Reality vs Counterfactual: Multi-World Contrastive Reinforcement Learning for Enhancing MLLM's Theory of Mind in Egocentric Videos

Guiyang Hou[1*], Yihui Fu[2*], Chen Wu[1], Xiang Huang[3], Zhe Zheng[1], Wenqi Zhang[1], Yongliang Shen[1], Weiming Lu[1†]

[1]Zhejiang University
[2]Northwest University
[3]Tongyi Lab
{gyhou, luwm}@zju.edu.cn

Abstract

Theory of Mind (ToM) refers to the ability to infer others' mental states, which is an essential capability for embodied AI agents to effectively collaborate and interact with humans. While improving Large Language Models' ability to reason about characters' mental states in text-based stories/dialogues has been extensively studied, enhancing Multimodal Large Language Models' ToM capabilities, particularly in egocentric video from an embodied perspective, remains unexplored. In this paper, we propose a contrastive Reinforcement Learning (RL) paradigm that explicitly encourages models to leverage temporal and causal evolutionary patterns in user action sequences to infer user's mental states (goals, beliefs, and potential next actions). Evaluation results on in-domain and out-of-domain demonstrate that our method achieves performance improvements of (+30.00%, +2.00%) and (+5.83%, +5.00%) compared to the backbone model and vanilla Group Relative Policy Optimization (GRPO) model, respectively. Additionally, we compare the performance of two post-training paradigms (Supervise Fine-Tuning and RL) and systematically analyze the reasoning trajectories across the base model, vanilla GRPO model, and our proposed method.

Introduction

Embodied AI agents need to construct representations of the external world, namely physical world model (Hu and Shu 2023; Xu et al. 2024; Ding et al. 2024; Ge et al. 2024; Bordes et al. 2025), to understand the laws of physics (for example, that a small ball will bounce after hitting the ground). This has become a widely accepted view. In addition, as illustrated in Figure 1 Top, we argue that they also need to construct representations of the human user's mental states, namely mental world model, in order to understand, when the user takes certain actions, their underlying goals, beliefs, and likely next actions, as well as how their emotions might change after experiencing an event. This representation of the human user's mental state is fundamentally rooted in Theory of Mind (ToM)—the ability to infer unobserved mental states that reflect what users want, think, or believe, etc (Premack and Woodruff 1978; Baron-Cohen, Leslie, and Frith 1985; Perner and Wimmer 1985; Perner, Leekam, and Wimmer 1987; Gandhi et al. 2023; Hou et al. 2024).

Prior work on enhancing Large Language Models' (LLMs) ToM capabilities has primarily focused on inferring character mental states within textual narratives and dialogues (see Figure 1, bottom panel). These methods can be broadly categorized into prompt-based methods (Wilf et al. 2024; Jung et al. 2024; Shinoda et al. 2025; Sarangi, Elgarf, and Salam 2025), tool-augmented methods (Huang et al. 2024; Sclar et al. 2023), and Bayesian model-based methods (Zhang et al. 2025; Shi et al. 2025; Jin et al. 2024). While these methods have achieved promising results, LLMs' ToM capabilities in text-based scenarios are misaligned with the application contexts of embodied AI agents. Enhancing Multimodal Large Language Models' (MLLMs) ToM abilities, particularly from egocentric video perspectives, would better align with embodied AI agent applications, enabling embodied agents to construct more accurate representations of human users' mental states and thereby achieve superior human-agent interaction. However, this direction remains largely unexplored in current research.

In this paper, we propose **M**ulti-**W**orld **C**ontrastive **R**einforcement **L**earning (**MWCRL**) to enhance MLLMs' ToM capabilities in egocentric videos from an embodied perspective. MWCRL extends the original Group Relative Policy Optimization (GRPO) algorithm (Guo et al. 2025) by explicitly encouraging the exploitation of temporal and causal evolutionary patterns in user action sequences to infer users' future goals, beliefs, and potential next actions. During training, the model is presented with two types of human action sequences: those following real-world operational rules and counterfactual sequences generated by shuffling the real-world human action sequences, producing two groups of responses. A positive reward is assigned only when the proportion of correct answers from the real-world human action sequences exceeds that from the counterfactual sequences.

In our experiments, we validate the effectiveness of our proposed MWCRL method through both In-Domain and Out-of-Domain (OOD) evaluations. Under In-Domain eval-

*These authors contributed equally.
†Corresponding author.

