Tuning Language Models with Spatial Logic for Complex Reasoning

Tanawan Premsri
Michigan State University
premsrit@msu.edu

Parisa Kordjamshidi
Michigan State University
kordjams@msu.edu

Abstract

Recent research shows that more data and larger models can provide more accurate solutions to natural language problems requiring reasoning. However, models can easily fail to provide solutions in unobserved complex input compositions due to not achieving the level of abstraction required for generalizability. To alleviate this issue, we propose training the language models with neuro-symbolic techniques that can exploit the logical rules of reasoning as constraints and provide additional supervision sources to the model. Training models to adhere to the regulations of reasoning pushes them to make more effective abstractions needed for generalizability and transfer learning. We focus on a challenging problem of spatial reasoning over text. Our results on various benchmarks using multiple language models confirm our hypothesis of effective domain transfer based on neuro-symbolic training.

1 Introduction

Large language models dramatically altered the world of natural language processing (NLP) research through their performance on various benchmarks (OpenAI, 2023; Brown et al., 2020). Nevertheless, several limitations have been addressed by many researchers. One of the significant limitations is complex reasoning (Valmeekam et al., 2023a; Hao et al., 2023). Reasoning plays a crucial role in human cognition. Therefore, reasoning abilities are essential for establishing more reliable human-like intelligent systems (Huang and Chang, 2023). The high performance of the LLMs (OpenAI, 2023; Brown et al., 2020; Touvron et al., 2023) promised the research community that they could achieve the level of abstraction required for the reasoning process and achieve a deeper understanding of natural language. Particularly, the recent results on QA tasks indicate that the LLMs are competitive with human performance, and LLMs also show significant improvements in various reasoning of LLMs, such as mathematical reasoning (Imani et al., 2023), and logical reasoning (Creswell et al., 2022). However, it is still unclear whether LLMs are capable of true reasoning or simply memorizing patterns from training data (Huang and Chang, 2023). Recent research also illustrates that LLMs lack fundamental properties for generalization and performing human-like interactions (Gendron et al., 2023). One type of reasoning where LLMs still lag significantly is spatial reasoning.

Spatial reasoning is essential for many applications, including language grounding (Zhang et al., 2021), computer vision (Zhang and Kordjamshidi, 2022; Liu et al., 2023), robotics (Sisbot et al., 2007; Yadollahi et al., 2023) and more specific fields such as medical domain (Atif et al., 2007; Datta et al., 2020; Gong et al., 2023). Recent works (Liu et al., 2023; Chen et al., 2024) on high-level domains highlight this significant shortcomings in the spatial reasoning abilities of LLMs. Moreover, LLMs alone occasionally falter in abstract reasoning when multiple hops of reasoning in basic domains (Yang et al., 2023; Mirzaee and Kordjamshidi, 2023a). This indicates the challenge of spatial reasoning that needs attention. In this paper, we focus on one basic domain, spatial reasoning over text. Improvements in this area could potentially encourage advancements in more complex domains.

In the context of utilizing explicit logic to address multi-hop spatial reasoning, Yang et al. 2023 incorporates LLMs in a neuro-symbolic framework to pass the formal extractions to the Answer Set Programming for solving the problem. This technique overcomes the lack of spatial reasoning of LLMs and achieves a very high accuracy over a synthetic benchmark. However, utilizing this technique requires human-engineered knowledge that often does not cover all possible scenarios and rules. In particular, formalizing all required commonsense knowledge for utilizing this framework is not possible in realistic domains.

We tackle the issue of spatial reasoning in LLMs and their difficulty in achieving the abstractions required for generalizability in unobserved complex situations with a more generic neuro-symbolic framework. The main idea is to exploit spatial knowledge with available data and knowledge. Exploiting knowledge can alleviate the need for huge amounts of data and provide more robust models in unobserved situations (Faghihi et al., 2023). The main idea is to exploit spatial knowledge with available data and knowledge. Exploiting knowledge can alleviate the need for huge amounts of data and provide more robust models in unobserved situations (Faghihi et al., 2023). We proposed to fine-tune the language models with a neuro-symbolic technique that leverages available spatial logical rules of reasoning to strengthen the level of abstraction obtained by the pre-trained language model. Particularly, we train the models to minimize not only the cross-entropy loss but also the violation of logical constraints. Our hypothesis is that obtaining supervision from logical knowledge enriches the models with levels of abstraction that improve generalizability. The advantage of our proposed approach is that it does not require full access to logical knowledge. Any partially available knowledge can be exploited during training without further use at inference time. This is crucial since inference-time symbolic reasoning can be problematic for real-time applications.

We select three benchmarks, SPARTQA-HUMAN (Mirzaee et al., 2021), ResQ (Mirzaee and Kordjamshidi, 2022), and STEPGAME (Shi et al., 2022) to evaluate our proposed method. Our improved results confirm our hypothesis about the impact of neuro-symbolic training on generalizability. The contribution[1] of this paper can be summarized as follows, **(1)** We propose to exploit the available logical spatial knowledge in tuning language models to deal with their lack of spatial reasoning indicated by previous research. **(2)** We provide extensive experimental results and analysis on the existing benchmarks. Our results consistently confirm the effectiveness of our proposed approach in both encoder-based and generative language models and their generalizability for transferring knowledge to different domains. **(3)** We show that the smaller models with our proposed method generalize better on out-of-domain and complex reasoning compared to using prompt-engineered larger models.

[1]all the code is publicly available https://github.com/premsrit/SPARTUNQChain.git

C: There are three blocks called A, B, and C. A is to the left of B and B is to the left of C. In A, there is a small yellow square near and to the left of a large yellow circle. The large circle is near and to the left of a large blue triangle. In B, there is a small yellow square near and below a large blue triangle. There is a medium black triangle near and above the large blue triangle. In C, there is a small black square far to the left and above a small blue circle. The blue circle is near and to the left of a large yellow square."

q1: Is the small black thing to the right of medium black thing?

a1:Yes

q2: What is the position of small black thing relative to medium black thing.

a2: Right

Figure 1: An example of story and questions of SQA. q_1 is Yes-No and q_2 is Find relation type of questions.

2 Related Works

Investigating the reasoning capability of NLP models has become a trending research topic given the instability in the performance of LLMs (Gendron et al., 2023; Valmeekam et al., 2023b; Feng et al., 2024; Chen et al., 2024). Usually, this ability is measured via question-answering benchmarks (Weston et al., 2015; Zhou et al., 2020; Tan et al., 2023). However, there are fewer studies focusing on spatial reasoning over text. Recent benchmarks, such as SPARTUN (Mirzaee and Kordjamshidi, 2022) and STEPGAME (Shi et al., 2022) datasets, are created to address this issue by providing evaluation resources.

Some studies based on the aforementioned benchmarks report the performance of LLMs and highlight their struggle even when reasoning based on synthetic benchmarks (Yang et al., 2023). This is problematic when multi-hop reasoning is involved in inferring the answer. Multiple research papers have tried to improve the spatial reasoning capability of LLMs. Mirzaee and Kordjamshidi 2022 utilizes fine-tuning on synthetic data and illustrates an improvement in multi-hop reasoning even when applied to realistic domains. Another approach was designing the specialized memory network based on the recurrence neural network to manipulate the deep reasoning questions (Shi et al., 2022). Nevertheless, it was less effective than tuning pre-trained language models (PLMs). Some papers also focus on improving in-context learning approaches. Sharma 2023 proposes a prefix-based prompting that retrieves specific few-shot examples, aiming to transfer knowledge from simple to more complex spatial relations. Meanwhile, Hu et al. presents the idea of modifying the commonly used in-context learning technique, Chain-of-Thought (CoT) (Wei et al., 2023). They replace the textual explanation of spatial relations with symbolic representation. The approach has significantly improved spatial understanding in simple environments but is less applicable to environments with more diverse and complex relations. Another notable methodology is the disentangling of relation extraction and reasoning. Yang et al. 2023 uses GPT to perform extraction and then applies Answer set programming (ASP) to perform the reasoning process. While Mirzaee and Kordjamshidi 2023b utilizes fine-tuned models for both extraction and reasoning procedures.

In this work, we take the spatial-logical knowledge into account of the language models, so-called a neuro-symbolic training approach to solve the reasoning problem. There are many studies that emphasize the usefulness of logical knowledge to solve question-answering tasks (Sun et al., 2022; Amizadeh et al., 2020; Prager et al., 2004). We utilize logical knowledge in the form of logical constraints. Similar approaches have been used in solving multiple NLP problems. For example, Lu et al. 2021 uses a heuristic function inspired by the A* algorithm to restrict the generator given a set of constraints. While Qin et al. 2022 uses a sampling method looking for the most optimal solution that satisfies the constraints. However, both of these techniques apply the constraints during inference, which is different from what we proposed for using them during the training. There is also the benchmark that investigates the benefits of logical constraints in several NLP problems (Faghihi et al., 2023) under the DomiKnowS framework (Faghihi et al., 2021). Though we use techniques previously proposed to integrate the logic, our approach utilizes constraints solely during training, eliminating the need for access to logical constraints during inference where using them might be problematic for real-time applications. As demonstrated in our experiments, leveraging logic can guide models to achieve higher levels of abstraction during training, resulting in better generalization at test time.

