MIND-VBENCH: A DIAGNOSTIC BENCHMARK FOR BDI-CONSISTENT BEHAVIOR IN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

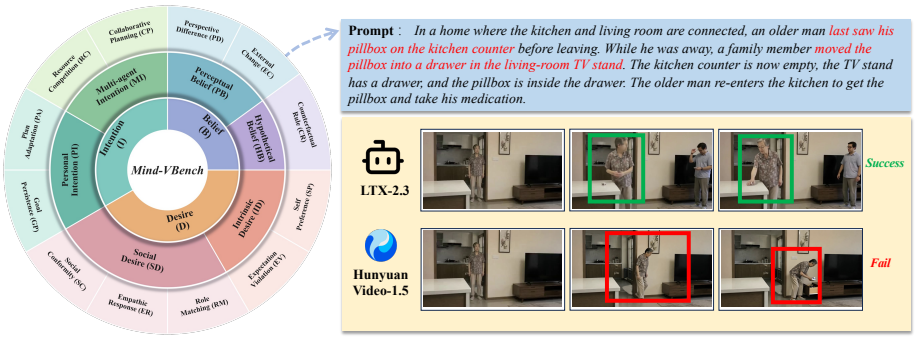
Recent video generators produce increasingly realistic and coherent videos, yet visually plausible videos may still depict characters acting inconsistently with the information available to them or with the goals they pursue. We study this discrepancy by evaluating whether generated character behavior remains consistent with the mental states established by a scenario. To make this question systematically evaluable in video generation, we introduce **Mind-VBench**, a diagnostic benchmark that operationalizes the Theory of Mind (ToM) question as observable behavioral constraints, using the Belief–Desire–Intention (BDI) framework from cognitive science and agent modeling. It contains 360 instances organized into three BDI components, six diagnostic dimensions, and twelve tasks. We evaluate representative video generation models under text-to-video and image-to-video settings. Current models frequently fail both to establish the situation required for behavioral judgment and to generate character behavior consistent with the specified BDI constraints. These findings show that current video generators do not reliably maintain mental-state consistency in character behavior, highlighting a limitation for human-centered world simulation. Our demo is available in the supplementary material.

1 INTRODUCTION

Recent video generators have made substantial progress in visual quality and temporal coherence. Yet visual realism does not ensure that characters act consistently with the information available to them or with the goals they pursue. A related gap between constructing a plausible scene and directing appropriate subsequent behavior has also been observed in image generation (Li et al., 2026). Consider the classic false-belief paradigm in Theory-of-Mind research, exemplified by the Sally–Anne task (Wimmer & Perner, 1983; Baron-Cohen et al., 1985). Figure 1 illustrates a representative instance of this type of task: an older man last saw his pillbox on the kitchen counter before leaving. While he was away, a family member moved it into a drawer in the living-room TV stand. When he returns to retrieve his medication, an information-consistent continuation should show him returning to the kitchen counter rather than going directly to the TV stand. A video generator may instead depict him walking directly to the TV stand, implicitly granting him access to information he never received. The resulting video may remain visually plausible, yet the character's action is inconsistent with his own informational perspective.

Existing video-generation benchmarks have expanded beyond visual quality to evaluate compositional dynamics, physical and implicit rules, verifiable outcomes, and social behavior (Sun et al., 2025; Luo et al., 2025; He et al., 2025; Liu et al., 2026a; Peng et al., 2026). Recent work further frames video generation as a multimodal reasoning paradigm rather than visual synthesis alone (Tong et al., 2026b). However, existing benchmarks have not systematically examined whether generated characters act according to their own understanding of a situation and the outcomes they seek. This question is closely related to Theory of Mind (ToM), which concerns how mental states shape behavior. Existing ToM benchmarks for video understanding, such as R^3-VQA (Niu et al., 2026), infer a character's mental states from observed actions. Video generation poses the reverse challenge: generating actions that remain consistent with the mental states established by a scenario. This capability is important for human-centered world simulation, where plausible scene dynamics must be accompanied by behavior consistent with characters' perspectives and goals.

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.



(a) Taxonomy of Mind-VBench (b) Example Prompt & Results

Figure 1: (a) Taxonomy. (b) False-belief Success and Fail examples.

To make this broad ToM question systematically evaluable in video generation, we adopt the Belief–Desire–Intention (BDI) framework and use it to construct a top-down evaluation taxonomy (Cohen & Levesque, 1990; Rao & Georgeff, 1991; Georgeff et al., 1999). Originating from theories of rational agency and practical reasoning, BDI models action selection in terms of what an agent believes, what it desires, and the goals or plans it intends to pursue. Belief represents the information available to a character, Desire represents the outcomes or states the character seeks, and Intention represents the goals or plans the character is committed to pursuing. This decomposition provides a principled structure for constructing diagnostic video-generation tasks, in which the mental states established by a scenario impose observable constraints on subsequent character behavior.

Building on this framework, we introduce **Mind-VBench**, a diagnostic benchmark for evaluating whether characters in generated videos act consistently with the mental states established by a scenario. The benchmark comprises three BDI components, six diagnostic dimensions, and twelve video-generation tasks, with 30 instances per task and 360 instances in total. As illustrated in Figure 1 and summarized in Table 1, the tasks cover character-specific information, preferences and expectations, social motivations, goal persistence, replanning, and multi-character coordination.

We evaluate representative video generation models under text-to-video (T2V) and image-to-video (I2V) settings, using the original prompts and an auxiliary Prompt Enhancement (PE) condition that explicitly provides the expected continuation. We treat PE as an assisted generation condition rather than evidence of stronger mental-state reasoning. Current models frequently fail to establish situations in which character behavior can be clearly judged, and even valid generations often contain actions inconsistent with the specified BDI constraints. First-frame conditioning mainly improves scene establishment, whereas explicitly specifying the expected continuation produces larger gains in behavioral success but does not eliminate the remaining failures. These results show that current video generators remain limited in producing mental-state-consistent character behavior, highlighting a limitation for human-centered world simulation.

Our main contributions are summarized as follows:

- We formulate ToM-oriented character behavior as an evaluation problem for video generation and introduce a BDI-grounded diagnostic framework that assesses whether generated character behavior is consistent with character-specific mental states, using a validity gate that separates scene establishment from behavioral consistency.
- We instantiate this framework as **Mind-VBench**, a 360-instance benchmark spanning three BDI components, six diagnostic dimensions, and twelve video-generation tasks covering informational perspectives, preferences, social motivations, planning, and multi-character coordination.
- Through extensive evaluations of representative video generation models under T2V and I2V settings, we show that current models struggle with both scene establishment and mental-state-consistent character behavior.

2 Related Work

Reasoning-oriented evaluation for video generation. Video-generation evaluation has expanded beyond perceptual quality to compositional fidelity, temporal consistency, physical plausibility, rule following, and socially coherent behavior. StoryBench (Bugliarello et al., 2023) studies continuous story visualization, while VBench (Huang et al., 2024) and VBench-2.0 (Zheng et al., 2025) assess general generation quality and intrinsic faithfulness, respectively. T2V-CompBench (Sun et al., 2025) evaluates compositional generation across attribute and motion binding, spatial relations, object interactions, and numeracy, and TC-Bench (Feng et al., 2025) focuses on temporal compositionality. VideoPhy (Bansal et al., 2024), PhyGenBench (Meng et al., 2025), and PhyWorldBench (Gu et al., 2026) examine physical commonsense and physical realism in generated videos. V-ReasonBench (Luo et al., 2025) and TiViBench (Chen et al., 2026) study verifiable and think-in-video reasoning, while RULER-Bench (He et al., 2025) evaluates rule following across diverse cognitive rules. Complementary video-understanding benchmarks study reasoning across time and broad video reasoning capabilities (Chen et al., 2024a; Wang et al., 2026), but do not evaluate whether a generator realizes such reasoning through character behavior. RISE-Video (Liu et al., 2026a) evaluates whether models can infer and realize implicit world rules from multimodal conditions. SVBench (Peng et al., 2026) extends video-generation evaluation to socially coherent behavior through psychology-grounded social cognition paradigms. Mind-VBench focuses on whether character actions are consistent with specified beliefs, desires, and intentions, using a BDI-derived taxonomy that separately evaluates scene establishment and behavioral consistency. Operationally, Mind-VBench terminates prompts at an unresolved decision point under explicit epistemic asymmetry and scores the action through a two-stage prerequisite gate. A detailed comparison with representative benchmarks is provided in Appendix Table 1.

Theory of Mind and mental-state reasoning. Theory of Mind (ToM) concerns reasoning about agents' mental states and their consequences for behavior (Premack & Woodruff, 1978; Dennett, 1989; Wellman, 1990; Malle, 2006b). Developmental research has further documented a broad range of ToM measures spanning emotions, desires, intentions, knowledge, and beliefs (Beaudoin et al., 2020). ToMBench (Chen et al., 2024b), Hi-ToM (Wu et al., 2023), CogToM (Tong et al., 2026a), and OmniToM (Bawatneh et al., 2026) evaluate complementary aspects of ToM reasoning in language models, while OpenToM (Xu et al., 2024) and ExploreToM (Sclar et al., 2025) construct naturalistic and controllable narratives for mental-state reasoning. These studies also expose persistent difficulty in separating beliefs from knowledge and facts (Suzgun et al., 2025). EgoToM (Li et al., 2025) evaluates goal inference, belief reasoning, and next-action prediction from egocentric videos, while recent video-focused work studies ToM in multimodal video models (Chen et al., 2025; Liu et al., 2026b). MMToM-QA (Jin et al., 2024) and MuMA-ToM (Shi et al., 2025) extend ToM evaluation to multimodal observations and embodied multi-agent interactions. Social Genome (Mathur et al., 2025) examines grounded social reasoning, while R^3-VQA (Niu et al., 2026) evaluates beliefs, desires, intentions, emotions, and their causal relations in socially complex videos. Accounts of intentional action connect an agent's reasons to its subsequent behavior (Davidson, 1963). The Belief–Desire–Intention (BDI) framework organizes mental states into beliefs as information available to an agent, desires as preferred outcomes, and intentions as committed goals or plans (Cohen & Levesque, 1990; Rao & Georgeff, 1991; 1995; Georgeff et al., 1999). BDI also builds on the broader agent-oriented programming tradition and has been extended to model conflicts among beliefs, obligations, intentions, and desires (Shoham, 1993; Broersen et al., 2001). BDIQA (Mao et al., 2024) applies this formulation to belief, desire, and intention reasoning in video question answering. Mind-VBench adopts BDI as the top-level structure for video-generation evaluation and derives a systematic taxonomy for assessing whether specified mental states are consistently reflected in observable character behavior.

3 Mind-VBench Benchmark

3.1 Task Formulation and Taxonomy

To translate the broad Theory of Mind (ToM) question introduced above into evaluable behavioral constraints, we adopt the Belief–Desire–Intention (BDI) framework (Cohen & Levesque, 1990; Rao & Georgeff, 1991; Georgeff et al., 1999). Mind-VBench evaluates whether video generation models

can produce character behavior consistent with these constraints. Specifically, characters should act according to the information available to them, pursue outcomes consistent with their preferences, and maintain or adapt their established goals and plans. We organize these constraints into a top-down taxonomy containing three BDI components, six diagnostic dimensions, and twelve tasks. The complete taxonomy and task definitions are summarized in Table 1.

Belief. Belief concerns which information can guide a character's behavior, following the broader distinction between an agent's representation of the world and objective reality in developmental ToM research (Gopnik & Wellman, 1992; Wellman & Liu, 2004). We consider personal belief, grounded in the character's own observations, and hypothetical belief, established through an explicitly specified nonstandard rule. Appendix Figure 1 illustrates a personal-belief example in which an object is moved while the focal character is absent. A belief-consistent continuation should show the character acting according to the most recent state it observed rather than information available only to the viewer.

Desire. Desire captures the outcomes or states preferred by a character. The individual dimension covers personal preferences and expectations, whereas the social dimension captures motivations arising from social roles, the needs of others, and group influence, including observable helping and cooperative responses (Malle, 2006a; Warneken & Tomasello, 2006; 2007). Appendix Figure 2 presents an individual-desire example in which the focal character has established a preference between two options. A desire-consistent continuation should show the character selecting the preferred option rather than making an arbitrary choice.

Intention. Intention concerns the goals or plans that a character remains committed to pursuing, consistent with planning accounts in which intentions organize action over time (Bratman, 1987). Personal intention focuses on goal persistence and replanning by a single character, while multi-agent intention considers competition and cooperation among multiple characters, drawing on the notion of shared intentionality in coordinated action (Tomasello & Carpenter, 2007). Appendix Figure 4 presents a personal-intention example in which the character's original route toward a target becomes blocked. An intention-consistent continuation should preserve the original goal while adopting an alternative route or action to achieve it.

3.2 Benchmark Construction

Building on the taxonomy defined in Section 3.1, we construct Mind-VBench for both text-to-video (T2V) and image-to-video (I2V) evaluation through the standardized pipeline shown in Figure 2. Each of the twelve tasks contains 30 instances, yielding 360 instances in total. A T2V instance contains a text prompt from which the model generates both the scenario setup and subsequent character behavior. An I2V instance additionally includes a first frame that establishes the initial scene, from which the model generates the continuation. In both settings, the expected action remains unspecified. Each instance is paired with an evaluation specification introduced in Section 3.3.

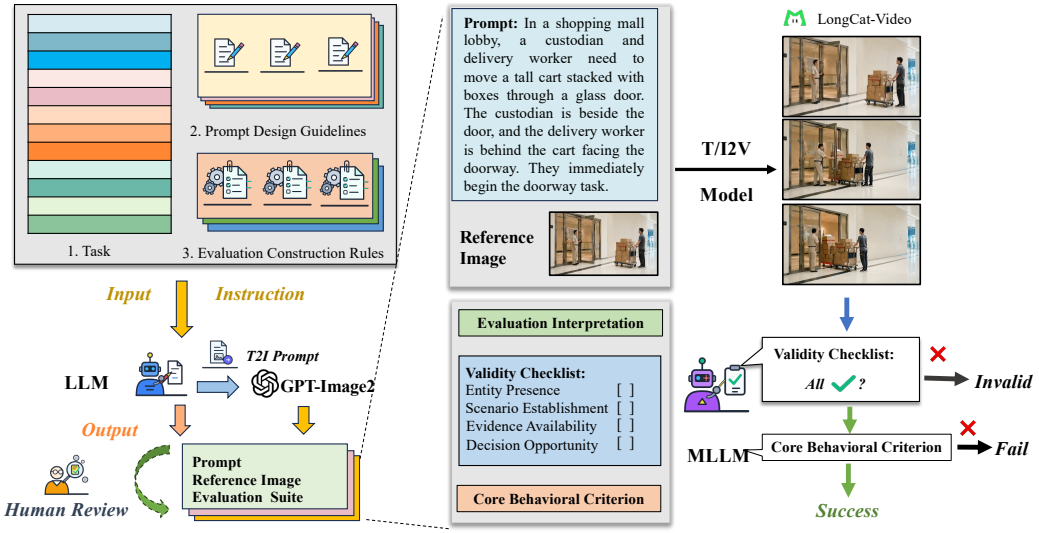
To construct instances consistently across tasks, we design a unified construction instruction consisting of three components: (1) a task definition specifying the target BDI constraint; (2) prompt design guidelines for constructing diverse scenarios that support unambiguous behavioral evaluation; and (3) task-specific requirements defining the information, preference, or goal relations that must be established. We use GPT-5.4 (OpenAI, 2026b) to generate candidate video prompts following these requirements. The complete construction instruction and task-specific templates are provided in Appendix B.

Each generated prompt specifies the relevant characters, objects, states, observations, preferences, and goals. The prompt establishes the behavioral decision point without specifying the expected continuation. The prompt design guidelines encourage variation in characters, environments, objects, and event structures while preserving the diagnostic condition of each task. We avoid explicit BDI terminology, unstable on-screen text, fine-grained identity requirements, and unnecessary visual details that may introduce generation failures unrelated to the target behavior.

For I2V, each prompt is paired with a corresponding first frame generated by gpt-image2 (OpenAI, 2026a). The frame establishes the characters, objects, spatial relations, and information boundaries required by the scenario while leaving the evaluated behavior unresolved. It represents the state be-

Table 1: Mind-VBench tasks grouped by diagnostic dimension and BDI component.

BDI	Dimension	Task	Definition
Belief (**B**)	*Perceptual Belief* (**PB**)	*Perspective Difference* (**PD**)	*Act according to information available from a currently restricted perspective.*
		External Change (**EC**)	*Act based on outdated information after an unobserved change.*
	Hypothetical Belief (**HB**)	*Counterfactual Rule* (**CR**)	*Follow the specified counterfactual rule.*
Desire (**D**)	*Intrinsic Desire* (**ID**)	*Self Preference* (**SP**)	*Select according to personal preference.*
		Expectation Violation (**EV**)	*Adapt behavior after unexpected outcomes.*
	Social Desire (**SD**)	*Role Matching* (**RM**)	*Differentiate actions in one event according to visible social roles.*
		Empathic Response (**ER**)	*Respond to another character's visible difficulty or need.*
		Social Conformity (**SC**)	*Align behavior with relevant actions of other agents.*
Intention (**I**)	*Personal Intention* (**PI**)	*Goal Persistence* (**GP**)	*Maintain an established goal after an external interruption.*
		Plan Adaptation (**PA**)	*Adapt the plan when the original course of action is blocked.*
	Multi-agent Intention (**MI**)	*Resource Competition* (**RC**)	*Compete among multiple agents for limited resources.*
		Collaborative Planning (**CP**)	*Coordinate complementary actions toward a shared task.*



(a) Data Generation (b) Evaluation

Figure 2: Mind-VBench construction and evaluation pipeline.

fore the focal character acts and must not initiate or complete the expected action. We further enforce task-specific conditions, such as keeping hidden information occluded, blocked routes inaccessible, and competitive or cooperative situations unresolved. For the PE condition, GPT-5.6 (OpenAI, 2026c) rewrites the prompt to state the expected continuation.