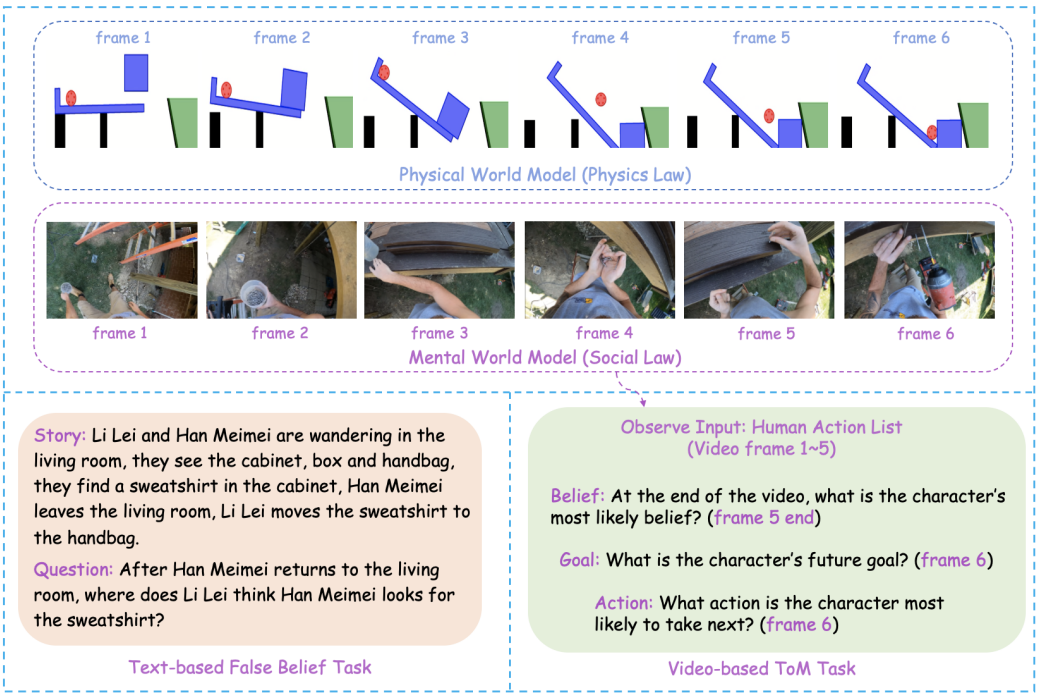

Figure 1: **Top**: Physical World Model and Mental World Model that reflect physical laws and social laws. **Bottom**: data sample of text-based false belief tasks and video-based ToM tasks.

uation, our method achieves a comprehensive performance of 84.67%, demonstrating improvements of +30.00% and +2.00% compared to the backbone model and Vanilla GRPO model, respectively. For OOD evaluation, we construct four categories of questions: action causal dependencies, humans' potential next actions, beliefs, and multi-hop reasoning (first inferring a human's potential next action and then its effect on object states) to test the model's generalization capability. Our method achieves improvements of +5.83% and +5.00% compared to the backbone model and Vanilla GRPO model, respectively.

Furthermore, we conduct comparative analyses between Supervised Fine-Tuning (SFT) and RL post-training paradigms and perform detailed comparisons of reasoning trajectories across the Base Model, Vanilla GRPO model, and our MWCRL method. We find that models trained with MWCRL: (1) genuinely learn to infer users' mental states based on temporal and causal evolutionary patterns in user action sequences, rather than relying on shortcuts derived from isolated actions as observed in Vanilla GRPO models; (2) exhibit novel reasoning behaviors in their inference trajectories that do not appear in the Base Model, such as "Let's verify" and "Oh, I know", commonly referred to as "aha moments" (Kounios and Beeman 2009; Yang et al. 2025; Zhou et al. 2025a; Feng et al. 2025a).

Our contributions include: (1) We argue that for Embodied AI Agents, beyond constructing internal representations of physical world laws (i.e., physical world model), it is equally important to build internal representations of human mental states (i.e., mental world model). (2) We propose a multi-world contrastive reinforcement learning approach that explicitly encourages models to leverage temporal and causal evolutionary patterns in user action sequences to infer users' mental states (goals, beliefs, and potential next actions). (3) Quantitative evaluation experiments on both in-domain and OOD settings demonstrate the effectiveness of our method, while qualitative case studies reveal various interesting behaviors exhibited in our model's reasoning trajectories. (4) To the best of our knowledge, we are the first to investigate enhancing MLLMs' ToM capabilities, particularly in egocentric videos from an embodied perspective. We highlight that our model is well-suited for applications in embodied wearable agents.

Background and Related Works

Strategies for Enhancing LLMs' ToM Capabilities in Text Domain

SimToM (Wilf et al. 2024) guides LLMs to adopt perspective-taking cognitive strategies, while PercepToM (Jung et al. 2024) enhances perception-to-belief inference by identifying salient contextual cues. Huang et al. (2024) employs an LLM as a world model to monitor dynamic changes in environmental entities and character belief states. To address complex social reasoning, Hou et al. (2024) introduces a belief solver that reduces higher-order ToM problems into lower-order ones through temporal set intersections. SymbolicToM (Sclar et al. 2023) leverages graphical structures to explicitly represent and update agents' belief states. Additionally, AutoToM, MMToM, and

