SSEToM: Spatially Guided Reasoning Enhances Theory of Mind in Large Language Models

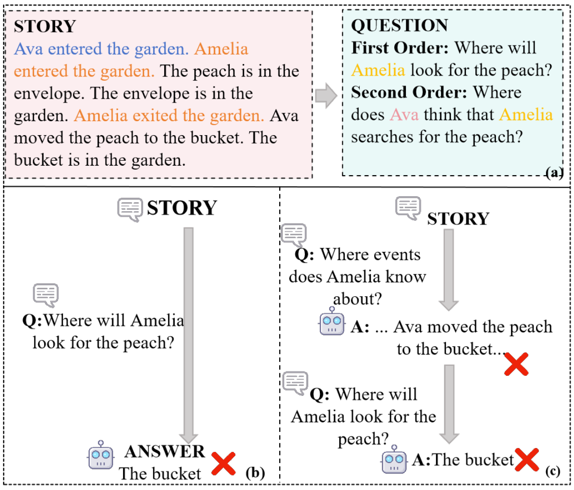
Anonymous ACL submission

Abstract

Theory of Mind (ToM) refers to an individual's ability to understand and infer the mental states of others. While this capability develops naturally in humans, equipping Large Language Models (LLMs) with similar abilities remains challenging. Some chain-of-thought (CoT) methods, such as SimToM, have improved LLM performance in ToM reasoning. However, existing methods often overlook the spatial dimension perception that humans utilize when solving ToM problems. To address this limitation, we propose SSEToM, a method inspired by the Event Segmentation Theory (EST) in psychology, which posits that individuals in different spatial locations perceive information about events within their respective environments. SSEToM segments ToM stories into discrete events based on spatial dimensions, enhancing the LLM's ability to perceive and reason about the mental states of the discussed characters. Experiments conducted on three datasets demonstrate that SSEToM significantly enhances LLMs' reasoning capabilities in ToM tasks, achieving state-of-the-art performance.

1 Introduction

Psychology and cognitive science have extensively studied Theory of Mind (ToM) (Premack and Woodruff, 1978) across various scenarios. This cognitive ability involves understanding the mental states of others, such as beliefs, desires, and thoughts. People utilize ToM in diverse social environments, and ToM makes communication and connection in social interactions more efficient (Zhang et al., 2024). ToM tasks translate abstract ToM abilities into quantifiable data, as shown in Figure.1(a) for ToM task forms. With Large Language Models (LLMs) playing an increasingly important role in our lives, developing ToM-capable LLMs can enable AI agents to better understand users' intentions, emotions, and potential misunderstandings (Wang et al., 2024). The AI community has also shown that ToM reasoning is crucial in dialogue (Kim et al., 2023), games (Zhou et al., 2022), and even multimodal settings (Jin et al., 2024).



STORY
Ava entered the garden. Amelia entered the garden. The peach is in the envelope. The envelope is in the garden. Amelia exited the garden. Ava moved the peach to the bucket. The bucket is in the garden.

QUESTION
First Order: Where will Amelia look for the peach?
Second Order: Where does Ava think that Amelia searches for the peach?

(a)
STORY
Q:Where will Amelia look for the peach?
ANSWER
The bucket
(b)
STORY
Q: Where events does Amelia know about?
A: ... Ava moved the peach to the bucket...
Q: Where will Amelia look for the peach?
A:The bucket
(c)

Figure 1: (a): Typical example of story and question for ToM task. (b): Answering ToM questions through a one-step reasoning process. (c): Answering ToM questions through a two-step reasoning process that involves extracting characters' perceptions.

Humans are innately equipped with the ToM ability, whereas LLMs still face challenges in this regard. Kosinski (2024) benchmarked LLM's ToM capabilities through stories from a mini-test set, suggesting that LLMs may develop ToM abilities, while Ullman (2023) made minor alterations to Kosinski (2024)'s mini-test set and found that LLMs fail when the ToM task undergoes slight changes, indicating that LLMs' performance in ToM reasoning is not stable. CoT (Wei et al., 2022) have difficulty significantly improving LLMs' performance in ToM tasks. To address this, several new methods (Wilf et al., 2023; Jung et al., 2024; Hou et al., 2024) for ToM reasoning have been proposed. These methods propose that a key step to better answer ToM questions is to understand what the discussed character knows before answering a question about their mental state. As illustrated in Figure.1(c)), in order to answer *"Where will Amelia look for the peach?"*, first understand Amelia's knowledge by asking *"What events does Amelia know about?"*. However, in this step, these methods directly guide LLMs through prompts to answer the character's perception, neglecting the use of spatial dimensions, which often results in the output lacking critical information for answering ToM questions. Specifically, when an event occurs, if the discussed character is in the same spatial location as the event, it is usually natural for that character to perceive the occurrence of the event.