3 Methodology

3.1 Problem Definition

For the spatial QA task, the input consists of textual context C and a textual question Q asking about spatial relations of objects within the scene description. The output is the answer(s) to the question, denoted as A. There are two types of questions/answers. The first type is Yes/No (YN) questions, and the other is find-relation (FR) questions. We restrict the answer domain of YN to $A = \{Yes, No\}$, while the answers for FR depend on the dataset. An example of answer domain for FR is $A = \{left, right, above, below, \dots\}$. More details on the variations of FR answer(s) are provided in the experimental section.

3.2 Backbone Language Models

We selected two types of language models, which are encoder-based models, BERT family (Devlin et al., 2019), and generative models, Flan-T5 (Chung et al., 2022), to evaluate the effectiveness of our proposed neuro-symbolic fine-tuning approach. Furthermore, we evaluate the performance of Large Language Models, GPT-family, and Llama3 on these tasks to compare them with our proposed approach.

3.2.1 Encoder-based Language Models

We utilize BERT as the backbone architecture, following previously reported results (Mirzaee and Kordjamshidi, 2022). We fine-tune BERT with an additional classification layer to perform QA tasks. For the YN domain, we use a binary classification layer, while for the FR domain, we use multiple binary classification layers to conduct multi-label/multi-class classification. The number of binary classification layers in the FR domain depends on the number of possible answers, $|A|$. The input to the BERT model is formed by concatenating the question, Q, and the context, C. After feeding this input to BERT, we use the $[CLS]$ token from the last layer as the input to the classification layer(s) for final answer prediction.

3.2.2 Generative Language Models

We use Flan-T5 (Chung et al., 2022) as the generative baseline model, which is an open-source model for our fine-tuning purposes. To reduce the intensive computational cost, we utilize the LoRA adapter (Hu et al., 2021), which decreases the number of training parameters.

We applied the YN setting to Flan-T5, where processing the output is more straightforward since the first token can simply represent the outputs and be accessible from the generation of the model. We select the highest probability between the Yes and No tokens from the first token of the Flan-T5 output as the answer to the input question. The input prompt to the Flan-T5 model has the following structure, "Answer based on the context: C Q."

For In-context Learning, we select Llama3-8B (AI@Meta, 2024), GPT3.5 (Brown et al., 2020), and GPT-4 (OpenAI, 2023) as the backbone LLMs for prompt engineering. Then, we apply *Zero_shot*, *Few_shot* techniques as the baseline of prompt-based learning.

Zero-shot. We give the prompt to LLMs for answers based on the scene description and the question without having any examples.

Few-shots. We randomly select four questions from the training set. Then, we add the information about the response format for each question with the corresponding label. We provide these examples along with the scene description and target question to GPT for querying the answer.

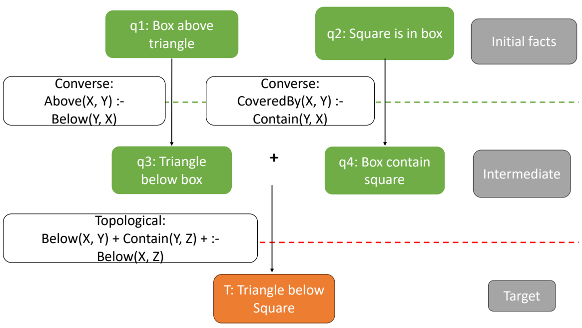

Figure 2: An example of the chain of reasoning questions (Q-chain). Note that the factual sentences will turn to questions like "Is triangle below square?"

3.3 Training with Spatial Logic

To address the challenge of multi-hop spatial reasoning, we utilize symbolic knowledge that expresses logical spatial reasoning rules during training. We assume a formal knowledge base (KB) of rules, referred to as *spatial logic*, is given during training time. This *spatial logic* does not need to be exhaustive or cover all reasoning aspects. Any available knowledge can be exploited to enhance the reasoning ability of the underlying Language Model. In this work, we use 79 rules of reasoning collected in (Mirzaee and Kordjamshidi, 2022). These rules are divided into four categories: converse, inverse, transitive, and transitive + topological. The KB covers rules between 15 spatial relations, including directional and topological relations, such as "*If above(x,y) then below (y,x).*" or "*If inside(x,y) and left(x,z) then left(y,z).*" Our main hypothesis is that providing supervision from high-level logical knowledge enables the model to capture higher levels of abstraction, improving generalization to other domains. To exploit the spatial logic, we follow two steps, 1) Translate spatial logic: we convert the *spatial logic* into example-based logical constraints, 2) Obtain the soft logic surrogate: we convert the logical constraints to differentiable soft logic, 3) Incorporate Constraints: we add the constraint violations as the part of training loss objective.

Translate spatial logic. As questions in the dataset require multiple hops of reasoning following spatial logic, we formulate the constraints to express consistency with this chain of reasoning. An example of a question is shown in Figure 2. The target question asks about the relation between a triangle and a square, *"Is the triangle below square?"*. To answer this question, we introduce the intermediate facts and turn them into a set of questions, denoted as $Q\text{-}chain$. An example of intermediate questions in $Q\text{-}chain$ is shown in the green boxes of Figure 2. Given the chain of reasoning, we observe that two initial facts, q_1 entails q_3, $q2$ entails $q4$, and $q4$ using the converse rule. Both $q3$ and $q4$ entail the target question using the transitivity rule. The corresponding logical constraint between q1 and q3 is $q_1 \Rightarrow q_3$. More constraints associated with this example can be found in Table 1. All conversions of spatial logic to constraint can be found in Appendix B.

Obtain the soft logic surrogate There are three commonly used types of conversions for mapping logical constraints into differentiable soft logic: t-norm product, t-norm ukasiewicz, and t-norm Gödel (Li et al., 2019). We use the t-norm product for our conversion, $\neg A$ is 1 - a, $A \wedge B$ is ab, $A \vee B$ is a + b - ab, and $A \Rightarrow B$ is min(1, $\frac{b}{a}$) where A, B are concepts with probabilities a, b, respectively.

Incorporate Constraints. Optimization of an objective that includes both task performance and soft constraints loss has been proposed in several previous research (Li et al., 2019; Asai and Hajishirzi, 2020; Étienne Bamas et al., 2020). Inspired by them, we use the following objective,

$$\nabla_w \mathcal{L}(w;\Lambda) = \nabla_w L(w) + \sum_{k=1}^{K} \lambda_k \nabla_w h_k(w) \quad (1)$$

where $L(w)$ is the task-performance loss function, i.e. Cross-Entropy loss, K is the number of logical constraints, h_k measures the violation from a given logical constraint in soft logic differentiable form, and λ_k is a learning parameter indicating the importance of obeying k^{th} rule.

In fact, in this work, we utilize the implementation in DominKnowS (Faghihi et al., 2021) for integration of the constraints. The DomiKnowS framework provides a declarative language to integrate symbolic knowledge as logical constraints. We provide the original logical forms, and it automatically converts them to the differentiable form and facilitates incorporating them in the loss function. We, specifically, use an implemented optimization that exploits a dual formulation of the original loss proposed in (Étienne Bamas et al., 2020), called the Primal-Dual(PD) program.

Creating the Q-chain. We automatically augmented the questions in the training data to include the $Q\text{-}chain$ for every question. To automatically create the $Q\text{-}chain$, we use the initial fact annotations provided in SpaRTUN to obtain the intermediate facts for concluding the target answer. We exhaustively search for finding the resolution tree for the target fact.

Rules	Constraints in YN	Constraints in FR
R_1	$q_1 \Rightarrow q_3$	$Above(q_1) \Rightarrow Below(q_3)$
R_2	$q_2 \Rightarrow q_4$	$CoveredBy(q_2) \Rightarrow Contain(q_4)$
R_3	$q_3 \wedge q_4 \Rightarrow t$	$Below(q_3) \wedge Contain(q_4) \Rightarrow Below(t)$

Table 1: Example of logical constraints presented in above example of $Q - chain$, where R_i refer to presented rule i used in the example.

3.4 In-context Learning with Spatial Logic

We utilize spatial logical reasoning to create few-shot examples in the in-context prompting approaches, including Chain-of-Thoughts (Wei et al., 2023), Chain-of-Symbols (Hu et al., 2023), and others. The intention of these experiments is to analyze and compare the LLMs' performance against our proposed fine-tuning method.

