. "O" denotes the omniscient perspective, where the reasoning agent (e.g., Bob) has full knowledge of all events, while "N" denotes the non-omniscient perspective, where reasoning is constrained by the agent's partial observability and thus involves false belief.

Performance of Closed-source Models. GPT-5 achieves the best performance among its peers on FWW-V with a significant performance advantage. It attains overall performance exceeding 90% on Memory, Reality, and first-order ToM tasks. However, its performance declines substantially on higher-order false belief (-N) ToM tasks, achieving only slightly above 70% and 55.56% on second-order false belief and third-order false belief, respectively. Under higher interaction complexity settings, it obtains merely 41.67% and 0.00% performance on higher-order questions and second-order questions in transparent-barrier scenarios, respectively. Other closed-source models demonstrate considerably worse performance across all tasks.

Performance of Open-source Models. Overall, open-source models exhibit satisfactory performance on memory, reality, and first-order omniscient perspective settings, yet demonstrate catastrophic failures across other task dimensions. On false belief (-N) ToM tasks, these models achieve only marginal performance or fail entirely. All open-source models score below 40.00% on second-order false belief ToM tasks, while 66.67% completely fail on third-order tasks (0.00% accuracy). Under higher interaction complexity settings, only Qwen3-VL-235B-A22B-Thinking completes a small subset of tasks; nearly all other open-source models fail almost completely.

3.3 Effectiveness Evaluation of FWW-EVAL

Validation of Evaluation Model Reliability. We manually examine the reasoning trajectories of all evaluated models and compute agreement with GPT-5's judgments. GPT-5 achieves a misjudg-

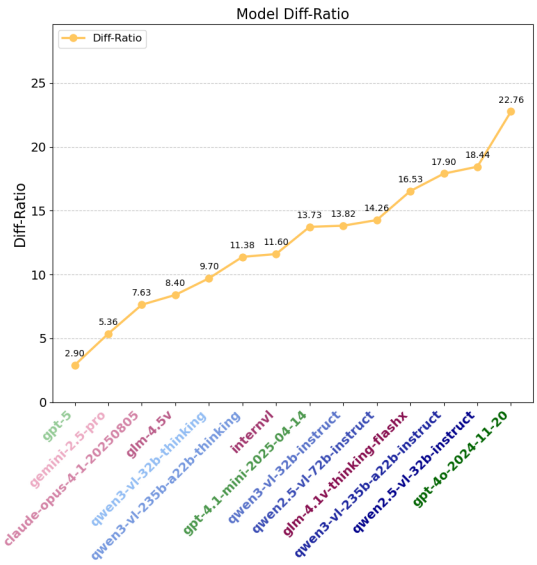

Figure 5: Diff-Ratio across different models.

Models	0-Order (Basic)			1st-Order (F-)			2nd-Order (S-)		
	Mem	Rel	Δ	F-O	F-N	Δ	S-O	S-N	Δ
Open-source MLLMs									
Qwen2.5-VL-32B-Instruct	90.91	81.82	9.48	38.89	0.00	-10.16	20.54	0.00	0.24
Qwen2.5-VL-72B-Instruct	100.00	72.73	-0.25	83.33	25.00	12.46	50.00	12.50	3.41
Qwen3-VL-32B-Instruct	100.00	81.82	-3.36	77.78	25.00	-2.40	37.50	12.50	-3.41
Qwen3-VL-32B-Thinking	100.00	81.82	-6.27	77.78	0.00	-9.94	33.93	75.00	26.43
GLM-4.1V-Thinking-Flash	100.00	72.73	-1.73	27.78	0.00	-19.83	31.25	12.50	-4.66
GLM-4.5V	100.00	90.91	-0.31	83.33	25.00	6.72	25.00	12.50	-11.25
Qwen3-VL-235B-A22B-Instruct	81.82	63.64	-11.08	72.22	25.00	-20.56	19.64	12.50	-19.42
Qwen3-VL-235B-A22B-Thinking	100.00	81.82	-4.16	83.33	37.50	9.27	29.46	62.50	9.16
Closed-source MLLMs									
GPT-4o-2024-11-20	90.91	45.45	0.58	55.56	37.50	-7.01	16.96	0.00	-15.75
GPT-4.1-mini-2025-04-14	81.82	63.64	-5.44	55.56	50.00	15.88	20.54	12.50	-9.15
Claude-Opus-4-1-20250805	90.91	54.55	-16.01	66.67	87.50	8.63	45.54	50.00	-9.62
Gemini-2.5-Pro	90.91	63.64	-12.16	77.78	87.50	-2.78	69.64	62.50	-8.34
GPT-5	90.91	81.82	-10.82	100.00	87.50	2.35	93.75	75.00	2.85

Table 3: Evaluation results of MLLMs on the continuous multi-room video category. Δ represents the performance difference from the single-room video category for the same question type.

ment rate below **2%** across all evaluated models. Specifically, for the best-performing model (GPT-5) and the worst-performing one (Qwen2.5-VL-32B-Instruct), GPT-5 achieves misjudgment rates of **1.12%** and **0.93%**, respectively. These results demonstrate strong alignment between human and automated evaluations. Furthermore, the evaluation model exhibits robust generalizability: it successfully identifies cases where answers are correct but reasoning is flawed in poorly-performing models, while avoiding false negatives in well-performing models. We provide further analysis of self-judging and misjudgment sources in Appendix E.1.