Inspired by the Event Segmentation Theory (EST) (Zacks and Swallow, 2007) in psychology, which posits that individuals in different spatial locations can perceive information about events in their respective locations and detect changes in events at those locations, we propose a new method—SSEToM. This enables LLMs to identify the perception of the characters discussed in ToM questions based on spatial locations within the ToM story. Specifically, SSEToM divides a continuous ToM story into discrete events, guiding LLMs to focus on the relationship between where events occur and the spatial locations of the characters, thereby helping LLMs better understand the characters' perceptions and infer their actions. The main contributions of this work are as follows:

(1) We propose a novel ToM reasoning method—SSEToM. This is the first approach to integrate ToM reasoning with Event Segmentation Theory in psychological, while emphasizing the utilization of spatial dimensions.

(2) We propose a method for identifying character perception through spatial location, effectively improving the accuracy of LLMs in extracting character perception, thereby enhancing the performance of LLMs in ToM reasoning.

(3) We conducted tests and analyses on three datasets. The experimental results indicate that SSEToM significantly enhances the LLMs' performance in ToM tasks. Furthermore, we demonstrate a positive correlation between the spatial location information and ToM reasoning.

2 Related Work

Theory of Mind has been extensively studied in psychology and cognitive science in a variety of scenarios. The AI community has also shown that ToM reasoning is useful in dialogue (Kim et al., 2023), games (Zhou et al., 2022), and even multimodal settings (Jin et al., 2024) are important. Kosinski (2024) benchmarked LLM's ToM abilities through the story of the mini-test set, suggesting that LLMs may develop ToM abilities, while Ullman (2023) made minor changes to Kosinski (2024)'s mini-test set with minor changes and found that LLMs failed when the ToM task changed slightly, suggesting that LLMs' performance in ToM reasoning is not stable.

2.1 Evaluating LLMs' ability of ToM

Existing generative datasets allow ToM studies to be conducted at scale. Le et al. (2019) generated stories in question-answer format by constructing automated templates for the adaptation of classic psychology tests such as Sally-Anne. Chen et al. (2024) considered 8 ToM tasks and 31 ToM abilities, adapted for everyday social scenarios, and manually created test samples. The above reading comprehension-based datasets basically have more obvious data patterns. Kim et al. (2023) proposes interactive conversation-based datasets, where each conversation revolves around a specific topic and characters join or leave the conversation to create information asymmetry, capable of measuring the LLMs' ability to track multi-party beliefs in conversations, especially when certain information is inaccessible to certain participants. In addition, Xu et al. (2024) proposes character interaction-based datasets for mental state inference by simulating real social scenarios.

2.2 Enhancing LLMs' ability of ToM

Despite the outstanding performance of LLMs across various task scenarios, they still face significant challenges in reasoning tasks, particularly in ToM tasks. Currently, various prompting methods have been developed to enhance the reasoning capabilities of LLMs (Wei et al., 2022; Wang et al., 2022; Chia et al., 2023). However, traditional prompting methods may not significantly improve LLMs' performance in ToM tasks (Moghaddam and Honey, 2023). As a result, a number of new strategies have been proposed, including approaches based on external tools or models, approaches based on cueing strategies, and approaches based on time-space construction. Wilf et al. (2023) proposes to perform a perspective shift that allows the LLMs to filter out the contextual information known to the characters in the question.

Jung et al. (2024) guides LLMs to infer the perception of the character from the input context and then isolate the context perceived by the target character through a Perspective-Taking step. Tang and Belle (2024) fine-tunes LLMs so that it can convert natural language problems into symbolic formulas before the SMCDEL model checker executes the formulas to arrive at the final result. Hou et al. (2024) constructs the temporal space and constructs a chain of temporal belief states for each character, combined with a time-aware belief solver, based on which the ToM question is answered.

3 Method

In this section, we introduce SSEToM, an approach designed to improve the performance of LLMs in ToM tasks. We uses spatial dimension analysis to enhance the LLMs' understanding of the character's perception by simulating the human perceptual process, thus answering ToM questions more accurately. The core idea of SSEToM is to break down the ToM story into discrete events and identify and organize these events through spatial location, which in turn keeps track of the character's perception in different spatial locations. This approach eliminates the need for additional training of pre-trained LLMs and can be seamlessly integrated with minimal prompt tuning of the different models. The overall framework of SSEToM is shown in Figure 2.

3.1 Event Segmentation

According to the framework of Event Segmentation Theory, we first identify and define each independent event in a story, decomposing a continuous story into a series of discrete events. In reading comprehension scenarios, each independent event usually consists of a dynamic action or a static description, specifically, each sentence in the story corresponds to an independent event. In interactive dialogue scenarios, a single utterance by a character corresponds to an independent event.

Given input story s, prompt p_{ES} and model M, SSEToM adds specific event numbers to input story s to form s_E:

$$s_E = M(p_{ES}||s). \quad (1)$$

$||$ denotes a connection and p_{ES} is shown in Appendix. For example, the model explicitly adds event numbers before each sentence of a given input story s, as shown in Figure.2(a).

3.2 Event Unit Formation

In this module, we annotate each event with a specific spatial location (i.e., we identify where each event occurred).

