CCL25-Eval任务1系统报告: 基于上下文学习与格式化约束的空间语义理解

郑奕扬
上海大学机电工程与自动化学院
query_zyy@outlook.com

摘要

本系统报告详细介绍了我们团队参加第五届中文空间语义理解评测（SpaCE2025）的方法和成果。SpaCE2025旨在评估大语言模型在空间语义理解和空间推理能力上的表现，涵盖空间信息正误判断、空间异形同义判断、空间参照实体判断、中文空间方位关系推理和英文空间方位关系推理五个子任务。针对不同任务，我们采用基于上下文的有监督微调和格式化约束的逻辑推理框架，结合LoRA高效微调Qwen2.5-7B-Instruct和DeepSeek-R1-Distill-Qwen-7B模型，设计了约束提取、排列遍历和求解器求解的推理流程。在测试集上，我们在信息正误判断、异形同义判断、参照实体判断、中文方位推理、英文方位推理分别取得0.6454、0.7082、0.7720、0.6254、0.5997的准确率，综合排名第二。

关键词： 空间语义理解；大语言模型

System Report for CCL25-Eval Task 1: Spatial Semantic Understanding Based on In-Context Learning and Structured Constraints

Yiyang Zheng
Shanghai University School of Mechanical Engineering and Automation
query_zyy@outlook.com

Abstract

This system report details our team's methods and results in the Fifth Chinese Spatial Cognition Evaluation (SpaCE2025). SpaCE2025 aims to assess large language models' capabilities in spatial semantic understanding and reasoning, covering five subtasks: judging spatial information, retrieving spatial referents, recognizing synonymous expression, and spatial position reasoning in Chinese and English. We employed supervised fine-tuning based on in-context learning and a structured logical reasoning framework, leveraging LoRA-efficient fine-tuning of Qwen2.5-7B-Instruct and DeepSeek-R1-Distill-Qwen-7B models, with a reasoning pipeline involving constraint extraction, permutation traversal, and solver-based resolution. On the test set, we achieved accuracies of 0.6454, 0.7082, 0.7720, 0.6254, and 0.5997 for judging spatial information, recognizing synonymous expression, retrieving spatial referents, Chinese spatial position reasoning, and English spatial position reasoning, respectively, securing an overall second-place ranking.

Keywords: Spatial semantic understanding , Large Language Model

第二十四届中国计算语言学大会论文集，第1页-第12页，济南，中国，2025年8月11日至14日。

1 引言

空间语义理解是自然语言处理（NLP）中的一个重要且具有挑战性的任务。空间表达描述了物体之间的空间方位关系，是自然语言中的高频现象。实现空间语义理解不仅依赖语言知识，还需要调用空间认知能力，准确构建文本表征的空间场景。第五届中文空间语义理解评测（SpaCE2025）旨在测试LLM 在空间语义和空间推理方面的能力，涵盖了从判断空间信息正确性到解决复杂排列问题的多个任务。SpaCE2025 相较于前几届（如SpaCE2024）进行了优化，去除了依赖简单语言标记的子任务，增加了需要更深层次认知处理的子任务，并引入了中英文双语推理子任务，以探索语言依赖性。

针对不同的子任务，我们采用了区别化的解决策略，包括基于上下文的微调,以及对复杂推理任务的统一的处理流程：约束提取、排列遍历和求解器求解，最终，我们在测试集的信息正误判断、异形同义判断、参照实体判断、中文方位推理、英文方位推理5个子任务上分别取得了0.6454、0.7082、0.7720、0.6254、0.5997的准确率，总结果排名第二。

2 相关研究

早期关于空间语义理解的评测工作主要集中在英语语境下展开。Pustejovsky等人 (2015) 进一步提出了ISO-Space标注体系。该体系细化了空间关系的表达方式，使得空间语义角色标注在精度和覆盖范围上均有显著提升。

自2021年起，SpaCE评测不断引入更具挑战性的任务以检验机器的中文空间语义理解。对于数据集本身，Yue等人 (2023) 针对SpaCE2021数据集展开了质量评估，从句子特征、可替换词结构类型、空间方位类型及样本正误分布等多个方面分析其特性，结果显示高质量的数据设计能够显著提升评测的可靠性。

在方法层面，Zhao和Wei (2023) 针对中文空间语义理解提出了基于思维链的整体空间推理方法，利用大语言模型应对多义性和同义词替换等挑战，在SpaCE数据集上展现出优于常规提示学习、接近监督模型的性能。Su和Zhan (2019) 则通过基于语料库的近义词驱动方法研究了中文词汇语义，为空间语义的细致分析奠定了理论基础。