Chain-of-Thought (CoT). To enable LLMs to provide reasoning explanations rather than a single answer, we use CoT. In this setting, we manually augment the answer response with the reasoning explanations based on spatial logical rules. Then, we give these CoT examples alongside prompt and target questions to generate the target answer with an explanation. An example of a reasoning explanation of CoT is given in Table 2.

Logical Representation (LR). Generating the chain of reasoning requires applying spatial logical rules such as symmetric and transitivity. In our problem setting, we further modify the CoT format to a first-order logical form. The predicate-argument form, $R(obj_i, obj_j)$ represents relation R holds between arguments obj_i and obj_j that denote object i and object j respectively. We follow the same pipeline for CoT by replacing the response format with this logical form. An example of CoT using formal rules format, denoted as LR, is provided in Table 2.

Chain of Symbol (CoS). This approach was introduced in (Hu et al., 2023). This method illustrates the advantage of symbolic representation over natural language on both number of tokens and performance effectiveness. We construct the symbolic explanations of our CoT with their proposed format. An example of the CoS counterpart of our CoT expression is shown in Table 2.

Step-by-Step Reasoning. Based on the few-shot examples of the SPARTQA-Human dataset, often scene descriptions are complex and long. Therefore, we decided to simplify the context. We take an additional prompting step to split the context using LLMs before asking the question. In this setting, we expect that each line of generated context contains only one simple spatial description. An example of the original context is "There exists a big red square, and medium red square in a block called A. The big one is touching the right edge of the block." The expected generated context from LLM should be in the form of "The big red square in block A. The big red square is touching the right edge of block A..." More examples can be found in Appendix A.2. We call this prompting approach *Step-by-Step Reasoning* in the experiments.

Domain	Text
Chain of Thought(CoT)	large red square is to the left of a small green square
Logical Representation(LR)	Left(large red square, small green square)
Chain of Symbol(CoS)	(large, red, square) < (small, green, square)

Table 2: An example of a spatial relation used in Chain of Thought, Logical Representation, and Chain of Symbol.

4 Experimental

The main focus of the experiments is evaluating the ability of LMs in spatial reasoning by looking into both fine-tuning and in-context learning. Moreover, we explore the impact of using spatial logic in both models, especially demonstrating the advantage of our proposed neuro-symbolic model.

4.1 Datasets

Our experiments are conducted on the following datasets: SpaRTUN, SPARTQA-Human, ResQ, and STEPGAME. However, we only evaluate models on three out of four datasets, which are SPARTQA-Human, ResQ, and STEPGAME. This decision is based on the previous results (Mirzaee and Kordjamshidi, 2022) that demonstrate the challenge of these datasets compared to testing on SpaRTUN which is likely to be solved with typical fine-tuning on SpaRTUN.

SpaRTUN (Mirzaee and Kordjamshidi, 2022) is a synthetic SQA dataset. This dataset contains both YN and FR types of questions that require multi-hop reasoning to answer. It covers a wide range of spatial relations.[2] The answer for YN is $A = \{Yes, No\}$. While, the answer for the FR is the subset of $A = \{left, right, above, below, behind, front, near, far, dc, ec, po, tpp, ntpp, tppi, ntppi\}$. The dataset provides the chain of reasoning annotations which we use to create the $Q-chains$ for our proposed training method.

SPARTQA-Human (Ver.1) (Mirzaee et al., 2021) is a small human-annotated SQA dataset. The dataset contains both YN and FR types of questions. We only use the YN portion of this dataset to fine-tune and evaluate our models.

SPARTQA-HUMAN (Ver.2), we follow the methodologies from (Mirzaee et al., 2021) to further extend the human annotation dataset with the new contexts and questions. The purpose of creation is to enrich the SPARTQA-HUMAN with a larger test for more diverse patterns and questions. Annotators included authors and an undergrad student paid as a research assistant.

ResQ (Kordjamshidi et al., 2017; Mirzaee and Kordjamshidi, 2022) is a small realistic domain SQA dataset that includes multi-hop reasoning questions. The depth of reasoning is smaller than other synthetic datasets, but it often requires commonsense knowledge to infer the answer. It includes the information about the depth of reasoning denoted as k that can be 1 or 2. For many questions, the reasoning is complex and mixed up with commonsense in which case this is annotated as *unclassified* depth. 12.30% of the test examples fall in the k=1 split, 23.93% fall in the k=2 split, and the rest are unclassified. This dataset contains only YN questions. We use this dataset to evaluate the performance of our models in realistic domains.

STEPGAME (Shi et al., 2022) is a synthetic SQA dataset containing extensive multi-hop reasoning questions over spatial relations until the depth of ten with annotation denoted as k in our tables of experimental results. This dataset contains only FR questions. The domain of answer is $a = \{left$, $right$, $above$, $below$, $overlap$, $lower\text{-}left$, $lower\text{-}right$, $upper\text{-}left$, $upper\text{-}right\}$. The advantage of utilizing this dataset is the possibility of evaluating the performance of the models at various depths of reasoning.

4.2 Experimental Models

4.2.1 Fine-tuning Models

BERT. This is the baseline architecture described in Section 3.2.1. We use bert-based-uncased as the initial checkpoint. This model only fine-tune with the target dataset in each experiment.

BERT-T. This is the baseline model for transfer learning. We fine-tune the model with SpaRTUN first. Then, we further fine-tune with the target dataset of each experiment.

BERT-T + Q-Chain. We follow the methodology explained in section 3.3 to inject logical knowledge into the BERT-T model via training with constraints using the DomiKnowS framework (Faghihi et al., 2021).

Flan-T5. We follow the methodology described in section 3.2.2. We selected the flan-t5-base in this experiment as the starting checkpoint.

Flan-T5-T. We utilize the same setting as the Flan-T5 model. However, we further fine-tuned the model with SpaRTUN before finally fine-tuning it with the target dataset.

Flan-T5-T + Q-Chain. We inject the logical knowledge into Flan-T5-T model using proposed method in section 3.3. The backbone is the same as the baseline model without any changes.

Experimental Setup. We use the same hyperparameters on all experiments. The learning rate was one of $\{1e-5, 8e-6, 1e-6\}$. For SPARTQA-

[2]The dataset also includes Don't Know questions but for simplicity we assume a closed world assumption and use No label for those questions.

	SPARTQA-Human		ResQ				
Model	Ver.1	Ver.2	k=1	k=2	unclassified	All	Line
BERT	54.54	53.57	70.67	56.85	60.66	60.98	1
RoBERTa	54.54	-	76.00	56.16	58.87	60.33	2
BERT-T	55.94	58.03	**76.00**	54.79	**61.18**	61.15	3
RoBERTa-T	49.65	-	64.67	57.87	55.78	56.72	4
BERT-T+Q-Chain (Our)	**59.44**	**58.92**	72.00	**58.90**	59.90	**61.31**	5
Flan-T5	54.54	60.71	74.67	56.16	61.44	61.80	6
Flan-T5-T	49.65	57.14	81.33	54.79	61.44	62.30	7
Flan-T5-T+Q-Chain (Our)	**55.94**	**61.61**	**81.33**	**57.53**	**63.75**	**64.43**	8
GPT3.5 (zero-shot)	58.04	58.03	74.67	60.95	66.58	66.22	9
GPT3.5 (few-shot)	62.23	58.92	84.00	68.49	68.12	70.16	10
GPT3.5 (CoT)	65.73	**71.43**	**86.67**	67.12	68.64	70.49	11
GPT-4 (zero-shot)	**77.62**	68.75	84.00	73.97	**76.86**	**77.05**	12
Llama-3 (zero-shot)	61.54	50.89	80.00	64.38	67.35	68.20	13
Llama-3 (few-shot)	62.94	60.71	82.67	69.86	71.46	72.46	14
Llama-3 (CoT)	67.83	70.54	82.76	**76.03**	67.10	71.15	15

Table 3: Accuracy of SPARTQA-Human and ResQ with various models. For ResQ, k is the number of the reasoning steps required for answering the questions. *Unclassified* indicates the cases in which k was a challenge for human annotators to decide.

Human and ResQ, the number of epochs was 100, whereas for STEPGAME, it was 30. For fine-tuning with SPARTUN, 12 epochs for BERT and 8 epochsfor Flan-T5 were used. The final hyper-parameters were selected using the development portion of the target dataset. The loss function in all experiments was Cross-Entropy Loss, and the optimizer was Adam with *amsgrad* parameters set to *True*. We run all experiments on 8 A6000 GPUs, taking roughly 100 GPU hours.