Given the story s_E with the event numbers added, prompt p_{EUF} and the model M, SSEToM identifies the spatial location l_i where each event s_E occurs:

$$l_i = M(p_{EUF}||s_E). \quad (2)$$

i represents the spatial location number (there may be more than one spatial location in a story) and p_{EUF} is shown in Appendix. Based on the spatial location l_i, we combine the events involved in the same l_i to form event units (i.e., each event unit contains a series of events occurring in the same location). For example, as shown in Figure.2(b), SSEToM divides the story into two spatial locations, *'garden'* and *'playroom'*, and all the events occurring in *'garden'* form a single event unit, including *E1, E3, E4, E5, E6, E7, E8*, and all the events in *'playroom'* form one event unit, including *E2*.

3.3 Character Perception Tracking

With the first two modules, each event in the story is organised into spatial location-based event units. By tracking the characters' appearances and departures in each spatial location, we are able to determine their perception (i.e., the events they are likely to perceive). Specifically, when a character appears in the corresponding spatial location l_i during an event, we assume that the character is able to perceive the event. In particular, in interactive dialogue scenarios, a character is considered to perceive the current event when the character is involved in the dialogue.

Given the story s_E with added event number, character c_j, spatial location l_i, prompt p_{CPT} and model M, SSEToM identifies whether character c_j is present at a certain spatial location l_i in each input event s_E, obtaining the corresponding triplet $(c_j, present/absent, l_i)$:

$$(c_j, present/absent, l_i) = M(p_{CPT}||s_E). \quad (3)$$

j represents the character number (there are multiple characters in a study) and p_{CPT} is displayed in Appendix. After getting triplet $(c_j, present/absent, l_i)$ for each event, we know whether the character c_j is at the position l_i where