随着数据集和建模方法的不断进步，空间语义理解的研究正逐步向更为复杂的空间推理任务拓展。这要求模型不仅具备扎实的语言知识，还要能够整合空间认知能力，对隐含的空间场景进行建模与推理。

针对模型的空间推理能力，Yamada等人 (2023) 在自然语言导航任务中测试了GPT-3.5、GPT-4与Llama2等模型在不同空间结构（如方形、六边形、三角形网格、环形与树形结构）中的表现。结果表明模型在简单结构下表现良好，但在复杂结构上性能波动较大，错误涉及空间与非空间因素。Greatrix (2024) 则进一步探索了模型在未见空间推理问题上的能力，发现Claude 3等大模型具备在新颖空间推理任务上展现出复杂推理能力。

与此同时，空间语义理解能力在跨语言、跨结构场景下也受到关注。Hu等人 (2024) 将跨语言推理过程细分为知识检索和无知识推理两阶段，分析其在中英之间的可迁移性，发现无知识推理几乎完全可迁移，而知识检索能力则因语言特异性受到限制。Hershcovich等人 (2022) 指出，跨语言推理还需关注文化共同知识的差异，这对空间表达的理解产生影响。

3 数据集

第五届空间语义理解评测（SpaCE2025）数据集涵盖两大任务类别（空间语言能力和空间推理能力），共五个子任务，包括空间信息正误判断、空间异形同义判断、空间参照实体判断、中文空间方位关系推理和英文空间方位关系推理。数据来源于报刊、文学作品、中小学课本、交通事故、人体动作、地理百科等真实语料的改写，以及基于知识库合成的空间推理数据，旨在考察大语言模型在空间语义理解和推理能力上的表现。表1展示了数据集分布，总数据量为18,423题，包含示例集、训练集、验证集和测试集，每条数据包括题目编号、文本、选项（推理任务）或答案（判断任务），评测形式包括判断题和选择题（选项固定为4个）。在数据分布上，空间方位关系推理题目占比最高（中英文各6,000题），空间异形同义判断题目最少（1,120题）。相较SpaCE2024，SpaCE2025舍弃依赖语言形式标记的任务，新增跨语言推理任

务，并提升数据多样性和平衡性。数据集的多样性和任务的认知复杂度为评测带来挑战，旨在更全面评估模型的空间认知能力。

序号	任务	任务要求	示例集	训练集	验证集	测试集	数据总量
1	空间信息正误判断	判断文本空间信息正误，正确文本可构造合乎常理的空间场景，错误文本则不能。回答“正确”或“错误”。	20	0	0	3,500	3,520
2	空间异形同义判断	判断两个空间表达描述的空间场景是否相同。回答“相同”或“不同”。	20	0	0	1,100	1,120
3	空间参照实体判断	确定空间方位依赖参照物，判断给定实体是否为正确参照物。回答“正确”或“错误”。	20	0	0	1,763	1,783
4	中文空间方位关系推理	根据情景与方位关系条件，推理未知的中文空间关系。选择题，有单选和多选。	0	2,000	500	3,500	6,000
5	英文空间方位关系推理	根据情景与方位关系条件，推理未知的英文空间关系。与中文推理为中英对照。选择题，有单选和多选。	0	2,000	500	3,500	6,000
合计			**60**	**4,000**	**1,000**	**13,363**	**18,423**

Table 1: SpaCE2025 评测任务及数据集统计概览

4 基于上下文学习的有监督微调

针对任务1——3，“任务信息正误判断”、“异形同义判断”和“参照实体判断”三类任务，我们主要采用了基于上下文的有监督微调方法。随着模型规模和训练数据增长，大型语言模型（LLMs）在上下文理解和推理能力上表现出色，能够通过少量上下文示例完成复杂任务。然而，仅依赖预训练和基本的上下文学习，模型的能力尚未得到充分发挥。为了进一步提升模型对上下文信息的理解和指令执行能力，我们采用**有监督指令微调的上下文学习**（Supervised Instruction Fine-tuning with In-Context Learning, ICL），具体流程如下所示。