3.3 Evaluation and Scoring

Evaluation Design. Each instance is paired with an evaluation specification containing four scene prerequisites and one core behavioral criterion. The scene prerequisites determine whether the generated video establishes the conditions required for behavioral judgment. They examine whether (1) the required entities and base scene are present, (2) the task-defining event or state is established, (3) the information and relations needed for judgment are observable, and (4) the focal character reaches a clear decision point with sufficient opportunity to act. Each prerequisite is evaluated as a binary decision, and all four must be satisfied for a video to be considered valid.

The core behavioral criterion specifies the observable behavior that satisfies the target BDI constraint and identifies alternative behaviors that violate it. Each generated video is labeled as *Invalid*, *Fail*, or *Success*. A video is labeled *Invalid* if any scene prerequisite is not satisfied. A valid video is labeled *Success* if the focal character's behavior satisfies the core behavioral criterion and *Fail* otherwise. The fourth prerequisite requires only that the character has a reasonable opportunity to act. Therefore, inaction, omission, or failed coordination after a valid decision point is treated as behavioral failure rather than invalidity.

All candidate generation inputs and their corresponding evaluation specifications are independently reviewed by two human annotators before inclusion in the benchmark. The annotators examine whether each instance matches its assigned task, whether the expected action is inadvertently revealed, whether the behavioral decision remains unresolved, and whether the prompt and, where applicable, the first frame correctly establish the intended information, preference, or goal relations. They also verify that the scene prerequisites and behavioral criterion are clear and consistent with the generation input. Disagreements are resolved through discussion. Instances containing scene ambiguity, target-action leakage, generation difficulty unrelated to the target task, or unclear evaluation specifications are revised or removed. Detailed evaluation rules and representative specifications are provided in Appendix C.

Metrics. For each task t, let N_t denote the total number of generated videos, V_t the number of valid videos, and S_t the number of successful videos. We compute

$$\mathrm{VR}_t = \frac{V_t}{N_t}, \qquad \mathrm{CSR}_t = \frac{S_t}{V_t}, \qquad \mathrm{OSR}_t = \frac{S_t}{N_t}. \tag{1}$$

Validity Rate (VR) measures whether the model successfully establishes the situation required for behavioral judgment. Conditional Success Rate (CSR) measures the proportion of valid videos whose character behavior satisfies the corresponding BDI constraint. Overall Success Rate (OSR) measures whether the model both establishes the required situation and produces BDI-consistent behavior, and is used as the primary metric.

To obtain component-level results, we average the task-level scores within each BDI component. The overall score is the macro-average across the twelve tasks; it therefore weights tasks equally rather than weighting the three BDI components equally. CSR is not reported for a task when no valid video is generated. The four scene-prerequisite pass rates are reported separately for diagnostic analysis and are not included in the overall score.

4 Experiments

4.1 Experimental Setup

Evaluated Models. We evaluate representative video generation models under text-to-video (T2V) and image-to-video (I2V) settings, covering both open-source and closed-source systems. The open-source models include Wan2.2-TI2V-5B, Wan2.2-T2V-A14B (Wan Team et al., 2025), LTX-2.3-22B (HaCohen et al., 2026), LongCat-Video-13.6B (Meituan LongCat Team et al., 2025), and HunyuanVideo-1.5-8.3B (Wu et al., 2025). We additionally evaluate the closed-source systems Kling-3.0 (Kling Team et al., 2025), HappyHorse-1.1 (Happy Horse, 2026), and Dreamina-Seedance-2.0 (Seedance Team et al., 2026) under their available generation settings (a few closed-source runs are blocked by provider safety checks; see Appendix E). All models are included in the original T2V benchmark comparison. Analyses involving additional generation conditions are restricted to models for which the corresponding settings can be applied consistently.

Table 2: T2V OSR without PE; shaded columns are BDI averages and Overall is the twelve-task average. Per-task VR and CSR breakdowns are provided in Appendix E.

Model	Belief				Desire						Intention					Overall
	PD	EC	CR	*Avg.*	SP	EV	RM	ER	SC	*Avg.*	GP	PA	RC	CP	*Avg.*	
Open-Source Models																
Wan2.2-TI2V-5B	**43.3**	13.3	0.0	18.9	23.3	6.7	23.3	16.7	10.0	16.0	13.3	3.3	36.7	46.7	25.0	19.7
Wan2.2-T2V-A14B	33.3	10.0	**3.3**	15.6	16.7	16.7	13.3	13.3	10.0	14.0	**23.3**	6.7	33.3	26.7	22.5	17.2
LTX-2.3-22B	36.7	**23.3**	**3.3**	**21.1**	**56.7**	**30.0**	**63.3**	**23.3**	**26.7**	**40.0**	13.3	**13.3**	**56.7**	**56.7**	**35.0**	**33.6**
LongCat-Video-13.6B	6.7	10.0	**3.3**	6.7	20.0	13.3	20.0	10.0	6.7	14.0	20.0	10.0	20.0	53.3	25.8	16.1
HunyuanVideo-1.5-8.3B	30.0	16.7	0.0	15.6	16.7	13.3	10.0	6.7	13.3	12.0	6.7	**13.3**	26.7	36.7	20.8	15.8
Closed-Source Models																
Kling-3.0	53.3	43.3	50.0	48.9	73.3	43.3	26.7	23.3	**30.0**	39.3	50.0	33.3	36.7	**83.3**	50.8	45.6
HappyHorse-1.1	53.3	**56.7**	56.7	**55.6**	**83.3**	76.7	**96.7**	**46.7**	**30.0**	**66.7**	**70.0**	30.0	**73.3**	76.7	**62.5**	**62.5**
Dreamina-Seedance-2.0	**63.3**	33.3	**62.1**	52.9	80.0	**83.3**	73.3	16.7	16.7	54.0	**70.0**	**36.7**	36.7	73.3	54.2	53.8

Table 3: VR/CSR/OSR under four conditions; parentheses show signed percentage-point changes from T2V. Closed-source systems use provider-default prompt engineering; full T2V/I2V results are in Appendix E.

	Model	Setting	VR	CSR	OSR
Open Source	LTX-2.3-22B	T2V	55.6	58.6	33.6
		PE-T2V	66.9 (+11.4)	79.7 (+21.1)	53.3 (+19.7)
		I2V	74.2 (+18.6)	52.5 (-6.1)	40.3 (+6.7)
		PE-I2V	83.6 (+28.1)	78.1 (+19.5)	65.3 (+31.7)
	HunyuanVideo-1.5-8.3B	T2V	28.1	60.4	15.8
		PE-T2V	36.4 (+8.3)	79.4 (+19.0)	28.6 (+12.8)
		I2V	66.1 (+38.1)	48.7 (-11.7)	32.8 (+16.9)
		PE-I2V	76.9 (+48.9)	80.4 (+20.0)	61.7 (+45.8)
Closed Source	Kling-3.0	T2V	76.4	60.6	45.6
		I2V	81.9 (+5.5)	60.9 (+0.3)	50.2 (+4.7)
	HappyHorse-1.1	T2V	83.9	74.4	62.5
		I2V	84.9 (+1.0)	76.3 (+1.9)	65.3 (+2.8)

Evaluation Protocol. For each model–condition–instance tuple, we generate one video and evaluate it using the corresponding instance-specific evaluation specification. Following the aggregation protocol defined in Section 3.3, we report component-level Validity Rate (VR), Conditional Success Rate (CSR), and twelve-task macro-averaged Overall Success Rate (OSR). OSR serves as the primary metric. The unassisted conditions are the primary benchmark setting and measure BDI-consistent behavior without an explicit behavioral target; the PE conditions are evaluated separately as an oracle-style diagnostic condition.

Implementation Details. All models are evaluated on the same benchmark instances with condition-specific inputs. The target video duration is 10 seconds. Detailed implementation information is provided in Appendix F.

VR/CSR Decomposition. We analyze OSR through VR and CSR to identify where failures occur. The analysis shows that current models fail at both stages: they often do not establish a judgeable situation, and even valid generations can still contain behavior inconsistent with the target BDI constraint. As shown in Table 3, some models achieve relatively high VR but low CSR, indicating that their main limitation lies in behavioral consistency after scene establishment. Other models show lower VR but relatively higher CSR, suggesting that their failures are dominated by missing entities, incorrect relations, incomplete information states, or unclear decision opportunities.



(a) HunyuanVideo-1.5 on MI-CP-02 (Collaborative Planning: move a washing machine over a threshold).



(b) LTX-2.3 on HB-CR-30 (Counterfactual Rule: drawn symbols become physical tokens).

Figure 3: Effects of I2V and PE. I2V makes MI-CP-02 valid but unsuccessful, whereas PE-I2V succeeds; for HB-CR-30, PE enables the counterfactual behavior.

These results show that improving Mind-VBench performance requires addressing both scene-level validity and behavior-level consistency. A model must first generate the required situation, but it must also maintain the character-specific mental-state constraint when producing the subsequent action.

Model Variants and Video Duration. We additionally examine model variants and video duration. Under the original T2V setting, Wan2.2-TI2V-5B and Wan2.2-T2V-A14B achieve OSR scores of 19.7% and 17.2%, respectively. This comparison does not reveal a monotonic association between nominal parameter count and benchmark performance, although the dense TI2V and MoE T2V variants also differ in architecture, active parameters, and conditioning objectives.

In additional duration experiments, LongCat-Video-13.6B exhibits a moderate decline as video duration increases from 5 to 15 seconds, with OSR decreasing from 16.7% to 14.4%. At 20 seconds, OSR further drops to 6.9%, accompanied by decreases in both VR and CSR. For this model, longer generation horizons make it more difficult to maintain both scene validity and behavioral consistency.

4.2 More Analysis

Impact of First-frame Conditioning. We examine whether first-frame conditioning improves BDI-consistent behavior by comparing T2V and I2V in Table 3. The main effect of I2V is to improve scene validity rather than behavioral success. Across the eight evaluated models, I2V increases average VR by 17.4 percentage points, with gains for seven models. The largest improvements are observed for HunyuanVideo-1.5, LongCat-Video-13.6B, and Wan2.2-A14B, whose VR scores increase by 38.1, 31.1, and 24.7 percentage points, respectively. However, this improvement in validity does not consistently translate into higher CSR. Average CSR decreases by 3.7 percentage points, and five of the eight models obtain lower CSR under I2V. For example, HunyuanVideo-1.5 improves in VR from 28.1% to 66.1%, while its CSR decreases from 60.4% to 48.7%. This conditional decrease may partly reflect denominator expansion, because I2V admits additional and potentially more difficult instances into the valid pool. Aggregate values are computed over the full eight-model suite; representative entries are shown in Table 3. These results show that first-frame conditioning helps establish judgeable situations, but does not by itself solve the problem of generating behavior consistent with the specified BDI constraints.

Impact of Prompt Enhancement. We next turn to the Prompt Enhancement (PE) condition. Unlike first-frame conditioning, PE improves both scene validity and behavioral success when the expected continuation is explicitly provided. As shown in Table 3, PE-T2V improves average VR, CSR, and OSR over T2V by 10.5, 19.3, and 16.9 percentage points, respectively, and PE-I2V improves the corresponding metrics over I2V by 11.9, 25.2, and 25.7 percentage points. PE still leaves behavioral failures on some instances, indicating that the failures of unassisted generation cannot be attributed solely to the absence of an explicit behavioral target.

The largest gains appear on Desire-related tasks, whose average OSR increases by 20.4 percentage points under PE-T2V and 31.1 percentage points under PE-I2V. Belief- and Intention-related tasks also improve, but the gains are less uniform across models.

Case Study. Figure 3 illustrates how first-frame conditioning and Prompt Enhancement affect different failure sources. In MI-CP-02, HunyuanVideo-1.5 remains invalid under T2V and PE-T2V, and only PE-I2V satisfies the collaborative-planning criterion; in HB-CR-30, LTX-2.3 produces valid scenes but completes the counterfactual continuation only under PE. These examples show that I2V can turn invalid generations into judgeable scenes, while PE is more directly associated with behavioral success in the illustrated cases.

4.3 Human Evaluation Alignment

Following recent work on scalable MLLM-based video evaluation (Waheed et al., 2026), we assess the reliability of our automatic evaluation protocol through a human alignment study on a randomly sampled subset of 20 videos covering all twelve tasks and different model groups. Each video is independently rated by 12 human annotators using the same evaluation specifications as the automatic evaluator, with final labels determined by majority voting.

We compare automatic and human judgments for scene validity and behavioral success using accuracy and Cohen's κ; detailed results are provided in Appendix D.

On the alignment subset, Qwen3.5-27B (Qwen Team, 2026) reaches 83.3% accuracy and Cohen's $\kappa = 0.571$ for behavioral success, the strongest behavioral alignment among the evaluated judges. Given both its alignment performance and inference cost, we adopt Qwen3.5-27B as the main evaluator.

5 Conclusion

We introduce Mind-VBench, a benchmark for evaluating mental-state-consistent character behavior generation in video generation models. Grounded in the Belief–Desire–Intention framework, Mind-VBench assesses whether generated characters act consistently with their available information, preferred outcomes, and committed goals, while separating scene establishment from behavioral consistency to enable more interpretable failure analysis. Experiments on representative video generation models show that current models still struggle both to establish judgeable situations and to produce actions consistent with specified BDI constraints. First-frame conditioning mainly improves scene establishment, whereas Prompt Enhancement substantially improves behavioral success, highlighting how much behavioral consistency can be recovered by explicitly specifying the expected continuation. Overall, Mind-VBench provides a foundation for studying more realistic, socially coherent, and goal-consistent character generation.

Limitations. Mind-VBench uses controlled scenarios and explicit task specifications, but does not cover open-ended social interactions or long-horizon planning. The benchmark measures observable behavior; BDI serves as an organizational and diagnostic framework, rather than evidence about internal mental-state representations. Each model–condition–instance tuple is generated once, so the reported rates do not characterize generation-level stochasticity. Because Prompt Enhancement supplies the expected continuation, its results describe an assisted condition and do not by themselves separate mental-state consistency from generic instruction following or action execution; we therefore do not attribute the reported failures to a single underlying mechanism. Future work may extend the benchmark to broader interaction settings, richer social dynamics, and more comprehensive human-aligned evaluation.

GENERATIVE AI DISCLOSURE

Generative AI tools were used during benchmark development for candidate scenario generation, instance expansion, first-frame prompt rewriting, Prompt Enhancement rewriting, and automatic video evaluation. They were also used for limited language editing and LaTeX formatting during manuscript preparation. All AI-assisted benchmark materials were reviewed through the human verification and evaluation procedures described in the paper and appendix. The authors verified the final manuscript, benchmark specifications, experimental results, and released artifacts, and take full responsibility for their accuracy, originality, and integrity.

REPRODUCIBILITY STATEMENT

Mind-VBench contains 360 instances organized into twelve tasks under three BDI components and six diagnostic dimensions. The main paper defines the task formulation, instance construction process, evaluation protocol, and the VR, CSR, and OSR metrics. The accompanying appendix provides the taxonomy and task details, construction and quality-control procedures, evaluation criteria, human-alignment protocol, model-variant and duration analyses, and qualitative results.

For each model–condition–instance tuple, the evaluation uses one generated video and an instance-specific evaluation specification. Our benchmark, code, automatic evaluation results, and example videos are available in the supplementary material. The release will include sufficient documentation to reproduce the benchmark construction and evaluation pipeline. Any unavailable closed-source model configuration will be identified explicitly, together with the provider and interface settings used for evaluation, so that the resulting limitations are clear.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. URL http://arxiv.org/abs/2502.13923.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating Physical Commonsense for Video Generation. In *Thirty-Eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://neurips.cc/virtual/2024/103544.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46, 1985. doi: 10.1016/0010-0277(85)90022-8. URL https://www.sciencedirect.com/science/article/pii/0010027785900228.

Adam Bawatneh, Sagar Sapkota, Amrit Singh Bedi, Santu Karmaker, and Mubarak Shah. OmniToM: Benchmarking Theory of Mind in LLMs via Explicit Belief Modeling, 2026. URL http://arxiv.org/abs/2605.26322.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Frontiers in Psychology*, 10:2905, 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.02905. URL https://www.frontiersin.org/article/10.3389/fpsyg.2019.02905/full.

Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987. ISBN 9780674458185.

Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the Fifth International Conference on Autonomous Agents*, AGENTS '01, pp. 9–16, New York, NY, USA, 2001. Association for Computing Machinery. doi: 10.1145/375735.375766. URL https://dl.acm.org/doi/10.1145/375735.375766.

Emanuele Bugliarello, Hernan Moraldo, Ruben Villegas, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Han Zhang, Dumitru Erhan, Vittorio Ferrari, Pieter-Jan Kindermans, and Paul Voigtlaender. StoryBench: A Multifaceted Benchmark for Continuous Story Visualization. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=AwhpBEqmyo.

Harold Haodong Chen, Disen Lan, Wen-Jie Shu, Qingyang Liu, Zihan Wang, Sirui Chen, Wenkai Cheng, Kanghao Chen, Hongfei Zhang, Zixin Zhang, Rongjin Guo, Yu Cheng, and Ying-Cong Chen. TiViBench: Benchmarking Think-in-Video Reasoning for Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11403–11413, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_TiViBench_Benchmarking_Think-in-Video_Reasoning_for_Video_Generation_CVPR_2026_paper.html.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. ReXTime: A Benchmark Suite for Reasoning-Across-Time in Videos. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=4Vhc7uPHjn#discussion.

Zhanwen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. Through the Theory of Mind's Eye: Reading Minds with Multimodal Video Large Language Models, 2025. URL http://arxiv.org/abs/2406.13763.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking Theory of Mind in Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983, Bangkok, Thailand, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.847. URL https://aclanthology.org/2024.acl-long.847/.

Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2–3):213–261, 1990. ISSN 00043702. doi: 10.1016/0004-3702(90)90055-5. URL https://linkinghub.elsevier.com/retrieve/pii/0004370290900555.

Donald Davidson. Actions, Reasons, and Causes. *The Journal of Philosophy*, 60(23):685–700, 1963. ISSN 0022-362X. doi: 10.2307/2023177. URL https://www.jstor.org/stable/2023177.