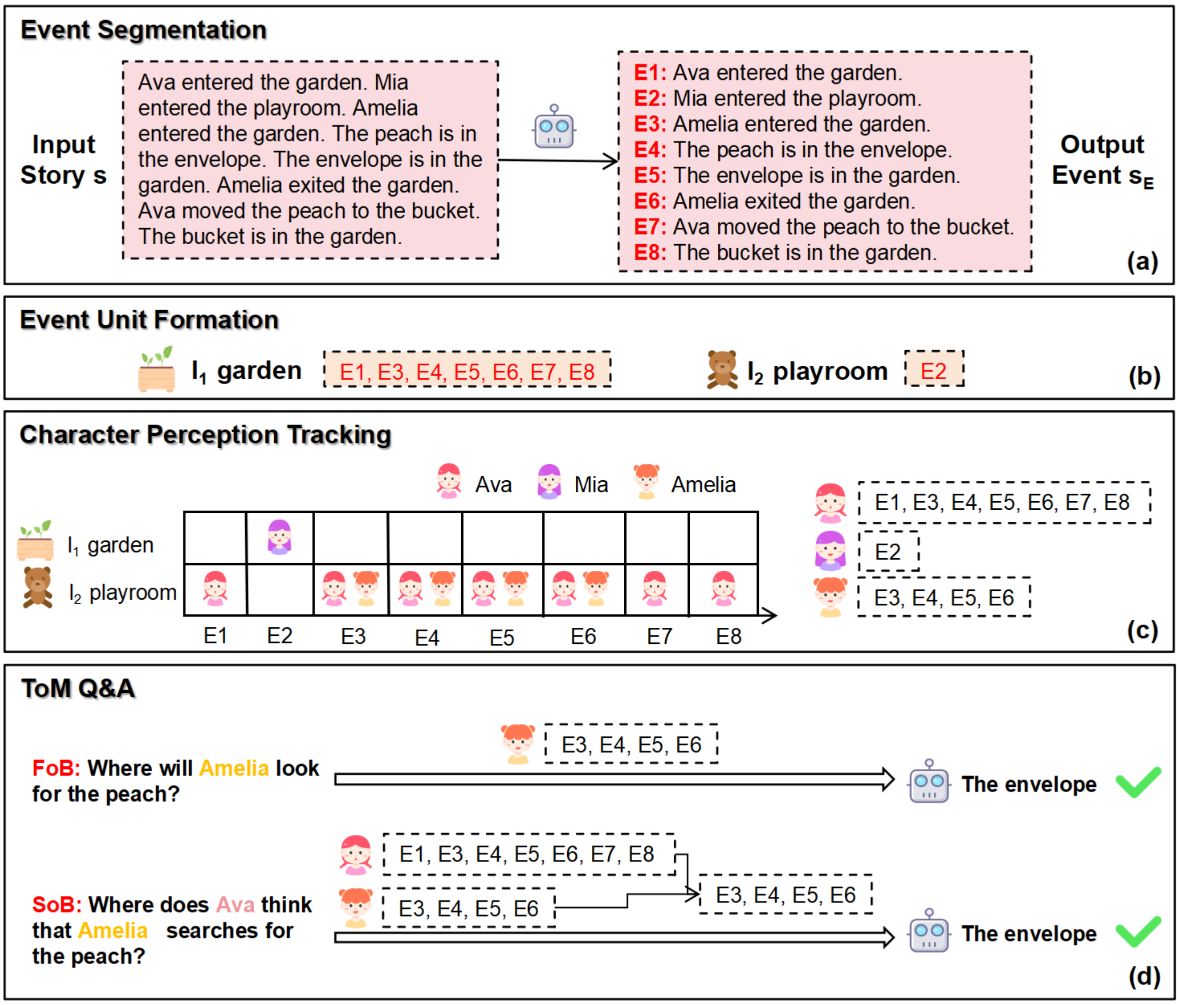

Event Segmentation
Input Story s
Ava entered the garden. Mia entered the playroom. Amelia entered the garden. The peach is in the envelope. The envelope is in the garden. Amelia exited the garden. Ava moved the peach to the bucket. The bucket is in the garden.
E1: Ava entered the garden.
E2: Mia entered the playroom.
E3: Amelia entered the garden.
E4: The peach is in the envelope.
E5: The envelope is in the garden.
E6: Amelia exited the garden.
E7: Ava moved the peach to the bucket.
E8: The bucket is in the garden.
Output Event s_E
(a)
Event Unit Formation
l1 garden
E1, E3, E4, E5, E6, E7, E8
l2 playroom
E2
(b)
Character Perception Tracking
Ava
Mia
Amelia
l1 garden
l2 playroom
E1
E2
E3
E4
E5
E6
E7
E8
E1, E3, E4, E5, E6, E7, E8
E2
E3, E4, E5, E6
(c)
ToM Q&A
FoB: Where will Amelia look for the peach?
E3, E4, E5, E6
The envelope
SoB: Where does Ava think that Amelia searches for the peach?
E1, E3, E4, E5, E6, E7, E8
E3, E4, E5, E6
E3, E4, E5, E6
The envelope
(d)

Figure 2: The overview of SSEToM, providing a detailed description of how event segmentation, event unit formation, and character perception tracking enable LLMs to better simulate the human process of understanding characters' mental states within stories.

the event occurs when each event occurs, and through string matching, we extract the events that match the character's perception. For example, as shown in Figure.2(c), Ava appears in the garden all the time starting from event *E1*, and then, except for event *E2* which occurs in the playroom, all the other events occur in garden, so *E1,E3,E4,E5,E6,E7,E8* are categorized as Ava's perception. Mia appears in the Playroom from event *E2*, so Mia only knows event *E2*. Amelia appears in the garden during events *E3,E4,E5,E7*, so *E3,E4,E5,E7* are categorized as Emily's perception.

3.4 ToM Q&A

In this module, we answer ToM questions. For first-order questions, we directly use the perception of a single character as input. For higher-order questions, we use the approach proposed in Hou et al. (2024), choosing the perceptions of multiple characters involved in the intersection as input. Given the perception p_C of a character, the ToM question q_{FO}/q_{SO}, the prompt p_{QA} and the model M, the answer $Answer_{FO}$/$Answer_{SO}$ to the ToM question is obtained:

$$Answer_{FO} = M(p_{FO}||q_{FO}, p_C), \qquad (4)$$

$$Answer_{SO} = M(p_{SO}||q_{SO}, p_{c1} \cap p_{c2}). \qquad (5)$$

$||$ denotes connections, p_{FO} and p_{SO} are shown in Appendix. For example, as shown in Figure. 2(d). For the First-Order ToM question *"Where will Amelia look for the peach?"*, there is only one character involved (i.e., Amelia), so we use Amelia's perception (*E3, E4, E5, E6*) to answer the question. However, for the Second-Order ToM question *"Where does Ava think that Amelia* *searches for the peach?"*, which involves two characters (i.e., Ava and Amelia), so we use the intersection (*E3,E4,E5,E6*) of Ava's perception (*E1,E3,E4,E5,E6,E7,E8*) and Amelia's perception (*E3,E4,E5,E6*), to answer the question.

4 Experiment

4.1 Settings

Models We use two LLMs for testing. An open source model, Mistral-7B-Instruct (Jiang et al., 2023), and a closed source model, GPT-4o-mini (Hurst et al., 2024). To ensure the accuracy of the answers to the questions, we set the temperature of to 0.

Datasets We conducted experiments on three datasets—**ToMI** (Le et al., 2019), **ToMBench** (Chen et al., 2024) and **FANToM** (Kim et al., 2023). The **ToMI** and **ToMBench** datasets contain reading comprehension scenarios. In comparison, the stories in **ToMBench** are more structured, while the stories in **ToMI** contain some distracting information. The **FANToM** dataset contain interactive dialogue scenarios, which is longer and more complex than the reading comprehension scenario. In order to have a clearer understanding of LLMs' perceptions level of different belief categories, we divided each dataset into four question types: First-Order True Belief, First-Order False Belief, Second-Order True Belief and Second-Order False Belief. "True Belief" means that the character in question knows all the information. "False Belief" means that the character in question knows only some of the information, and the LLMs need to analyze which parts of the information are known.