4.1 基于上下文的学习

针对具体下游任务，设有数据集$D = \{(x_i, Y_i)\}_{i=1}^{n}$，其中$x_i$为输入文本，$Y_i$为候选答案集合$\{y_1^{(i)}, y_2^{(i)}, \ldots, y_m^{(i)}\}$。我们首先随机采样同类型的$K$条示例$\{(x_j, y_j)\}_{j=1}^{K}$，将其作为上下文示例嵌入到提示词（prompt）中，为每个测试样本x构建如下输入：

$$C = \{I,\, s(x_1, y_1),\, s(x_2, y_2),\, \ldots,\, s(x_k, y_k)\} \tag{1}$$

其中I为任务指令，$s(x_j, y_j)$为自然语言形式的示例，最终的模型输入为(C, x)。

4.2 有监督微调

为进一步提升模型对上下文及任务指令的理解能力，我们对模型进行了有监督微调。针对训练样本(x, Y, y_{gt})，其中y_{gt}为正确答案，采用交叉熵损失进行优化：

$$\mathcal{L} = -\log P(y_{gt} \mid x, C) \tag{2}$$

考虑到全参数微调的资源消耗巨大，我们采用了**LoRA**（Low-Rank Adaptation）进行参数高效微调。LoRA通过引入可训练的低秩矩阵(A, B)，仅对部分层的权重进行如下更新，有效减少了所需微调参数量：

$$W' = W + \Delta W = W + BA \tag{3}$$

其中W为原有模型权重，A, B为低秩矩阵，满足$rank(BA) \ll rank(W)$。

4.3 标签归一化与概率计算方法

为统一模型训练与推理流程，我们对原始标签进行了归一化处理：将**正确/相同/无异常**标签统一映射为0，将**错误/不同/有异常**标签映射为1，并以0或1作为有监督微调的输出目标。

在推理阶段，假设模型输出的token序列包含字符'0'与'1'对应的token，分别记为token_0和token_1。我们从模型输出的logits中提取这两个token的对数概率（$\log p_0$和$\log p_1$），据此计算概率。具体过程如下：

设logits中包含的token集合为S，对于$i \in \{0, 1\}$，令token_i分别表示字符'0'和'1'的token，对应的对数概率为$\log p_i$，则有：

$$p_i = \begin{cases} \exp(\log p_i), & \text{若} \text{token}_i \in S \\ 0, & \text{若} \text{token}_i \notin S \end{cases} \tag{4}$$

为保证概率归一化，进一步计算：

$$T = p_0 + p_1 \tag{5}$$

$$\widetilde{p}_1 = \begin{cases} \dfrac{p_1}{T}, & T > 0 \\ 0, & T = 0 \end{cases} \tag{6}$$

其中$\widetilde{p}_1$即为推理阶段"错误/不同/有异常"标签（1）对应的概率，用于最终决策或评测分析。

4.4 任务指令设计与提示词模版

为充分发挥模型的上下文学习能力，我们设计了明确的任务指令和提示词（prompt）模版,并按上述流程微调**Qwen2.5-7B-Instruct**(2024)，具体如下：

空间信息正误判断提示词

请分析下列内容空间信息是否正确，请先阅读示例，随后逐步分析，最后将结果标注于□中，$\boxed{1}$代表有异常，$\boxed{0}$代表无异常。
例1: 碰撞发生后，"VANMANILA"轮船长迅速...在后面看见"XIANGZHOU"轮甲板以下已没入水上。分析：【没入水上】。"没入"指在水面之下，即水中，与"水上"矛盾。故$\boxed{1}$。

<其余3个例子在此省略>

你的任务：

<task>

例题分析只是简略版概括，以展示常见的问题。请先简要研读例题，理解任务核心，随后进行分析。

空间异形同义判断提示词

请详细推理，判断text1 和text2 描述的空间场景是否相同。请注意，中文中可能会用不同的空间词汇表达相同的意思，请根据常理判断两情景是否一致。最后将结果标注于□中，$\boxed{0}$代表一致，$\boxed{1}$代表不一致。
以下为例题，请在回答时参考例题

<例子在此省略>

以下是你的任务

`text1: <text1>`
`text2: <text2>`

再次强调，中文中可能会用不同的空间词汇表达相同的意思，请根据常理判断两情景是否一致。判断结果:

空间参照实体判断提示词

请详细推理，判断句中指代是否正确，最后将结果标注于□中，$\boxed{1}$代表错误，$\boxed{0}$代表正确。
以下为例题，请在回答时参考例题

<例子在此省略>

以下是你的任务

`<text>`
`<interpretation>`

判断结果:

