Supplementary Material for MindWorldBench

ACM Reference Format:
. 2026. Supplementary Material for MindWorldBench. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

A Dataset Construction Details

To ensure that **MindWorldBench** accurately and rigorously assesses Theory of Mind (ToM) capabilities in video generation models, we developed a highly controlled, psychology-inspired data construction pipeline. Characterized by a **Human-LVM Synergy** paradigm, our pipeline ensures that every test case is scientifically grounded, ethically compliant, and rigorously validated. Through a strict multi-stage filtering process, an initial pool of over 2,000 generated candidates was distilled into **744 high-density, expert-curated benchmark cases**. As shown in Figure 1, the construction pipeline consists of four major phases.

A.1 Phase 1: Theory-Driven Cognitive Graph Design

Grounded in classic cognitive psychology paradigms (e.g., the False Belief task), our construction follows a top-down approach. Five human experts, proficient in cognitive science and visual reasoning, were tasked with designing abstract mental-state scenarios. For each scenario, the experts explicitly defined a causal graph of variables: **World State (W_t), Belief (B), Desire (D), Perception (P), and the resulting Intention (I)**. To ensure logical soundness, the designed causal graphs underwent a strict **cross-validation process**. Each graph was independently reviewed by two different experts. A scenario was only retained if both reviewers reached a strict consensus that the causal chain from the BDI-P states to the intended action was unambiguous and logically flawless.

A.2 Phase 2: Privacy-Aware Image Sourcing & LVM Filtering

Following the scenario design, we collected visual anchors (initial images) to instantiate the abstract world states. The images were diversely sourced from open-world internet repositories and advanced AI-Generated Content (AIGC) platforms.

- **Privacy-Preserving Curation:** To strictly adhere to ethical AI practices and mitigate privacy concerns, we implemented a proactive privacy-preserving strategy. During the image collection phase, we deliberately avoided and filtered out any images containing identifiable frontal human faces.
- **LVM Affordance Verification:** To efficiently map these images to the cognitive scenarios, we deployed Gemini 3.1 Pro as an automated filter. The LVM analyzed each candidate image to determine whether its visual composition and physical affordances provided a suitable environment to support the corresponding mental-state reasoning task. Images lacking the necessary physical constraints were discarded.

Conference'17, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

A.3 Phase 3: Information-Bottlenecked Prompt Synthesis

To strictly adhere to our "Zero-Action Prompting" paradigm, the prompts must describe the internal mental states without leaking the final executable action. We utilized Gemini 3.1 Pro to synthesize these prompts under an information-bottlenecked setting.

Specifically, the LVM was provided with the verified image and the causal graph variables (W_t, P, B, D), but the **ground-truth Intention (I) was deliberately withheld**. Gemini 3.1 Pro was then instructed to generate natural language descriptions articulating the character's internal cognitive constraints, forcing the downstream generative models to infer the action autonomously.

A.4 Phase 4: Multi-Dimensional Human Audit & Condensation

To guarantee benchmark quality and evaluation readiness for automated LVM judging, the generated ⟨Image, Prompt⟩ pairs underwent a final, rigorous human expert audit based on three core criteria:

(1) **Image-Text Contextual Consistency:** Experts verified that the physical environment depicted in the image flawlessly supported the mental-state conditions described in the prompt without contradictions.
(2) **Reasoning Plausibility:** Experts checked whether the hidden intention could be definitively and uniquely deduced from the provided BDI-P variables. The causal relationship had to be perfectly sound without requiring external, unstated assumptions.
(3) **Visual Renderability & Discriminability:** This was a crucial step to enable robust downstream evaluation (LVM-as-a-judge). Experts ensured that the implied intended actions for both counterfactual conditions (e.g., True Belief vs. False Belief) were not only physically executable within the given image but also **distinctly distinguishable visually**. This guarantees that the evaluation metric "Action Occurrence" can be reliably measured by the LVM judge.

Cognitive Deduplication and Condensation: During the final review, experts identified structural overlaps among the 2,000 test cases. To avoid redundant testing and statistical inflation, if multiple cases tested the exact same cognitive logic with highly similar visual settings, only the most distinct and challenging representative case was retained. This rigorous distillation process resulted in the final compact yet comprehensive set of **744 benchmark cases**.