Baselines We test the models separately using six methods, including **Vanilla**, **CoT** (Wei et al., 2022), **SiMToM** (Wilf et al., 2023), **TimeToM** (Hou et al., 2024), **PercepToM** (Jung et al., 2024), and **SSEToM**. **Vanilla** uses the content of the original dataset directly as input. **CoT** adds "think step by step" to the end of each question. We directly used the open source code of **SiMToM**. And we used the instructions given in the **TimeToM** and **PercepToM** papers to reimplement these two methods. Due to the different classification criteria of question types of the dataset, we re-tested all the methods, using each method to test the same question type five times and taking the average of the five test results to ensure accuracy.

4.2 Main Results

We find that SSEToM leads to significant performance improvements compared to the five methods—Vanilla, CoT, SimToM, TimeToM, and PercepToM. The results in Table 1 reflect these improvements in the ToMI, ToMBench, and FANToM benchmarks.

ToMI Results In the ToMI benchmark, the SSEToM method showed significant performance improvements in two models. Compared to Vanilla, GPT-4o-mini improved accuracy by 5.2%, 77.8%, 24.2%, and 54%, respectively, and Mistral-7B-Instruct improved accuracy by 7%, 14%, 53.6%, and 42.2%, respectively. It is worth noting that in First-Order False Beliefs, GPT-4o-mini saturated human performance using the SSEToM method. Four methods—CoT, SimToM, TimeToM, and PercepToM—performed better than Vanilla in the False Belief question type, and played the opposite role in the True Belief question type, which may be due to the True Belief question type can be answered by understanding the true state of things without understanding the mental state of the character.

ToMBench Results In the ToMBench benchmark, the SSEToM method performs almost best in the False Belief question type. In First-Order and Second-Order question types, the accuracy of GPT-4o-mini improves by 88% and 74% and the accuracy of Mistral-7B-Instruct improves by 48.8% and 41% compared to Vanilla, respectively. The SSEToM method is even more useful for Second-Order question type in True Belief question type, with the accuracy of GPT-4o-mini and Mistral-7B-Instruct have improved accuracy by 54.4% and 45.8%, respectively. It is worth noting that two models performed particularly poorly in the Vanilla method when dealing with the False Belief question type, suggesting that it is almost impossible to infer a character's belief state without any prompts.

FANToM Results In the FANToM benchmark test, accuracy was generally low on False Belief questions. Mistral-7B-Instruct failed to reach the random level (50%) in almost half of the results, while using the SSEToM method exceeded the random level on all four question types, demonstrating the robustness of SSEToM. TimeToM generally performed best on False Belief questions. PercepToM generally performs best on True Belief questions. While SSEToM performs best on average

Method	ToMI				ToMBench				FANToM			
	Fo-true	***Fo-false***	***So-true***	***So-false***	***Fo-true***	***Fo-false***	***So-true***	***So-false***	***Fo-true***	***Fo-false***	***So-true***	***So-false***
GPT-4o-mini												
Vanilla	$90.0_{0.01}$	$22.2_{0.01}$	$72.6_{0.02}$	$41.0_{0.02}$	$\mathbf{98.0_{0.00}}$	$9.2_{0.02}$	$35.2_{0.03}$	$19.6_{0.03}$	$86.8_{0.01}$	$30.1_{0.01}$	$\mathbf{94.2_{0.01}}$	$43.9_{0.01}$
CoT	$82.6_{0.01}$	$48.4_{0.03}$	$70.4_{0.03}$	$53.2_{0.02}$	$97.6_{0.01}$	$48.4_{0.02}$	$76.8_{0.07}$	$86.0_{0.09}$	$77.6_{0.2}$	$52.2_{0.02}$	$89.3_{0.02}$	$62.3_{0.03}$
SimToM	$76.4_{0.01}$	$84.8_{0.05}$	$64.2_{0.01}$	$61.6_{0.01}$	$96.0_{0.00}$	$\mathbf{98.0_{0.00}}$	$81.6_{0.04}$	$89.6_{0.01}$	$84.8_{0.03}$	$63.9_{0.02}$	$83.8_{0.03}$	$53.3_{0.04}$
TimeToM	$78.6_{0.44}$	$72.2_{0.06}$	$46.6_{0.03}$	$92.2_{0.01}$	$96.0_{0.00}$	$96.0_{0.00}$	$\mathbf{91.6_{0.03}}$	$92.8_{0.01}$	$84.3_{0.02}$	$77.0_{0.02}$	$62.0_{0.03}$	$\mathbf{80.2_{0.03}}$
PercepToM	$58.8_{0.00}$	$94.4_{0.01}$	$13.0_{0.01}$	$91.8_{0.01}$	$91.6_{0.01}$	$97.2_{0.01}$	$88.0_{0.00}$	$90.0_{0.00}$	$92.1_{0.02}$	$\mathbf{77.1_{0.09}}$	$91.4_{0.04}$	$54.6_{0.03}$
SSEToM	$\mathbf{95.2_{0.01}}$	$\mathbf{100.0_{0.00}}$	$\mathbf{96.8_{0.01}}$	$\mathbf{95.0_{0.01}}$	$83.2_{0.08}$	$97.2_{0.01}$	$89.6_{0.01}$	$\mathbf{93.6_{0.01}}$	$\mathbf{92.7_{0.02}}$	$74.2_{0.03}$	$92.7_{0.01}$	$75.9_{0.01}$
Mistral-7B-Instruct												
Vanilla	$68.0_{0.02}$	$62.8_{0.01}$	$41.0_{0.02}$	$42.8_{0.01}$	$92.4_{0.08}$	$24.4_{0.06}$	$29.6_{0.04}$	$31.2_{0.03}$	$84.4_{0.02}$	$17.9_{0.02}$	$89.9_{0.02}$	$36.3_{0.05}$
CoT	$74.8_{0.04}$	$51.2_{0.05}$	$45.6_{0.03}$	$49.4_{0.01}$	$\mathbf{94.4_{0.02}}$	$22.4_{0.01}$	$32.8_{0.03}$	$26.0_{0.05}$	$81.8_{0.03}$	$19.9_{0.02}$	$84.4_{0.02}$	$38.4_{0.05}$
SimToM	$47.8_{0.03}$	$60.0_{0.03}$	$30.6_{0.02}$	$27.2_{0.01}$	$79.6_{0.04}$	$73.2_{0.06}$	$49.2_{0.08}$	$44.0_{0.05}$	$71.1_{0.05}$	$42.2_{0.04}$	$61.7_{0.07}$	$53.8_{0.04}$
TimeToM	$60.6_{0.02}$	$54.4_{0.03}$	$48.8_{0.01}$	$52.4_{0.01}$	$42.8_{0.08}$	$48.8_{0.05}$	$\mathbf{77.6_{0.03}}$	$66.4_{0.02}$	$30.2_{0.08}$	$\mathbf{59.2_{0.03}}$	$37.9_{0.02}$	$\mathbf{80.0_{0.05}}$
PercepToM	$47.2_{0.01}$	$49.4_{0.01}$	$42.8_{0.02}$	$37.2_{0.03}$	$45.6_{0.04}$	$21.6_{0.03}$	$50.4_{0.04}$	$52.0_{0.02}$	$\mathbf{85.4_{0.02}}$	$20.6_{0.04}$	$\mathbf{90.4_{0.02}}$	$40.2_{0.03}$
SSEToM	$\mathbf{75.0_{0.01}}$	$\mathbf{76.8_{0.02}}$	$\mathbf{94.6_{0.02}}$	$\mathbf{85.0_{0.01}}$	$84.6_{0.02}$	$\mathbf{73.2_{0.04}}$	$75.4_{0.03}$	$\mathbf{72.2_{0.09}}$	$80.2_{0.05}$	$52.3_{0.09}$	$67.9_{0.05}$	$55.6_{0.06}$