4.2.2 In-context Learning with LLMs.

As explained in Section 3.2.2 and Section 3.4, we evaluate the performance of all selected LLMs, GPT3.5, GPT-4, and Llama3-8B, using ResQ and SPARTQA-Human to compare with our fine-tuning approach. The selected examples and prompts for each in-context learning method can be found in the Appendix A.2.

Zero-shot. We directly ask LLMs to answer the question given the corresponding context.

Few-shot. We provide LLMs with four randomly selected examples from the training data.

CoT. We augment the *few-shot* setting with reasoning explanations along with the *CoT* examples.

CoT-Formal. We provide the few-shot examples with the described logical format in Section 3.4.

CoS. We further alter the *few-shot* examples with the CoS symbolic forms explained in Section 3.4.

4.3 Results

Realistic Domain. ResQ is our realistic domain. As observed in Table 3, using the *Q-chain* demonstrates its effectiveness on both fine-tuning models (BERT and Flan-T5) with more significant improvement on Flan-T5. Specifically, Flan-T5-T+*Q-chain* (line 8) shows a 2% improvement over Flan-T5-T (line 7). For a deeper analysis of this result, we evaluated the performances of the three different splits of ResQ. Based on our observations in Table 3, it revleals that our model consistently imporve on $k = 2$, but adversely affects BERT's performance on $k = 1$ and the *unclassified* categories. We emphasize that the $k = 2$ split requires more hops of reasoning, while the unclassified portion requires commonsense knowledge. It is expected that our method enhances deeper reasoning steps while it does not address the lack of commonsense knowledge in the model. This hypothesis is confirmed by results with LLMs. We can observe that LLMs, on average, achieve higher performance on this dataset, especially in *unclassified* category (lines 9 to 15). LLMs consistently show around 2% to 13% improvement over Flan-T5+T+Q-Chain. This indicates that most LLMs' improvement is mainly due to their commonsense knowledge rather than their complex reasoning capability, the main objective of our proposed method. We tried few-shot and CoT prompting, and the results did not significantly vary the results. Therefore, we believe integrating our method into the baselines with a stronger commonsense capability will increase the overall performance. This improvement can be seen when comparing Flan-T5 to BERT, where the larger generative model (Flan-T5 here) shows a larger improvement on the *unclassified* category. Consequently enhancing its performance in other sub-categories as well.

Model	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
BERT	98.51	95.53	91.68	66.71	49.11	41.47	41.47	32.09	28.94	28.16
BERT-T	98.50	95.32	**93.26**	**76.78**	**66.36**	58.76	53.70	46.27	42.71	40.12
BERT-T+Q-Chain (Our)	**98.70**	**96.45**	93.03	74.58	64.95	**59.04**	**54.38**	**49.23**	**45.36**	**44.05**
GPT3 (few-shot)	55.00	37.00	25.00	30.00	32.00	29.00	21.00	22.00	34.00	31.00
GPT3 (CoT)	61.00	45.00	30.00	35.00	35.00	27.00	22.00	24.00	23.00	25.00
Llama-3 (few-shot)	38.01	27.87	24.15	21.27	19.75	18.03	16.88	15.52	15.17	14.70

Table 4: Accuracy of STEPGAME on several models including results of GPT3 reported in (Yang et al., 2023).

Synthetic Domain with More Complex Logical Reasoning. We evaluated SPARTQA-Human and STEPGAME for this analysis. We consistently observe improvement with our proposed *Q-chain* in this domain, which requires multiple hops of reasoning. As observed in Table 3, *Q-chain* consistently shows improvement in both Flan-T and BERT compared to fine-tuning without it. Moreover, the gap between small PLMs and LLMs is much less on this dataset compared to the realistic domain (ResQ). This is expected, as LLMs are better at commonsense rather than complex reasoning, as previously explained. This result is further supported when evaluating the model on STEPGAME. As seen in Table 4, the fine-tuning method consistently demonstrates significant differences in all steps of reasoning compared to LLMs. The struggle of GPT3 on reasoning on this dataset is also investigated in (Yang et al., 2023). We took the reported results from this paper in Table 4. Looking into the details of STEPGAME, we notice that our proposed method consistently gains an improvement of 1% - 4% on high hops of reasoning ($k = 6$ to $k = 10$), similar to the observation in ResQ. Overall, these results confirm our main hypothesis that our proposed method equips the models with a higher level of logical abstraction to conduct deeper steps of reasoning.

Model	Raw Context	Step by step
GPT3.5 (zero-shot)	58.04	63.64
GPT3.5 (few-shot)	62.23	64.33
GPT3.5 (CoT)	65.73	67.83
GPT3.5 (LR)	64.33	59.44
GPT3.5 (CoS)	60.14	58.74
GPT-4 (zero-shot)	**77.62**	**78.32**

Table 5: The accuracy of LLMs on SPARTQA-HUMAN.

In-context Learning. For comparison, we also experimented with variations of in-context learning and prompt engineering. According to Table 3, Table 5, we found that CoT is the most effective way to conduct spatial reasoning in both LLMs (GPT and Llama). We also investigated whether using Formal Representations in the context would make a difference in the performance. However, our results in Table 5 (line 4), show that the formal representation slightly worsened the performance, and using natural language in CoT was more effective. We further evaluated a CoS symbolic representation alternative in our experiments. Our results show that CoT and LR outperform the CoS while using the same in-context examples. This demonstrates the advantage of using natural language text in the prompt compared to the symbolic representation proposed in CoS. To evaluate more sophisticated prompting techniques, we ran the step-by-step reasoning explained in Section 3.4 on LLMs. As shown in Table 5, step-by-step reasoning achieves higher results compared to the majority of natural prompting techniques including zero-shot, few-shot, and CoT. The improvement ranges from 2% to 5%, indicating a simpler and shorter context can lead to a better understanding of the spatial relations in LLMs. Lastly, to compare the opened-source, Llama, and closed-source models, GPT. Our experiments indicate that Llama-3 is competitive with the closed-source GPT3.5, with only a small difference in the Yes/No domain as shown in Table 3. Therefore, we prioritize utilizing open-source models in our future research.

5 Conclusion

Given the importance of spatial reasoning in many real-world applications, we focus on improving this reasoning skill in language models. We equip LLMs with neuro-symbolic supervision to leverage logical knowledge during fine-tuning. This practice boosts their capability of capturing the necessary level of abstraction for spatial reasoning and enhances the generalizability for transferring knowledge across domains. We demonstrate that our constraint-based training technique achieves higher accuracy compared to other competitive Spatial Question-answering baselines across various benchmarks. Furthermore, the results indicate that our model performs better in scenarios requiring more reasoning steps. Lastly, we compare our models with state-of-the-art LLMs. Experimental comparisons show that while large LLMs like GPT3.5 perform better in commonsense reasoning, we achieve higher performance in multi-hop spatial question-answering with comparatively smaller language models like BERT.

6 Limitations

While we improve the reasoning capability of the models, our approach by no means solves the general reasoning problem. The trustworthiness and reliability of the LLMs are still a research challenge. Our models will need GPU resources to run which might be expensive. Our research is based on open source data and code and the results will be reproducible. We do not see any ethical concerns in our research approach and in the impact of our results. Our data, being limited to the spatial reasoning domain, does not include any specific type of bias that can harm minority people belonging to any specific gender or race.

References

AI@Meta. 2024. Llama 3 model card.

Saeed Amizadeh, Hamid Palangi, Oleksandr Polozov, Yichen Huang, and Kazuhito Koishida. 2020. Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning". *Preprint*, arXiv:2006.11524.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. *Preprint*, arXiv:2004.10157.

Jamal Atif, Céline Hudelot, Geoffroy Fouquier, Isabelle Bloch, and Elsa D Angelini. 2007. From generic knowledge to specific reasoning for medical image interpretation using graph based representations. In *IJCAI*, pages 224–229.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *Preprint*, arXiv:2401.12168.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *Preprint*, arXiv:2205.09712.

Surabhi Datta, Yuqi Si, Laritza Rodriguez, Sonya E Shooshan, Dina Demner-Fushman, and Kirk Roberts. 2020. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest x-ray reports using deep learning. *Journal of Biomedical Informatics*, 108:103473.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Hossein Rajaby Faghihi, Quan Guo, Andrzej Uszok, Aliakbar Nafar, Elaheh Raisi, and Parisa Kordjamshidi. 2021. Domiknows: A library for integration of symbolic domain knowledge in deep learning. *Preprint*, arXiv:2108.12370.

Hossein Rajaby Faghihi, Aliakbar Nafar, Chen Zheng, Roshanak Mirzaee, Yue Zhang, Andrzej Uszok, Alexander Wan, Tanawan Premsri, Dan Roth, and Parisa Kordjamshidi. 2023. Gluecons: A generic benchmark for learning under constraints. *Preprint*, arXiv:2302.10914.