Daniel C. Dennett. *The Intentional Stance*. The MIT Press, 1989. ISBN 9780262540537. URL https://mitpress.mit.edu/9780262540537/the-intentional-stance/.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. TC-Bench: Benchmarking Temporal Compositionality in Conditional Video Generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4638–4662, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.241. URL https://aclanthology.org/2025.findings-acl.241/.

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The Belief-Desire-Intention Model of Agency. In Jörg P. Müller, Anand S. Rao, and Munindar P. Singh (eds.), *Intelligent Agents V: Agents Theories, Architectures, and Languages*, volume 1555, pp. 1–10. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999. doi: 10.1007/3-540-49057-4_1. URL http://link.springer.com/10.1007/3-540-49057-4_1.

Google. Gemini 3.1 Pro: Announcing our latest Gemini AI model. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026. Accessed: 2026-07-31.

Alison Gopnik and Henry M. Wellman. Why the Child's Theory of Mind Really Is a Theory. *Mind & Language*, 7(1–2):145–171, 1992. ISSN 1468-0017. doi: 10.1111/j.1468-0017.1992.tb00202.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0017.1992.tb00202.x.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. PhyWorldBench: A Comprehensive Evaluation of Physical Realism in Text-to-Video Models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rlZeILv3fm.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient Joint Audio-Visual Foundation Model, 2026. URL http://arxiv.org/abs/2601.03233.

Happy Horse. Happy horse: The simple, all-in-one ai video platform. https://www.happyhorse.com/, 2026. Accessed: 2026-07-29.

Xuming He, Zehao Fan, Hengjia Li, Fan Zhuo, Hankun Xu, Senlin Cheng, Di Weng, Haifeng Liu, Can Ye, and Boxi Wu. RULER-Bench: Probing Rule-based Reasoning Abilities of Next-level Video Generation Models for Vision Foundation Intelligence, 2025. URL http://arxiv.org/abs/2512.02622.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal Theory of Mind Question Answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16077–16102, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.851. URL https://aclanthology.org/2024.acl-long.851/.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-Omni Technical Report, 2025. URL http://arxiv.org/abs/2512.16776.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier Painting Than Thinking: Can Text-to-Image Models Set the Stage, but Not Direct the Play? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iqAFhWistW.

Yuxuan Li, Vijay Veerabadran, Michael L. Iuzzolino, Brett D. Roads, Asli Celikyilmaz, and Karl Ridgeway. EgoToM: Benchmarking Theory of Mind Reasoning from Egocentric Videos, 2025. URL http://arxiv.org/abs/2503.22152.

Mingxin Liu, Shuran Ma, Shibei Meng, Xiangyu Zhao, Zicheng Zhang, Shaofeng Zhang, Zhihang Zhong, Peixian Chen, Haoyu Cao, Xing Sun, Haodong Duan, and Xue Yang. RISE-Video: Can

Video Generators Decode Implicit World Rules?, 2026a. URL http://arxiv.org/abs/2602.05986.

Siqi Liu, Xinyang Li, Bochao Zou, Junbao Zhuo, Huimin Ma, and Jiansheng Chen. Video-Only ToM: Enhancing Theory of Mind in Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19208–19218, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Video-Only_ToM_Enhancing_Theory_of_Mind_in_Multimodal_Large_Language_CVPR_2026_paper.html.

Yang Luo, Xuanlei Zhao, Baijiong Lin, Lingting Zhu, Liyao Tang, Yuqi Liu, Ying-Cong Chen, Shengju Qian, Xin Wang, and Yang You. V-ReasonBench: Toward Unified Reasoning Benchmark Suite for Video Generation Models, 2025. URL http://arxiv.org/abs/2511.16668.

Bertram F. Malle. Folk Theory of Mind: Conceptual Foundations of Human Social Cognition. In Ran R. Hassin, James S. Uleman, and John A. Bargh (eds.), *The New Unconscious*, pp. 225–255. Oxford University Press, 2006a. doi: 10.1093/acprof:oso/9780195307696.003.0010. URL https://doi.org/10.1093/acprof:oso/9780195307696.003.0010.

Bertram F. Malle. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT Press, 2006b.

Yuanyuan Mao, Xin Lin, Qin Ni, and Liang He. BDIQA: A New Dataset for Video Question Answering to Explore Cognitive Reasoning through Theory of Mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 583–591, 2024. doi: 10.1609/aaai.v38i1.27814. URL https://doi.org/10.1609/aaai.v38i1.27814.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social Genome: Grounded Social Reasoning Abilities of Multimodal Models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24868–24891, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1264. URL https://aclanthology.org/2025.emnlp-main.1264/.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. LongCat-Video Technical Report, 2025. URL http://arxiv.org/abs/2510.22200.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards World Simulator: Crafting Physical Commonsense-Based Benchmark for Video Generation. In *Forty-Second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dIjMswSzgF.

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.

OpenAI. Introducing ChatGPT Images 2.0. https://openai.com/index/introducing-chatgpt-images-2-0/, 2026a. Accessed: 2026-07-31.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026b. Accessed: 2026-07-31.

OpenAI. GPT-5.6 System Card. https://deploymentsafety.openai.com/gpt-5-6, 2026c. Accessed: 2026-07-31.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32872–32881, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Peng_SVBench_Evaluation_of_Video_Generation_Models_on_Social_Reasoning_CVPR_2026_paper.html.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978. doi: 10.1017/S0140525X00076512. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0.

Qwen Team. Qwen3.5-Omni Technical Report, 2026. URL https://arxiv.org/abs/2604.15804.

Anand S. Rao and Michael P. Georgeff. Modeling rational agents within a BDI-architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, KR'91, pp. 473–484, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

Anand S. Rao and Michael P. Georgeff. BDI Agents: From Theory to Practice. In Victor R. Lesser and Les Gasser (eds.), *Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995, San Francisco, California, USA*, pp. 312–319. The MIT Press, 1995.

Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. Explore Theory of Mind: Program-guided adversarial data generation for theory of mind reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=246rHKUnnf.

Seedance Team, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing Video Generation for World Complexity, 2026. URL http://arxiv.org/abs/2604.14148.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. MuMA-ToM: Multi-modal multi-agent theory of mind. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, volume 39 of *AAAI'25/IAAI'25/EAAI'25*, pp. 1510–1519. AAAI Press, 2025. doi: 10.1609/aaai.v39i2.32142. URL https://doi.org/10.1609/aaai.v39i2.32142.

Yoav Shoham. Agent-oriented programming. *Artificial Intelligence*, 60(1):51–92, 1993. ISSN 0004-3702. doi: 10.1016/0004-3702(93)90034-9. URL https://www.sciencedirect.com/science/article/pii/0004370293900349.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-video_Generation_CVPR_2025_paper.html.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E. Ho, Thomas Icard, Dan Jurafsky, and James Zou. Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence*, 7(11):1780–1790, 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01113-8. URL https://www.nature.com/articles/s42256-025-01113-8.

Michael Tomasello and Malinda Carpenter. Shared intentionality. *Developmental Science*, 10(1): 121–125, 2007. ISSN 1467-7687. doi: 10.1111/j.1467-7687.2007.00573.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7687.2007.00573.x.

Haibo Tong, Zeyang Yue, Feifei Zhao, Erliang Lin, Lu Jia, Ruolin Chen, Yinqian Sun, Qian Zhang, and Yi Zeng. CogToM: A Comprehensive Theory of Mind Benchmark inspired by Human Cognition for Large Language Models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31383–31425, San Diego, California, United States, 2026a. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1448. URL https://aclanthology.org/2026.acl-long.1448/.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, Xinchi Chen, Jun Zhao, Xuanjing Huang, and Xipeng Qiu. Thinking with Video: Video Generation as a Promising Multimodal Reasoning Paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41121–41129, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Tong_Thinking_with_Video_Video_Generation_as_a_Promising_Multimodal_Reasoning_CVPR_2026_paper.html.

Abdul Waheed, Zhen Wu, Dareen Safar Alharthi, Seungone Kim, and Bhiksha Raj. VideoJudge: Bootstrapping Enables Scalable Supervision of MLLM-as-a-Judge for Video Understanding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=31CznLfRIS.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models, 2025. URL http://arxiv.org/abs/2503.20314.

Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, Jiahui Ge, Qianli Ma, Hang He, Yifan Zhou, Lingzi Guo, Lantao Mei, Jiachen Li, Hanwen Xing, Tianqi Zhao, Yu Fengyuan, Weihang Xiao, Yizheng Jiao, Jianheng Hou, Danyang Zhang, Pengcheng Xu, Boyang Zhong, Zehong Zhao, Gaoyun Fang, John Kitaoka, Xu Yile, Hua Xu, Kenton Blacutt, Tin Nguyen, Siyuan Song, Haoran Sun, Shaoyue Wen, Linyang He, Runming Wang, Yanzhi Wang, Mengyue Yang, Ziqiao Ma, Raphaël Millière, Freda Shi, Nuno Vasconcelos, Daniel Khashabi, Alan Yuille, Yilun Du, Ziming Liu, Dahua Lin, Ziwei Liu, Vikash Kumar, Yijiang Li, Lei Yang, Zhongang Cai, and Hokin Deng. A Very Big Video Reasoning Suite. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=AwC77yHpP6.

Felix Warneken and Michael Tomasello. Altruistic Helping in Human Infants and Young Chimpanzees. *Science*, 311(5765):1301–1303, 2006. doi: 10.1126/science.1121448. URL https://www.science.org/doi/10.1126/science.1121448.

Felix Warneken and Michael Tomasello. Helping and Cooperation at 14 Months of Age. *Infancy*, 11(3):271–294, 2007. ISSN 1532-7078. doi: 10.1111/j.1532-7078.2007.tb00227.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-7078.2007.tb00227.x.

Henry M. Wellman. *The Child's Theory of Mind*. The MIT Press, 1990. ISBN 9780262231534. doi: 10.7551/mitpress/1811.001.0001. URL https://mitpress.mit.edu/9780262231534/the-childs-theory-of-mind/.

Henry M. Wellman and David Liu. Scaling of theory-of-mind tasks. *Child Development*, 75(2): 523–541, 2004. ISSN 0009-3920. doi: 10.1111/j.1467-8624.2004.00691.x.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983. ISSN 0010-0277. doi: 10.1016/0010-0277(83)90004-5. URL https://www.sciencedirect.com/science/article/pii/0010027783900045.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. HunyuanVideo 1.5 Technical Report, 2025. URL http://arxiv.org/abs/2511.18870.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.717. URL https://aclanthology.org/2023.findings-emnlp.717/.

xAI. Introducing Grok 4.5. https://x.ai/news/grok-4-5, 2026. Accessed: 2026-07-31.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8593–8623, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.466. URL https://aclanthology.org/2024.acl-long.466/.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness, 2025. URL http://arxiv.org/abs/2503.21755.

A Benchmark Taxonomy and Task Details

Mind-VBench is organized around three BDI components: Belief, Desire, and Intention. These components are organizational and diagnostic categories, not stages of model inference and not claims about internal mental states in a video generation model. Each component contains two dimensions, and each dimension contains one or more tasks. The benchmark contains twelve tasks in total, with 30 instances per task and 360 instances overall.

A.1 Comparison with Existing Benchmarks

Table 4 compares Mind-VBench with representative video-generation and video-understanding benchmarks.

Table 4: Comparison of representative video-generation and video-understanding (Video Und.) benchmarks in terms of evaluation focus, task organization, and character-specific mental states.

Benchmark	Modality	Primary Target	Reasoning Focus	Mental-State Usage
T2V-CompBench (Sun et al., 2025)	T2V	Compositional generation	Object and attribute binding	Not modeled
V-ReasonBench (Luo et al., 2025)	T2V/I2V	Verifiable reasoning	Structured, spatial, pattern, and physical reasoning	Not modeled
RULER-Bench (He et al., 2025)	T2V/I2V	Rule-based generation	Explicit rule consistency	Not modeled
RISE-Video (Liu et al., 2026a)	TI2V	World-rule realization	Implicit physical and commonsense reasoning	Not modeled
SVBench (Peng et al., 2026)	T2V	Social behavior generation	Social interaction reasoning	Annotation labels
R^3-VQA (Niu et al., 2026)	Video Und.	Mental-state inference	Mental-physical causal reasoning	Annotation labels
Mind-VBench (**Ours**)	T2V/I2V	Character behavior consistency	Belief–Desire–Intention constraints	Behavioral constraints

A.2 Belief

Belief tasks evaluate whether a character acts according to information available from the character's own observations rather than privileged information available to the viewer. The benchmark distinguishes differences in current perspective, unobserved changes over time, and explicitly stated counterfactual rules.

A.2.1 Perceptual Belief (PB)

Perceptual Belief concerns information grounded in what the focal character can currently observe or has previously observed. It tests whether generated behavior respects character-specific access to scene information. The two tasks distinguish a restriction in the character's current perspective from a change that occurs after the character's last observation.

Perspective Difference (PD). This task evaluates whether a character acts according to a restricted visual perspective. The character lacks access to an event, object, or state that is available to another character or to the viewer. A successful behavior is one that is consistent with the focal character's observed information and does not rely on privileged information.

External Change (EC). This task evaluates whether a character preserves the last state that the character observed after an unobserved external change. An object or scene state changes while the focal character is absent or unable to observe the change. A successful behavior should be based on the previously witnessed state, unless the character subsequently obtains evidence that updates its information.

A.2.2 HYPOTHETICAL BELIEF (HB)

Hypothetical Belief concerns information established by an explicitly given nonstandard rule rather than by ordinary perception. It tests whether the generated scene and behavior remain consistent with the rule treated as valid within the instance. Unlike Perceptual Belief, the diagnostic condition is not who observed an event, but whether an explicitly specified alternative rule is followed.

Counterfactual Rule (CR). This task evaluates whether a character acts consistently with an explicitly established nonstandard rule. The rule defines a condition under which the relevant scene or object behaves differently from ordinary expectations. A successful video must preserve the rule as the operative condition and show behavior that is compatible with it. This task does not test general physical plausibility and does not require the model to possess an internal symbolic representation of the rule.

A.3 DESIRE

Desire tasks evaluate whether behavior is directed toward outcomes preferred by the character or shaped by relevant social conditions. The target is an observable choice or response, rather than an inference about an unobserved mental state.

A.3.1 INTRINSIC DESIRE (ID)

Intrinsic Desire concerns preferences and expectations established for an individual character. It tests whether subsequent behavior is consistent with the character's own preferred option or adapts when an anticipated outcome is violated. The relevant motivation is defined at the individual level rather than by a social role, another character's need, or group behavior.

Self Preference (SP). This task evaluates whether a character selects an option that matches an established personal preference when multiple alternatives are available. The preference must be established before the decision point, and the evaluated behavior is the character's subsequent choice.

Expectation Violation (EV). This task evaluates whether a character responds appropriately when an expected outcome does not occur. The scenario establishes an expectation and then presents an observable outcome that violates it. The success criterion is the task-specific, observable adaptation after the violation, rather than the mere occurrence of surprise or a generic emotional reaction.

A.3.2 SOCIAL DESIRE (SD)

Social Desire concerns behavior shaped by relations between the focal character and other agents. It covers differentiated role obligations, responses to another character's visible need, and alignment with a relevant group tendency. Unlike Intrinsic Desire, the criterion depends on an observable social relation rather than only on the focal character's personal preference or expectation.

Role Matching (RM). This task evaluates whether characters perform differentiated actions that match their visible social roles in a shared event. The roles, relevant responsibilities, and decision point must be observable. Merely appearing in the same scene or performing identical actions is not sufficient when the task requires role-specific behavior.

Empathic Response (ER). This task evaluates whether a character responds to another character's visible difficulty or need. The need must be observable, and the expected response must be defined by the instance specification. The label concerns the observable response to the other character, not a claim about empathy as an internal psychological property.

Social Conformity (SC). This task evaluates whether a character aligns its behavior with relevant actions of other agents when the scenario establishes a meaningful group signal or social tendency. The criterion is instance-specific: the group action must be relevant to the focal character's decision, and incidental co-occurrence does not constitute conformity.

A.4 Intention

Intention tasks evaluate whether a character maintains an established goal, adapts a plan when necessary, or coordinates with other agents toward a competitive or shared objective. They focus on observable temporal behavior and do not imply that the generator has an internal planning process.

A.4.1 Personal Intention (PI)

Personal Intention concerns the continuation of an established goal by a single focal character. It tests whether the goal survives an interruption and whether the character can change its means when the original plan becomes unavailable. The two tasks distinguish resuming a goal after an unrelated interruption from adapting the plan after a goal-relevant blockage.

Goal Persistence (GP). This task evaluates whether a character maintains an established goal after an external interruption. The interruption should not change the target goal, and the character should resume or continue the unfinished objective when an appropriate opportunity is available.

Plan Adaptation (PA). This task evaluates whether a character preserves the same goal while adapting its plan after the original course of action is blocked. A successful behavior changes the route, method, or sequence of actions while maintaining the established objective.

A.4.2 Multi-agent Intention (MI)

Multi-agent Intention concerns goal-directed behavior whose outcome depends on the actions of multiple agents. It distinguishes agents pursuing competing claims over a limited opportunity from agents coordinating complementary actions toward a shared objective. In both cases, success requires an observable interaction structure rather than the mere co-presence of several characters.

Resource Competition (RC). This task evaluates observable competition among multiple agents for a limited resource, space, or opportunity. The instance must establish the relevant scarcity and the competing objectives. Success is determined by the task-specific competitive behavior, not by the mere presence of multiple characters near the resource.

Collaborative Planning (CP). This task evaluates whether multiple characters coordinate complementary actions toward a shared task. Successful behavior may involve division of labor, complementary actions, or an appropriate temporal sequence. Merely appearing together or moving independently is insufficient when no functional coordination is visible.

A.5 Twelve-Task Showcase

The twelve-task showcase provides one benchmark example for every task across five figures organized by component or diagnostic dimension (Figures 4–8). Each example presents the generation input, scene prerequisites, and behavioral criterion corresponding to the task definition above.

B Benchmark Construction and Quality Control

B.1 Instance Construction

Each benchmark instance separates the generation input from the information used to evaluate the generated video. An instance contains a generation prompt, scene prerequisites, and a core behavioral criterion. The generation prompt establishes the relevant characters, objects, states, observations, preferences, or goals, but stops at the behavioral decision point. It does not directly specify the expected continuation that will later be scored.