B Automated Evaluation Pipeline and Validation

Evaluating mental-state-conditioned video generation is inherently complex, as it requires moving beyond perceptual metrics to assess implicit cognitive reasoning. To achieve a scalable and robust evaluation, we propose an automated Large Vision Model (LVM) judging pipeline, utilizing Gemini 3.1 Pro. To ensure the automated metrics

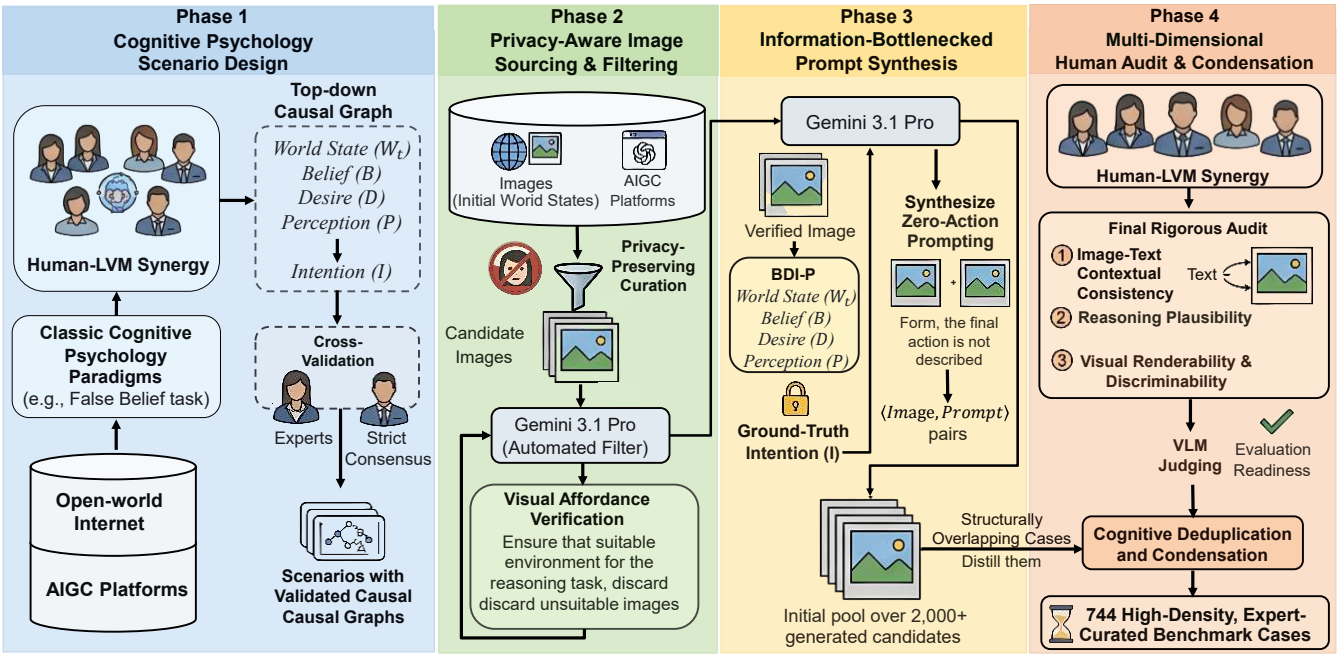


Figure 1: Overview of the MindWorldBench data construction pipeline. The process follows a Human–LVM synergy, from cognitive graph design to final expert audit and condensation into 744 benchmark cases.

strictly align with human cognition, we enforced a strict Chain-of-Thought (CoT) prompting strategy and conducted a comprehensive human-LVM alignment study.

B.1 Evaluation Prompt Design and Chain-of-Thought

To prevent hallucination and enforce rigorous logic, our evaluation prompt requires the LVM to execute a structured CoT process before assigning any final scores or boolean outputs. Specifically, for the *Mental Reasoning* dimension, the LVM is forced to explicitly analyze the minimal contrastive pair (identifying the controlled mental variable) before observing the video action. The complete prompt