Diff-Ratio: Measuring Reasoning-Answer Alignment. To quantify the discrepancy between answer accuracy and reasoning integrity, we introduce Diff-Ratio, a metric representing the proportion of samples where the correctness of the answer and the reasoning are inconsistent. Formally, it is defined as:

$$\text{Diff-Ratio} = \frac{\sum_{i=1}^{N} \mathbb{1}[\hat{a}_i \neq \hat{r}_i]}{N}, \qquad (1)$$

where N is the total sample size, $\hat{a}_i, \hat{r}_i \in \{0, 1\}$ denote the correctness of the i-th answer and its corresponding reasoning process, respectively, and $\mathbb{1}[\cdot]$ is the indicator function. As illustrated in Figure 5 and Table 2, Diff-Ratio is inversely correlated with model capability; specifically, weaker models exhibit a significantly higher prevalence of such inconsistencies. This suggests that less capable models often arrive at correct answers through flawed or incoherent logic, failing to demonstrate genuine task mastery, as illustrated in Figure 7. This phenomenon underscores the **value of FWW-EVAL, which exposes "superficial correctness"** that conventional answer-only metrics would fail to detect.

Models	Higher-Order (H-)		Transparent-Barrier 2nd-Order (Bonus)
	H(3,4)-O	H(3,4)-N	
Open-source MLLMs			
Qwen2.5-VL-32B-Instruct	0.00	0.00	0.00
Qwen2.5-VL-72B-Instruct	0.00	0.00	0.00
Qwen3-VL-32B-Instruct	10.00	8.33	0.00
Qwen3-VL-32B-Thinking	3.34	0.00	0.00
GLM-4.1V-Thinking-Flash	1.67	0.00	0.00
GLM-4.5V	3.34	0.00	0.00
Qwen3-VL-235B-A22B-Instruct	6.67	0.00	0.00
Qwen3-VL-235B-A22B-Thinking	31.67	29.17	0.00
Closed-source MLLMs			
GPT-4o-2024-11-20	1.67	0.00	0.00
GPT-4.1-mini-2025-04-14	15.00	4.17	0.00
Claude-Opus-4-1-20250805	26.67	20.83	0.00
Gemini-2.5-Pro	48.33	29.17	0.00
GPT-5	51.67	41.67	0.00
Human Performance	89.33		80.00

Table 4: Evaluation results of MLLMs on the single-room with higher interaction complexity video category. H(3,4) denotes higher-order problems at orders 3 and 4.

Evaluation Model's Superiority over Pattern-matching Assessment The majority of instances are answer (correct) - reasoning(incorrect) (denoted as 1-0), where models' reasoning violates criteria described in Section 2.3. The correctness of their answers is incidental rather than reflecting a genuine mastery of the problem-solving methodology. A minority of cases are incorrect - correct (denoted as 0-1). As illustrated in Figure 6, these cases

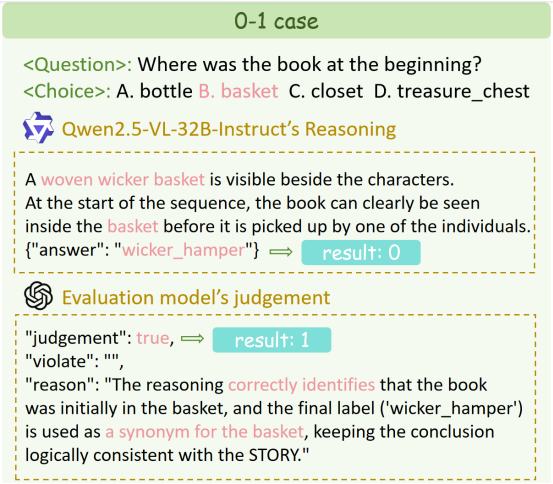

Figure 6: Example of a 0-1 case where pattern matching fails but FWW-EVAL succeeds. The model's reasoning correctly identifies the book in the basket, but answers with "wicker" (a synonym).

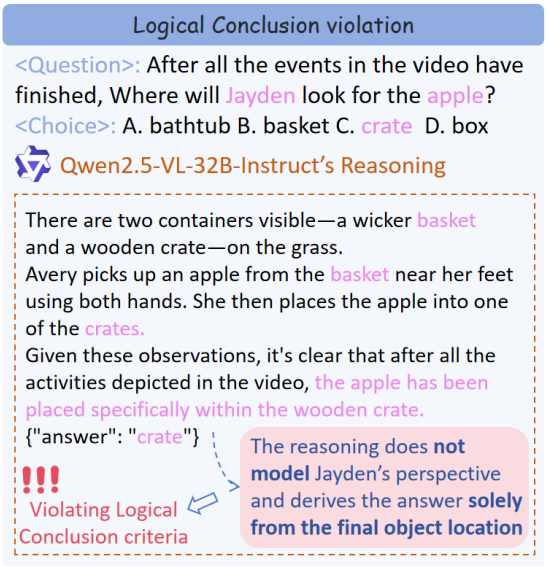

Figure 7: Example of Logical Conclusion violation by Qwen2.5-VL-32B-Instruct. It fails to model Jayden's mental state, a key requirement for this ToM question. The model's answer reflects intuitive reasoning based on object location rather than logical inference about another agent's perspective.

typically occur when models provide **synonyms of the correct answer** that are marked incorrect by the matching of rigid patterns. The evaluation model's judgment mitigates this rigidity inherent in pattern-matching-based correctness assessment.

3.4 In-Depth Analysis

Performance Gap between Open-source and Closed-source Models. A critical finding of our benchmark reveals a substantial performance disparity between closed-source and open-source models. When comparing the top-performing closed-source model GPT-5 against the top-performing open-source model Qwen3-VL-235B-A22B-Thinking, we observe a substantial performance gap exceeding 20% overall. This finding contrasts sharply with existing mainstream MLLM benchmarks, where open-source models demonstrate performance comparable to their closed-source counterparts.