Chun Feng, Joy Hsu, Weiyu Liu, and Jiajun Wu. 2024. Naturally supervised 3d visual grounding with language-regularized concept learners. *Preprint*, arXiv:2404.19696.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2023. Large language models are not abstract reasoners. *Preprint*, arXiv:2305.19555.

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 2023. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *Preprint*, arXiv:2306.13465.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023.

Reasoning with language model is planning with world model. *Preprint*, arXiv:2305.14992.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. 2023. Chain-of-symbol prompting elicits planning in large langauge models. *Preprint*, arXiv:2305.10276.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. *Preprint*, arXiv:2212.10403.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *Preprint*, arXiv:2303.05398.

Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017. Clef 2017: Multimodal spatial role labeling (msprl) task overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 367–376, Cham. Springer International Publishing.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. *Preprint*, arXiv:1909.00126.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2021. Neurologic a*esque decoding: Constrained text generation with lookahead heuristics. *Preprint*, arXiv:2112.08726.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. Spartqa: : A textual question answering benchmark for spatial reasoning. *Preprint*, arXiv:2104.05832.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. *Preprint*, arXiv:2210.16952.

Roshanak Mirzaee and Parisa Kordjamshidi. 2023a. Disentangling extraction and reasoning in multi-hop spatial reasoning. *Preprint*, arXiv:2310.16731.

Roshanak Mirzaee and Parisa Kordjamshidi. 2023b. Disentangling extraction and reasoning in multi-hop spatial reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3379–3397, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

John Prager, Jennifer Chu-Carroll, and Krzysztof Czuba. 2004. Question answering using constraint satisfaction: QA-by-dossier-with-contraints. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 574–581, Barcelona, Spain.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *ArXiv*, abs/2202.11705.

Manasi Sharma. 2023. Exploring and improving the spatial reasoning abilities of large language models. *Preprint*, arXiv:2312.01054.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. *Preprint*, arXiv:2204.08292.

Emrah Akin Sisbot, Luis F Marin, and Rachid Alami. 2007. Spatial reasoning for human robot interaction. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2281–2287. IEEE.

Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2022. Reasoning over logically interacted conditions for question answering. *Preprint*, arXiv:2205.12898.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *Preprint*, arXiv:2306.08952.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023a. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Preprint*, arXiv:2206.10498.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023b. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *Preprint*, arXiv:2206.10498.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *Preprint*, arXiv:1502.05698.

Elmira Yadollahi, Miguel Alexandre Monteiro, and Ana Paiva. 2023. Learning spatial reasoning in virtual vs. physical games with robots. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, HAI '23, page 162170, New York, NY, USA. Association for Computing Machinery.

Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. Coupling large language models with logic programming for robust and general reasoning from text. *Preprint*, arXiv:2307.07696.

Yue Zhang, Quan Guo, and Parisa Kordjamshidi. 2021. Towards navigation by reasoning over spatial configurations. *Preprint*, arXiv:2105.06839.

Yue Zhang and Parisa Kordjamshidi. 2022. Lovis: Learning orientation and visual signals for vision and language navigation. *Preprint*, arXiv:2209.12723.

Ming Zhou, Nan Duan, Shujie Liu, and Heung-Yeung Shum. 2020. Progress in neural nlp: Modeling, learning, and reasoning. *Engineering*, 6(3):275–290.

Étienne Bamas, Andreas Maggiori, and Ola Svensson. 2020. The primal-dual method for learning augmented algorithms. *Preprint*, arXiv:2010.11632.

A Appendix

A.1 Data Statistic

The data statistic of each dataset is shown in Table 6.

Dataset	Train	Test	Dev
SpaRTUN (YN)	20633	3232	3152
SpaRTUN (FR)	18400	2818	2830
Q-Chain (YN)	56987	-	-
Q-Chain (FR)	46750	-	-
SPARTQA-Human(Ver.1)	161	143	51
SPARTQA-Human(Ver.2)	200	112	60
ResQ	1008	610	333
STEPGAME	50000	5000	100000

Table 6: Size of each SQA benchmark used in experiments

A.2 In-context Learning Examples

The example of each in-context learning prompt and example is indicated below.

A.2.1 Few-Shot

system : You will be given story and question. Then, you will answer either only Yes or No based on given story. Candidate answer: [Yes, No]

user : There exist a big red square, a big red triangle, a medium red square, and a small green circle in a block called A. The triangle and medium square are touching the bottom edge of the block. The big and medium square are touching the right edge of the block. And the circle is above the big square which is above the medium square. There is another block called B to the left of block A. A medium green square is touching the left edge of block B and is below a medium red square. The medium red square is above and to the left of a small green square. Also a medium red triangle is below and to the left of the small square. Is the medium red triangle below the red square in block B?

assistant : Yes

user : There are three blocks A, B and C. Block A has a medium white triangle and a small red rectangle. There exists a medium white rectangle to the left of the triangle and to the right of the red rectangle. Aove block A there is block B. It has a small white oval touching the bottom edge of the block and a medium white triangle to the right of it. Below block A there is block C. Its has a small green triangle touching the top edge of the block. A medium white triangle is below a small green rectangle. Another small green triangle is below and to the right of the white triangle. Are all rectangles above a white triangle?'

assistant : Yes

user : There is a block called A. Inside A, there is a black triangle that is above and to the left of a black circle. There is another block called B which is to the right of A. In B, there are two yellow circles, a blue circle, and a blue square. The blue circle is touching the left edge of B. One of the yellow circles is to the right of the blue circle and to the left of the blue square. The blue square is touching the right edge of B and is above the other yellow circle, Which is also touching the right edge of B. Does block B have all of the circles inside it?

assistant : No

user : There is a block called A. In A, there is a large blue square near and below a medium blue triangle. Far to the left of the medium blue triangle is a small yellow square. Far above the medium blue triangle is a small blue triangle. There is another block called B, which is to the right of A. In B, There is a small yellow circle touching the left edge of B. Far to the right of the circle is a large yellow square. Near and above the large square is a small yellow square. There is another block called C, which is to the right of B. In C, there is a small blue circle near and to the left of a large yellow triangle. The large yellow triangle is near and to the left of a large blue triangle. The large blue triangle is far above a large blue circle. The large blue circle is touching the bottom edge of C. Are the large blue things in C to the left of the large blue thing in A?

assistant : No

user : There is a block called A. In A there is a red triangle Which is above a black circle. The black circle is touching the bottom of A. A yellow square is touching the top edge of A and below that there is a red oval is touching the right edge of A. There is a second block call B Which is to the left of A. Inside B there are 3 squares, a brown one, a green one, and a red one. The green square is above the red square and touching the left edge of B. The brown square is touching both the bottom edge and the right edge of B. The red square is touching the left edge of the brown square. Is the oval to the right of the brown thing?

A.2.2 Chain of Thought

system : You will be given story and question. Then, you will answer either Yes or No with explanation based on given story step by step. Candidate answer: [Yes, No]

user : There exist a big red square, a big red triangle, a medium red square, and a small green circle in a block called A. The triangle and medium square are touching the bottom edge of the block. The big and medium square are touching the right edge of the block. And the circle is above the big square which is above the medium square. There is another block called B to the left of block A. A medium green square is touching the left edge of block B and is below a medium red square. The medium red square is above and to the left of a small green square. Also a medium red triangle is below and to the left of the small square. Is the medium red triangle below the red square in block B?

assistant : From information provided:The medium red square is above a small green square in B.Therefore, the small green square is below the medium red square in B. The medium red triangle is below the small square in B, and the small green square is below the medium red square in B.Therefore, the medium red triangle is below the medium red square in B. Answer: Yes

user : There are three blocks A, B and C. Block A has a medium white triangle and a small red rectangle. There exists a medium white rectangle to the left of the triangle and to the right of the red rectangle. Aove block A there is block B. It has a small white oval touching the bottom edge of the block and a medium white triangle to the right of it. Below block A there is block C. Its has a small green triangle touching the top edge of the block. A medium white triangle is below a small green rectangle. Another small green triangle is below and to the right of the white triangle. Are all rectangles above a white triangle?'

assistant : From given information: A medium white triangle is below a small green rectangle in C.Then, small green rectangle is above medium white triangle in C.A small red rectangle in block A, a medium white triangle in C, and block C is below block A.Then, a medium white triangle is below a A small red rectangle. A medium white triangle is below a A small red rectangle. Therefore, a small red rectangle is above a medium white triangle.A medium white rectangle is in A,a medium white triangle in C, and block C is below block A.Then, a medium white triangle is below a A medium white rectangle. A medium white triangle is below a A medium white rectangle. Therefore, a medium white rectangle is above a medium white triangle.A medium white rectangle is above a medium white triangle.Thus, all rectangles from given story are above a white triangle. Answer: Yes'

user : There is a block called A. Inside A, there is a black triangle that is above and to the left of a black circle. There is another block called B which is to the right of A. In B, there are two yellow circles, a blue circle, and a blue square. The blue circle is touching the left edge of B. One of the yellow circles is to the right of the blue circle and to the left of the blue square. The blue square is touching the right edge of B and is above the other yellow circle, Which is also touching the right edge of B. Does block B have all of the circles inside it?