Table 1: SSEToM results on ToMI, ToMBench, FANToM for four question types: First-Order True Belief, First-Order False Belief, Second-Order True Belief and Second-Order False Belief. We compare SSEToM with five other methods: Vanilla, CoT, SimToM, TimeToM, and PercepToM, showing the differences in accuracy. Each method was tested five times, and the results are the averages of five test results. The best results for each question type are bolded, with subscripts representing the standard deviation of the five test results.

across the four question types. By comparing the results in Table 1, we found that SSEToM is more suitable for reading comprehension scenarios than interactive dialogue scenarios.

Overall Results The results in Table 1 show lower standard deviations in all cases, indicating less susceptibility to random fluctuations. Moreover, the GPT-4o-mini model generally showed higher accuracy compared to the Mistral-7B-Instruct model in almost all methods and question types. In addition, interactive dialogue-based datasets generally showed lower accuracy compared to reading comprehension-based datasets, suggesting that interactive dialogue-based datasets are more difficult for LLM.

5 Disscussion

5.1 True Belief & False Belief analysis

Figure 3 illustrates the trend of accuracy rates across three datasets and four question types for the Vanilla, SimToM, and SSEToM methods. The Vanilla method demonstrates a consistent pattern where the accuracy rate for True Belief questions is significantly higher than that for False Belief questions of the same order, with a substantial gap between them. Within the True Belief questions, First-Order questions outperform Second-Order questions. However, in the False Belief questions, the accuracy does not reach the level of random chance. This may be because True Belief questions can be answered by understanding the factual state of affairs without the need to infer the characters' beliefs. The accuracy trends for both SimToM and SSEToM methods across all question types also exhibit similar consistency, but the performance gap between First-Order and Second-Order questions is reduced, indicating that the improved method is more suitable for handling False Belief questions. This trend is particularly pronounced in reading comprehension scenarios, suggesting that interactive dialogue scenarios are more challenging. It is noteworthy that the SSEToM method outperforms the SimToM method in most results.

	ToMI	**ToMBench**	**FANToM**
Prompt	38.0%	57.9%	5.4%
Spatial	96.0%	76.0%	86.5%

Table 2: The accuracy of the characters' perceptions across three datasets for two approaches-guiding LLMs through prompts, guiding LLMs through spatial dimension analysis.