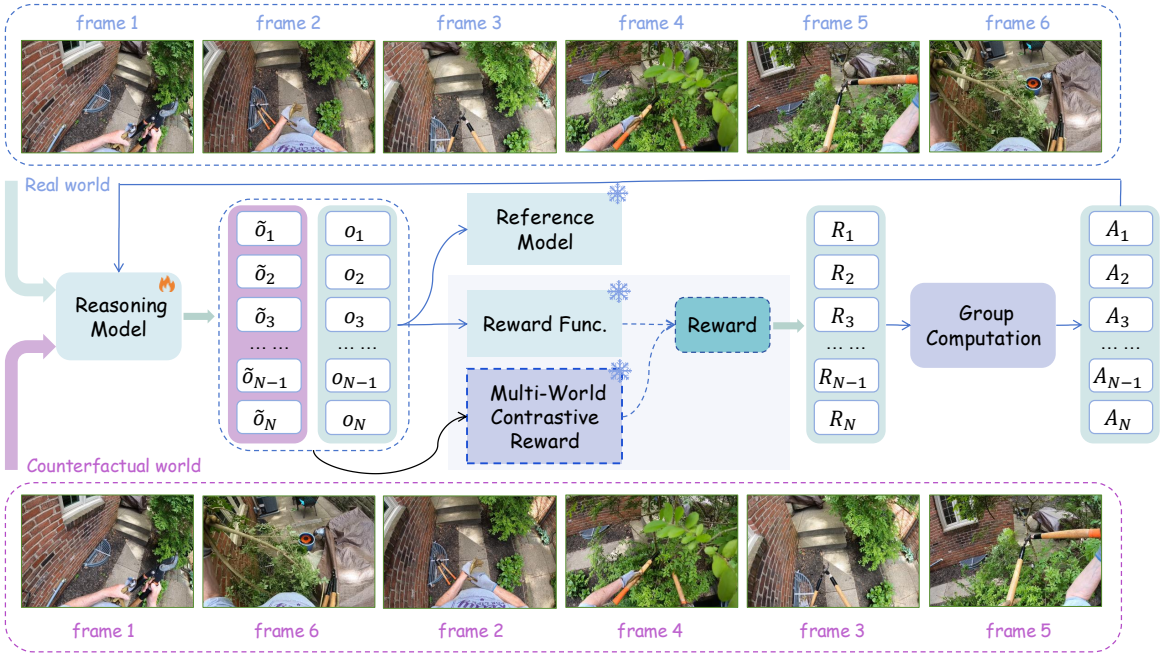

Figure 2: An illustration of multi-world contrastive reinforcement learning via GRPO.

MuMA-ToM (Zhang et al. 2025; Shi et al. 2025; Jin et al. 2024) propose Bayesian model methods. Recently, Hou et al. (2025) systematically explored improving LLMs' ToM capabilities in the text domain from the perspectives of post-training and test-time intervention. However, **a notable gap remains with respect to the modalities studied**.

LLM Post-Training

SFT and RL have emerged as predominant paradigms for LLM post-training to improve performance on specific tasks. There are also multiple studies exploring and analyzing these two different post-training paradigms, such as "What LLMs Can—and Still Can't—Solve after SFT?" (Chen et al. 2025; Sun et al. 2025) and "SFT Memorizes, RL Generalizes" observations (Chu et al. 2025). RL-based post-training has been applied to multiple domains, such as image classification, emotion classification (Li et al. 2025b; He et al. 2025), search engine calling (Jin et al. 2025; Song et al. 2025), video reasoning (Feng et al. 2025a), logic puzzles (Xie et al. 2025), machine translation (Feng et al. 2025b), gui agents (Luo et al. 2025), text to sql (Ma et al. 2025), and more (Li et al. 2025a; Xia and Luo 2025; Zhou et al. 2025b). However, **RL-based post-training to enhance MLLMs' ToM capabilities remains unexplored**.

MWCRL Method for Embodied Application

Wearable Agents, as one type of embodied agents (Dong 2024; Waisberg et al. 2024; Fung et al. 2025), are capable of capturing the physical environment around human users from a first-person perspective, seeing what the user sees, thus creating a "shared perceptual field." By sharing the user's first-person perspective, our MWCRL-trained MLLM adapts well to this agent application scenario, can infer the user's mental states, such as goals and beliefs, enabling the wearable agent system to provide guidance, support, and personalized experiences across diverse tasks.

Multi-World Contrastive Reinforcement Learning via GRPO

GRPO based solely on outcome and format rewards lacks explicit reward signals for learning the temporal and causal evolutionary patterns within human user action sequences. To address this limitation, we propose **M**ulti-**W**orld **C**ontrastive **R**einforcement **L**earning (**MWCRL**), which introduces a contrastive reward mechanism that explicitly encourages the exploitation of temporal and causal evolutionary patterns in user action sequences to infer users' future goals, beliefs, and potential next actions, as illustrated in Figure 2.

Training Template

We follow Guo et al. (2025)'s paradigm, encouraging models to engage in a reasoning (thinking) process before producing the final answer. The prompt is defined as follows: Please output your answer in the format: `<think>` reasoning process here `</think>` `<answer>` answer here `</answer>`.

Reward Modeling

Following Guo et al. (2025) and Feng et al. (2025a), the reward function consists of four components: format reward, outcome reward, length-based reward, and multi-world contrastive reward. The format reward validates that responses

adhere to the required structural format, ensuring that all elements appear in the correct sequence and are enclosed within appropriate tags:

$$r_{\text{format}} = \begin{cases} 1, & \text{if format is correct} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The outcome reward is a rule-based metric that verifies whether the content enclosed in the <answer> </answer> tags exactly matches the ground truth (gt) label, which is designed as follows:

$$r_{\text{accuracy}} = \begin{cases} 1, & \text{if answer tag exist and answer match} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The length-based reward aims to strike a balance between encouraging deeper reasoning and preventing overthinking. For each reasoning path o_i, an additional reward $r_{length} = \omega$ is assigned when two conditions are simultaneously satisfied: (1) the predicted answer is correct, and (2) the response length lies within the predefined bounds $[l_{\min}, l_{\max}]$. Formally:

$$r_{length} = \begin{cases} \omega, & \text{if } o_i \text{ is correct and } l_{\min} \leq \text{len}(o_i) \leq l_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

This reward encourages the model to think deeply without overthinking.