Single-room vs Continuous Multi-room. From single-room to continuous multi-room scenarios illustrates more naturalistic, everyday social situations, with correspondingly longer video durations. We specifically analyze the MLLMs' performance differences between single-room and continuous multi-room settings for question types of equivalent difficulty. As presented in Table 3, the substantial proportion of pink-shaded cells indicates an overall trend of performance degradation.

Transparent-Barrier Second-order Tasks Cause all MLLMs to Fail. For human experts, reasoning about how visual accessibility induces changes in cognitive states is relatively straightforward. This is exemplified by the 80.00% accuracy (Table 4) achieved by humans, who correctly account for visual access through transparent glass even when positioned outside the room. In contrast, all evaluated MLLMs fail this task. An analysis of the models' reasoning processes reveals a total absence of references to transparent barriers, highlighting the substantial gap that remains between current MLLMs and expert-level understanding.

4 Conclusion

In this paper, we introduce FWW, a generative framework that bridges text-based ToM (words) to video-based ToM (worlds), and construct FWW-V, a benchmark that aligns with classical ToM task paradigms in cognitive science while incorporating video categories that more closely reflect everyday situations, including continuous multi-room scenarios and videos with higher interaction complexity. We further design a rubric-based evaluation framework, FWW-EVAL, to rigorously assess MLLMs' ToM capabilities. Experimental results reveal that current state-of-the-art MLLMs still struggle significantly with ToM reasoning. We hope that FWW-V will serve as a catalyst for future research aimed at advancing MLLMs' ToM capabilities.

Limitations

To our knowledge, this work has the following limitations:

- We use Google's Veo 3 Series model for video generation, which imposes constraints on video duration. Additionally, generating usable data requires multiple model calls to produce candidate videos, from which only a subset of suitable ones are selected. This process incurs additional token consumption and increases the overall generation cost.
- The validity of FWW-EVAL also depends on the judge model. In this work, we employ the relatively powerful GPT-5 as the judge model, which demonstrates strong comprehensive capabilities such as instruction following.

Ethical considerations

The research is performed entirely independently of any external funding sources that might bias the results or conclusions. When constructing the human performance baseline, we collect data on human responses to the questions; this data is used solely for establishing the human performance baseline and serves no other purpose. We construct FWW-V and propose FWW-EVAL to evaluate the ToM capabilities of MLLMs, highlighting substantial room for improvement in ToM capabilities to achieve more socially friendly, nuanced, and effective human-AI interaction and collaboration. AI assistants are used solely for polishing and improving the clarity of this paper.

References

Anthropic. 2025. Claude opus 4.1.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15959–15983. Association for Computational Linguistics.

Ronghao Dang, Yuqian Yuan, Wenqi Zhang, Yifei Xin, Boqiang Zhang, Long Li, Liuyi Wang, Qinyang Zeng, Xin Li, and Lidong Bing. 2025. Ecbench: Can multi-modal foundation models understand the egocentric world? A holistic embodied cognition benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24593–24602. Computer Vision Foundation / IEEE.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. 2025. Somi-tom: Evaluating multi-perspective theory of mind in embodied social interactions. *CoRR*, abs/2506.23046.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hervé Jégou, Alessandro Lazaric, Arjun Majumdar, Andrea Madotto, Franziska Meier, Florian Metze, Théo Moutakanni, Juan Pino, Basile Terver, Joseph Tighe, and Jitendra Malik. 2025. Embodied AI agents: Modeling the world. *CoRR*, abs/2506.22355.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. 2025. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 18959–18969. Computer Vision Foundation / IEEE.

Google DeepMind. 2025. Google veo 3.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Vincent Cartillier, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Devansh Kukreja, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, and 67 others. 2025. Ego4d: Around the world in 3,600 hours of egocentric video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(11):9468–9509.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *CoRR*, abs/2410.13648.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *CoRR*, abs/2507.17746.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025a. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 8450–8460. Computer Vision Foundation / IEEE.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, and 58 others. 2025b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *CoRR*, abs/2507.01006.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *CoRR*, abs/2501.13826.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer D. Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 16077–16102. Association for Computational Linguistics.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5871–5876. Association for Computational Linguistics.

Yuxuan Li, Vijay Veerabadran, Michael L. Iuzzolino, Brett D. Roads, Asli Celikyilmaz, and Karl Ridgeway. 2025. Egotom: Benchmarking theory of mind reasoning from egocentric videos. *CoRR*, abs/2503.22152.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

OpenAI. 2025a. Introducing GPT-4.1 in the API.

OpenAI. 2025b. Introducing GPT-5.

Chirag Parikh, Deepti Rawat, Rakshitha R. T, Tathagata Ghosh, and Ravi Kiran Sarvadevabhatla. 2025. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19002–19011. Computer Vision Foundation / IEEE.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E. Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Kenji Toyama, Robert James Berry, Divya Tyamagundlu, Timothy P. Lillicrap, and Oriana Riva. 2025. Androidworld: A dynamic benchmarking environment for autonomous agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2257–2273. Association for Computational Linguistics.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 1510–1519. AAAI Press.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *Thirty-Ninth AAAI*

Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025, pages 1520–1528. AAAI Press.

Gemini Team. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Qwen Team. 2025b. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

Tomer D. Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *CoRR*, abs/2302.08399.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. 2025a. Lvbench: An extreme long video understanding benchmark. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 22958–22967. IEEE.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8292–8308. Association for Computational Linguistics.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 10691–10706. Association for Computational Linguistics.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. 2025. Towards dynamic theory of mind: Evaluating LLM adaptation to temporal evolution of human states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24036–24057. Association for Computational Linguistics.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8593–8623. Association for Computational Linguistics.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, Lecture Notes in Computer Science, pages 169–186. Springer.

Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. 2025. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 20626–20636. IEEE.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. 2025. MMVU: measuring expert-level multi-discipline video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 8475–8489. Computer Vision Foundation / IEEE.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. MLVU: benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*

2025, Nashville, TN, USA, June 11-15, 2025, pages 13691–13701. Computer Vision Foundation / IEEE.