5.2 The accuracy of extracting characters' perceptions analysis

We compared two distinct approaches to enhance the model's ability to identify the perception of discussed characters. The first method involves directly using prompts to guide LLMs in extracting the target character's perception from the in-

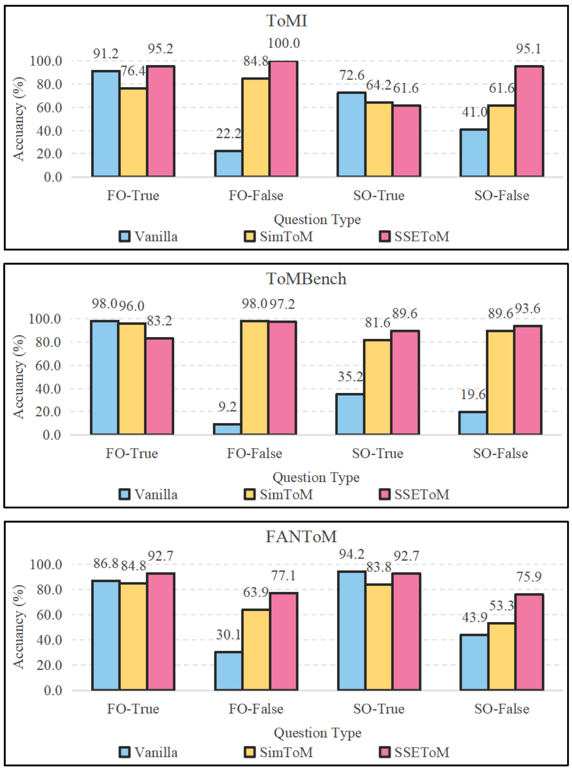


Figure 3: The trend of accuracy across three datasets for four question categories for three methods—Vanilla, SiMToM, and SSEToM.

put story. The second approach leverages spatial dimension analysis to guide LLMs. To evaluate the effectiveness of these methods, we conducted experiments on three different datasets—ToMI, ToMBench, and FANToM. The experimental results are presented as percentages of manually judged accuracy, as shown in Table 2. The results indicate that the spatial dimension analysis-guided method significantly outperforms the direct prompting method across all three datasets. Specifically, compared to the direct prompting method, the spatial method increased accuracy by 58.0%, 18.1%, and 81.1% on the ToMI, ToMBench, and FANToM datasets, respectively. This suggests that by utilizing spatial dimension analysis, LLMs can more accurately identify and comprehend the perception of target characters, thereby demonstrating higher accuracy when addressing ToM questions.

5.3 Correlation coefficients analysis

We computed the accuracy of extraction of character's perception and the accuracy in ToM problems for two methods——direct prompting (Prompt) and spatial information guidance (Spatial)——and assessed the correlation between these accuracies using the Matthews Correlation Coefficient (MCC). The MCC is a metric used to evaluate the quality of predictions for binary classification problems, ranging from -1 to 1, with values close to 1 indicating a high consistency between predictions and actual outcomes.

Experiments were conducted across three dataset—ToMI, ToMBench, and FANToM, with results presented in Table 3, which lists the correlation coefficients for these two methods across First-Order True Belief and First-Order False Belief question types. The results indicate that the trend of correlation coefficients calculated by these two methods shows consistency. Furthermore, the correlation coefficients are lower in interactive dialogue scenarios compared to reading comprehension scenarios. These findings support our hypothesis that enhancing the accuracy of spatial information can effectively improve the performance of LLMs on ToM tasks. This trend is more pronounced in reading comprehension scenarios than in interactive dialogue scenarios. Additionally, within the same dataset, this trend is more significant for First-Order questions than for Second-Order questions.

5.4 Reasons for error in extracting characters' perception

Figure 4 illustrates the reasons for errors in the extraction of character perception within the ToMI, ToMBench, and FANToM datasets when using two distinct methods—Prompt and Spatial for LLMs. These errors are classified into three categories: wide range, lack of irrelevant information, and lack of relevant information. "Relevant information" pertains to events that directly impact the answer to the ToM question.

The outcomes presented in Figure 4 demonstrate that the error cause distributions derived from employing both methods are fundamentally analogous. Within the ToMI and ToMBench datasets, errors attributed to wide range are virtually non-existent. The principal error source is identified as the lack of irrelevant information, which constitutes as high as 77.27% and 79% for True Beliefs, and 80% and 92.86% for False Beliefs, respectively. However, this irrelevant information does not appear to influence the ultimate answer, as the model is capable of delivering the correct response even in the lack of such information. In contrast, within the FANToM dataset, the primary cause of errors in False Belief questions is the lack of relevant information.