The construction process follows three requirements. First, each instance is assigned one target task and must establish the corresponding BDI constraint. Second, the scenario must provide enough visual information for an evaluator to determine whether the prerequisite conditions are present. Third, the behavioral criterion must distinguish the target behavior from relevant failure behaviors without depending on unstable text, hidden author intent, or fine-grained identity cues.

For each task, Gemini 3.1 (Google, 2026) first generates 50 candidate seed scenarios together with their underlying task mechanisms. Human reviewers screen these candidates for task correspondence, visualizability, behavioral evaluability, ambiguity, redundancy, and generation difficulty, retaining 30 seed scenarios per task. GPT-5.4 (OpenAI, 2026b) then expands each retained seed according to the unified construction guidance and task-specific rules. The resulting instance includes the complete generation prompt, a first-frame description, four scene prerequisites, and a core behavioral criterion.

We vary characters, objects, environments, and event structures while preserving the diagnostic relation required by each task. We avoid explicit BDI terminology in generation prompts, unnecessary visual details, and requirements that are unrelated to the target behavior. The benchmark is constructed for both text-to-video (T2V) and image-to-video (I2V) evaluation. For I2V, gpt-image2 (OpenAI, 2026a) generates a corresponding reference image from the first-frame description. This image establishes the characters, objects, and initial scene before the focal behavior is resolved; it must not initiate or complete the expected continuation.

B.2 Data Organization

The benchmark contains 360 instances, with 30 instances assigned to each of the twelve tasks. Each instance is associated with an original text prompt, an optional first-frame input for I2V evaluation, and a Prompt Enhancement variant for the assisted conditions. These inputs remain aligned with the same instance-specific scene prerequisites and core behavioral criterion. The same benchmark instances are used across the evaluated models and applicable generation conditions; each model–condition–instance tuple produces one video rather than multiple random-seed repetitions. The submitted materials preserve these instance-level correspondences so that generation inputs, evaluation specifications, and generated results can be matched without relying on local filesystem paths.

B.3 Prompt Enhancement

Prompt Enhancement (PE) is an assisted generation condition in which the original prompt is augmented with a description of the expected continuation. It makes the otherwise implicit behavioral requirement more explicit to the video generator. PE is a prompt condition, not a new video modality and not evidence that the video model itself performs stronger mental-state reasoning. The PE condition should therefore be interpreted as a controlled study of prompt explicitness.

The rewrite is intended to preserve the original scene and task relation while converting the implicit psychological situation into an explicit action sequence. GPT-5.6 (OpenAI, 2026c) receives the original prompt and its behavioral criterion and produces the corresponding PE prompt. The generated PE outputs are retained without subsequent human rewriting or filtering. The rewritten prompt must not be treated as an unassisted benchmark input, and comparisons with the original T2V condition are interpreted as assisted-versus-unassisted generation comparisons.

B.4 Human Verification

After GPT-5.4 expansion and first-frame generation, the complete generation inputs and their evaluation specifications are independently reviewed by two human annotators before inclusion. The review checks whether each instance matches its assigned task, whether the target action is inadvertently revealed, whether the behavioral decision remains unresolved, and whether the prompt and first frame establish the intended information, preference, or goal relation. Reviewers also check that the scene prerequisites and core behavioral criterion are clear and mutually consistent.

Disagreements are resolved through discussion. Instances with scene ambiguity, target-action leakage, generation difficulty unrelated to the target task, or unclear evaluation specifications are revised or removed.

C Evaluation Protocol

C.1 Instance-Specific Evaluation Specification

Each generated video is evaluated against the specification attached to its benchmark instance. The specification consists of two complementary parts: scene validity, which determines whether behavioral evaluation is applicable, and behavioral success, which measures whether the generated behavior satisfies the target BDI constraint.

Scene Validity. Scene validity is determined by four binary prerequisites:

1. **Entity Presence (EP):** the required entities and base scene are present;
2. **Scenario Establishment (SE):** the task-defining event or state is established;
3. **Evidence Availability (EA):** the information and relations required for judgment are observable;
4. **Decision Opportunity (DO):** the focal character reaches a clear decision point with sufficient opportunity to act.

A video is considered valid only if all four prerequisites are satisfied.

Behavioral Success. For each benchmark instance, a single instance-specific behavioral criterion defines the observable behavioral outcome required by the target BDI constraint. Behavioral success is evaluated only on valid videos. A valid video is labeled as *Success* if the criterion is satisfied and *Fail* otherwise.

The DO prerequisite only requires that the focal character reaches a meaningful opportunity to act; it does not require the target behavior itself. Therefore, inaction, omission, or failed coordination after a valid decision point is treated as a behavioral failure rather than scene invalidity. Together, these criteria assign each generated video one of three final labels: **Invalid**, **Fail**, or **Success**.

C.2 Automatic Evaluator Instructions

The main automatic evaluator is Qwen3.5-27B (Qwen Team, 2026). It receives the generated video and the corresponding instance-specific evaluation specification. The evaluator first determines the four prerequisite decisions and then scores the core behavioral criterion only in the context of the resulting scene.

The expected machine-readable output contains the four validity decisions, a joint validity result, and the core behavioral decision. The output is parsed into the labels Invalid, Fail, and Success used for aggregation.

C.3 Metrics and Aggregation

For task t, let N_t be the number of generated videos, V_t the number of valid videos, and S_t the number of successful videos. We compute

$$\mathrm{VR}_t = \frac{V_t}{N_t}, \qquad \mathrm{CSR}_t = \frac{S_t}{V_t}, \qquad \mathrm{OSR}_t = \frac{S_t}{N_t}. \tag{2}$$

Validity Rate (VR) measures whether a model establishes the situation needed for behavioral judgment. Conditional Success Rate (CSR) measures behavioral success among valid videos. Overall Success Rate (OSR) measures the joint outcome of establishing a valid situation and producing the target behavior, and is the primary metric.

Task-level scores are macro-averaged within each dimension and component. The overall score is the macro average over the twelve tasks. CSR is not defined for a task with no valid videos and is not replaced by zero in macro-averaging. The four prerequisite pass rates are reported as diagnostic quantities and are not included as separate terms in OSR.

D HUMAN EVALUATION ALIGNMENT

D.1 ANNOTATION PROTOCOL

We conduct a human evaluation alignment study on a randomly sampled subset of 20 generated videos covering all twelve tasks and different model groups. Each video is independently evaluated by 12 human annotators using the same instance-specific evaluation specifications as the automatic evaluator. Final human labels are determined by majority voting.

Agreement is examined at the level of scene validity and behavioral success, and the automatic evaluator is compared with the human majority label. The human evaluation is an alignment study for the observable labeling protocol; it does not make human judgments a proof that a model possesses a belief, desire, intention, or Theory of Mind.

D.2 ALIGNMENT RESULTS

The main paper reports accuracy and Cohen's kappa for automatic–human agreement. Table 5 reports the same metrics for scene validity and behavioral success, together with the final-label agreement used in the analysis. Qwen3.5-27B is used as the main evaluator because it provides the strongest observed behavioral alignment among the evaluated automatic judges while maintaining the intended evaluation cost.

E ADDITIONAL QUANTITATIVE RESULTS

This section provides fine-grained evidence that complements, rather than duplicates, the model-level results in the main paper. In particular, the additional analyses separate task-level variation from component averages and make the validity bottleneck directly inspectable.

E.1 TASK-LEVEL VR, CSR, AND OSR

The main paper emphasizes task-level OSR and component-level aggregates. Table 6 additionally reports VR, CSR, and OSR for all twelve tasks under the original T2V condition. This view makes it possible to distinguish a component-level average from a task-specific difficulty and to identify tasks where scene validity and behavioral consistency fail at different rates.

For the three closed-source models, base (non-PE) runs do not always return all 360 requested videos because provider-side safety checks or request failures can block generation. The returned JSON/video counts are Kling-3.0: 360 T2V and 359 I2V (719 total); HappyHorse-1.1: 360 T2V and 357 I2V (717 total); and Dreamina-Seedance-2.0: 359 T2V and 183 I2V (542 total). The complete per-instance counts and evaluation results are provided in the supplementary materials.

E.2 TASK-LEVEL CONDITION COMPARISON

The controlled conditions are T2V, I2V, PE-T2V, and PE-I2V where the corresponding setting can be applied consistently. Table 7 reports task-level OSR across these conditions. This analysis complements the main paper's model-level condition summary by showing whether a condition improves all tasks or only selected behavioral relations.

The current interpretation is that I2V primarily improves scene validity, whereas PE can improve both scene validity and behavioral success. These are average tendencies, not universal task-level effects; individual tasks may regress under a condition.

The four open-source models are reported under all four conditions. The three closed-source models are reported under T2V and I2V only because PE is not available for them; the two source groups are not combined into a single condition average.

E.3 VALIDITY CHECKLIST BREAKDOWN

The joint VR metric can be decomposed into the four prerequisite checks EP, SE, EA, and DO. This breakdown diagnoses which scene requirements prevent a video from reaching behavioral evalua-

tion. It should be read as a diagnostic decomposition rather than as four independent substitutes for VR.

Table 8 reports the four checklist pass rates and their joint VR for the original T2V condition. Individual prerequisite rates are diagnostic and do not replace the joint validity definition.

E.4 Model Variants and Video Duration

We report two additional controlled analyses: model-variant comparison within the Wan2.2 family and video-duration comparison for LongCat-Video-13.6B. Both analyses use the original T2V condition and can be read with the task-level T2V results in Table 6.

Under the original T2V setting, Wan2.2-TI2V-5B and Wan2.2-T2V-A14B reach OSR scores of 19.7% and 17.2%, respectively, in Table 9. The two variants show no monotonic association between nominal parameter count and benchmark performance: the larger variant achieves a higher CSR but a lower VR, and its higher CSR is concentrated in Belief- and Intention-related tasks. This suggests that increasing model capacity does not by itself resolve the character-specific information-state constraints that the benchmark isolates.

Table 10 reports LongCat-Video-13.6B under target durations of 5, 10, 15, and 20 seconds. OSR declines as the target duration increases, and the drop is strongest at 20 seconds, where both VR and CSR decrease. The component-level OSR shows that the decline affects all three BDI components, with the largest relative drop on Intention-related tasks. These results indicate that longer generation horizons make it more difficult to maintain both scene validity and behavioral consistency.

E.5 Overall Capability Visualization

Radar plots provide a complementary view of task-level performance profiles. Unlike the preceding tables, which support exact numerical comparison, these plots are intended to highlight differences in how performance is distributed across the twelve tasks. They should not be used as substitutes for the reported numerical results or as evidence of statistical significance.

All radar panels report task-level OSR on a common scale and use the same task ordering. The axes group PD, EC, and CR under Belief; SP, EV, RM, ER, and SC under Desire; and GP, PA, RC, and CP under Intention. Using a fixed ordering and axis range makes model and condition profiles comparable without implying that the enclosed polygon area is itself an evaluation metric.

Figures 9 and 10 compare model profiles under T2V and I2V in separate open-source and closed-source figures. Each figure uses a T2V/I2V dual-panel layout with task-level OSR on a shared 0–100 scale and the same twelve-task ordering. Splitting by release type limits curve overlap while preserving a common visual encoding. These views complement the numerical tables by showing whether models with similar aggregate OSR exhibit different patterns across the taxonomy. Condition-level differences (including PE) are left to Table 7; we do not add crowded all-model condition radar overlays.

F Implementation Details

All evaluated models are tested on the same benchmark instances with condition-specific inputs. The target video duration is 10 seconds, and each model–condition–instance tuple contributes one generated video. The evaluated open-source models are Wan2.2-TI2V-5B, Wan2.2-T2V-A14B, LTX-2.3-22B, LongCat-Video-13.6B, and HunyuanVideo-1.5-8.3B. The evaluated closed-source systems are Kling-3.0, HappyHorse-1.1, and Dreamina-Seedance-2.0. All models are included in the original T2V comparison; Wan2.2-TI2V-5B is reported separately in Table 9 because it is a lightweight variant of the same family rather than a primary benchmark model. Additional conditions are restricted to models for which the setting can be applied consistently.

Open-source models are evaluated using the default configurations provided by their official repositories, and all open-source generation runs are conducted on eight NVIDIA A100 GPUs. Closed-source systems are evaluated through their official interfaces using the provider-default settings.

The automatic evaluation is performed by Qwen3.5-27B (Qwen Team, 2026). In the alignment study, we additionally compared the candidate judges GPT-5.6 (OpenAI, 2026c), Qwen2.5-VL-72B (Bai et al., 2025), Grok-4.5 (xAI, 2026), and Qwen3.5-35B-A3B (Qwen Team, 2026) before selecting the main evaluator.

G Qualitative Results

G.1 Cross-Model T2V/I2V Examples

Figures 11–22 provide one qualitative example for each of the twelve tasks. Every figure uses the same layout: the generation prompt at the top, followed by T2V and I2V panels in which each row shows ten frames from one model. The comparison is restricted to the non–Prompt-Enhanced condition so that T2V and I2V differ only in whether a first-frame reference is provided. These visualizations complement the numerical tables and radar plots by showing how models behave on the same instance under the two modalities; they are not intended as an additional aggregate metric.

Perspective Difference (PD)



Prompt: In a living room, a young man carries a large cardboard box taller than his upper body toward a sofa, preparing to place it beside the sofa. A brightly colored chair stands in his path, and the large box blocks his view directly ahead and downward.

Checklist:
1. The video must show an identifiable young man, a living room with a sofa, a large cardboard box, and a chair in his path, and his goal of carrying the box beside the sofa must be clear.
2. The man must walk toward the sofa while holding a large cardboard box taller than his upper body. The chair must be directly ahead in his path, and the box must block his forward and downward view of the chair. These relationships m...
3. The box must already block the man's view toward the chair at the start of the video, and the opening must not clearly indicate that he has already seen the chair. There must be no external changes such as the chair being placed, moving on...
4. The man must reach a position where he can continue carrying the box along the current route. The large box, sofa, chair ahead, and walking area must all be identifiable, with enough space and continuous footage for him to initiate the fi...

Core Score:
Occlusion-Boundary Alignment: Does the young man first initiate forward movement with the box along the apparently clear route, following the preset direction to act directly on currently visible information, before seeing the chair by lowering the box, turning sideways, peeking around it, or changing position? Maintaining his pace, moving his body forward, or stepping toward the chair is sufficient; he does not need to reach it. If he precisely avoids the chair's location, turns early, or changes his carrying route without gaining a new view, this criterion is not satisfied. If he first proceeds normally and then adjusts after contacting the chair, lowering the box, or changing position and gaining new information, that does not affect evaluation of the initial action.

External Change (EC)



Prompt: In the living room of a rented apartment, a man last saw his keys on the table before leaving the living room. While he was away, his roommate moved the keys into a box on the shoe cabinet by the door. The table is now empty, there is a box on the shoe cabinet, and the keys are inside the box. The man re-enters the living room to get his keys.

Checklist:
1. The man, keys, table, and box on the shoe cabinet are clearly visible in the video.
2. The keys are no longer on the table and are inside the box on the shoe cabinet, clearly establishing the location change.
3. The man re-enters from outside the living room, and no one points out or tells him the keys' new location.
4. The video begins after the keys have been moved, and the man's first action to get the keys after re-entry can be clearly judged.

Core Score:
Change-Knowledge Alignment: The man's first judgeable goal-directed action after re-entry must clearly target the table where he last saw the keys before leaving. It is sufficient for his initial direction of travel, gaze and search location, or reaching direction to clearly point toward the table; exact consistency among minor movements is not required. After actually seeing, touching, or checking that the keys are not on the table, he may naturally turn to search the box on the shoe cabinet. If he re-enters and immediately walks directly to the shoe cabinet or targets the box on it without any visible information-acquisition process, this criterion is not satisfied. It is also not satisfied if he merely stops, wanders randomly, or moves ambiguously between the table and shoe cabinet with...

Counterfactual Rule (CR)



Prompt: In this counterfactual world, the red and blue edges of every ordinary sheet of white paper are two magnetic poles: edges of the same color repel, and edges of different colors attract. A boy needs to connect two red-and-blue-edged worksheets on the table end to end into a long strip and place it in the transparent long document sleeve in front of him. There are no fastening tools on the table.

Checklist:
1. Are the boy, two red-and-blue-edged worksheets, and a transparent long document sleeve present?
2. Before the boy acts, are the two worksheets separate and the document sleeve empty?
3. Is the boy positioned in front of the two worksheets, with the paper edges and sleeve opening within his direct reach?
4. Are the boy's hands, the paper edges, and the sleeve opening clearly visible, and have the sheets not been replaced with magnets or magnetic material?

Core Score:
Belief-consistent Behavior: Does the boy naturally use the attraction between opposite poles so that the two central task sheets actually attract and firmly connect into a long strip, then move the strip toward the document sleeve or align it with the sleeve opening? Actual attachment is required; a clear insertion attempt is sufficient, and complete insertion is not required. Merely overlapping the sheets by hand, using incorrect polarity or real-world fastening tools, reacting with surprise or testing the effect, or showing no insertion attempt after connecting them does not satisfy this criterion. The case must not be rated Fail merely because insertion is incomplete.

Figure 4: Belief tasks: Perspective Difference (PD), External Change (EC), and Counterfactual Rule (CR). Each card shows the generation input, scene prerequisites, and behavioral criterion.

Self Preference (SP)

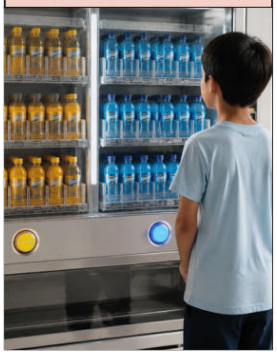


Prompt: In a video arcade aisle, the boy prefers racing machines when he visits. He holds a bucket of tokens, a steering-wheel keychain swings from his backpack, and several colorful tickets dispensed by a racing machine are clipped to the bucket rim. A claw machine is on the left and a racing simulator is on the right. He must choose one machine, insert a token, and start playing.