A Data Release and Licensing

According to the latest terms of service of Google Vertex AI, we are permitted to release videos generated by Veo 3 for non-commercial research purposes under the CC BY-NC-SA 4.0 license. In addition to the video data, we will release the data generation prompt, evaluation suite, metadata, and QA pairs, ensuring the benchmark is fully accessible to the research community.

B Related Work

B.1 Benchmarking ToM Capabilities

A widely adopted approach to evaluating ToM abilities is the Sally–Anne false-belief test. In this paradigm, one agent (Sally) possesses certain knowledge about the environment and then temporarily leaves. During her absence, another agent (Anne) alters the state of the world. For instance, Sally places a ball in a basket and leaves the room, after which Anne moves the ball to a box (Wilf et al., 2024). ToMi (Le et al., 2019), HiToM (Wu et al., 2023), BigToM (Gandhi et al., 2023), and OpenToM (Xu et al., 2024) are all Sally-Anne false belief-style ToM benchmarks designed to evaluate the ToM capabilities of LLMs. Text-based ToM benchmarks have been continuously evolving, with multiple benchmarks such as ToMBench (Chen et al., 2024), ToMATO (Shinoda et al., 2025), SimpleToM (Gu et al., 2024), and DynToM (Xiao et al., 2025) successively proposed to address richer social contexts and capture more dimensions of mental states, including emotion and desire, in addition to belief. However, multimodal ToM benchmarks remain relatively scarce, and existing ones, such as MMToM-QA (Jin et al., 2024), which is built on simulation environments, and EgoToM (Li et al., 2025), which employs limited observation perspectives, suffer from significant limitations.

B.2 Video Understanding Benchmarks

Video-MME (Fu et al., 2025) is a comprehensive evaluation benchmark for MLLMs in video understanding, encompassing a full spectrum of video lengths, highly diverse video types, and covering different core abilities of MLLMs. MLVU (Zhou et al., 2025) focuses on evaluating the long video understanding capabilities of MLLMs and proposes multiple evaluation tasks tailored for long video understanding. It similarly incorporates various video genres, reflecting model performance in long video understanding across different scenarios. Both benchmarks have been widely adopted to evaluate mainstream MLLMs such as Qwen3-VL (Team, 2025b) and InternVL3.5 (Wang et al., 2025b), as presented in their technical reports. Additionally, MotionBench (Hong et al., 2025a) is designed to evaluate fine-grained video motion understanding, ECBench (Dang et al., 2025) to evaluate embodied cognitive abilities based on egocentric videos, RoadSocial (Parikh et al., 2025) to evaluate road event understanding from social video narratives, and Video-TT (Zhang et al., 2025) to evaluate both the correctness and robustness of complex visual narrative video understanding. Despite the extensive evaluation dimensions currently available for MLLMs, their ToM capabilities, though fundamentally important (Fung et al., 2025), have not yet received adequate attention.

C Additional Experiments

C.1 Compare the performance of GPT-5.2 and GPT-5.

We select GPT-5.2-2025-12-11, a newer version of the model, and compare its performance against GPT-5-2025-08-07 in the single-room video category to verify whether a newer model version necessarily performs better. The experimental results are shown in Table 5. GPT-5.2-2025-12-11 significantly underperforms GPT-5-2025-08-07 despite being a newer version, demonstrating that our FWW-V benchmark continues to pose substantial challenges even for updated models.

C.2 Model Performance on A New Video Data Variant

As shown in Figure 8, the female Isla turns her back (becoming visually unreachable) while the male Jackson moves the basketball from the bucket to the suitcase. Under this video data variant, we select the Gemini-2.5-Pro model to explore its performance on first-order and second-order questions, and compare it with its previous performance. We generate a total of 15 videos of this type, each corresponding to two questions. The comparison between Gemini-2.5-Pro's performance on the original FWW-V video data and on this new video data variant is shown in Figure 9. Its performance on first-order questions drops from 78.05 to 60.00, and on second-order questions from 70.59 to 33.33, both exhibiting substantial declines.

Models	0-Order (Basic)		1st-Order (F-)		2nd-Order (S-)		3rd-Order (T-)	
	Mem	Rel	F-O	F-N	S-O	S-N	T-O	T-N
GPT-5-2025-08-07	98.59	95.77	92.79	90.00	90.13	72.92	77.78	55.56
GPT-5.2-2025-12-11	98.59	84.51	85.71	60.00	46.52	51.02	66.67	33.33
Δ	0.00	-11.26	-7.08	-30.00	-43.61	-21.90	-11.11	-22.23

Table 5: Evaluation results of GPT-5 and GPT-5.2 on the single room video category. Δ represents the performance gap by which GPT-5.2 falls behind GPT-5.

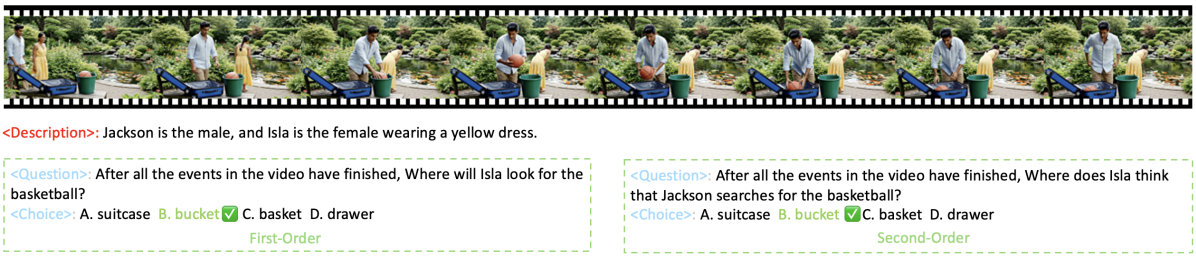

Figure 8: A new video data variant: The female appears in the scene, but when the male moves the basketball, she has her back turned away (with no line of sight).