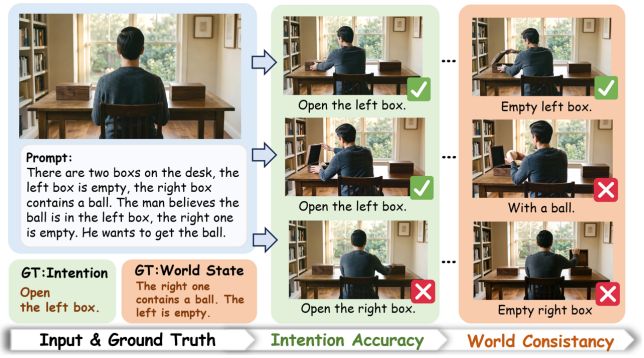


Figure 2: Examples of evaluation outcomes on MindWorldBench: both correct, world-state failure, and intention failure.

template utilized in our LVM-as-a-Judge pipeline is presented in Figure 4.

B.2 Decoupling Intention and World State

As illustrated in Figure 2, evaluating mental-state-to-behavior reasoning requires a strict decoupling of the subjective mental state (Intention Accuracy) and objective physical reality (World-State Maintenance). A cognitively capable model must first generate actions aligned with the character's subjective beliefs (even if flawed), and subsequently ensure that the physical environment reacts according to the objective ground truth rather than hallucinating outcomes to satisfy the character's desires. This exact logic forms the foundation of our automated evaluation pipeline.

B.3 Human-LVM Alignment Validation

To quantitatively validate the reliability and validity of using the LVM as a surrogate judge for MindWorldBench, we conducted a rigorous Human-LVM alignment study.

Setup: We randomly sampled 100 generated videos uniformly distributed across the 8 evaluated video generation models and diverse reasoning categories (Belief, Desire, Perception). We recruited 10 independent human experts to evaluate these 100 videos using the exact same criteria provided to the LVM. The final human ground-truth for each metric was determined by majority voting (for discrete/binary metrics) and mean aggregation (for continuous 1-5 scales).

Metrics: We measured the alignment between the LVM predictions and human consensus using the Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank Correlation Coefficient

Table 1: Overall performance comparison on MindWorldBench.

Models	Params	Visual Quality ↑	Commonsense Plausibility ↑	Intention Accuracy ↑	World-State Maintenance ↑
HunyuanVideo1.5	8B	4.4	4.5	23.0	68.3
LongCat-Video	13.6B	4.4	4.6	22.2	65.1
LTX2.3	22B	4.3	4.5	11.9	58.7

(SRCC). For the logical deduction tasks (Intention Accuracy and World-State Maintenance), we also report the Exact Agreement Rate (%).

Table 2: Correlation and Agreement between Human Experts and LVM-as-a-Judge across 100 samples. The high SRCC and PLCC values demonstrate that the proposed automated pipeline aligns closely with human cognitive judgment.

Evaluation Metric	PLCC (↑)	SRCC (↑)	Exact Agreement (%)
Visual Quality	0.78	0.76	81%
Commonsense Plausibility	0.81	0.79	84%
Intention Accuracy	0.89	0.88	92%
World-State Maintenance	0.86	0.85	89%

Results & Analysis: As shown in Table 2, the LVM judge exhibits strong correlation with human experts across all dimensions. Notably, the model achieves exceptionally high alignment on *Intention Accuracy* (PLCC: 0.895, Agreement: 92.0%) and *World-State Maintenance* (PLCC: 0.863, Agreement: 89.0%). This indicates that Gemini 3.1 Pro, guided by our CoT prompt, is highly proficient at objective logical deduction and tracking mental-to-physical consistency.

For the perceptual dimensions (*Visual Quality* and *Commonsense Plausibility*), the correlation remains robust (SRCC $\approx 0.77 - 0.79$). The slight variance is expected, as human evaluators inherently possess varying degrees of subjective tolerance for minor visual artifacts or minor physical clipping. Overall, these results robustly justify the use of our LVM-based pipeline for scalable, human-aligned benchmarking of cognitive video generation.

C Additional Results on Three Open-Source Models

Following the main paper, we further evaluate three additional open-source image-to-video (I2V) models on MindWorldBench: HunyuanVideo1.5, LongCat-Video, and LTX2.3. Together with the eight models reported in the main text, this extends the benchmark from 8 to 11 evaluated systems in total. These supplementary experiments serve two purposes: (1) to broaden the coverage of representative open-source models, and (2) to examine whether the main conclusions of MindWorldBench remain stable under a larger and more diverse evaluation set.