The multi-world contrastive reward is a mechanism that explicitly encourages the exploitation of temporal and causal evolutionary patterns in user action sequences to infer user's mental states (goals, beliefs, and potential next actions). The core idea involves comparing the model's performance on the same ToM question when human action sequence are provided in two different patterns: (1) following real-world operation patterns, and (2) shuffled conterfactual world patterns. For each input question, we generate two groups of responses $\{o_i\}_{i=1}^{N}$ and $\{\tilde{o_i}\}_{i=1}^{\tilde{N}}$ using the real-world and counterfactual world inputs, respectively. Let p and $\tilde{p}$ denote the proportion of correct answers in each group. We then define a multi-world contrastive reward coefficient r_{mwc} as:

$$r_{mwc} = \begin{cases} \alpha, & \text{if } p > \mu \cdot \tilde{p} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where α and μ are hyper-parameters. *We conduct a systematic exploration of how to set these two hyperparameters in the Experiments section and propose some key insights*.

This contrastive design incentivizes the model to leverage temporal and causal evolutionary patterns in user action sequences to infer user's mental states (goals, beliefs, and potential next actions). The model receives this positive reinforcement only when its response strategy for a specific question demonstrates clear dependence on action sequence evolution patterns. The multi-world contrastive reward r_{mwc} is selectively applied only to correct responses, when the model successfully leverages action sequence evolution patterns, correct responses receive enhanced reinforcement through this higher reward, while incorrect responses remain unaffected.

The final reward function r is a combination of the four rewards and is defined as:

$$r = r_{\text{format}} + r_{\text{accuracy}} + r_{\text{length}} + r_{\text{mwc}} \tag{5}$$

Reinforcement Learning Algorithm: GRPO

Unlike SFT, which optimizes models through token-level losses, RL-based methods like GRPO utilize policy gradients, calculated from the reward loss, for optimization (Li et al. 2025b). This encourages exploring a much larger solution space (Guo et al. 2025).

Let Q be the question set, $\pi_{\theta_{\text{old}}}$ be the policy model and $\{o_1, o_2, \cdots, o_N\}$ be a group of responses from $\pi_{\theta_{\text{old}}}$ for a question q. Let $\pi_{\theta_{\text{ref}}}$ denote the frozen reference model. The GRPO algorithms aim to optimize model π_θ by the following objective:

$$\begin{aligned} J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim Q, \{o_i\}_{i=1}^{N} \sim \pi_{\theta_{\text{old}}}} \\ \Bigg[\frac{1}{N} \sum_{i=1}^{N} \min \Bigg(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \\ \text{clip} \left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon \right) A_i \Bigg) \\ - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Bigg] \end{aligned} \tag{6}$$

where ϵ and β are clipping hyper-parameter and the coefficient controlling the Kullback–Leibler (KL) penalty, respectively. Here,

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_N\})}{\text{std}(\{r_1, r_2, \ldots, r_N\})} \tag{7}$$

is the advantage using the group reward $\{r_1, r_2, \ldots, r_N\}$, and

$$D_{KL}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - \log \left(\frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} \right) - 1 \tag{8}$$

is the KL divergence loss.

GRPO eliminates the critic model in PPO (Schulman et al. 2017) by estimating the relative advantage by sampling a group of responses $\{o_i\}_{i=1}^{N}$ and normalizing their rewards within the group to compute a relative advantage, which is more computationally efficient (Shao et al. 2024).

Experiments

Baseline Methods

To evaluate the effectiveness of our proposed MWCRL method, we provide the performance of various baseline methods for comparison. These include the performance of six foundation models: Gemini-2.5-pro-preview-0506-thinking/no-thinking (Comanici et al. 2025), VideoLLaMA2-7B/72B (Cheng et al. 2024), InternVL3-8B (Zhu et al. 2025), and CogVLM2 (Hong et al. 2024); the performance of direct SFT on In-Domain data as well as SFT on mixed multi-domain data (Math, World Knowledge, Chart, Spatial, etc.); the performance of vanilla GRPO models that only apply format rewards and outcome rewards. In addition, human performance is also included.

Model	α	μ	Frames	Belief	Goal	Next Action	AVG
Backbone Models							
Qwen2.5-VL-7B-Instruct	-	-	32	56.00%	76.00%	32.00%	54.67%
Foundation Models							
VideoLLaMA2-7B	-	-	8	40.00%	76.00%	36.00%	50.67%
VideoLLaMA2-72B	-	-	8	44.00%	84.00%	40.00%	56.00%
VideoLLaMA2-7B	-	-	16	36.00%	72.00%	36.00%	48.00%
CogVLM2	-	-	24	40.00%	78.00%	40.00%	52.67%
InternVL3-8B	-	-	32	46.00%	84.00%	34.00%	54.67%
Gemini-2.5-pro-preview-0506-thinking	-	-	32	66.00%	90.00%	56.00%	70.67%
Gemini-2.5-pro-preview-0506	-	-	32	62.00%	88.00%	52.00%	67.33%
Gemini-2.5-pro	-	-	32	62.00%	90.00%	68.00%	73.33%
Human Performance							
Human	-	-	-	78.00%	90.00%	82.00%	83.33%
Our SFT-Based Models							
Direct SFT	-	-	32	46.00%	84.00%	38.00%	56.00%
Mixed Multi-domain SFT	-	-	32	52.00%	82.00%	36.00%	56.67%
Our RL-Based Models							
Vanilla GRPO	-	-	32	74.00%	98.00%	76.00%	82.67%
MWCRL GRPO	0.3	0.8	32	82.00%	94.00%	78.00%	84.67%
	0.3	1.0	32	80.00%	98.00%	74.00%	84.00%
	0.3	0.6	32	76.00%	98.00%	82.00%	85.33%
	0.6	0.8	32	76.00%	96.00%	74.00%	82.00%
	0.6	1.0	32	76.00%	92.00%	70.00%	79.33%
Δ_{Backbone}	-	-	-	+26.00%	+18.00%	+46.00%	+30.00%
$\Delta_{\text{Vanilla GRPO}}$	-	-	-	+8.00%	- 4.00%	+ 2.00%	+2.00%