Checklist:
1. The boy holding the token bucket must be the single central character, and the video arcade aisle must be shown so that his goal of choosing a machine, inserting a token, and playing is understandable.
2. The claw machine on the left and the racing simulator on the right must both appear and be usable before the choice occurs, with their left-right positions clear.
3. The boy's racing preference cues must be established. At least a steering-wheel keychain, racing-machine tickets, or similar racing-related personal traces must be visible. If these cues are entirely absent, this item is not satisfied.
4. The video must show the process from both options remaining available to the boy forming a discernible inclination or making a clear choice. The process may be brief, but this item is not satisfied if the video cuts directly from no visi...

Core Score:
Preference-Consistent Choice: The boy must subsequently show a discernible preference-consistent inclination toward the racing simulator on the right, such as approaching it, inserting a token into it, sitting in its seat, or beginning to grip its steering wheel. Any clear indication of this inclination is sufficient. If he clearly chooses the claw machine on the left, or his choice direction remains indiscernible through the end, the sample must be rated Fail.

Expectation Violation (EV)



Prompt: A boy wants to buy a bottle of yellow-colored orange juice to drink and stands in front of a glass-front vending machine. On the shelves inside the vending machine, yellow drinks are on the left with a yellow button below them, and blue drinks are on the right with a blue button below them. The boy presses the yellow button on the left, waits for the drink, and prepares to pick it up and drink it. However, the machine dispenses a bottle of blue-colored drink into the pickup tray.

Checklist:
1. The boy and the glass-front vending machine must appear in the video, and the boy, both types of drinks, both buttons, and the pickup tray must all be clearly visible.
2. Yellow drinks must be placed on the left side of the shelves inside the vending machine, with a yellow button directly below them; blue drinks must be placed on the right, with a blue button directly below them. The left-right corresponde...
3. The video must continuously and clearly show the boy pressing the yellow button on the left, followed by the machine dispensing the product into the pickup tray. If the button color, left-right position, or dispensing action is unclear,...
4. The machine must dispense a bottle of blue-colored drink into the pickup tray, not yellow-colored orange juice. If any key product, button, or resulting object is missing, replaced, or unrecognizable, the sample is rated Invalid.

Core Score:
Response Consistency: After the blue-colored drink is dispensed into the pickup tray, the boy must interrupt his original action of taking and drinking the yellow-colored orange juice and show a clearly visible brief response, such as stopping his reaching hand, looking at the blue-colored drink, or looking back at the yellow button on the left. If there is no behavioral change after the result appears, or if the boy directly takes and drinks the blue-colored drink without hesitation, S1 is not satisfied.

Figure 5: Intrinsic Desire: Self Preference (SP) and Expectation Violation (EV).

Role Matching (RM)

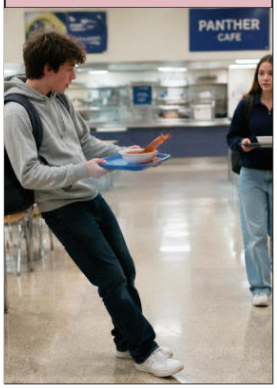


Prompt: In a brightly lit pastry shop, a child slaps the glass cake display and points to a cream cake on the upper shelf; an apron-wearing clerk stands beside the child.

Checklist:
1. The pastry shop's glass cake display, the child slapping the glass, the apron-wearing clerk beside the child, and the cake on the upper shelf inside the display must be clearly visible.
2. The triggering action of the child slapping the glass and pointing to the upper-shelf cake must be visible, with the clerk beside the child and able to respond immediately.
3. The child must be identifiable by their small stature and the clerk by the apron; they must stand side by side before the glass display, with the upper-shelf cake within both people's visible range.
4. The video must first clearly show continuous footage after the child has slapped the glass and pointed to the cake, while the cake remains inside the display and the clerk stands beside the child; it must not begin directly with the cake alre...

Core Score:
Role-Action Alignment: The child may continue slapping or pointing to the upper-shelf cake, lean close to the glass, or stand on tiptoe; the clerk may turn toward the cake display beside the child, raise a hand toward the cake the child is pointing to, approach the child, or reach out to signal the child to stop slapping. The clerk continuously looking at the cake the child indicates while also turning toward it may count as an early response. S1 is not satisfied if the clerk gives no visible response or leaves the child's side, shows only an undirected gaze, the child opens the display and retrieves an item, or both people only watch motionlessly.

Empathic Response (ER)

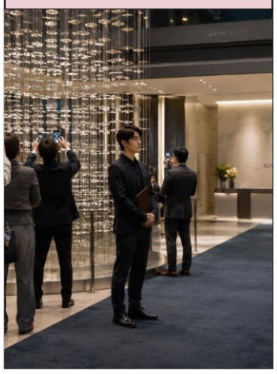


Prompt: In a school cafeteria seating area, a male student carries a tray with hot soup through an aisle between tables and chairs. His backpack strap suddenly catches on the corner of a chair at the edge of the aisle, pulling the tray backward and tilting it. The soup sloshes violently and is about to spill. A female student carrying her own tray and looking for a seat approaches from the opposite direction in the same aisle. The male student and the chair on which his backpack strap is caught are directly ahead of her.

Checklist:
1. The school cafeteria seating area and two distinguishable students must be clearly visible: a male student carrying a tray of hot soup and wearing a backpack, and a female student carrying her own tray while looking for a seat along the ais...
2. The key difficulty and its causal relationship must remain continuously visible: the backpack strap is actually caught on the chair corner, which pulls the tray backward and off balance, while the hot soup visibly sloshes and is close to spilli...
3. The observer's original task and opportunity to observe must both be established: the female student is originally carrying her own tray along the same aisle while looking for a seat, and the male student and the chair on which his backp...
4. Before the female student responds, the male student must still be simultaneously controlling the hot-soup tray and the backpack caught on the chair corner. The difficulty must not resolve itself or be resolved early by a third person.

Core Score:
Response Direction Consistency: The female student must clearly notice that the male student's backpack strap is caught on the chair corner and is tilting his hot-soup tray, and she must show actual help or a clear intention to help that is relevant to the immediate need. Acceptable behaviors include stopping and turning toward him, freeing one hand, reaching to steady the tray or move the chair, or beginning any of these actions. She does not have to completely free the strap before the video ends. This criterion is not satisfied if she does not notice at all, continues looking for a seat or walks around him and leaves, gives only a brief glance without approaching or preparing to help, or performs an action unrelated to the current difficulty.

Social Conformity (SC)



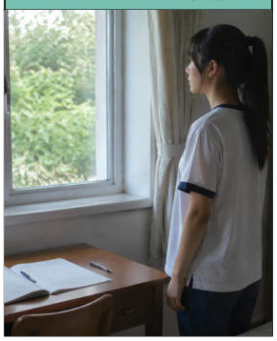
Prompt: In an office building lobby, a circle of people face the atrium and hold up their phones to film a light installation, leaving an open carpeted path toward the reception desk. A young male visitor carrying a folder is walking toward the reception desk on the other side and reaches the space between the filming crowd and the carpeted path.

Checklist:
1. The scene must simultaneously show an office building lobby, a circle of people holding up phones to film toward the atrium, a young male visitor carrying a folder, and a visible reception counter or clearly indicated direction to reception.
2. The group trend must be clearly visible: multiple people remain facing the same atrium light installation and hold up phones to film, rather than standing separately or performing unrelated actions.
3. The subject's original goal of registering at reception must be visually readable; for example, he approaches in the direction of reception while carrying a folder and reaches the open carpeted path.
4. The visitor's identity and spatial relationship must remain continuously distinguishable before and after he reaches the path, and his subsequent direction and actions must be clear enough to determine whether he turns toward the atrium to f...

Core Score:
Group-Choice Consistency: The group holds up phones around the atrium to film the light installation, while the subject was originally carrying a folder to reception to register. This criterion is satisfied only if the same subject follows his established route to and through the carpeted path without turning toward the atrium, raising a phone, or stopping to join the onlookers, then continuously walks along the path to reception and places his identification on the counter, clearly continuing his registration goal. It is not satisfied if the subject joins the filming or onlookers, stalls or wanders without reason after the choice point, loses the direction of reception, or fails to form a continuous action sequence from the path to the counter.

Figure 6: Social Desire: Role Matching (RM), Empathic Response (ER), and Social Conformity (SC).

Goal Persistence (GP)



Prompt: At a desk, a girl is copying into an exercise book. The homework is unfinished, and the open exercise book and pen are on the desk. Halfway through, the window is suddenly blown open and the pages flutter. She gets up to close the window; the desk is near the window, and the exercise book and pen remain in their original places on the desktop.

Checklist:
1. The video must clearly show the girl, desk, open exercise book, and pen, and the exercise book must appear partially written but clearly unfinished.
2. The interruption caused by the window being blown open and its handled state must be clearly visible, such as her closing the window or the window already being closed or latched with the pages no longer moving in the wind. The entire process...
3. After the window is closed, no new arrival, event, or continuing demand may appear immediately, and the unfinished exercise book and pen must remain on the desk. The girl may take her next action immediately without pausing, but it m...
4. The first clear action after she finishes closing the window must be clearly visible and sufficient for scoring. The sample cannot be scored if there is only looking, a brief pause, an ambiguous turn, or the video ends immediately.

Core Score:
Resumption and Goal-Chain Continuity: Starting when the window is closed and the pages no longer flutter, the girl must clearly reconnect to the original unfinished task, such as returning to the same desk, turning the exercise book back to the unfinished page, picking up the pen again, and continuing to copy. It is also acceptable for her to press down the pages, straighten page corners, adjust the book, or find the line where she stopped before continuing to write. This criterion is not satisfied if she remains by the window, begins an unrelated action, leaves the desk area, or only turns pages, organizes the book, or touches stationery without actually resuming writing.

Plan Adaptation (PA)



Prompt: In a wide shopping mall corridor, an adult woman carrying shopping bags was going to enter the restroom ahead through its main entrance, but an overturned metal display rack completely blocks the main entrance passage. A service corridor on the right leading to the restroom's rear entrance is open, with a light-colored doorframe visible at the end.

Checklist:
1. The video must clearly show the woman carrying shopping bags, the restroom's main entrance, and her movement toward it, establishing the specific goal of entering that restroom.
2. After the woman reaches the main-entrance direction, the overturned metal display rack must continue occupying the entire passage, and she must be unable to pass through the main entrance. Signs, text, or sound alone are insufficient.
3. When scoring begins, the woman must still be outside the restroom, the service corridor on the right must be open and continuously lead to the rear entrance, and she must not yet have entered that corridor.
4. The first goal-related action after the rack blocks the way must remain continuously visible from start to outcome. The subsequent footage must be sufficient to determine whether she makes an effective detour, repeats an ineffective attem...

Core Score:
Planning Direction Consistency: After the main entrance is completely blocked by the overturned display rack, the subject must stop advancing along the failed main-entrance route, clearly enter the open service corridor on the right, and continue along it toward the rear entrance of the same restroom. This criterion is satisfied only if the change in direction, the corridor used, and the original goal of entering the restroom form a continuous and observable action chain. It is not satisfied if she remains in place, turns around and leaves, continues forcing her way toward the blocked main entrance, moves right without entering the service corridor, or goes to an unrelated location such as a store.

Figure 7: Personal Intention: Goal Persistence (GP) and Plan Adaptation (PA).

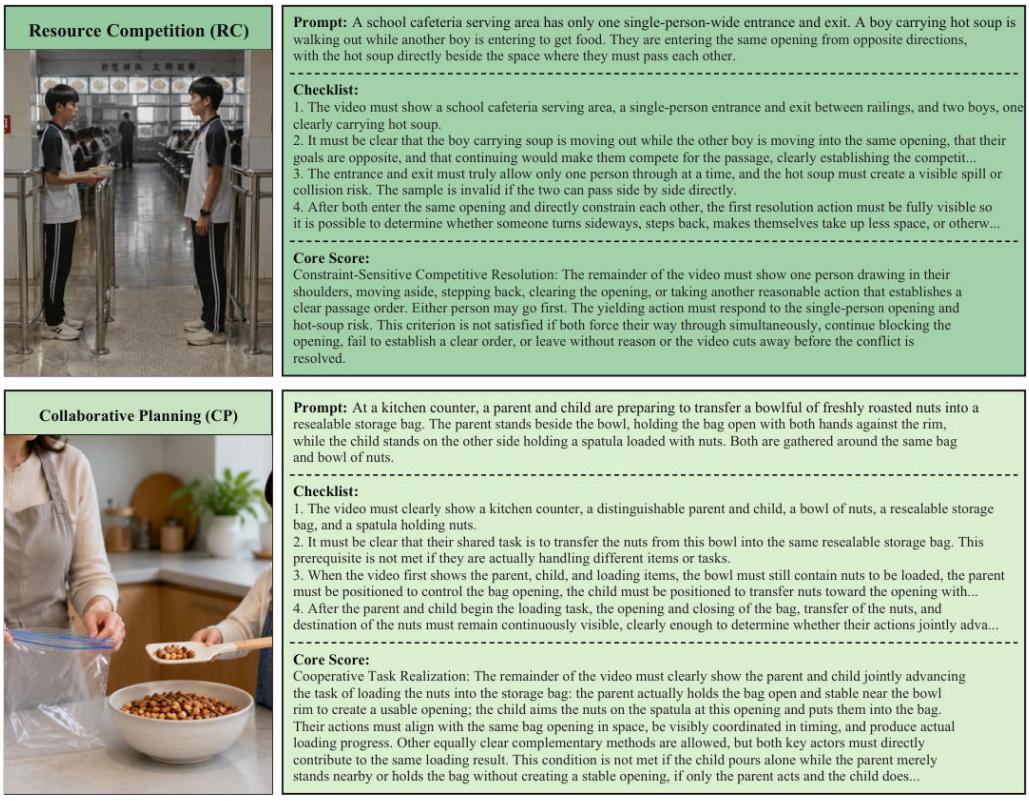


Figure 8: Multi-agent Intention: Resource Competition (RC) and Collaborative Planning (CP).

Table 5: Agreement between human majority labels and automatic judges on a randomly sampled video subset. Each video receives one of three final labels: *Invalid*, *Fail*, or *Success*. **Final label** compares these three-way outcomes (Agree, Pearson r, Cohen's κ, and QWK). **Scene validity** compares whether the video is valid (not Invalid) on all sampled videos. **Behavioral success** compares Success versus Fail on videos judged valid by the automatic evaluator. Best values in each column are in bold.

	Final label				**Scene validity**		**Behavioral success**	
Judge	Agree	Pearson r	κ	QWK	Acc. (%)	κ	Acc. (%)	κ
Qwen3.5-27B	0.550	**0.537**	**0.274**	0.400	55.0	0.211	**83.3**	**0.571**
GPT-5.6	**0.700**	0.484	0.221	**0.471**	75.0	0.306	66.7	0.250
Qwen2.5-VL-72B	0.350	0.402	0.161	0.219	35.0	0.085	75.0	0.385
Grok-4.5	**0.700**	0.361	0.137	0.357	**80.0**	**0.385**	40.0	−0.364
Qwen3.5-35B-A3B	0.250	0.100	0.013	0.043	25.0	−0.056	37.5	0.091

Table 6: T2V (non-PE): VR, CSR, and OSR with tasks as rows and models as columns. Left columns mark BDI components and diagnostic dimensions: PB (Perceptual Belief), HB (Hypothetical Belief), ID (Intrinsic Desire), SD (Social Desire), PI (Personal Intention), MI (Multi-agent Intention). Overall is the macro average over the twelve tasks. Within the open-source group and within the closed-source group, the best value in each row is in **bold** and the second-best is underlined.

				Open-Source				Closed-Source		
			Metric	Wan2.2 T2V-A14B	LTX-2.3 22B	LongCat 13.6B	Hunyuan 1.5-8.3B	Kling-3.0	HappyHorse 1.1	Dreamina Seedance-2.0
Belief	*PB*	**PD**	VR	56.7	**60.0**	26.7	46.7	**86.7**	83.3	83.3
			CSR	58.8	61.1	25.0	**64.3**	61.5	64.0	**76.0**
			OSR	33.3	**36.7**	6.7	30.0	53.3	53.3	**63.3**
		EC	VR	10.0	**40.0**	13.3	16.7	60.0	**73.3**	50.0
			CSR	**100.0**	58.3	75.0	**100.0**	72.2	**77.3**	66.7
			OSR	10.0	**23.3**	10.0	16.7	43.3	**56.7**	33.3
	HB	**CR**	VR	53.3	**60.0**	46.7	50.0	**80.0**	**80.0**	79.3
			CSR	6.2	5.6	**7.1**	0.0	62.5	70.8	**78.3**
			OSR	**3.3**	**3.3**	**3.3**	0.0	50.0	56.7	**62.1**
Desire	*ID*	**SP**	VR	23.3	**70.0**	23.3	20.0	76.7	86.7	**90.0**
			CSR	71.4	81.0	**85.7**	83.3	95.7	**96.2**	88.9
			OSR	16.7	**56.7**	20.0	16.7	73.3	**83.3**	80.0
		EV	VR	36.7	**53.3**	16.7	23.3	80.0	76.7	**83.3**
			CSR	45.5	56.2	**80.0**	57.1	54.2	**100.0**	**100.0**
			OSR	16.7	**30.0**	13.3	13.3	43.3	76.7	**83.3**
	SD	**RM**	VR	33.3	**66.7**	33.3	30.0	90.0	**96.7**	90.0
			CSR	40.0	**95.0**	60.0	33.3	29.6	**100.0**	81.5
			OSR	13.3	**63.3**	20.0	10.0	26.7	**96.7**	73.3
		ER	VR	**73.3**	63.3	30.0	20.0	86.7	**90.0**	**90.0**
			CSR	18.2	**36.8**	33.3	33.3	26.9	**51.9**	18.5
			OSR	13.3	**23.3**	10.0	6.7	23.3	**46.7**	16.7
		SC	VR	**63.3**	53.3	33.3	30.0	83.3	**90.0**	76.7
			CSR	15.8	**50.0**	20.0	44.4	**36.0**	33.3	21.7
			OSR	10.0	**26.7**	6.7	13.3	**30.0**	**30.0**	16.7
Intention	*PI*	**GP**	VR	30.0	26.7	**33.3**	16.7	66.7	73.3	**83.3**
			CSR	**77.8**	50.0	60.0	40.0	75.0	**95.5**	84.0
			OSR	**23.3**	13.3	20.0	6.7	50.0	**70.0**	**70.0**
		PA	VR	26.7	**46.7**	36.7	16.7	56.7	**70.0**	53.3
			CSR	25.0	28.6	27.3	**80.0**	58.8	42.9	**68.8**
			OSR	6.7	**13.3**	10.0	**13.3**	33.3	30.0	**36.7**
	MI	**RC**	VR	33.3	**56.7**	26.7	30.0	53.3	**93.3**	70.0
			CSR	**100.0**	**100.0**	75.0	88.9	68.8	**78.6**	52.4
			OSR	33.3	**56.7**	20.0	26.7	36.7	**73.3**	36.7
		CP	VR	43.3	**70.0**	63.3	36.7	**96.7**	93.3	80.0
			CSR	61.5	81.0	84.2	**100.0**	86.2	82.1	**91.7**
			OSR	26.7	**56.7**	53.3	36.7	**83.3**	76.7	73.3
		Overall	VR	40.3	**55.6**	31.9	28.1	76.4	**83.9**	77.4
			CSR	51.7	58.6	52.7	**60.4**	60.6	**74.4**	69.0
			OSR	17.2	**33.6**	16.1	15.8	45.6	**62.5**	53.8

Table 7: Task-level OSR under T2V, I2V, PE-T2V, and PE-I2V. Shaded average columns report BDI macro averages; Overall is the macro average over the twelve tasks. The four open-source models are reported under all four conditions; the three closed-source models are reported under T2V and I2V only because PE is unavailable for them. Within each source and condition block, the best value in each column is in **bold** and the second-best is underlined.