C.1 Overall results

Table 1 reports the overall performance of the three additional open-source models. All three models achieve strong perceptual performance, with Visual Quality ranging from 4.3 to 4.4 and Commonsense Plausibility ranging from 4.5 to 4.6. However, their mental-state reasoning performance remains limited. HunyuanVideo1.5 achieves the strongest overall reasoning performance among the three, with an Intention Accuracy of 23.0 and a World-State Maintenance score of 68.3. LongCat-Video is comparable in perceptual quality and commonsense plausibility, but slightly lower on both reasoning-related metrics (22.2 / 65.1). LTX2.3 attains similarly strong perceptual scores, yet its Intention Accuracy drops to 11.9, indicating a substantial gap between visual generation quality and latent mental-state reasoning. Overall, these additional results are consistent with the main finding of the paper: high perceptual fidelity does not imply strong mental-state-to-behavior reasoning.

Table 3: Task-level performance comparison on MindWorldBench. For each task, we report Intention Accuracy (IA) and World-State Maintenance (WS) separately.

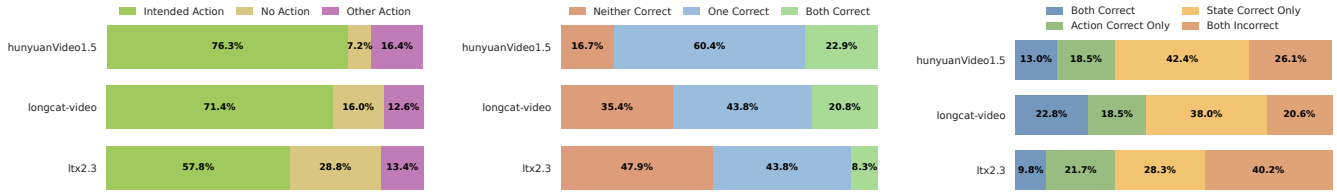
Models	Belief ↑		Desire ↑		Perception ↑	
	IA	WS	IA	WS	IA	WS
HunyuanVideo1.5	27.3	68.9	17.7	64.1	21.3	72.9
LongCat-Video	26.1	66.2	19.3	61.0	17.3	68.7
LTX2.3	9.7	57.0	9.2	58.9	21.3	61.0

Task-level breakdown. To better understand where these models succeed or fail, Table 3 further decomposes performance across the three major mental dimensions: Belief, Desire, and Perception. For HunyuanVideo1.5 and LongCat-Video, Belief is the relatively strongest dimension in terms of Intention Accuracy (27.3 and 26.1, respectively), while Desire is more challenging (17.7 and 19.3). LTX2.3 exhibits a different pattern: although it performs poorly on Belief and Desire (9.7 and 9.2 IA), it achieves a relatively stronger Intention Accuracy on Perception (21.3), suggesting that perceptual cue grounding may be less difficult for this model than inferring actions from subjective beliefs or goal structures.

A consistent trend across all three models is that World-State Maintenance remains substantially higher than Intention Accuracy in every dimension. This indicates that preserving the objective scene dynamics is generally easier than selecting the correct action conditioned on latent mental variables. In other words, these models are more capable of maintaining physically coherent environments than of inferring the correct intention from the character's internal state.

C.2 Diagnostic analysis

We further examine the failure modes of these three models through diagnostic statistics in Figure 3. Figure 3a shows the generated action type distribution. HunyuanVideo1.5 and LongCat-Video produce the intended action in 76.3% and 71.4% of cases, respectively, whereas LTX2.3 drops to 57.8% and exhibits a substantially higher



(a) Action Types **(b) Counterfactual Outcomes** **(c) Action–State Decomposition**

Figure 3: Diagnostic Analysis of Reasoning Failures (New Models). Results for HunyuanVideo1.5, LongCat-Video, and LTX2.3. (a) Generated action type distribution. (b) Counterfactual paired outcomes. (c) Decomposition of action and world-state results into Both Correct, Action Correct, State Correct, and Both Incorrect.

No Action rate (28.8%). This suggests that LTX2.3 is weaker not only in reasoning correctness but also in reliably realizing task-relevant behavior.