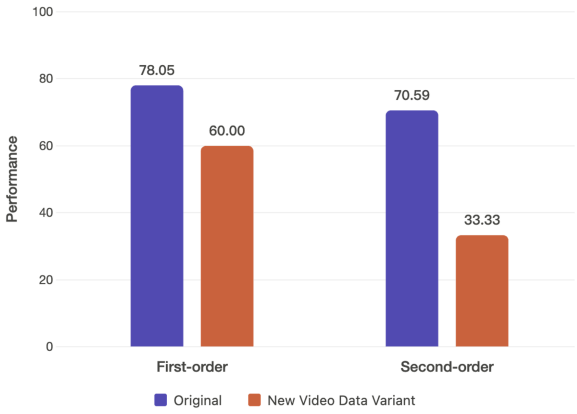

Figure 9: Comparison of the Gemini-2.5-Pro model's performance on first-order and second-order questions under the original FWW-V video data versus the New Video Data Variant.

D Detailed Experimental Settings

A common practice is to extract a sequence of frames from the video and interpret the resulting multi-image inputs. To estimate the upper-bound performance of advanced models such as GPT-5 and Gemini-2.5-Pro, we sample one frame every 0.5 seconds to ensure that no critical visual details are missed. For the InternVL3.5-241B-A28B model, we do not conduct evaluations on videos with relatively long durations, due to constraints on the number of input frames imposed by its API interface and deployment challenges arising from its large model size. For all experiments, we set the temperature to 0.2 (with exceptions for certain models where this parameter cannot be adjusted, such as GPT-5, which uses a default temperature of 1). The prompt used to query the models, along with the complete video frames as input, is as follows:

```
====== QUERING EACH MODEL ======
Prompt: In the video, {description}. Select
 the best answer to the following
multiple-choice question based on you
observation of the video.

Question: {question}
{choices_text}

give me your analysis and put your
answer in the format {{"answer": "
Your_Answer_Here"}}, where
Your_Answer_Here is the option text (not
 the letter).
```

We use description to specify each character's gender and clothing (e.g., Jack is a male wearing a blue T-shirt), thereby enabling the model to identify which character in the video corresponds to Jack. We evaluate the model's analysis using FWW-EVAL to determine whether it correctly solves the given problem.

E Additional Analysis of FWW-EVAL

We apply FWW-EVAL to evaluate the analysis process of MLLMs by leveraging GPT-5 as the judge.

E.1 Self-judging and Misjudgment Sources

Sources of misjudgment include: (1) when multi-scene scripts serve as references, the strict matching of perceptual signals occasionally causes misattribution errors at the conclusion level, primarily in 0th-order and 1st-order questions; (2) for nested belief judgments, GPT-5 occasionally conflates real-world states with character belief states, primarily in higher-order questions. To mitigate self-judging bias and validate judge consistency, we replace GPT-5 with Gemini-3-Pro-Preview as the judge model and reevaluate outputs from Qwen2.5-VL-32B and GPT-5. Divergent judgment rates are **1.31%** and **0.93%**, respectively.

E.2 Multi-scene Script as Reference

We emphasize that FWW-EVAL enables effective synergy between the multi-scene script as reference and the evaluation criteria at the perception, cognition, and conclusion levels. As illustrated by the purple text at the bottom of Figure 10, when significant perception errors influence question solving, the conclusion level criteria can still capture these issues due to the multi-scene script as reference. This synergy between the multi-scene script as reference and conclusion level criteria further reduces the false negative rate.

E.3 Genuine Reasoning or Spurious Correlation?

As shown in Figure 11, GPT-5 establishes a genuine problem-solving process, whereas the Qwen2.5-VL-32B-Instruct model merely describes basic visual and commonsense-related information. The ToM task indeed poses a significant challenge for the less capable Qwen2.5-VL-32B-Instruct model, which fails to identify an appropriate entry point for task resolution. Consequently, it frequently violates the conclusion level criteria. In contrast, although GPT-5 engages in genuine reasoning about the problem, it occasionally makes errors at the perception level.

E.4 Highly Capable GPT-5 as the Judge

Figure 12 presents the complete prompt employed in our FWW-EVAL framework, utilizing GPT-5 as the judge. As illustrated in the prompt, a model's analysis is classified as erroneous if it violates criteria at either the cognition or conclusion level, with perception-level criteria serving an auxiliary role. Given that perfect correspondence between video frames and script descriptions cannot be guaranteed during the script-to-video generation process, this may potentially introduce misjudgments at the perception level. Remarkably, we observe that under the collaboration of multiple components and empowered by the strong intelligence level of GPT-5, even when the judge's output indicates violations only at the perception level, it still returns a negative judgment (false), which is indeed aligned with the ground truth. This behavior aligns with our ultimate objective of minimizing the error rate in our evaluation framework.