Model	Belief				Desire						Intention					Overall
	PD	EC	CR	*Avg.*	SP	EV	RM	ER	SC	*Avg.*	GP	PA	RC	CP	*Avg.*	
T2V / Open-Source																
Wan2.2-T2V-A14B	33.3	10.0	**3.3**	15.6	16.7	16.7	13.3	13.3	10.0	14.0	**23.3**	6.7	33.3	26.7	22.5	17.2
LTX-2.3-22B	**36.7**	**23.3**	**3.3**	**21.1**	**56.7**	**30.0**	**63.3**	**23.3**	**26.7**	**40.0**	13.3	**13.3**	**56.7**	**56.7**	**35.0**	**33.6**
LongCat-Video-13.6B	6.7	10.0	**3.3**	6.7	20.0	13.3	20.0	10.0	6.7	14.0	20.0	10.0	20.0	53.3	25.8	16.1
HunyuanVideo-1.5-8.3B	30.0	16.7	0.0	15.6	16.7	13.3	10.0	6.7	13.3	12.0	6.7	**13.3**	26.7	36.7	20.8	15.8
T2V / Closed-Source																
Kling-3.0	53.3	43.3	50.0	48.9	73.3	43.3	26.7	23.3	**30.0**	39.3	50.0	33.3	36.7	**83.3**	50.8	45.6
HappyHorse-1.1	53.3	**56.7**	56.7	**55.6**	**83.3**	76.7	**96.7**	**46.7**	**30.0**	**66.7**	**70.0**	30.0	**73.3**	76.7	**62.5**	**62.5**
Dreamina-Seedance-2.0	**63.3**	33.3	**62.1**	52.9	80.0	**83.3**	73.3	16.7	16.7	54.0	**70.0**	**36.7**	36.7	73.3	54.2	53.8
I2V / Open-Source																
Wan2.2-I2V-A14B	46.7	30.0	3.3	26.7	36.7	23.3	16.7	13.3	20.0	22.0	**16.7**	3.3	43.3	56.7	30.0	25.8
LTX-2.3-22B	**60.0**	**36.7**	**13.3**	**36.7**	**53.3**	**40.0**	**46.7**	**23.3**	30.0	**38.7**	6.7	**26.7**	**80.0**	66.7	**45.0**	**40.3**
LongCat-Video-13.6B	46.7	23.3	6.7	25.6	33.3	23.3	23.3	16.7	10.0	21.3	**16.7**	16.7	46.7	53.3	33.3	26.4
HunyuanVideo-1.5-8.3B	56.7	13.3	10.0	26.7	46.7	13.3	30.0	**23.3**	**36.7**	30.0	**16.7**	20.0	56.7	**70.0**	40.8	32.8
I2V / Closed-Source																
Kling-3.0	**63.3**	20.0	50.0	44.4	86.2	63.3	46.7	13.3	33.3	48.6	36.7	40.0	73.3	**76.7**	56.7	50.2
HappyHorse-1.1	60.0	33.3	**69.0**	**54.1**	93.1	**80.0**	**90.0**	**44.8**	**56.7**	**72.9**	**63.3**	36.7	**80.0**	**76.7**	**64.2**	**65.3**
Dreamina-Seedance-2.0	58.3	**38.1**	33.3	43.3	**100.0**	33.3	56.2	30.0	29.2	49.7	38.5	**57.9**	41.2	73.7	52.8	49.1
PE-T2V / Open-Source																
Wan2.2-T2V-A14B	56.7	30.0	6.7	31.1	86.7	23.3	**56.7**	33.3	20.0	44.0	36.7	23.3	**66.7**	43.3	42.5	40.3
LTX-2.3-22B	**70.0**	**36.7**	**23.3**	**43.3**	**96.7**	**50.0**	**56.7**	**43.3**	**40.0**	**57.3**	50.0	**60.0**	60.0	**53.3**	**55.8**	**53.3**
LongCat-Video-13.6B	33.3	23.3	10.0	22.2	70.0	10.0	26.7	**43.3**	10.0	32.0	33.3	6.7	23.3	30.0	23.3	26.7
HunyuanVideo-1.5-8.3B	43.3	26.7	16.7	28.9	43.3	13.3	26.7	13.3	20.0	23.3	**60.0**	16.7	33.3	30.0	35.0	28.6
PE-I2V / Open-Source																
Wan2.2-I2V-A14B	**70.0**	36.7	20.0	42.2	90.0	50.0	70.0	66.7	33.3	62.0	**73.3**	33.3	90.0	66.7	65.8	58.3
LTX-2.3-22B	**70.0**	50.0	33.3	51.1	**93.3**	**63.3**	**80.0**	**70.0**	**43.3**	**70.0**	70.0	**43.3**	**93.3**	73.3	70.0	**65.3**
LongCat-Video-13.6B	60.0	33.3	16.7	36.7	90.0	6.7	53.3	53.3	23.3	45.3	50.0	30.0	70.0	56.7	51.7	45.3
HunyuanVideo-1.5-8.3B	63.3	**60.0**	**36.7**	**53.3**	86.7	43.3	66.7	63.3	36.7	59.3	**73.3**	**43.3**	86.7	**80.0**	**70.8**	61.7

Table 8: Validity-checklist pass rates (%) under the original T2V condition. V1–V4 denote EP (Entity Presence), SE (Scenario Establishment), EA (Evidence Availability), and DO (Decision Opportunity), respectively. VR denotes the percentage of videos for which all four checks pass. Within each task and check, the best and second-best values among open-source models and among closed-source models are shown in bold and underlined, respectively.

		Check	Open-Source: Wan2.2 T2V-A14B	Open-Source: LTX-2.3 22B	Open-Source: LongCat 13.6B	Open-Source: Hunyuan 1.5-8.3B	Closed-Source: Kling-3.0	Closed-Source: HappyHorse 1.1	Closed-Source: Dreamina Seedance-2.0
Belief	***PB***	**PD** EP	80.0	**96.7**	76.7	56.7	90.0	**96.7**	93.3
		PD SE	**66.7**	63.3	40.0	50.0	**90.0**	83.3	86.7
		PD EA	**100.0**	96.7	86.7	80.0	**100.0**	96.7	**100.0**
		PD DO	**66.7**	**66.7**	43.3	50.0	86.7	86.7	**90.0**
		PD **VR**	56.7	**60.0**	26.7	46.7	**86.7**	83.3	83.3
		EC EP	**80.0**	**80.0**	73.3	66.7	**86.7**	83.3	63.3
		EC SE	30.0	**46.7**	26.7	26.7	70.0	**76.7**	70.0
		EC EA	50.0	**63.3**	33.3	43.3	90.0	**100.0**	93.3
		EC DO	50.0	**66.7**	36.7	46.7	96.7	**100.0**	96.7
		EC **VR**	10.0	**40.0**	13.3	16.7	60.0	**73.3**	50.0
	HB	**CR** EP	**83.3**	80.0	80.0	73.3	96.7	**100.0**	96.6
		CR SE	**70.0**	66.7	53.3	60.0	**90.0**	86.7	86.2
		CR EA	73.3	80.0	**86.7**	66.7	**100.0**	**100.0**	96.6

Continued on next page

				Open-Source				Closed-Source		
			Check	Wan2.2 T2V-A14B	LTX-2.3 22B	LongCat 13.6B	Hunyuan 1.5-8.3B	Kling-3.0	HappyHorse 1.1	Dreamina Seedance-2.0
			DO	63.3	70.0	**80.0**	66.7	90.0	90.0	**96.6**
			VR	53.3	**60.0**	46.7	50.0	**80.0**	**80.0**	79.3
Desire	*ID*	**SP**	EP	83.3	**90.0**	86.7	73.3	93.3	96.7	**100.0**
			SE	**93.3**	**93.3**	70.0	73.3	**100.0**	**100.0**	96.7
			EA	86.7	**93.3**	73.3	66.7	**96.7**	**96.7**	**96.7**
			DO	26.7	**70.0**	26.7	26.7	83.3	86.7	**90.0**
			VR	23.3	**70.0**	23.3	20.0	76.7	86.7	**90.0**
		EV	EP	**96.7**	**96.7**	86.7	73.3	93.3	**100.0**	**100.0**
			SE	**80.0**	**80.0**	56.7	46.7	**100.0**	93.3	93.3
			EA	46.7	**86.7**	53.3	46.7	**90.0**	86.7	86.7
			DO	**66.7**	63.3	33.3	36.7	86.7	86.7	**93.3**
			VR	36.7	**53.3**	16.7	23.3	80.0	76.7	**83.3**
	SD	**RM**	EP	66.7	**80.0**	63.3	53.3	96.7	**100.0**	**100.0**
			SE	53.3	**73.3**	46.7	46.7	96.7	**100.0**	96.7
			EA	70.0	**80.0**	70.0	46.7	93.3	**100.0**	**100.0**
			DO	33.3	**70.0**	40.0	40.0	93.3	**96.7**	90.0
			VR	33.3	**66.7**	33.3	30.0	90.0	**96.7**	90.0
		ER	EP	**100.0**	96.7	90.0	60.0	90.0	**100.0**	93.3
			SE	**76.7**	66.7	40.0	43.3	93.3	90.0	**96.7**
			EA	**96.7**	90.0	90.0	43.3	96.7	**100.0**	93.3
			DO	**83.3**	70.0	50.0	53.3	**96.7**	90.0	93.3
			VR	**73.3**	63.3	30.0	20.0	86.7	**90.0**	**90.0**
		SC	EP	**83.3**	70.0	60.0	40.0	**96.7**	**96.7**	90.0
			SE	**93.3**	90.0	73.3	80.0	**100.0**	96.7	96.7
			EA	**83.3**	73.3	63.3	46.7	96.7	**100.0**	93.3
			DO	**66.7**	63.3	56.7	36.7	86.7	**96.7**	76.7
			VR	**63.3**	53.3	33.3	30.0	83.3	**90.0**	76.7
Intention	*PI*	**GP**	EP	**86.7**	73.3	**86.7**	76.7	**90.0**	86.7	**90.0**
			SE	**46.7**	30.0	40.0	23.3	**90.0**	73.3	**90.0**
			EA	**56.7**	30.0	50.0	26.7	**96.7**	90.0	93.3
			DO	36.7	30.0	**43.3**	20.0	73.3	**93.3**	90.0
			VR	30.0	26.7	**33.3**	16.7	66.7	73.3	**83.3**
		PA	EP	**86.7**	76.7	76.7	66.7	**100.0**	93.3	**100.0**
			SE	**63.3**	50.0	50.0	33.3	86.7	80.0	**90.0**
			EA	**80.0**	66.7	73.3	43.3	**100.0**	96.7	96.7
			DO	30.0	**53.3**	36.7	20.0	63.3	**76.7**	60.0
			VR	26.7	**46.7**	36.7	16.7	56.7	**70.0**	53.3
	MI	**RC**	EP	**73.3**	**73.3**	63.3	46.7	96.7	**100.0**	**100.0**
			SE	60.0	**80.0**	46.7	46.7	93.3	**100.0**	**100.0**
			EA	63.3	**76.7**	53.3	56.7	93.3	**100.0**	**100.0**
			DO	33.3	**63.3**	33.3	30.0	63.3	**93.3**	70.0
			VR	33.3	**56.7**	26.7	30.0	53.3	**93.3**	70.0

Continued on next page

		Check	Open-Source				Closed-Source		
			Wan2.2 T2V-A14B	LTX-2.3 22B	LongCat 13.6B	Hunyuan 1.5-8.3B	Kling-3.0	HappyHorse 1.1	Dreamina Seedance-2.0
Intention MI	**CP**	EP	83.3	90.0	**93.3**	66.7	**96.7**	93.3	**96.7**
		SE	73.3	83.3	**86.7**	43.3	96.7	96.7	**100.0**
		EA	63.3	**80.0**	76.7	46.7	**96.7**	**96.7**	93.3
		DO	46.7	**80.0**	73.3	40.0	**96.7**	**96.7**	80.0
		VR	43.3	**70.0**	63.3	36.7	**96.7**	93.3	80.0
	Overall	EP	**83.6**	**83.6**	78.1	62.8	93.9	**95.6**	93.6
		SE	67.2	**68.6**	52.5	47.8	**92.2**	89.7	91.9
		EA	72.5	**76.4**	67.5	51.1	95.8	**96.9**	95.3
		DO	50.3	**63.9**	46.1	38.9	84.7	**91.1**	85.5
		VR	40.3	**55.6**	31.9	28.1	76.4	**83.9**	77.4

Table 9: Wan2.2 model-variant comparison under original T2V. Belief, Desire, and Intention report component means of task-level VR, CSR, and OSR; Overall is the macro mean over the twelve tasks. Best / second-best in **bold** / underlined.

Model	Belief			Desire			Intention			Overall		
	VR	CSR	OSR	VR	CSR	OSR	VR	CSR	OSR	VR	CSR	OSR
Wan2.2-TI2V-5B	**41.1**	49.3	**18.9**	44.7	**40.3**	**16.0**	**38.3**	55.9	**25.0**	**41.7**	47.7	**19.7**
Wan2.2-T2V-A14B	40.0	**55.0**	15.6	**46.0**	38.2	14.0	33.3	**66.1**	22.5	40.3	**51.7**	17.2

Table 10: LongCat-Video-13.6B duration comparison under original T2V (5 / 10 / 15 / 20 s). Aggregation matches Table 9. Best / second-best in **bold** / underlined.

Duration	Belief			Desire			Intention			Overall		
	VR	CSR	OSR	VR	CSR	OSR	VR	CSR	OSR	VR	CSR	OSR
5s	28.9	**46.5**	11.1	**31.3**	45.1	14.0	30.8	**72.5**	24.2	30.6	**54.6**	**16.7**
10s	28.9	35.7	6.7	27.3	**55.8**	**14.0**	**40.0**	61.6	**25.8**	**31.9**	52.7	16.1
15s	**30.0**	41.5	**11.1**	28.7	44.9	10.7	30.0	67.3	21.7	29.4	51.5	14.4
20s	26.7	32.4	5.6	18.7	35.0	6.7	15.8	50.0	8.3	19.7	39.4	6.9

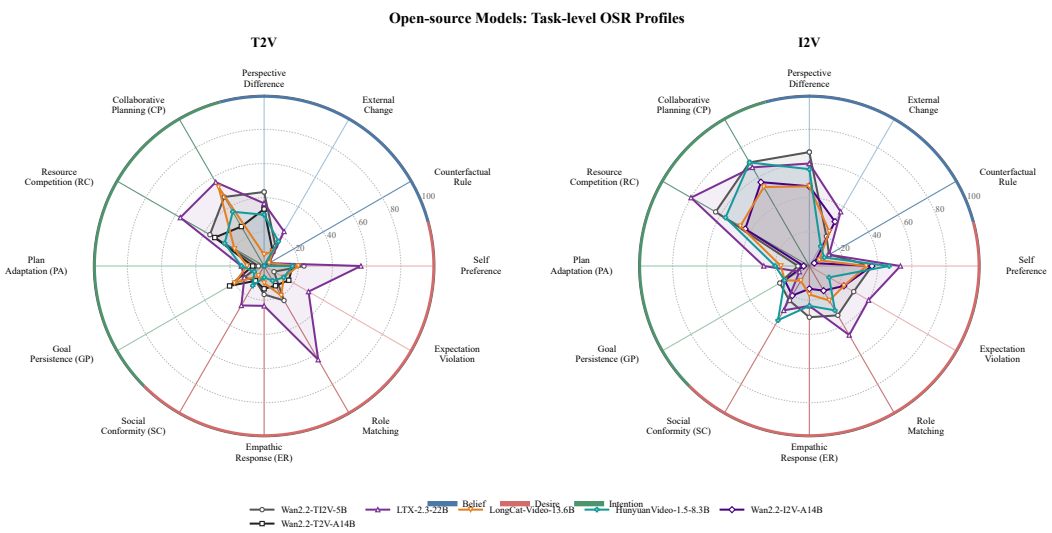


Figure 9: Open-source models: task-level OSR profiles under T2V (left) and I2V (right). Axes follow the twelve-task taxonomy (PD, EC, CR, SP, EV, RM, ER, SC, GP, PA, RC, CP), one axis per task. Wan 14B uses the mode-matched variant (`Wan2.2-T2V-A14B` / `Wan2.2-I2V-A14B`). Outer arcs mark Belief, Desire, and Intention. Values are percentages on a shared 0–100 scale.