5 基于格式化约束的逻辑推理

我们针对中文空间方位关系推理（任务4）及其英文对应任务（任务5），提出了一套处理与建模框架，涉及空间排列与约束条件的综合推理。整体流程包括约束提取、排列遍历与求解器推理三个核心步骤，涵盖环形、架子和方桌三类典型子问题。所有训练环节采用基于**LoRA**的有监督指令微调DeepSeek-R1-Distill-Qwen-7B(2025)。各环节输入输出与模型算法如表2所示。

流程	模型/算法	输入	输出
约束提取	解析器$\mathcal{P}$	text	constraint
排列遍历	遍历程序	constraint	solution(s)
求解器	$\mathcal{S}_k$	(text, question, options, solution)	answer
备用求解器	$\mathcal{S}_{\text{aux}}$	(text, question, options)	answer

Table 2: 各环节输入输出与模型/算法

5.1 解析器与求解器的训练

训练阶段主要包含如下步骤:

(1) **约束提取**：采用大模型（o3-mini-high）结合提示词工程，从原始题干文本自动抽取结构化约束，记为$o3-mini-high$: text $\rightarrow$ constraint。

(2) **约束验证**：利用遍历程序检验约束合法性，仅保留唯一解的样本作为训练集。

(3) **解析器训练**：以题干text为输入，标准格式化约束constraint为输出进行解析器$\mathcal{P}$微调。

(4) **求解器训练**：针对每类子问题，将(text, question, options, solution)作为输入，answer为输出微调$\mathcal{S}_k$，即

$$\mathcal{S}_k : (\text{text, question, options, solution}) \rightarrow \text{answer} \tag{7}$$

(5) **备用求解器训练**：用于约束提取失败场景，仅以(text, question, options)为输入，answer为输出得到$\mathcal{S}_{\text{aux}}$:

$$\mathcal{S}_{\text{aux}} : (\text{text, question, options}) \rightarrow \text{answer} \tag{8}$$

5.2 解析器与求解器的推理

推理流程如下:

(1) **约束抽取**：用解析器$\mathcal{P}$对测试样本抽取结构化约束。

(2) **排列验证**：遍历约束，判断是否有唯一合法解:

若唯一解时，将题目信息及解组合输入$\mathcal{S}_k$预测答案。

若多解/无解或约束异常时，直接用$\mathcal{S}_{\text{aux}}$以题目信息推理答案。

(3) **结果整合**：最终合并标准求解与备用求解各自输出，获得整体推理结果。

5.3 格式化约束定义

本节我们给出了对于架子约束、环形约束与方桌约束三种问题的约束定义。解析器，遍历程序与求解器均采用这组约束定义。

架子约束

$$Q = (E, G, \mathcal{R}_{\text{rel}}, \mathcal{R}_{\text{abs}})$$

其中:

- $E = \{e_1, e_2, \ldots, e_n\}$：实体集合，包含$n$ 个实体;
- $G = (n_x, n_y)$：网格结构，n_x 表示每行实体数，n_y 表示行数;
- $\mathcal{R}_{\text{rel}} = \left\{ (e_i, e_j, (r_x, t_x), (r_y, t_y)) \;\middle|\; \begin{array}{l} e_i, e_j \in E,\ r_x, r_y \in \mathbb{Z} \cup \{\emptyset\}, \\ t_x, t_y \in \{0,1,2,3,4\} \end{array} \right\}$

  ：相对位置约束集合。其中每个元素$(e_i, e_j, (r_x, t_x), (r_y, t_y))$ 表示两个实体间在x 轴和y 轴上的约束，含义如下:

  - r_x, r_y：e_i 相对于e_j 在x 轴和y 轴上的位置数值（如无具体数值则记为$\emptyset$）;
  - $t_x, t_y \in \{0,1,2,3,4\}$：位置关系类型，意义如下:
    - 0：无约束（即对应的r_x 或r_y 为$\emptyset$）;
    - 1：等于指定数值（即$e_i - e_j = r_x$ 或$e_i - e_j = r_y$）;
    - 2：距离为指定绝对值，方向未知（即$|e_i - e_j| = |r_x|$ 或$|e_i - e_j| = |r_y|$）;
    - 3：大于指定数值（即$e_i - e_j > r_x$ 或$e_i - e_j > r_y$）;
    - 4：小于指定数值（即$e_i - e_j < r_x$ 或$e_i - e_j < r_y$）。
- $\mathcal{R}_{\text{abs}} = \{(e_i, p_x, p_y) \mid e_i \in E, p_x, p_y \in \mathbb{Z} \cup \{\emptyset\}\}$：绝对位置约束集合。每一约束$(e_i, p_x, p_y)$ 表示实体e_i 在x 轴与y 轴上的绝对位置（若未指定则为$\emptyset$）。