assistant : From given relation, A black triangle is above a black circle in A.Then, there is black circle in A.Therefore, not all circle are in B. Answer: No

user : There is a block called A. In A, there is a large blue square near and below a medium blue triangle. Far to the left of the medium blue triangle is a small yellow square. Far above the medium blue triangle is a small blue triangle. There is another block called B, which is to the right of A. In B, There is a small yellow circle touching the left edge of B. Far to the right of the circle is a large yellow square. Near and above the large square is a small yellow square. There is another block called C, which is to the right of B. In C, there is a small blue circle near and to the left of a large yellow triangle. The large yellow triangle is near and to the left of a large blue triangle. The large blue triangle is far above a large blue circle. The large blue circle is touching the bottom edge of C. Are the large blue things in C to the left of the large blue thing in A?

assistant : Based on the provided content. Block B is to the right if block A.Block C is to the right of block B.Therefore, block C is to the right of block A.In block C, the yellow triangle is near and to the left of large blue triangle. Therefore, there is large blue triangle in C.In block A, there is a large blue square near and below a medium blue triangle.Therefore, there is large blue square in A.Then, large blue triangle in C is to the right of large blue square in A.So, large blue triangle in C is not to the left of large blue square in A. Answer: No

user : There is a block called A. In A there is a red triangle Which is above a black circle. The black circle is touching the bottom of A. A yellow square is touching the top edge of A and below that there is a red oval is touching the right edge of A. There is a second block call B Which is to the left of A. Inside B there are 3 squares, a brown one, a green one, and a red one. The green square is above the red square and touching the left edge of B. The brown square is touching both the bottom edge and the right edge of B. The red square is touching the left edge of the brown square. Is the oval to the right of the brown thing?

A.2.3 Formal Language

system : You will be given story and question. Then, you will answer either Yes or No with explanation based on given story step by step. Candidate answer: [Yes, No]

user : There exist a big red square, a big red triangle, a medium red square, and a small green circle in a block called A. The triangle and medium square are touching the bottom edge of the block. The big and medium square are touching the right edge of the block. And the circle is above the big square which is above the medium square. There is another block called B to the left of block A. A medium green square is touching the left edge of block B and is below a medium red square. The medium red square is above and to the left of a small green square. Also a medium red triangle is below and to the left of the small square. Is the medium red triangle below the red square in block B?

assistant : Below(Medium red triangle in B, red square in block B) Answer: Yes

user : There are three blocks A, B and C. Block A has a medium white triangle and a small red rectangle. There exists a medium white rectangle to the left of the triangle and to the right of the red rectangle. Above block A there is block B. It has a small white oval touching the bottom edge of the block and a medium white triangle to the right of it. Below block A there is block C. Its has a small green triangle touching the top edge of the block. A medium white triangle is below a small green rectangle. Another small green triangle is below and to the right of the white triangle. Are all rectangles above a white triangle?'

assistant : Below(C, A) => Above(A, C) Contain(A, Small red rectangle) + Above(A, C) + Contain(C, Medium white triangle)=>Above(Small red rectangle, Medium White triangle) Above(A, C) & Contain(A, Medium white rectangle) & Contain(C, Medium white triangle)=>Above(Medium white rectangle, Medium White triangle) Below(Medium white triangle in C, small green rectangle in C) =>Above(small green rectangle, Medium white triangle)Answer: Yes

user : There is a block called A. Inside A, there is a black triangle that is above and to the left of a black circle. There is another block called B which is to the right of A. In B, there are two yellow circles, a blue circle, and a blue square. The blue circle is touching the left edge of B. One of the yellow circles is to the right of the blue circle and to the left of the blue square. The blue square is touching the right edge of B and is above the other yellow circle, Which is also touching the right edge of B. Does block B have all of the circles inside it?

assistant : Contain(A, Black Circle) => ¬Contain(B, Black Circle) ¬Contain(B, Black Circle) => Some x, ¬Contain(B, x) where x is Circle. Answer: No

user : There is a block called A. In A, there is a large blue square near and below a medium blue triangle. Far to the left of the medium blue triangle is a small yellow square. Far above the medium blue triangle is a small blue triangle. There is another block called B, which is to the right of A. In B, There is a small yellow circle touching the left edge of B. Far to the right of the circle is a large yellow square. Near and above the large square is a small yellow square. There is another block called C, which is to the right of B. In C, there is a small blue circle near and to the left of a large yellow triangle. The large yellow triangle is near and to the left of a large blue triangle. The large blue triangle is far above a large blue circle. The large blue circle is touching the bottom edge of C. Are the large blue things in C to the left of the large blue thing in A?

assistant : Question: Left(large blue thing in C, large blue thing in A)Related relation from story: Right(B, A), Right(C, B), Contain(C, large blue triangle), Contain(A, large blue square)Steps: Right(C, B) + Right(B, A) => Right(C, A) Contain(C, large blue triangle) + Right(C, A) + Contain(A, large blue square) => Right(large blue triangle in C, large blue square in A)Right(large blue triangle in C, large blue square in A) => !Left(large blue thing in C, large blue thing in A) Answer: No

user : There is a block called A. In A there is a red triangle Which is above a black circle. The black circle is touching the bottom of A. A yellow square is touching the top edge of A and below that there is a red oval is touching the right edge of A. There is a second block call B Which is to the left of A. Inside B there are 3 squares, a brown one, a green one, and a red one. The green square is above the red square and touching the left edge of B. The brown square is touching both the bottom edge and the right edge of B. The red square is touching the left edge of the brown square. Is the oval to the right of the brown thing?

A.2.4 Chain of Symbol

system : You will be given story and question. Then, you will answer either Yes or No with explanation based on given story step by step. Candidate answer: [Yes, No]

user : There exist a big red square, a big red triangle, a medium red square, and a small green circle in a block called A. The triangle and medium square are touching the bottom edge of the block. The big and medium square are touching the right edge of the block. And the circle is above the big square which is above the medium square. There is another block called B to the left of block A. A medium green square is touching the left edge of block B and is below a medium red square. The medium red square is above and to the left of a small green square. Also a medium red triangle is below and to the left of the small square. Is the medium red triangle below the red square in block B?

assistant : Let analyze the relations given from context to answer question.B: 1. (medium, red, triangle) , < (small, red, square)Based on the information, B: 1. (medium, red, triangle) , < (small, red, square)2. (medium, red, triangle) (small, red, square)Answer: Yes

user : There are three blocks A, B and C. Block A has a medium white triangle and a small red rectangle. There exists a medium white rectangle to the left of the triangle and to the right of the red rectangle. Above block A there is block B. It has a small white oval touching the bottom edge of the block and a medium white triangle to the right of it. Below block A there is block C. Its has a small green triangle touching the top edge of the block. A medium white triangle is below a small green rectangle. Another small green triangle is below and to the right of the white triangle. Are all rectangles above a white triangle?'

assistant : Let consider the related relations to answer the questions,A: 1. (small, red, rectangle) 2. (medium, white, rectangle) C: 1. (small, green, rectangle) = edge 2. (medium, white, triangle)) (small, green, rectangle) C A Based on the information, let determined the relation of all rectangle and white triangles, 1. C A. Therefore, A C 2. A C, A: (small, red, rectangle), and C: (medium, white, triangle). Therefore, A: (small, red, rectangle) C: (medium, white, triangle) 3. A C, A: (medium, white, rectangle), and C: (medium, white, triangle). Therefore, A: (medium, white, rectangle) C: (medium, white, triangle). 4. C: (medium, white, triangle)) (small, green, rectangle). Therefore, (small, green, rectangle) (medium, white, triangle).5. A: (small, red, rectangle) C: (medium, white, triangle), A: (medium, white, rectangle) C: (medium, white, triangle), and C: (small, green, rectangle) (medium, white, triangle), therefore,all rectangle above a white triangle. Answer: Yes

user : There is a block called A. Inside A, there is a black triangle that is above and to the left of a black circle. There is another block called B which is to the right of A. In B, there are two yellow circles, a blue circle, and a blue square. The blue circle is touching the left edge of B. One of the yellow circles is to the right of the blue circle and to the left of the blue square. The blue square is touching the right edge of B and is above the other yellow circle, Which is also touching the right edge of B. Does block B have all of the circles inside it?