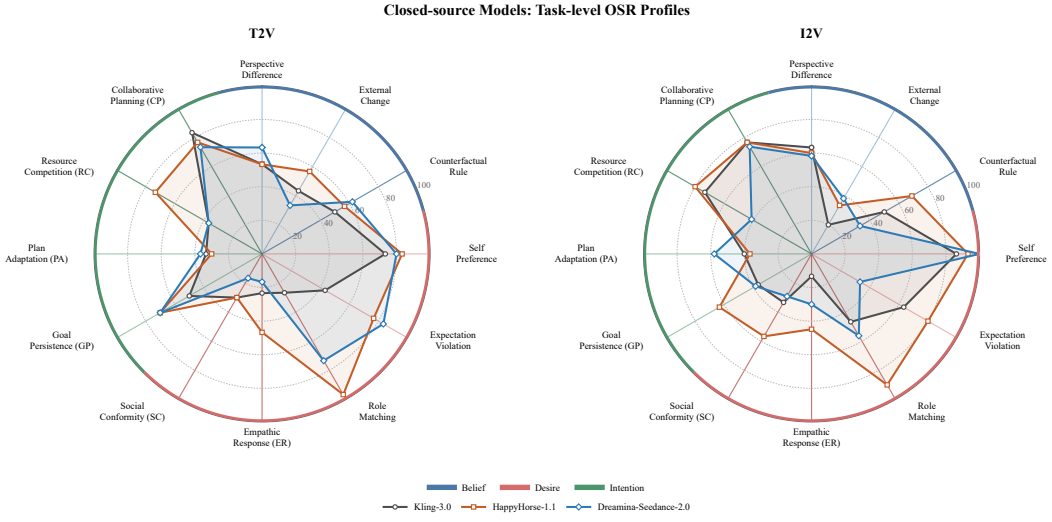


Figure 10: Closed-source models: task-level OSR profiles under T2V (left) and I2V (right). Axis ordering, scale, and taxonomy encoding match Figure 9.

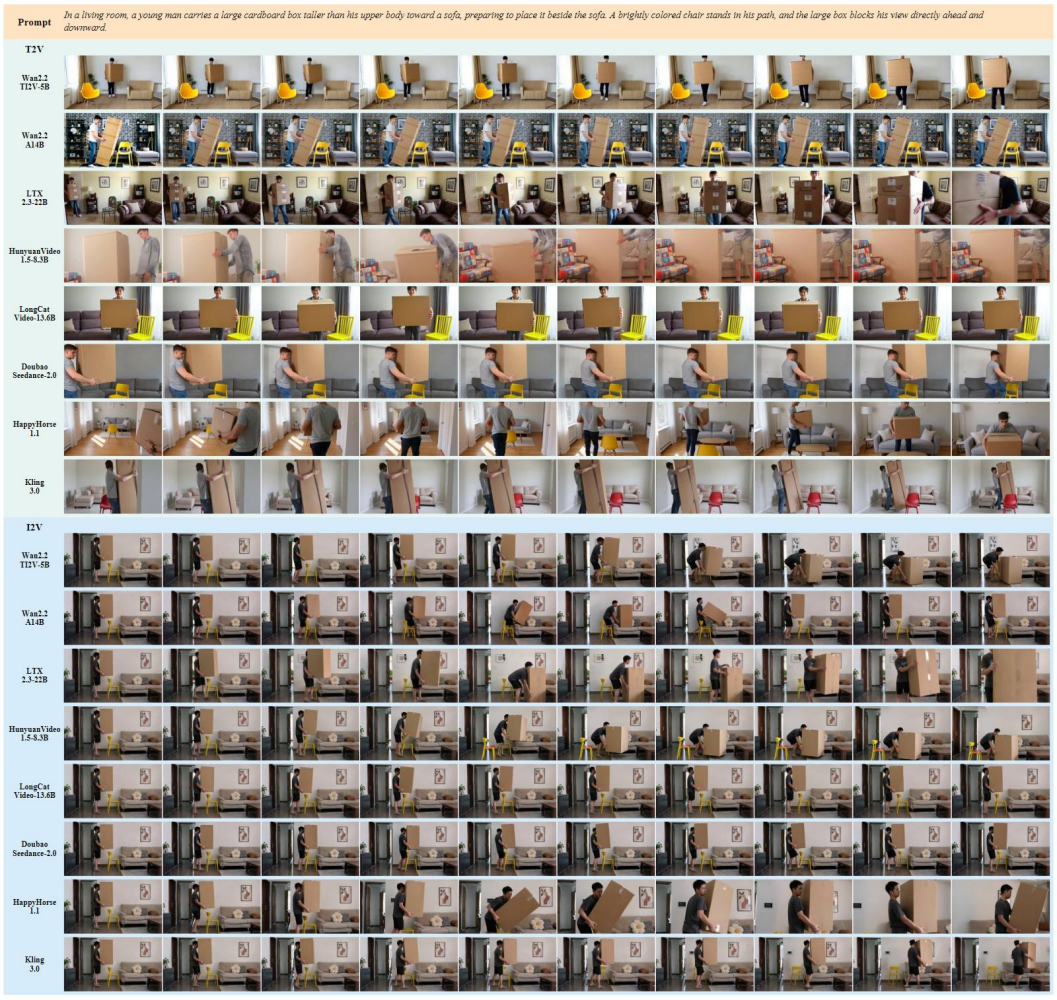


Figure 11: Perspective Difference (PD): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

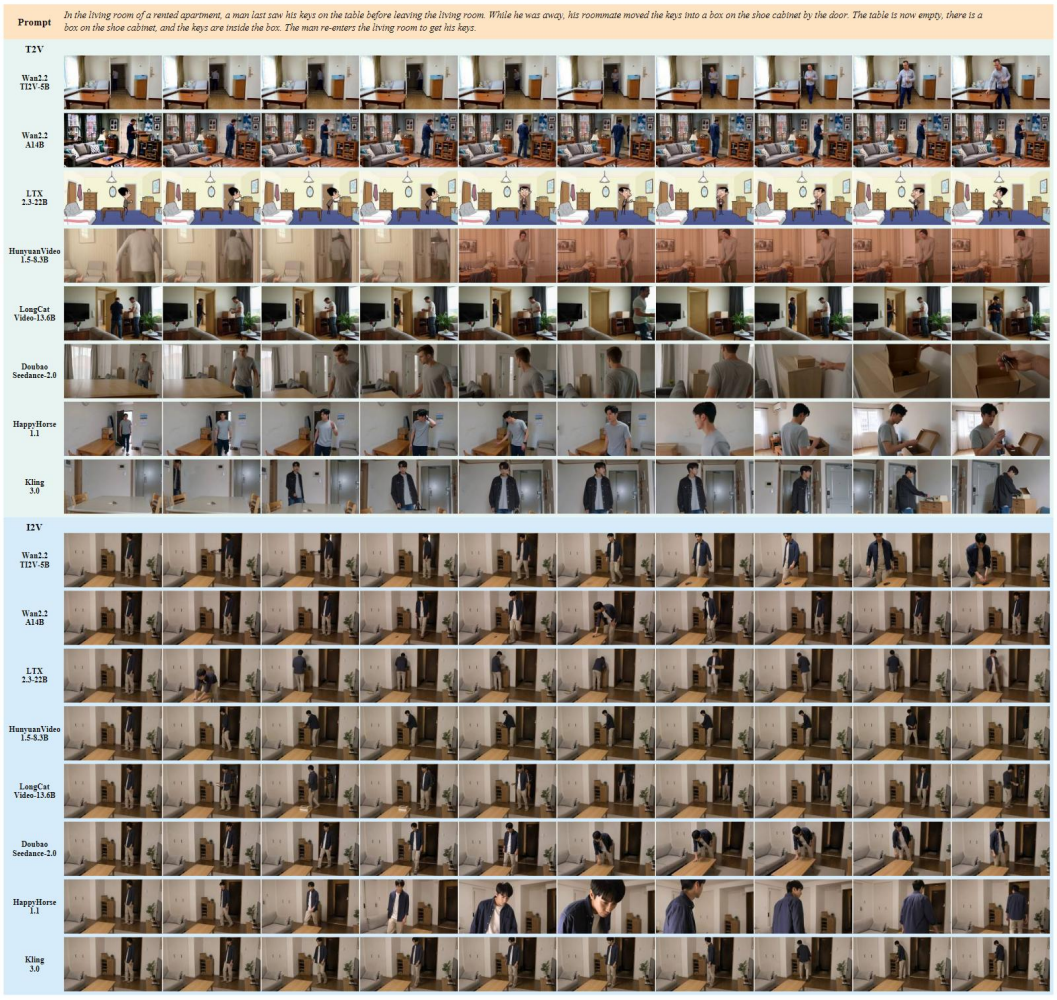


Figure 12: External Change (EC): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

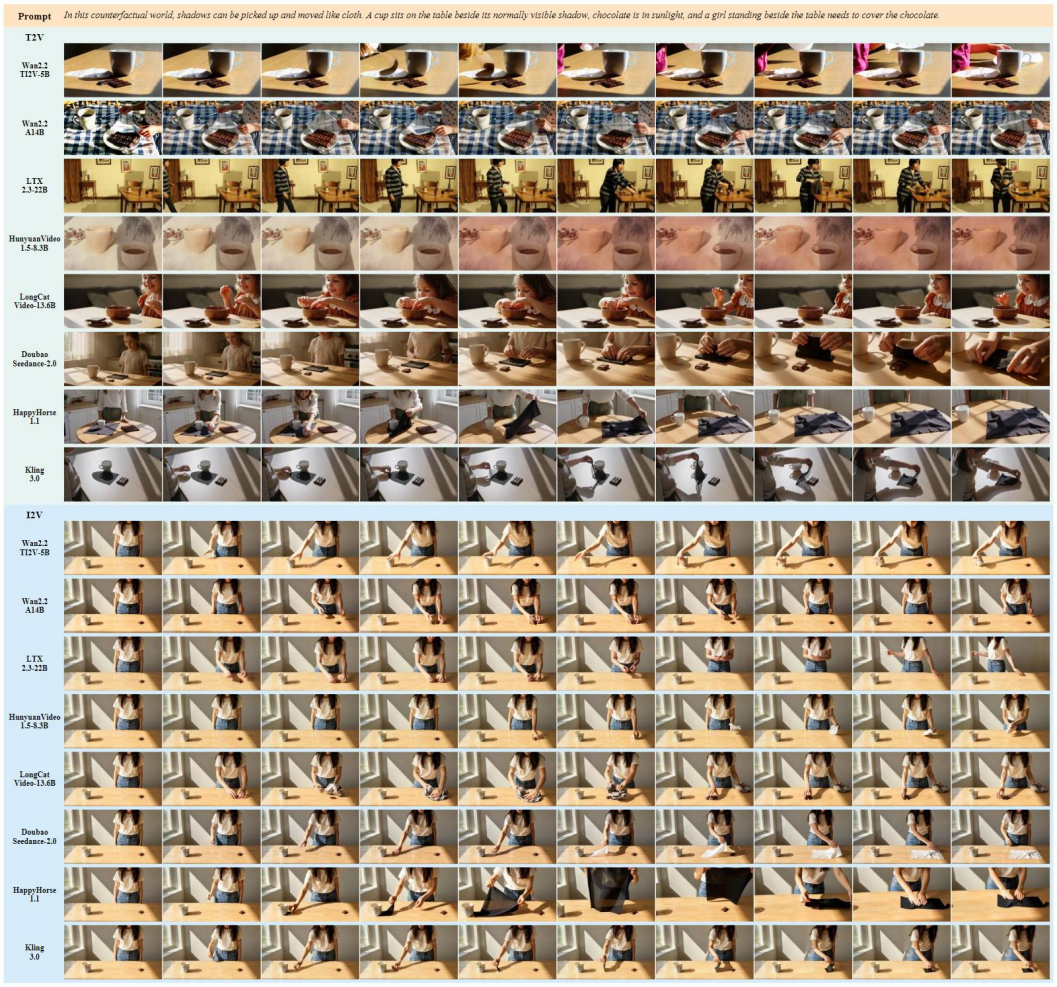


Figure 13: Counterfactual Rule (CR): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

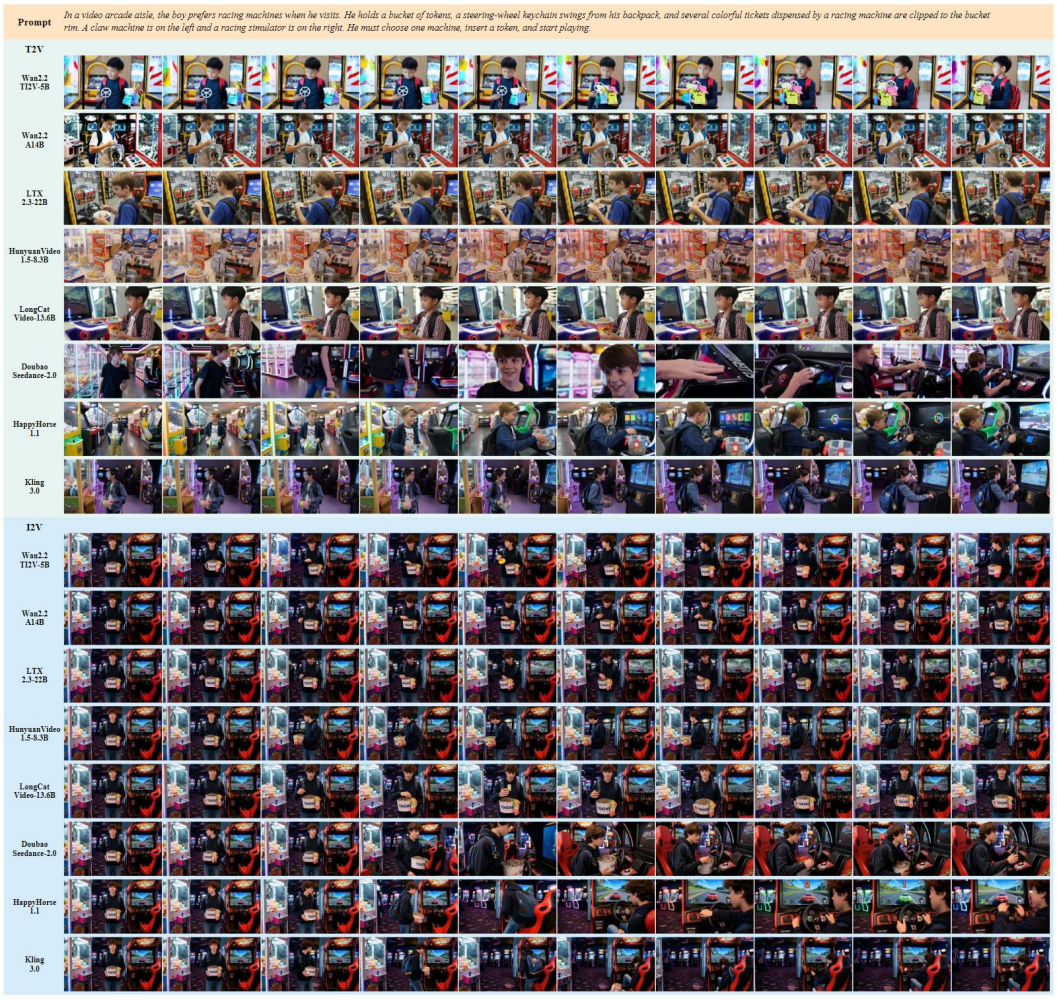


Figure 14: Self Preference (SP): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

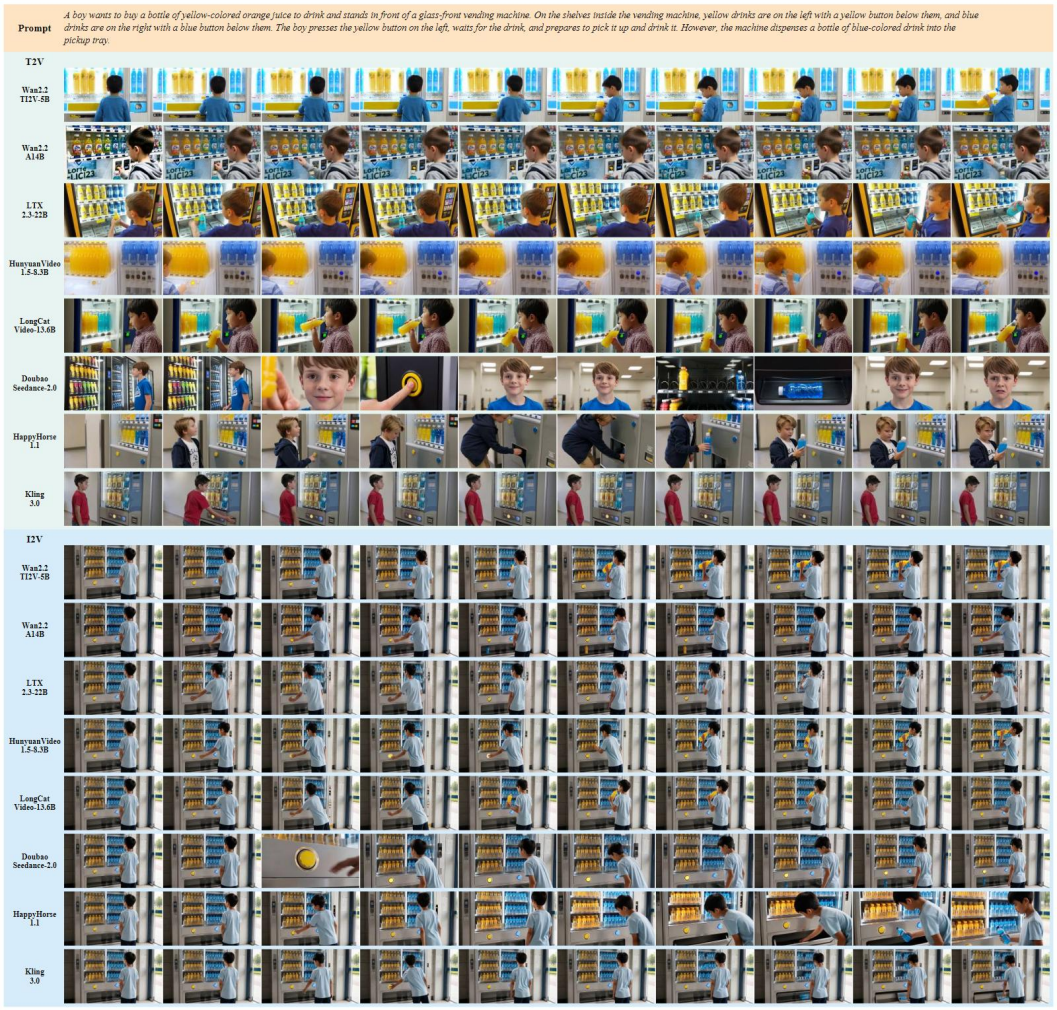


Figure 15: Expectation Violation (EV): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.