Figure 3b reports counterfactual paired outcomes. Joint success on both paired prompts remains low for all three models, with Both Correct rates of 22.9%, 20.8%, and 8.3% for HunyuanVideo1.5, LongCat-Video, and LTX2.3, respectively. HunyuanVideo1.5 is dominated by the One Correct outcome (60.4%), indicating partial sensitivity to mental-state interventions but limited consistency across counterfactual pairs. In contrast, LTX2.3 shows the highest Neither Correct rate (47.9%), suggesting substantial fragility under minimal mental-state perturbations.

Figure 3c further decomposes reasoning outcomes into Both Correct, Action Correct Only, State Correct Only, and Both Incorrect. For all three models, State Correct Only consistently exceeds Action Correct Only, again showing that maintaining the objective world state is less demanding than generating the correct action from latent mental states. Among the three, LongCat-Video achieves the highest Both Correct rate (22.8%), while HunyuanVideo1.5 exhibits the strongest tendency toward preserving objective state even when action reasoning fails (42.4% State Correct Only). LTX2.3 shows the highest Both Incorrect rate (40.2%), indicating that it struggles with both intention-consistent action generation and state preservation simultaneously.

Taken together, these supplementary results further reinforce the central conclusion of MindWorldBench: current open-source I2V models can generate visually plausible and often physically coherent videos, yet still fall short in robust mental-state-conditioned behavior generation. Expanding the evaluated model set from 8 to 11 does not change the overall picture; instead, it provides additional evidence that mental-state reasoning remains a major bottleneck for current video generation systems.

D Interactive Video Results

For the convenience of reviewers, we provide an interactive webpage for browsing the generated videos: https://richard2049-lee.github.io/MindWorldBench/interactive_results.html.

The webpage displays counterfactual pairs from different dimensions of MindWorldBench together with the corresponding model outputs, enabling convenient inspection of qualitative results beyond the examples included in the paper.

System Prompt: You are an expert video analyst and cognitive scientist. Your task is to evaluate a generated video based on a specific psychological reasoning benchmark (MindWorldBench).

Inputs provided to the LVM: 1. A generated video. 2. A pair of controlled experimental cases (Case 1 and Case 2), which include the `world_state`, the character's `perception`, `desire`, `belief`, and the ground-truth `intention`. 3. The `target_case` that this specific video was generated to depict.

EVALUATION DIMENSIONS

Dimension 1: Visual Stability (Score: 1-5)

Evaluate the base generation quality and consistency. *Criteria:* Image Consistency (no hallucinated items popping into existence) and Perceptual Quality (stable temporal motion without severe morphing).

Dimension 2: Commonsense & Physics (Score: 1-5)

Evaluate the physical simulation and commonsense rationality of the action itself, regardless of the mental intention. *Criteria:* Physics (no severe physical violations like clipping) and Commonsense (behavior executed naturally).

Dimension 3: Mental Reasoning (Analysis & 3 Specific Questions)

Before evaluating the specific actions, you MUST analyze the psychological setup.

- **Step 3.0 (Mental Dynamics Analysis):** Explain what the controlled variable is and logically deduce why this specific difference leads to the different Intentions in Case 1 vs. Case 2.
- **Step 3.1 (Action Execution):** Did the character actually perform an action? (*Case_1_Intention / Case_2_Intention / Other_Action / No_Action*).
- **Step 3.2 (Intention Accuracy):** Does the executed action match the specific `intention` of the `target_case`? (*True / False*).
- **Step 3.3 (World State Maintenance):** Evaluate whether the video correctly upholds the objective physical reality during and after the interaction, strictly independent of the character's subjective beliefs. (*True / False / N/A*).

OUTPUT FORMAT (Strict JSON Enforcement):

```
{
  "visual_quality": { "reasoning": "...", "score": <1-5> },
  "commonsense_plausibility": { "reasoning": "...", "score": <1-5> },
  "mental_reasoning": {
    "dynamics_analysis": "Based on the cases, the controlled variable is...",
    "action_execution": { "reasoning": "...", "result": "..." },
    "intention_accuracy": { "reasoning": "...", "result": <boolean> },
    "world_state_maintenance": { "reasoning": "...", "result": "<true/false/NA>" }
  }
}
```

Figure 4: The structured Chain-of-Thought (CoT) Prompt utilized for the LVM-as-a-Judge pipeline. The prompt enforces sequential reasoning (analyzing the contrastive pair first) and structured JSON outputs.