	ToMI		ToMBench		FANToM	
	FO-True	**FO-False**	**FO-True**	**FO-False**	**FO-True**	**FO-False**
Prompt + ToM Question	0.6824	0.6162	0.8246	0.7662	0.4214	0.3044
Spatial + ToM Question	0.7179	0.5846	0.7579	0.6996	0.4406	0.2652

Table 3: The Matthews Correlation Coefficient between the accuracy of characters' perceptions and the accurancy of ToM tasks for two methods-using prompt to guide LLM output of characters' perceptions and answering ToM questions based on this, using spatial dimension analysis to guide the model and answering ToM questions based on thison three datasets

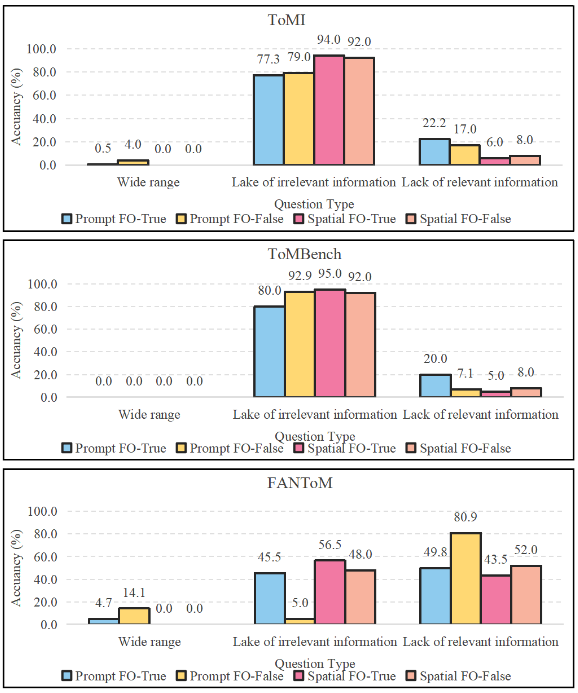


Figure 4: The potential reasons for error in extracting characters' perceptions for two methods-using prompt to guide the model, using spatial dimension analysis to guide the model. Reasons are categorized into three types: wide range, lack of irrelevant information, lack of relevant information.

This directly affects the precision of the final answer, as relevant information is indispensable for resolving the issue. In FANToM dataset, the error proportions resulting from the lack of irrelevant and relevant information are comparable in True Belief questions, indicating that when processing such stories, the model must concurrently consider both relevant and irrelevant information to ensure accuracy.

5.5 Case Study

In this section, we present a case study involving a specific scenario to better understand the process of SSEToM. The story is shown in Figure.2(a).

Event Segmentation The story is broken down into discrete events, as shown in Figure.2(a).

Event Unit Formation Spatial locations of all enents are identified:
Events are grouped into spatial units, as shown in Figure.2(b).

Character Perception Tracking Tracking each characters' presence in each location. Ava is present in the garden at the time of all events. Mia is present in the playroom after Event 2. Amelia is present in the garden during Events 3 and Event 6. On this basis, extracting perceptual ranges of each characters, as Figure.2(c).

ToM Q&A Answer the ToM question based on the extracted perceptual range of the character in question. As shown in Figure.2(d).

This case study highlights the importance of spatial perception in ToM reasoning and demonstrates how SSEToM enhances the LLMs' ability to understand and predict mental states in complex social scenarios.

6 Conclusion

In this paper, we propose SSETOM based on Event Segmentation Theory to improve the ability of LLMs to extract information known by the characters under discussion, thus improving the inference performance of LLMs in ToM tasks. Specifically, the original story is first divided into events according to rules and each event is annotated using spatial locations, and then the perceptions of the characters are tracked based on their presence at these locations, and the original ToM questions are answered on the basis of these perceptions. Extensive experimental results show that SSETOM significantly improves the ToM abilities of LLMs in reading comprehension scenarios and interactive dialogue scenarios, while making significant progress in coherent and robust ToM reasoning.

7 Limitations

Although the SSEToM method has made significant progress in improving the ToM abilities of LLMs, there are still some limitations in practical applications. Specifically, the SSEToM approach has achieved significant performance gains on specific datasets, but real-world ToM tasks may involve more complex contexts and richer background information, so future research needs to further explore the applicability of the approach in a wider range of scenarios. In addition, as the complexity of ToM tasks increases, the role of multimodal information (e.g., images, audio, etc.) in ToM reasoning becomes increasingly important. However, the SSEToM method currently focuses on textual information, and future research could explore how to combine spatial location information with other modal information to further enhance the performance of LLM in ToM tasks.

References

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2311.09277*.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. *arXiv preprint arXiv:2407.01455*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. *arXiv preprint arXiv:2407.06004*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Shima Rahimi Moghaddam and Christopher J Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

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

Weizhi Tang and Vaishak Belle. 2024. Tom-lm: Delegating theory of mind reasoning to external symbolic executors in large language models. In *International Conference on Neural-Symbolic Learning and Reasoning*, pages 245–257. Springer.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Qiaosi Wang, Sarah Walsh, Mei Si, Jeffrey Kephart, Justin D Weisz, and Ashok K Goel. 2024. Theory of mind in human-ai interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*.

Jeffrey M Zacks and Khena M Swallow. 2007. Event segmentation. *Current directions in psychological science*, 16(2):80–84.

Shao Zhang, Xihuai Wang, Wenhao Zhang, Yongshan Chen, Landi Gao, Dakuo Wang, Weinan Zhang, Xinbing Wang, and Ying Wen. 2024. Mutual theory of mind in human-ai collaboration: An empirical study with llm-driven ai agents in a real-time shared workspace task. *arXiv preprint arXiv:2409.08811*.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2022. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. *arXiv preprint arXiv:2212.10060*.