第二十四届中国计算语言学大会论文集，第1页-第12页，济南，中国，2025年8月11日至14日。

环形约束

$$Q = (E, \mathcal{R}_{rel})$$

其中：

- $E = \{e_1, e_2, \ldots, e_n\}$：实体集合，包含$n$ 个实体。
- $\mathcal{R}_{rel} = \{(e_i, e_j, r) \mid e_i, e_j \in E, r \in \mathbb{Z}\}$：相对位置约束集合，每个约束$(e_i, e_j, r)$ 表示：
  - e_i, e_j：两个实体。
  - $r \in \mathbb{Z}$：e_i 相对于e_j 的相对位置，顺时针为正（$r > 0$），逆时针为负（$r < 0$）。

方桌约束

$$Q = (E, \mathcal{R}_{rel}, \mathcal{R}_{abs}, D)$$

其中：

- $E = \{e_1, e_2, \ldots, e_n\}$：实体集合，包含$n$ 个实体。
- $\mathcal{R}_{rel} = \{(e_i, e_j, t) \mid e_i, e_j \in E, t \in \{1, 2, 3, 4\}\}$：相对位置约束集合，每个约束$(e_i, e_j, t)$ 表示：
  - e_i, e_j：两个实体。
  - $t \in \{1, 2, 3, 4\}$：约束种类，分别表示：
    - 1：e_i 在e_j 的左边；
    - 2：e_i 在e_j 的右边；
    - 3：e_i 在e_j 的对面；
    - 4：e_i 和e_j 在同侧，但左右关系无法确定。
- $\mathcal{R}_{abs} = \{(e_i, d_i) \mid e_i \in E, d_i \in \{N, S, W, E\}\}$：绝对位置约束集合，每个约束$(e_i, d_i)$ 表示：
  - e_i：实体。
  - $d_i \in \{N, S, W, E\}$：实体e_i 的朝向（北、南、西、东）。
- $D \in \{\text{true}, \text{false}\}$：布尔值，表示题目是否涉及朝向信息。

6 实验结果与分析

本节报告我们的系统在SpaCE2025测试集各任务的实验表现与分析。

6.1 总体任务表现与能力分项得分

我们首先对SpaCE2025全体任务进行归纳评估，并与基线模型（仅用提示词的**DeepSeek-R1-Distill-Qwen-7B**）进行对比。我们将其拆分为**语言能力**（信息正误、异形同义与参照实体）与**推理能力**两大能力维度。所有下列数据均基于准确率（**Accuracy, Acc**）指标。具体各子任务的准确率如下表所示。

模型	语言能力Acc			推理能力Acc	
	信息正误	异形同义	参照实体	中文推理	英文推理
基线	0.6274	0.5809	0.6375	0.2266	0.2977
本系统	**0.6454**	**0.7082**	**0.7720**	**0.6254**	**0.5997**

Table 3: 子任务测试集准确率对比

从表3中可以看出，我们的系统在所有子任务上均超越了基线表现，其中推理任务的提升尤为显著。以“中文推理”子任务为例，准确率从0.2266提升至0.6254，“英文推理”从0.2977提升至0.5997，这种大幅提升主要归功于模型在格式化约束策略的应用。语言能力类任务同样获得了明显提升，显示了我们方法的全面有效性。

6.2 推理路径成分分析

为了验证我们提出的格式化约束推理框架中各个模块的有效性，我们针对推理任务（任务4和5）设计了消融实验，结果如表4所示。实验配置包括：

- **本系统（完整）**：采用“解析器-求解器”主路径与“备用求解器”路径结合的完整框架。
- **本系统(除备用求解器)**：移除备用求解器。当主路径（解析、遍历）失败时，系统无法作答，计为错误。
- **本系统(仅备用求解器)**：不使用格式化约束框架，所有题目直接由经过微调的备用求解器模型进行端到端推理。
- **基线（仅提示词）**：使用基础大模型进行零样本推理，不进行任何微调。

模型配置	中文推理Acc	英文推理Acc	推理能力平均Acc
本系统（完整）	**0.6254**	**0.5997**	**0.6126**
本系统(除备用求解器)	0.5047	0.4686	0.4866
本系统(仅备用求解器)	0.3931	0.3934	0.3932
基线（仅提示词）	0.2266	0.2977	0.2622