Table 1: Performance evaluation of multiple methods and foundation models in In-Domain scenarios. Δ_{Backbone} represents the performance improvement brought by our proposed MWCRL method when applied to the backbone model, and $\Delta_{\text{Vanilla GRPO}}$ represents the performance advantage of our proposed MWCRL method compared to the widely adopted GRPO (only apply format reward and outcome reward). The results for the VideoLLaMA2-7B/72B, CogVLM2, and human performance are from (Li et al. 2025c).

Implementation Details

We adopt the Qwen2.5-VL-7B-Instruct model (Bai et al. 2025) as the backbone and implement our RL-based methods using the VeRL framework (Sheng et al. 2025). We set $\omega = 0.1$, $\alpha = 0.3$, $\mu = 0.8$ and fix the random seed to 42. DeepSpeed ZeRO-3 is employed for parameter sharding and memory optimization. Training is performed with a per-device batch size of 1 and gradient accumulation steps of 1, resulting in a global batch size of 4. We set the maximum input prompt length to 16,384 tokens and the maximum generation length to 768 tokens. The visual input resolution is limited to 401,408 pixels. The optimizer is initialized with a learning rate of 1×10^{-6} and employs cosine learning rate scheduling with a weight decay of 0.01. All models are trained using bf16 precision on 4 × NVIDIA A800 (80GB) GPUs, for a total of 3 epochs. Each RL training run takes approximately 8∼10 hours to complete, model checkpoints are saved every 100 training steps. Additional training diagnostics, including the dynamics of response lengths and reward trends over time, are provided in the *supplementary material*.

Data and Metrics

We split the data from EgoToM (Li et al. 2025c) for post-training and in-domain evaluation. EgoToM contains a total of 774 30-second videos and 1,022 questions, of which 872 are allocated to the training set and 150 to the test set. For OOD evaluation data, we select from EgoTaskQA (Jia et al. 2022), including four categories of questions: action causal dependencies, humans' potential next actions, beliefs, and multi-hop reasoning (first inferring a human's potential next action and then its effect on object states), with 30 videos per category (120 videos in total). We use the accuracy of question answering as a metric to measure performance.

Main Results

As shown in Table 1, in the In-Domain evaluation, MWCRL achieves substantial performance gains, outperforming the backbone model by 30.00% and the widely-adopted vanilla GRPO method by 2.00%. Compared to advanced foundation

models such as Gemini-2.5-pro-preview and InternVL3, our method demonstrates significant performance advantages. For the post-training paradigm of SFT, Direct SFT on the in-domain dataset and Mixed Multi-domain SFT achieve 56.00% and 56.67% performance respectively, still exhibiting a considerable performance gap compared to RL-based methods. Notably, our method achieves a comprehensive performance score of 84.67% with only 7B parameters, approaching human-level performance.

As shown in Table 2, in the OOD evaluation, we carefully select four categories of questions to assess the model's ability generalization. The experimental results demonstrate that MWCRL outperforms the backbone model by 5.83% and the widely-adopted vanilla GRPO method by 5.00%. Notably, on basic action dependency problems and next likely action prediction problems, MWCRL achieve significant improvements compared to the vanilla GRPO method, with increases of 10.00% and 6.66% respectively. On belief and action-object state multi-hop problem types, it achieve substantial improvements compared to the backbone model, with increases of 13.34% and 6.66% respectively. We provide specific data examples of these four question types in the *supplementary material*.

The Impact of Hyperparameters α and μ on Model Performance

Regarding the hyperparameter α, as shown in Table 1, our experimental results indicate that it should not be set too large. When increasing α from 0.3 to 0.6 while keeping μ constant, we observe a 2.67% $\sim$ 4.67% decrease in model performance. We argue that overly large α values may weaken and interfere with the driving force of the accuracy reward signal. The multi-world contrastive reward signal should serve as an auxiliary signal to assist the model's learning process.

As for the hyperparameter μ, although intuitively setting it to 1.0 would seem more reasonable, our experimental results show that moderate relaxation can lead to better model performance. When setting μ from 1.0 to 0.8 or 0.6, we observe a 0.67% $\sim$ 1.33% improvement in model performance. We believe this represents a trade-off between noisy signals and obtaining denser learning signals. Additionally, due to random sampling, the performance of shuffle groups also exhibits uncertainty. The choice of μ reflects a balanced decision under the interaction of multiple factors.