Figure 16: Role Matching (RM): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

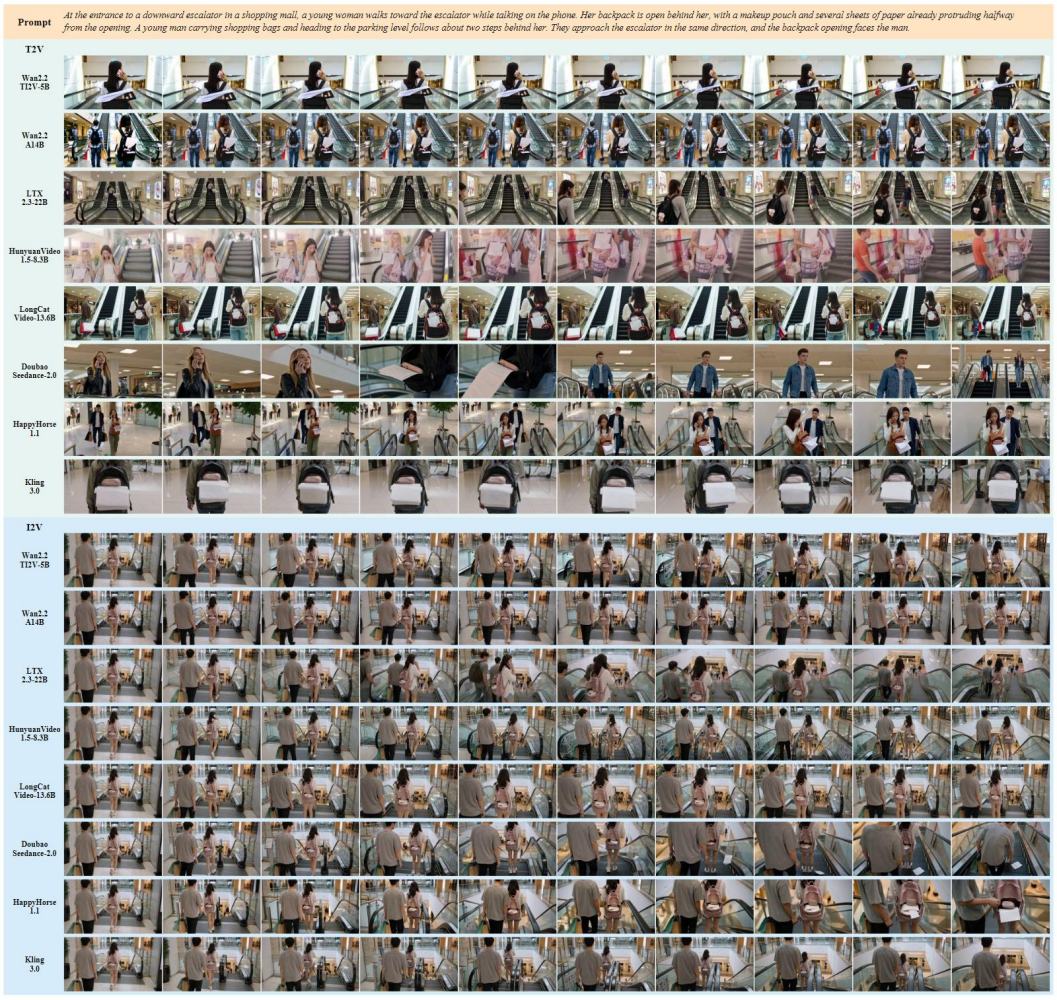


Figure 17: Empathic Response (ER): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

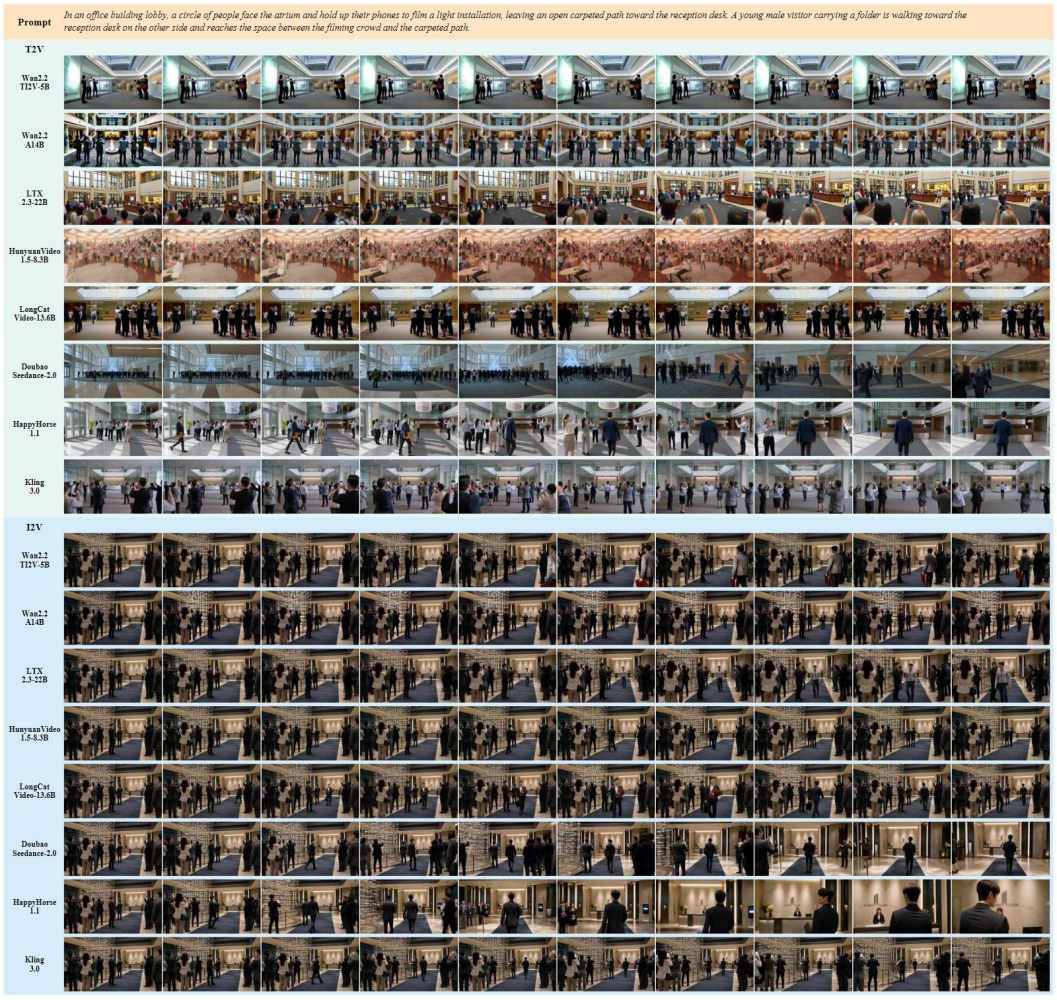


Figure 18: Social Conformity (SC): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

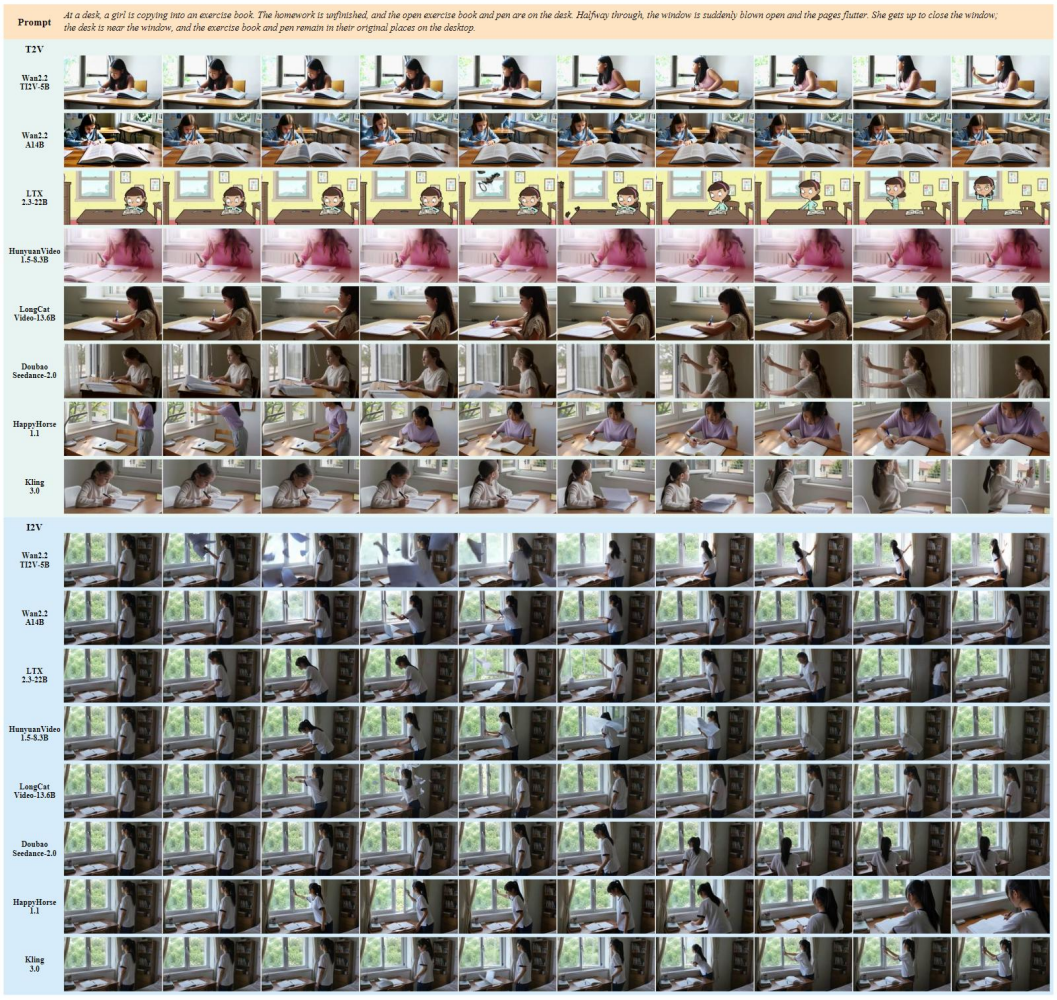


Figure 19: Goal Persistence (GP): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.



Figure 20: Plan Adaptation (PA): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

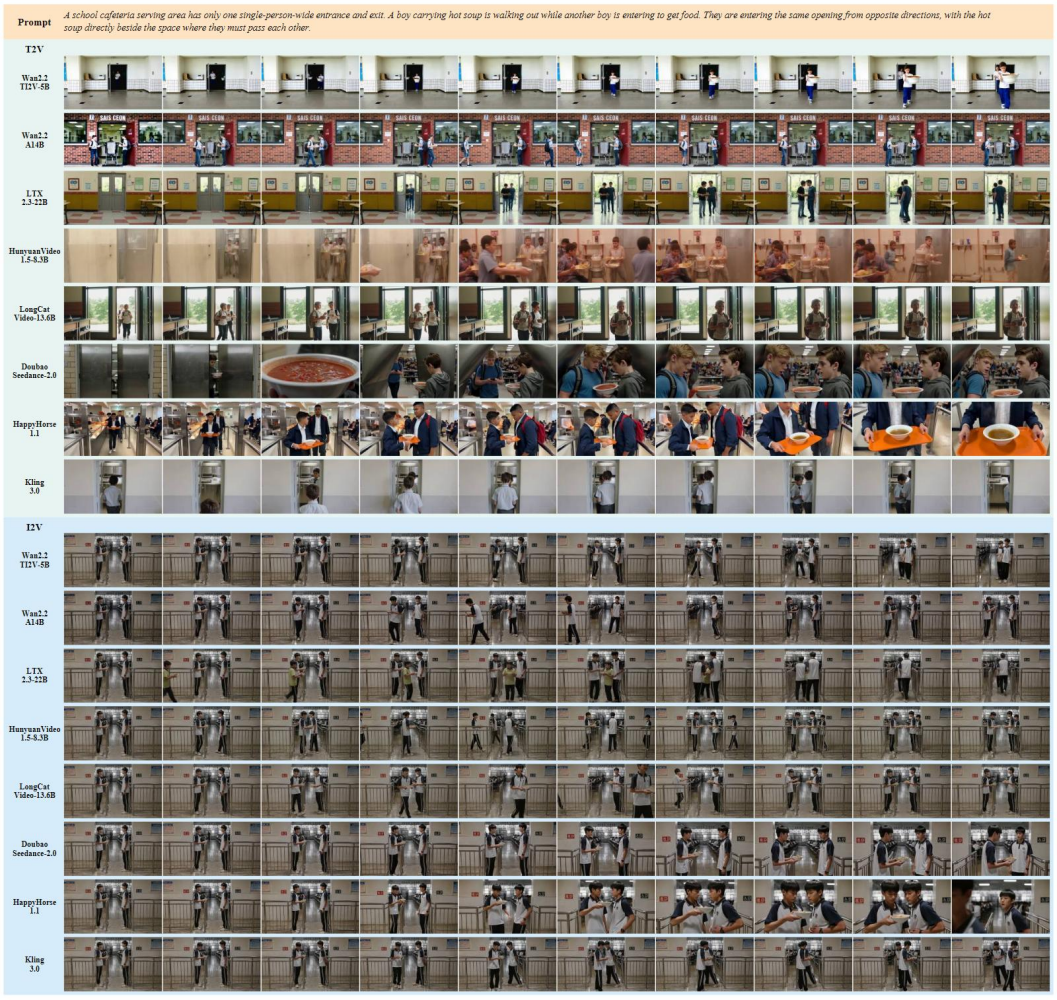


Figure 21: Resource Competition (RC): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.

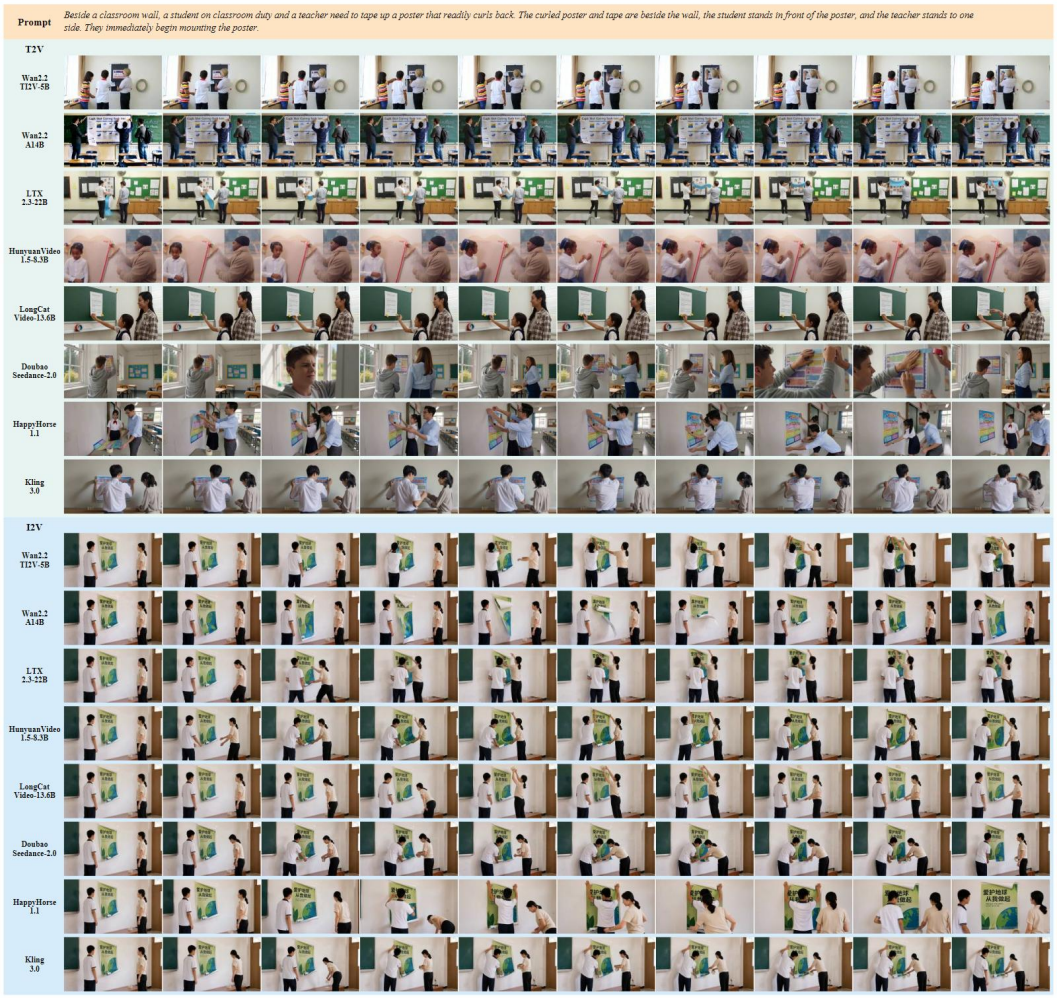


Figure 22: Collaborative Planning (CP): cross-model qualitative comparison on the same instance under T2V (top) and I2V (bottom), without Prompt Enhancement. Each row shows ten frames from one model.