assistant : Let consider the related relations to answer the questions,A: 1. (black, triangle) < (black, circle)B: 1. (yellow, circle1) right (blue, circle)2. (blue, square) (yellow, circle2) 3.(blue, circle) =left edge of B Based on the information, let determined the relation of all rectangle and white triangles, 1. A: 1. (black, triangle) < (black, circle). Therefore, A: (black, circle)2. A: (black, circle). Therefore, B doesn't have all circle. Answer: No

user : There is a block called A. In A, there is a large blue square near and below a medium blue triangle. Far to the left of the medium blue triangle is a small yellow square. Far above the medium blue triangle is a small blue triangle. There is another block called B, which is to the right of A. In B, There is a small yellow circle touching the left edge of B. Far to the right of the circle is a large yellow square. Near and above the large square is a small yellow square. There is another block called C, which is to the right of B. In C, there is a small blue circle near and to the left of a large yellow triangle. The large yellow triangle is near and to the left of a large blue triangle. The large blue triangle is far above a large blue circle. The large blue circle is touching the bottom edge of C. Are the large blue things in C to the left of the large blue thing in A?

assistant : Let consider the related relations to answer the questions,B > A,C > BA: 1. (large, blue square) (medium, blue square)C: 1. (large, yellow, triangle)Based on the information, let determined the relation large blue things in C and the large blue thing in A,1. C > B and B > A. Therefore, C > A2. A: 1. (large, blue, square) (medium, blue square). Therefore, A: (large, blue, square)3. C > A, A: (large, blue, square), and C: (large, yellow, triangle).Therefore, C: (large, yellow, triangle) > A: (large, blue, square).4. C: (large, yellow, triangle) > A: (large, blue, square).Therefore, C: (large, yellow, triangle) !< A: (large, blue, square)large blue things in C is not to the left of the large blue thing in A. Answer: No

user : There is a block called A. In A there is a red triangle Which is above a black circle. The black circle is touching the bottom of A. A yellow square is touching the top edge of A and below that there is a red oval is touching the right edge of A. There is a second block call B Which is to the left of A. Inside B there are 3 squares, a brown one, a green one, and a red one. The green square is above the red square and touching the left edge of B. The brown square is touching both the bottom edge and the right edge of B. The red square is touching the left edge of the brown square. Is the oval to the right of the brown thing?

A.2.5 Clarify Sentence Prompting

Raw context : There exist a big red square, a big red triangle, a medium red square, and a small green circle in a block called A. The triangle and medium square are touching the bottom edge of the block. The big and medium square are touching the right edge of the block. And the circle is above the big square which is above the medium square. There is another block called B to the left of block A. A medium green square is touching the left edge of block B and is below a medium red square. The medium red square is above and to the left of a small green square. Also a medium red triangle is below and to the left of the small square.

Spliting context : A big red square in block A.

A big red triangle in block A.

A medium red square in block A.

A small green circle in block A.

The triangle is touching the bottom edge of block A.

The medium square is touching the bottom edge of block A.

The big square is touching the right edge of block A.

The medium square is touching the right edge of block A.

The circle is above the big square in A.

The big square is above the medium square in A.

Block B is to the left of block A.

A medium green square is touching the left edge of block B.

A medium green square is below a medium red square in B.

The medium red square is above a small green square in B.

The medium red square is left of a small green square in B.

The medium red triangle is below the small square in B.

The medium red triangle is left the small square in B.

B Spatial Logical Rules

The conversion from spatial logical rules proposed in (Mirzaee and Kordjamshidi, 2022) to logical constraints used in our experiment is shown in Table 7.