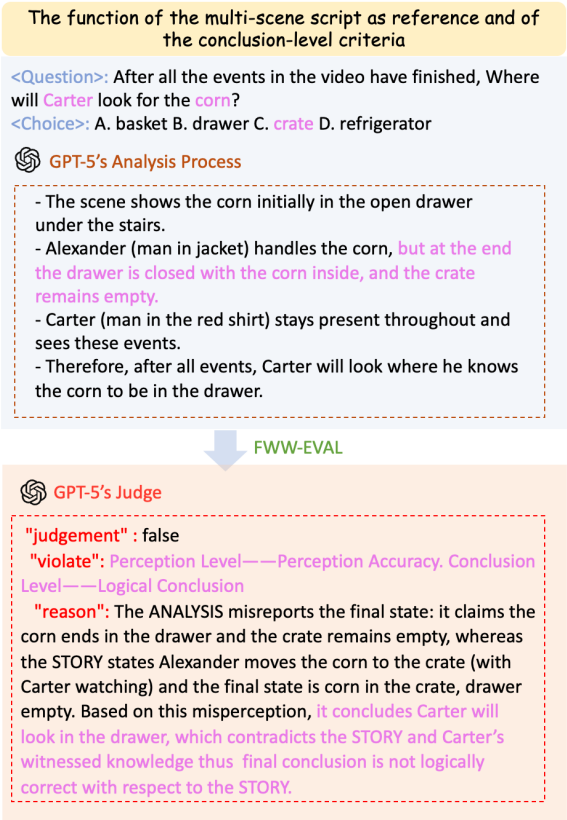

Figure 10: A case study illustrating the role of multi-scene script as reference and conclusion level criteria.

F Details on FWW-V Generation

In this section, we present the multi-scene script templates used to generate three video categories: single-room, single-room with higher interaction complexity, and continuous multi-room, along with the prompts used for video generation. We also report the human hours required for the entire FWW-V curation process.

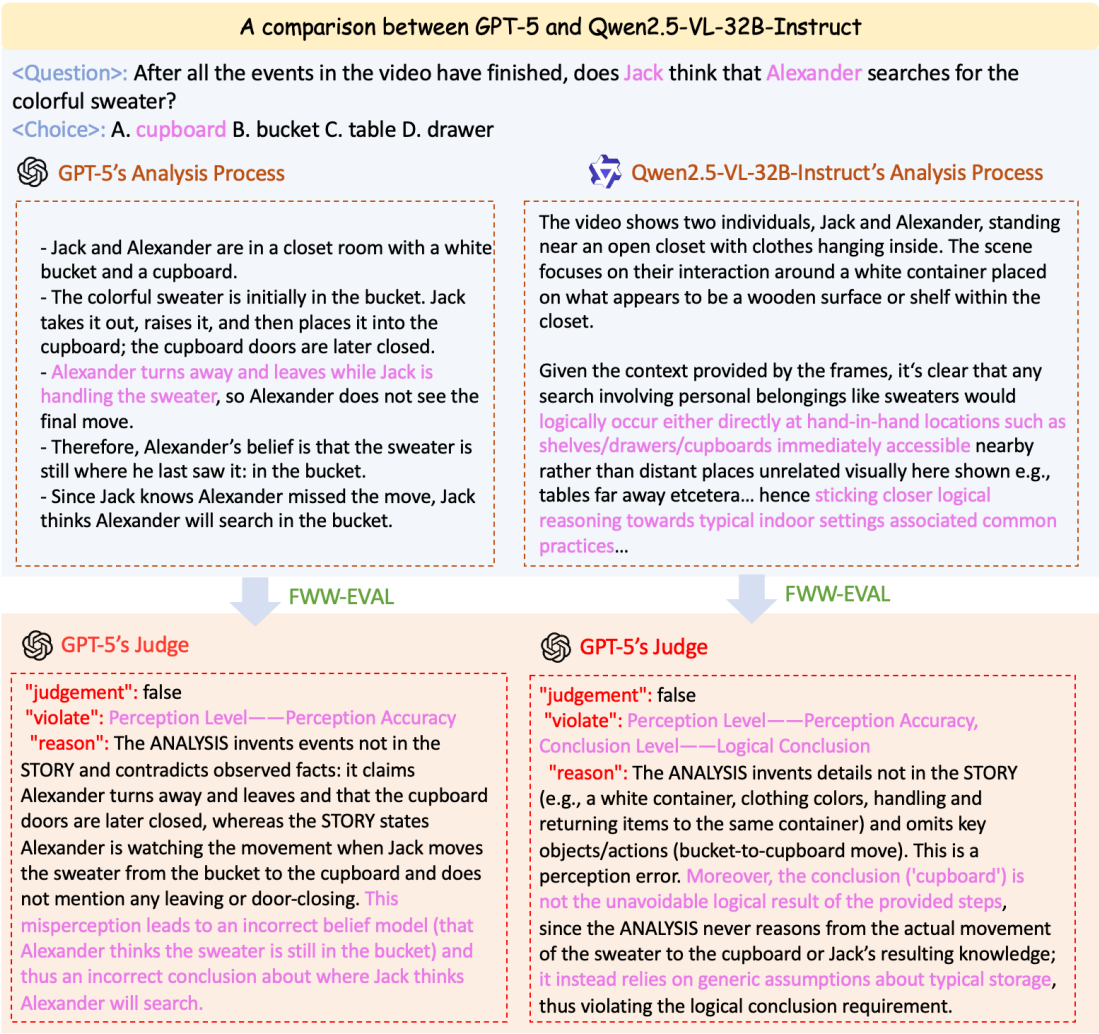

Figure 11: A comparison of the analysis process of GPT-5 and Qwen2.5-VL-32B-Instruct when tackling the same task, along with an illustration of the strong intelligence level GPT-5 demonstrates when acting as the judge.