SFT Has Very Little Effect, and Mixed Multi-domain SFT Is No Exception

Direct SFT on in-domain data and Mixed Multi-domain SFT achieve only modest performance gains of +1.33% and +2.00% respectively compared to the backbone model, demonstrating very limited effectiveness. Marginal improvements suggest that simply exposing models to task-specific examples through SFT is insufficient to develop robust mental state reasoning capabilities. From a capability transfer perspective, the performance remained generally poor even after SFT with mixed data from multiple domains including Math, Knowledge, Spatial, and Chart, highlighting the failure of the SFT-based post-training paradigm to transfer capabilities from knowledge-intensive tasks to mental state reasoning.

Model	Action Deps	Belief	Action	Multi-hop Act-Obj State	AVG
Qwen2.5-VL-7B	43.33%	53.33%	63.33%	46.67%	51.67%
Vanilla GRPO	36.67%	63.33%	56.67%	53.33%	52.50%
MWCRL GRPO α=0.3 μ=0.8	46.67%	66.67%	63.33%	53.33%	57.50%
Δ_{Backbone}	+3.34%	+13.34%	+0.00%	+6.66%	+5.83%
$\Delta_{\text{Vanilla GRPO}}$	+10.00%	+3.34%	+6.66%	+ 0.00%	+5.00%

Table 2: Performance Evaluation of Multiple Methods in OOD Scenarios. Δ_{Backbone} represents the performance improvement brought by our proposed MWCRL method when applied to the backbone model, and $\Delta_{\text{Vanilla GRPO}}$ represents the performance advantage of our proposed MWCRL method compared to the widely adopted GRPO (only apply format reward and outcome reward). Deps is short for Dependency.

Goal Inference Is Easy; Belief and Action Inferences Are Hard

Examining the performance of the backbone model reveals that it achieves a relatively high accuracy of 76.00% on goal inference, while belief and action inference demonstrate performance of only 56.00% and 32.00%, respectively. This substantial performance gap, with belief inference lagging nearly 20 percentage points behind goal inference, reveals fundamental differences in how MLLMs process various aspects of mental state. Interestingly, this pattern emerges consistently across all foundation models we evaluate. Even after undergoing SFT or RL-based post-training, this trend persists. We believe that addressing this performance disparity represents a promising direction for further enhancing MLLMs' ToM capabilities.

Case Study

Through comparative analysis of inference trajectories from the backbone model, vanilla GRPO model, and our MWCRL method, we observe that under the influence of RL post-training paradigm, models exhibit reasoning behaviors commonly referred to as "aha moments." In the inference trajectories of models trained with our MWCRL method, we discover model reasoning behaviors such as **"Oh! I know", "Oh, I see", "Let me think about this differently", "Let's verify", "No wait, it should be...", and "Let me think about it again."** The example in the upper half of Figure 3 demonstrates how the model first arrives at an initial answer through perception and reasoning, then proceeds to verify that answer in a subsequent step.

Furthermore, we find that through the introduction of multi-world contrastive reward signals, the model can **actively capture temporal and causal patterns in human action sequences**, rather than relying on shortcuts derived from isolated actions as observed in vanilla GRPO models, thereby enabling better inference of human mental states, as illustrated in the example shown in the lower half of Figure 3. We present more interesting examples in the *supple-*