Rule Type	Rule	Constraints in YN	Constraints in FR
Converse	$above(x,y) \Rightarrow below(y,x)$	$q_0 \Rightarrow q_1$	$above(q_0) \Rightarrow below(q_1)$
	$below(x,y) \Rightarrow above(y,x)$		$below(q_0) \Rightarrow above(q_1)$
	$left(x,y) \Rightarrow right(y,x)$		$left(q_0) \Rightarrow right(q_1)$
	$right(x,y) \Rightarrow left(y,x)$		$right(q_0) \Rightarrow left(q_1)$
	$front(x,y) \Rightarrow behind(y,x)$		$front(q_0) \Rightarrow behind(q_1)$
	$behind(x,y) \Rightarrow front(y,x)$		$behind(q_0) \Rightarrow front(q_1)$
	$coveredby(x,y) \Rightarrow cover(y,x)$		$coveredby(q_0) \Rightarrow cover(q_1)$
	$cover(x,y) \Rightarrow coveredby(y,x)$		$cover(q_0) \Rightarrow coveredby(q_1)$
	$inside(x,y) \Rightarrow contain(y,x)$		$inside(q_0) \Rightarrow contain(q_1)$
	$contain(x,y) \Rightarrow inside(y,x)$		$contain(q_0) \Rightarrow inside(q_1)$
Symmetric	$near(x,y) \Rightarrow near(y,x)$	$q_0 \Rightarrow q_1$	$near(q_0) \Rightarrow near(q_1)$
	$far(x,y) \Rightarrow far(y,x)$		$far(q_0) \Rightarrow far(q_1)$
	$touch(x,y) \Rightarrow touch(y,x)$		$touch(q_0) \Rightarrow touch(q_1)$
	$disconnected(x,y) \Rightarrow disconnected(y,x)$		$disconnected(q_0) \Rightarrow disconnected(q_1)$
	$overlap(x,y) \Rightarrow overlap(y,x)$		$overlap(q_0) \Rightarrow overlap(q_1)$
Transitivity	$left(x,y) \wedge left(y,z) \Rightarrow left(x,z)$	$q_0 \wedge q_1 \Rightarrow q_2$	$left(q_0) \wedge left(q_1) \Rightarrow left(q_2)$
	$right(x,y) \wedge right(y,z) \Rightarrow right(x,z)$		$right(q_0) \wedge right(q_1) \Rightarrow right(q_2)$
	$above(x,y) \wedge above(y,z) \Rightarrow above(x,z)$		$above(q_0) \wedge above(q_1) \Rightarrow above(q_2)$
	$below(x,y) \wedge below(y,z) \Rightarrow below(x,z)$		$below(q_0) \wedge below(q_1) \Rightarrow below(q_2)$
	$behind(x,y) \wedge behind(y,z) \Rightarrow behind(x,z)$		$behind(q_0) \wedge behind(q_1) \Rightarrow behind(q_2)$
	$front(x,y) \wedge front(y,z) \Rightarrow front(x,z)$		$front(q_0) \wedge front(q_1) \Rightarrow front(q_2)$
	$inside(x,y) \wedge inside(y,z) \Rightarrow inside(x,z)$		$inside(q_0) \wedge inside(q_1) \Rightarrow inside(q_2)$
	$contain(x,y) \wedge contain(y,z) \Rightarrow contain(x,z)$		$contain(q_0) \wedge contain(q_1) \Rightarrow contain(q_2)$
	$inside(x,y) \wedge coveredby(y,z) \Rightarrow inside(x,z)$		$inside(q_0) \wedge coveredby(q_1) \Rightarrow inside(q_2)$
	$contain(x,y) \wedge cover(y,z) \Rightarrow contain(x,z)$		$contain(q_0) \wedge cover(q_1) \Rightarrow contain(q_2)$
	$inside(x,y) \wedge left(y,z) \Rightarrow left(x,z)$		$inside(q_0) \wedge left(q_1) \Rightarrow left(q_2)$
	$inside(x,y) \wedge right(y,z) \Rightarrow right(x,z)$		$inside(q_0) \wedge right(q_1) \Rightarrow right(q_2)$
	$inside(x,y) \wedge above(y,z) \Rightarrow above(x,z)$		$inside(q_0) \wedge above(q_1) \Rightarrow above(q_2)$
	$inside(x,y) \wedge below(y,z) \Rightarrow below(x,z)$		$inside(q_0) \wedge below(q_1) \Rightarrow below(q_2)$
	$inside(x,y) \wedge behind(y,z) \Rightarrow behind(x,z)$		$inside(q_0) \wedge behind(q_1) \Rightarrow behind(q_2)$
	$inside(x,y) \wedge front(y,z) \Rightarrow front(x,z)$		$inside(q_0) \wedge front(q_1) \Rightarrow front(q_2)$
	$inside(x,y) \wedge near(y,z) \Rightarrow near(x,z)$		$inside(q_0) \wedge near(q_1) \Rightarrow near(q_2)$
	$inside(x,y) \wedge far(y,z) \Rightarrow far(x,z)$		$inside(q_0) \wedge far(q_1) \Rightarrow far(q_2)$
	$inside(x,y) \wedge disconnected(y,z) \Rightarrow disconnected(x,z)$		$inside(q_0) \wedge disconnected(q_1) \Rightarrow disconnected(q_2)$
	$coveredby(x,y) \wedge left(y,z) \Rightarrow left(x,z)$		$coveredby(q_0) \wedge left(q_1) \Rightarrow left(q_2)$
	$coveredby(x,y) \wedge right(y,z) \Rightarrow right(x,z)$		$coveredby(q_0) \wedge right(q_1) \Rightarrow right(q_2)$
	$coveredby(x,y) \wedge above(y,z) \Rightarrow above(x,z)$		$coveredby(q_0) \wedge above(q_1) \Rightarrow above(q_2)$
	$coveredby(x,y) \wedge below(y,z) \Rightarrow below(x,z)$		$coveredby(q_0) \wedge below(q_1) \Rightarrow below(q_2)$
	$coveredby(x,y) \wedge behind(y,z) \Rightarrow behind(x,z)$		$coveredby(q_0) \wedge behind(q_1) \Rightarrow behind(q_2)$
	$coveredby(x,y) \wedge front(y,z) \Rightarrow front(x,z)$		$coveredby(q_0) \wedge front(q_1) \Rightarrow front(q_2)$
	$coveredby(x,y) \wedge near(y,z) \Rightarrow near(x,z)$		$coveredby(q_0) \wedge near(q_1) \Rightarrow near(q_2)$
	$coveredby(x,y) \wedge far(y,z) \Rightarrow far(x,z)$		$coveredby(q_0) \wedge far(q_1) \Rightarrow far(q_2)$
	$coveredby(x,y) \wedge disconnected(y,z) \Rightarrow disconnected(x,z)$		$coveredby(q_0) \wedge disconnected(q_1) \Rightarrow disconnected(q_2)$
Transitivity + Topological	$inside(x,y) \wedge inside(h,z) \wedge left(y,z) \Rightarrow left(x,h)$	$q_0 \wedge q_1 \wedge q_2 \Rightarrow q_3$	$inside(q_0) \wedge inside(q_1) \wedge left(q_2) \Rightarrow left(q_3)$
	$inside(x,y) \wedge inside(h,z) \wedge right(y,z) \Rightarrow right(x,h)$		$inside(q_0) \wedge inside(q_1) \wedge right(q_2) \Rightarrow right(q_3)$
	$inside(x,y) \wedge inside(h,z) \wedge above(y,z) \Rightarrow above(x,h)$		$inside(q_0) \wedge inside(q_1) \wedge above(q_2) \Rightarrow above(q_3)$
	$inside(x,y) \wedge inside(h,z) \wedge below(y,z) \Rightarrow below(x,h)$		$inside(q_0) \wedge inside(q_1) \wedge below(q_2) \Rightarrow below(q_3)$
	$inside(x,y) \wedge inside(h,z) \wedge behind(y,z) \Rightarrow behind(x,h)$		$inside(q_0) \wedge inside(q_1) \wedge behind(q_2) \Rightarrow behind(q_3)$
	$inside(x,y) \wedge inside(h,z) \wedge front(y,z) \Rightarrow front(x,h)$		$inside(q_0) \wedge inside(q_1) \wedge front(q_2) \Rightarrow front(q_3)$
	$inside(x,y) \wedge inside(h,z) \wedge near(y,z) \Rightarrow near(x,h)$		$inside(q_0) \wedge inside(q_1) \wedge near(q_2) \Rightarrow near(q_3)$
	$inside(x,y) \wedge inside(h,z) \wedge far(y,z) \Rightarrow far(x,h)$		$inside(q_0) \wedge inside(q_1) \wedge far(q_2) \Rightarrow far(q_3)$
	$inside(x,y) \wedge inside(h,z) \wedge disconnected(y,z) \Rightarrow disconnected(x,h)$		$inside(q_0) \wedge inside(q_1) \wedge disconnected(q_2) \Rightarrow disconnected(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge left(y,z) \Rightarrow left(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge left(q_2) \Rightarrow left(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge right(y,z) \Rightarrow right(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge right(q_2) \Rightarrow right(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge above(y,z) \Rightarrow above(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge above(q_2) \Rightarrow above(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge below(y,z) \Rightarrow below(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge below(q_2) \Rightarrow below(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge behind(y,z) \Rightarrow behind(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge behind(q_2) \Rightarrow behind(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge front(y,z) \Rightarrow front(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge front(q_2) \Rightarrow front(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge near(y,z) \Rightarrow near(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge near(q_2) \Rightarrow near(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge far(y,z) \Rightarrow far(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge far(q_2) \Rightarrow far(q_3)$
	$coveredby(x,y) \wedge coveredby(h,z) \wedge disconnected(y,z) \Rightarrow disconnected(x,h)$		$coveredby(q_0) \wedge coveredby(q_1) \wedge disconnected(q_2) \Rightarrow disconnected(q_3)$
	$left(x,y) \wedge left(h,z) \wedge contain(y,z) \Rightarrow left(x,h)$		$left(q_0) \wedge left(q_1) \wedge contain(q_2) \Rightarrow left(q_3)$
	$left(x,y) \wedge left(h,z) \wedge cover(y,z) \Rightarrow left(x,h)$		$left(q_0) \wedge left(q_1) \wedge cover(q_2) \Rightarrow left(q_3)$
	$right(x,y) \wedge right(h,z) \wedge contain(y,z) \Rightarrow right(x,h)$		$right(q_0) \wedge right(q_1) \wedge contain(q_2) \Rightarrow right(q_3)$
	$right(x,y) \wedge right(h,z) \wedge cover(y,z) \Rightarrow right(x,h)$		$right(q_0) \wedge right(q_1) \wedge cover(q_2) \Rightarrow right(q_3)$
	$above(x,y) \wedge above(h,z) \wedge contain(y,z) \Rightarrow above(x,h)$		$above(q_0) \wedge above(q_1) \wedge contain(q_2) \Rightarrow above(q_3)$
	$above(x,y) \wedge above(h,z) \wedge cover(y,z) \Rightarrow above(x,h)$		$above(q_0) \wedge above(q_1) \wedge cover(q_2) \Rightarrow above(q_3)$
	$below(x,y) \wedge below(h,z) \wedge contain(y,z) \Rightarrow below(x,h)$		$below(q_0) \wedge below(q_1) \wedge contain(q_2) \Rightarrow below(q_3)$
	$below(x,y) \wedge below(h,z) \wedge cover(y,z) \Rightarrow below(x,h)$		$below(q_0) \wedge below(q_1) \wedge cover(q_2) \Rightarrow below(q_3)$
	$behind(x,y) \wedge behind(h,z) \wedge contain(y,z) \Rightarrow behind(x,h)$		$behind(q_0) \wedge behind(q_1) \wedge contain(q_2) \Rightarrow behind(q_3)$
	$behind(x,y) \wedge behind(h,z) \wedge cover(y,z) \Rightarrow behind(x,h)$		$behind(q_0) \wedge behind(q_1) \wedge cover(q_2) \Rightarrow behind(q_3)$
	$front(x,y) \wedge front(h,z) \wedge contain(y,z) \Rightarrow front(x,h)$		$front(q_0) \wedge front(q_1) \wedge contain(q_2) \Rightarrow front(q_3)$
	$front(x,y) \wedge front(h,z) \wedge cover(y,z) \Rightarrow front(x,h)$		$front(q_0) \wedge front(q_1) \wedge cover(q_2) \Rightarrow front(q_3)$
	$near(x,y) \wedge near(h,z) \wedge contain(y,z) \Rightarrow near(x,h)$		$near(q_0) \wedge near(q_1) \wedge contain(q_2) \Rightarrow near(q_3)$
	$near(x,y) \wedge near(h,z) \wedge cover(y,z) \Rightarrow near(x,h)$		$near(q_0) \wedge near(q_1) \wedge cover(q_2) \Rightarrow near(q_3)$
	$far(x,y) \wedge far(h,z) \wedge contain(y,z) \Rightarrow far(x,h)$		$far(q_0) \wedge far(q_1) \wedge contain(q_2) \Rightarrow far(q_3)$
	$far(x,y) \wedge far(h,z) \wedge cover(y,z) \Rightarrow far(x,h)$		$far(q_0) \wedge far(q_1) \wedge cover(q_2) \Rightarrow far(q_3)$
	$disconnected(x,y) \wedge disconnected(h,z) \wedge contain(y,z) \Rightarrow disconnected(x,h)$		$disconnected(q_0) \wedge disconnected(q_1) \wedge contain(q_2) \Rightarrow disconnected(q_3)$
	$disconnected(x,y) \wedge disconnected(h,z) \wedge cover(y,z) \Rightarrow disconnected(x,h)$		$disconnected(q_0) \wedge disconnected(q_1) \wedge cover(q_2) \Rightarrow disconnected(q_3)$

Table 7: The conversion from spatial logical rules proposed in (Mirzaee and Kordjamshidi, 2022) to logical constraints used in our experiment.