F.1 Script Templates and Video Generation Prompt

Single-room. For this video category, we provide three multi-scene script templates and one video generation prompt, as detailed below:

True Belief

```
Scene 1: {Character1} and {Character2} are
both shown entering {Location}. In the {
Location}, There is only a {Object} inside
 the {Character1}, An empty {Character2} is
 visible.

Scene 2: Both {Character1} and {Character2}
are looking at the {Container1},
establishing their shared belief about
the {Object}'s location.

Scene 3: {Character1} moves the {Object}
from the {Container1} to the {Container2}.
 {Character2} is watching the movement.

Scene 4: The {Object} is now in the {
Container2} and the {Container1} is empty.
```

False Belief

```
Scene 1: {Character1} and {Character2} are
both shown entering the {Location}. In
the {Location}, There is only a {Object}
inside the {Container1}. An empty {
Container2} is visible.

Scene 2: {Character2} leaves the {Location},
 turning [his\her] back to the {
Container1} and the {Object} inside.

Scene 3: After {Character2} left, {
Character1} moves the {Object} from the {
Container1} to the {Container2}.

Scene 4: The {Object} is now in the {
Container2} and the {Container1} is empty.
```

Three-Person False Belief

```
Scene 1: {Character1}, {Character2} and {
Character3} are all shown entering the {
Location}. In the {Location}, There is a {
Object} inside the {Container1}, An empty {
Container2} is visible.
```

```
Scene 2: {Character3} exits the {Location},
turning [his\her] back at the {Container1}
 and the {Object} inside, leaving {
Character1} and {Character2} in the {
Location}.

Scene 3: After {Character3}'s leaving, {
Character1} moves the {Object} from the {
Container1} to the {Container2}. {Character2}
 stays here watching the action.

Scene 4: The {Object} is now in the {
Container2} and the {Container1} is empty.
```

```
====== VIDEO GENERATION PROMPT ======
You are an insightful illustrator
specializing in creating visual stories
that depict complex psychological states
.

Important Rules (Strictly Follow)

1. Core Rule: Follow the input order of
scenes to generate video frames.
Clearly show a character leaving a room,
 looking away, or being blocked from
seeing an event
and key actions made by characters. This
 establishes why they might hold a false
 belief.

2. Style Requirement: Each panel must
accurately reflect its described scene
without adding or omitting details.
Strictly follow the number of objects
described!

3. Think Step-by-Step: Please carefully
complete all the above requirements
step-by-step.
```

Single-room with Higher Interaction Complexity. For this video category, we provide one multi-scene script template and one video generation prompt, as detailed below:

```
Higher Interaction Complexity
Scene 1: {Character1}, {Character2} and {
Character3} are all shown entering the {
Location}. In the {Location}, there is a {
Object} inside the {Container1}; an empty {
Container2} and {Container3} are visible.

Scene 2: {Character2} exits the {Location},
turning [his/her] back at the {Container1}
 and the {Object} inside, leaving {
Character1} and {Character3} in the {
Location}.

Scene 3: After {Character2}'s leaving, {
Character1} moves the {Object} from the {
Container1} to the {Container2}. {Character3}
 is watching this movement.

Scene 4: {Character1} exits the {Location},
turning [his/her] back at the {Container2}
 and the {Object} inside, leaving {
Character3} in the {Location}.

Scene 5: After {Character1}'s leaving, {
Character3} moves the {Object} from the {
Container2} to the {Container3}.

Scene 6: The {Object} is now in the {
Container3} and the {Container1} and {
Container2} are empty.
```

```
====== VIDEO GENERATION PROMPT ======
You are an insightful illustrator
specializing in creating visual stories
that depict complex psychological states
.

Important Rules (Strictly Follow)

1. Core Rule: Follow the input order of
scenes to generate video frames.
Clearly show a character leaving a room,
 looking away, or being blocked from
seeing an event
and key actions made by characters. This
 establishes why they might hold a false
 belief.

2. Style Requirement: Each panel must
accurately reflect its described scene
without adding or omitting details.
Strictly follow the number of object
described!

3. Character Constraints:
Ensure the accuracy of the number of
characters in the video. When a
character leaves, the number of
characters in video decreases.

4. Logical Coherence: Please ensure the
logical coherence and character
consistency of the video. Use smooth and
 realistic transitions that are
consistent with natural
real-world dynamics.

5. Think Step-by-Step: Please carefully
complete all the above requirements
step-by-step.
```

Continuous Multi-room. For this video category, we provide one multi-scene script template and two video generation prompts, as detailed below:

```
Continuous Multi-room
Scene 1: {Character1}, {Character2} enter the
 {Location1}. In the {Location1}, there is
a {Object1} inside the {Container1}, an
empty {Container2} is visible.

Scene 2: {Character1} moves the {Object1}
from the {Container1} to the {Container2}.
{Character2} is watching this movement.

Scene 3: After a smooth and realistic
transition consistent with natural real-
world dynamics, {Character1} and {
Character2} enter {Location2}.
```

```
Scene 4: In the {Location2}, there is a {
Object2} inside {Container3}, and an empty
{Container4} is visible. There is also a
cute little dog.

Scene 5: {Character1} leaves the {
Location2}, turning his/her back to the {
Container3} and the {Object2} inside.

Scene 6: After {Character1} left, {
Character2} moves the {Object2} from the {
Container3} to the {Container4}.

Scene 7: {Object2} is now in {Container4}
and {Container3} is empty.
```

```
====== VIDEO GENERATION PROMPT1 ======
You are an insightful illustrator
specializing in creating visual stories
that depict complex psychological states
.

Important Rules (Strictly Follow)

1. Core Rule:
Follow the input order of scenes to
generate video frames. Clearly show key
actions
performed by the characters.

2. Character Visible and Consistency:
Keep both characters {Character1} (female)
 and {Character2} (male)  visible
throughout the entire video. Maintain
strict character consistency for the
entire sequence.

3. Style Requirement:
Each panel must accurately reflect its
described scene without adding or
omitting any details.
Strictly follow the exact number of
objects mentioned in the input.

4. Think Step-by-Step:
Please carefully complete all the above
requirements step-by-step.
```

```
====== VIDEO GENERATION PROMPT2 ======
You are an insightful illustrator
specializing in creating visual stories
that depict complex psychological states
.

Important Rules (Strictly Follow)

1. Core Rule:
Follow the input order of scenes to
generate video frames. Clearly show key
actions
made by the characters.

2. Character Consistency:
Maintain character consistency
throughout the entire video.

3. Style Requirement:
Each panel must accurately reflect its
described scene without adding or
omitting details.
Strictly follow the number of objects
described.

4. Think Step-by-Step:
Please carefully complete all the above
requirements step-by-step.
```