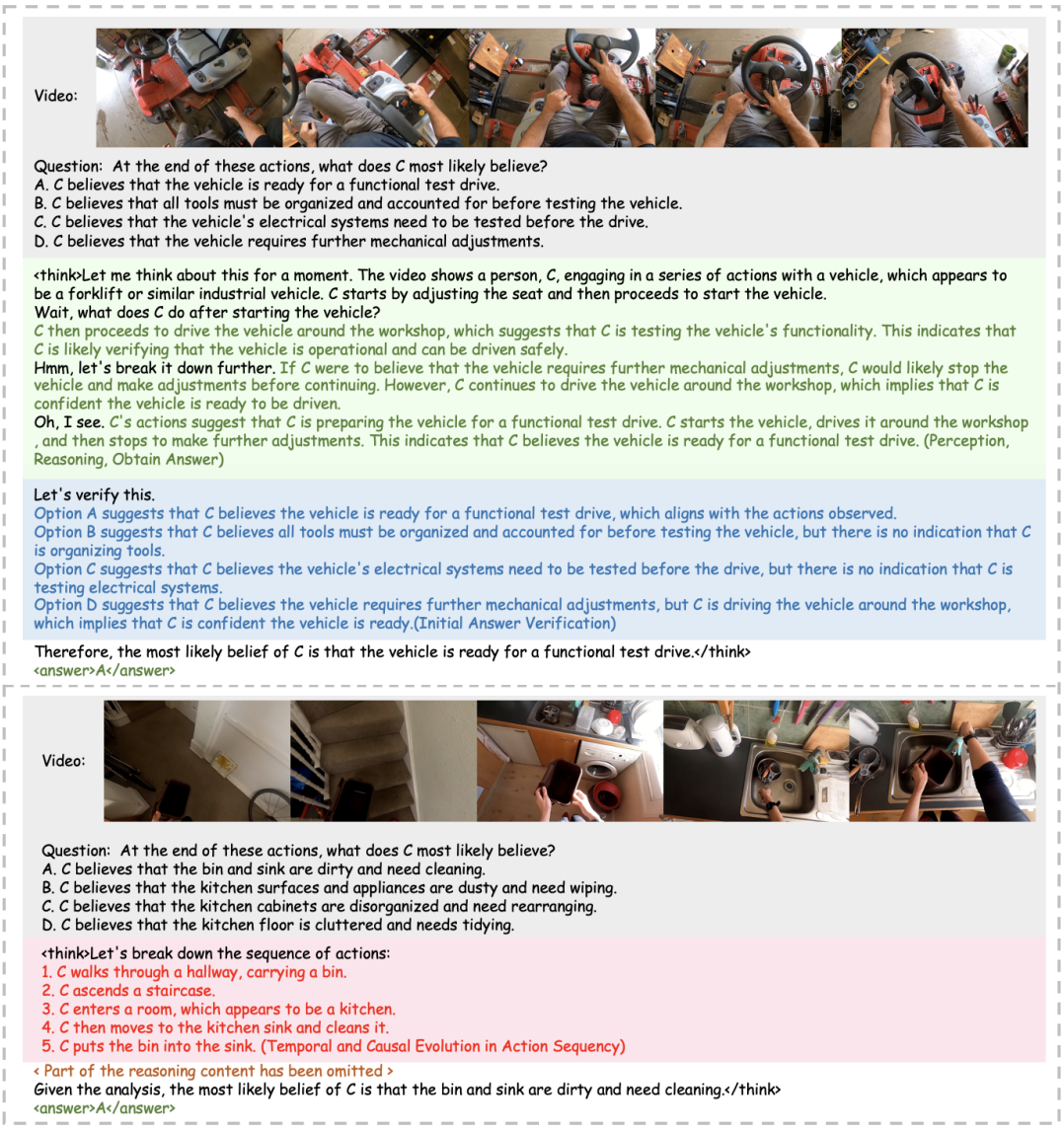

Figure 3: Two examples of MWCRL method inference outputs are presented. The above example demonstrates the verify behavior that emerges in its reasoning trajectory, commonly referred to as "aha moments." The below example illustrates its actively capture of temporal and causal patterns in human action sequences.

mentary material.

Conclusion

In this paper, motivated by the goal of developing more capable Embodied AI agents and recognizing that most existing work on enhancing ToM capabilities centers around the text domain, we investigate methods to improve the ToM abilities of MLLMs. We propose a multi-world contrastive reinforcement learning approach that explicitly encourages models to exploit the temporal and causal evolution patterns in user action sequences for inferring users' mental states—such as goals, beliefs, and potential next actions. Quantitative and qualitative analyses validate the effectiveness of our approach from multiple perspectives. By grounding inference in embodied action sequences, our method is naturally suited for downstream applications in embodied wearable agents.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62376245), the Key Research and Development Program of Zhejiang Province, China (No. 2024C03255), National Key Research and Development Project of China (No. 2018AAA0101900), and MOE Engineering Research Center of Digital Library.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

Bordes, F.; Garrido, Q.; Kao, J. T.; Williams, A.; Rabbat, M.; and Dupoux, E. 2025. IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments. *arXiv preprint arXiv:2506.09849*.

Chen, H.; Tu, H.; Wang, F.; Liu, H.; Tang, X.; Du, X.; Zhou, Y.; and Xie, C. 2025. SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models. *arXiv preprint arXiv:2504.11468*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ding, J.; Zhang, Y.; Shang, Y.; Zhang, Y.; Zong, Z.; Feng, J.; Yuan, Y.; Su, H.; Li, N.; Sukiennik, N.; et al. 2024. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*.

Dong, X. L. 2024. Next-generation Intelligent Assistants for Wearable Devices. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4735–4735.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wang, B.; and Yue, X. 2025a. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Feng, Z.; Cao, S.; Ren, J.; Su, J.; Chen, R.; Zhang, Y.; Xu, Z.; Hu, Y.; Wu, J.; and Liu, Z. 2025b. MT-R1-Zero: Advancing LLM-based Machine Translation via R1-Zero-like Reinforcement Learning. *arXiv preprint arXiv:2504.10160*.

Fung, P.; Bachrach, Y.; Celikyilmaz, A.; Chaudhuri, K.; Chen, D.; Chung, W.; Dupoux, E.; Jégou, H.; Lazaric, A.; Majumdar, A.; et al. 2025. Embodied AI Agents: Modeling the World. *arXiv preprint arXiv:2506.22355*.

Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; and Goodman, N. D. 2023. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 13518–13529.

Ge, Z.; Huang, H.; Zhou, M.; Li, J.; Wang, G.; Tang, S.; and Zhuang, Y. 2024. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7346–7355.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, M.; Zhao, F.; Lu, C.; Liu, Z.; Wang, Y.; and Qian, H. 2025. GenCLS++: Pushing the Boundaries of Generative Classification in LLMs Through Comprehensive SFT and RL Studies Across Diverse Datasets. *arXiv preprint arXiv:2504.19898*.

Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Hou, G.; Gao, X.; Wu, Y.; Huang, X.; Zhang, W.; Zheng, Z.; Shen, Y.; Du, J.; Huang, F.; Li, Y.; et al. 2025. TimeHC-RL: Temporal-aware Hierarchical Cognitive Reinforcement Learning for Enhancing LLMs' Social Intelligence. *arXiv preprint arXiv:2505.24500*.

Hou, G.; Zhang, W.; Shen, Y.; Wu, L.; and Lu, W. 2024. TimeToM: Temporal Space is the Key to Unlocking the Door of Large Language Models' Theory-of-Mind. In *Findings of the Association for Computational Linguistics: ACL 2024*, 11532–11547.