You are a meticulous Auditor specializing in Theory of Mind (ToM) validation.Your sole purpose is to evaluate a given ANALYSIS (a thought process) against a STORY (***reference***).
You must determine if the ANALYSIS is 100% sound, logical, and correctly models the target character's belief states to reach a correct conclusion. ***You are not answering the question yourself; you are only validating the provided analysis (a thought process).***
You are provided with:
- **STORY:** A sequence of scenes describing characters, objects, and actions. This is the absolute basis, ***reference.***
- **ANALYSIS:** A step-by-step reasoning trace that attempts to deduce a character's belief (or a nested belief) based on the STORY.

Critical Evaluation Criteria:
Perception Level:
- **Perception Accuracy:** ANALYSIS must correctly perceive the objects, containers in the STORY scene (perceiving "basket" as "basketball hoop" is considered correct perception, as these are synonyms). However, perceiving "basket" as "bucket" would constitute a perceptual error. The analysis also must not include any scenes that never appear in the STORY.
- **Temporal Accuracy:** The scenes in the STORY unfold in temporal order. In the ANALYSIS, the order of whether a character enters (leaves) first or an object is moved first should not be confused. Only need to focus the relative temporal order of [character enter], [exit], and [object movements]. The character's sequential entries or exits (one after another) are considered equivalent to "entering together" as described in the STORY.

Cognition Level:
- **No Perspective "Leakage"**(The "Curse of Knowledge" Fallacy): The ANALYSIS must not attribute knowledge of the actual, current state of the world to a character who could not possibly know it. A character can only know something if they have perceived it.
- **Nested Belief Accuracy:** For 2nd or 3rd-order ToM questions (e.g., "Where does A think B thinks...?"), the ANALYSIS must be evaluated from the correct perspective. The reasoning must deduce what "Character A thinks about Character B's..."

Conclusion Level:
- **Logical Conclusion:** The final conclusion of the ANALYSIS (the answer to the implied question) must be the unavoidable logical result of the preceding steps.
Output Format:
Respond with the lowercase string true or false and the reason for the decision. Return the specific rule item that was violated ("Perception Level——Perception Accuracy", "Perception Level——Temporal Accuracy", "Cognition Level——No Perspective Leakage", "Cognition Level——Nested Belief Accuracy" and "Conclusion Level——Logical Conclusion")

Return all violated rule items. （if the violation of one rule item causes the violation of another, only need to return the fundamental(root) rule item. For example, if the violations include **Perception Level — Perception Accuracy**, **Cognition Level — No Perspective Leakage**, and **Cognition Level — Nested Belief Accuracy**, but **"No Perspective Leakage" causes the violation of "Nested Belief Accuracy"**, then you only need to return: **Perception Level — Perception Accuracy**, **Cognition Level — No Perspective Leakage**).
- true: The ANALYSIS follows the rules of the **Perception Level**, **Cognition Level**, and **Conclusion Level**, or, even if there are violations at the Perception Level, they do not affect the Cognition or Conclusion Levels (i.e., the rules at the Cognition and Conclusion Levels are fully satisfied).
- false: The ANALYSIS violates the rules at the Cognition Level or the Conclusion Level.

EXAMPLE:
STORY:
description: Alexander is the man in blue shirt, Jayden is the man in orange shirt, Evelyn is the woman.
- Scene 1: Alexander, Jayden and Evelyn are all shown entering the sunroom. In the sunroom, There is a basketball inside the basket, An empty bucket is visible.
- Scene 2: Evelyn exits the sunroom, turning her back at the basket and the basketball inside, leaving Alexander and Jayden in the sunroom.
- Scene 3: After Evelyn's leaving, Alexander moves the basketball from the basket to the bucket. Jayden stays here watching the action.
- Scene 4: The basketball is now in the bucket and the basket is empty.

QUESTION: where does Evelyn think that Alexander thinks Jayden thinks the basketball is?
ANALYSIS: 1. Alexander (man in blue) and Jayden (man in orange) enter a room with a basketball hoop, a basketball placed on the hoop, and a bucket near the hoop.
2. Evelyn (woman in white) enters the room later, and all three observe the basketball.
3. Jayden attempts to move the basketball but does not put it back in the bucket.
4. Alexander picks up the basketball from the hoop and places it into the bucket.
5. The question is about Evelyn's belief regarding where Alexander thinks that Jayden thinks the basketball is.
6. Since the basketball is currently physically in the bucket, Evelyn can only infer that Alexander thinks Jayden thinks the basketball is in the bucket, even though the basketball is initially on the hoop.
7. This implies a nested belief about the location of the basketball from Evelyn's perspective, given the observed actions of Alexander and Jayden.
Response:
{{Judgement: false,
Violate: Cognition Level——No Perspective Leakage. Perception level——Perception Accuracy.
Reason: The ANALYSIS injects actions not in the STORY, then commits perspective leakage by attributing knowledge to Evelyn that she could not have. Specifically, it says Jayden tried to move the ball (never happens), fails to note that Evelyn leaves before the move, and concludes based on the actual location rather than nested beliefs from Evelyn's perspective. }}

Please evaluate the following ANALYSIS, QUESTION and STORY:
STORY:{story}
QUESTION:{question}
ANALYSIS:{analysis}

Figure 12: The complete prompt employed in our FWW-EVAL framework for judging models' analysis process.