Hu, Z.; and Shu, T. 2023. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*.

Huang, X.; La Malfa, E.; Marro, S.; Asperti, A.; Cohn, A.; and Wooldridge, M. 2024. A Notion of Complexity for Theory of Mind via Discrete World Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2964–2983.

Jia, B.; Lei, T.; Zhu, S.-C.; and Huang, S. 2022. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Jin, C.; Wu, Y.; Cao, J.; Xiang, J.; Kuo, Y.-L.; Hu, Z.; Ullman, T.; Torralba, A.; Tenenbaum, J.; and Shu, T. 2024. MMToM-QA: Multimodal Theory of Mind Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16077–16102.

Jung, C.; Kim, D.; Jin, J.; Kim, J.; Seonwoo, Y.; Choi, Y.; Oh, A.; and Kim, H. 2024. Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19794–19809.

Kounios, J.; and Beeman, M. 2009. The Aha! moment: The cognitive neuroscience of insight. *Current directions in psychological science*, 18(4): 210–216.

Li, L.; Chen, W.; Li, J.; and Chen, L. 2025a. Relation-r1: Cognitive chain-of-thought guided reinforcement learning for unified relational comprehension. *arXiv preprint arXiv:2504.14642*.

Li, M.; Zhong, J.; Zhao, S.; Lai, Y.; and Zhang, K. 2025b. Think or Not Think: A Study of Explicit Thinking in Rule-Based Visual Reinforcement Fine-Tuning. *arXiv preprint arXiv:2503.16188*.

Li, Y.; Veerabadran, V.; Iuzzolino, M. L.; Roads, B. D.; Celikyilmaz, A.; and Ridgeway, K. 2025c. Egotom: Benchmarking theory of mind reasoning from egocentric videos. *arXiv preprint arXiv:2503.22152*.

Luo, R.; Wang, L.; He, W.; and Xia, X. 2025. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*.

Ma, P.; Zhuang, X.; Xu, C.; Jiang, X.; Chen, R.; and Guo, J. 2025. Sql-r1: Training natural language to sql reasoning model by reinforcement learning. *arXiv preprint arXiv:2504.08600*.

Perner, J.; Leekam, S. R.; and Wimmer, H. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2): 125–137.

Perner, J.; and Wimmer, H. 1985. "John thinks that Mary thinks that. . . " attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3): 437–471.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526.

Sarangi, S.; Elgarf, M.; and Salam, H. 2025. DecomposeToM: Enhancing Theory of Mind Reasoning in Large Language Models through Simulation and Task Decomposition. In *Proceedings of the 31st International Conference on Computational Linguistics*, 10228–10241.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sclar, M.; Kumar, S.; West, P.; Suhr, A.; Choi, Y.; and Tsvetkov, Y. 2023. Minding Language Models'(Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13960–13980.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 1279–1297.

Shi, H.; Ye, S.; Fang, X.; Jin, C.; Isik, L.; Kuo, Y.-L.; and Shu, T. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1510–1519.

Shinoda, K.; Hojo, N.; Nishida, K.; Yamazaki, Y.; Suzuki, K.; Sugiyama, H.; and Saito, K. 2025. Let's Put Ourselves in Sally's Shoes: Shoes-of-Others Prefixing Improves Theory of Mind in Large Language Models. *arXiv preprint arXiv:2506.05970*.

Song, H.; Jiang, J.; Min, Y.; Chen, J.; Chen, Z.; Zhao, W. X.; Fang, L.; and Wen, J.-R. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Sun, Y.; Zhou, G.; Wang, H.; Li, D.; Dziri, N.; and Song, D. 2025. Climbing the Ladder of Reasoning: What LLMs Can-and Still Can't-Solve after SFT? *arXiv preprint arXiv:2504.11741*.

Waisberg, E.; Ong, J.; Masalkhi, M.; Zaman, N.; Sarker, P.; Lee, A. G.; and Tavakkoli, A. 2024. Meta smart glasses—large language models and the future for assistive glasses for individuals with vision impairments. *Eye*, 38(6): 1036–1038.

Wilf, A.; Lee, S.; Liang, P. P.; and Morency, L.-P. 2024. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8292–8308.

Xia, X.; and Luo, R. 2025. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*.

Xie, T.; Gao, Z.; Ren, Q.; Luo, H.; Hong, Y.; Dai, B.; Zhou, J.; Qiu, K.; Wu, Z.; and Luo, C. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Xu, H.; Han, L.; Yang, Q.; Li, M.; and Srivastava, M. 2024. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, 1–7.

Yang, S.; Wu, J.; Chen, X.; Xiao, Y.; Yang, X.; Wong, D. F.; and Wang, D. 2025. Understanding aha moments: from external observations to internal mechanisms. *arXiv preprint arXiv:2504.02956*.

Zhang, Z.; Jin, C.; Jia, M. Y.; and Shu, T. 2025. AutoToM: Automated Bayesian Inverse Planning and Model Discovery for Open-ended Theory of Mind. *arXiv preprint arXiv:2502.15676*.

Zhou, H.; Li, X.; Wang, R.; Cheng, M.; Zhou, T.; and Hsieh, C.-J. 2025a. R1-Zero's" Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. *arXiv preprint arXiv:2503.05132*.

Zhou, Y.; Jiang, S.; Tian, Y.; Weston, J.; Levine, S.; Sukhbaatar, S.; and Li, X. 2025b. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.