Table 4: 推理任务消融实验结果

消融实验结果清晰地揭示了各模块的贡献：

(1) **格式化约束框架的核心作用**：对比“本系统（完整）”与“仅备用求解器”，准确率从0.3932大幅提升至0.6126。这表明，虽然端到端微调（仅备用求解器）已能超越基线，但我们设计的“约束提取-排列遍历-求解器”的逻辑推理框架是性能提升的核心驱动力。它将复杂的自然语言问题分解为机器可处理的结构化任务，极大地提升了推理的准确性。

(2) **备用求解器的关键容错能力**：对比“本系统（完整）”与“本系统（除备用求解器)”，移除备用求解器后，平均准确率从0.6126骤降至0.4866。这证明了备用求解器作为一种关键的容错机制，有效处理了主路径因解析失败或约束歧义而无法求解的情况，显著增强了系统的鲁棒性和整体性能。

格式化约束主路径与备用求解器路径的结合，是我们的系统在推理任务上取得优异表现的关键。

6.3 推理任务子类型分布与难度

针对任务4与任务5（空间推理），我们进一步将测试集分为“架子约束”、“环形约束”、“方桌约束”三大子类型，统计各自准确率，并对比解析器性能。

子类型	总体正确率	解析器Acc	解析器/求解器路径Acc
架子约束	0.6884	0.5278	0.6280
环形约束	0.5328	0.8008	0.5995
方桌约束	**0.8360**	**1.0000**	**0.8360**

Table 5: 空间推理各子类型准确率及解析器表现

从表5可以看出，模型在三类约束问题上的表现存在显著差异。方桌约束的总体正确率最高（0.8360），且解析器准确率达到100%，表明模型对此类规范化问题的理解与推理能力极强。

相比之下，环形约束的整体表现最弱（0.5328），尽管其解析器准确率较高（0.8008），但后续求解环节表现不佳，这可能源于环形排列固有的对称性与歧义性，导致求解困难。这揭示了模型在处理高歧义性空间推理时的瓶颈。

6.4 推理路径影响分析

为了更深入地理解主路径和备用路径的协同作用，我们统计了两条路径在各子类型下的准确率。

子类型	解析器/求解器路径Acc	备用求解器路径Acc
架子约束	0.6280	**0.7560**
环形约束	**0.5995**	0.2650
方桌约束	**0.8360**	/

Table 6: 不同推理路径下各子类型准确率对比

从上述结果可以看出，解析器/求解器路径通常是主要的推理通路。对于大多数子类型，该路径准确率较高，但在架子约束问题上，备用求解器路径的准确率（0.7560）反而超过了解析器/求解器路径（0.6280），显示出在解析失败时备用路径能够对性能提供较好的补偿。相较之下，环形约束问题的备用路径表现较弱，仅为0.2650，远低于解析器/求解器路径，说明该类型题目的解析器和求解器联动可有效提升解题效果。值得注意的是，方桌约束全部通过主路径解决，备用路径未触发，显示出模型在此类型场景下极高的稳定性。总体来看，解析器的性能提升对于推理效果至关重要，而备用路径的补偿能力在个别子类型下表现突出，未来可进一步关注模型推理鲁棒性的整体提升。

6.5 典型错误案例分析

尽管本系统在各项任务中取得了较好的性能，但在处理部分复杂或模糊的空间语言现象时仍存在局限性。通过对测试集中的错误案例进行分析，我们识别出两类主要的失败模式。

6.5.1 复杂位置表述导致的解析失败

部分案例中的空间关系描述涉及多重参照或视角转换，超出了当前解析器的处理能力。例如，在“环形约束”子任务的`zh-test-1280`中，存在如下约束：

(3)赵志敬背对的位置是王处一右边第二个位置；

该表述要求模型进行两步推理：首先，确定“王处一”的“背对位置”（即其对面）；然后，以此位置为新参照物，找到其“右边第二个位置”作为“赵志敬”的位置。我们的解析器在处理这种嵌套的、动态变化的参照系时出现困难，未能正确分解推理步骤，导致约束提取错误，进而影响后续的排列求解。

6.5.2 自然语言模糊性引发的约束混淆

自然语言中的描述在精确度上存在差异，模型未能准确捕捉这些细微差别，从而导致约束定义不当。例如，在“架子约束”子任务的`zh-test-2779`中，题干包含两个看似相似但强度不同的约束：

(3)茉莉在月季正下方且二者不隔层；
(4)牡丹在月季正下方；

在此案例中，“正下方且二者不隔层”是一个强约束，明确了实体在Y轴上坐标相差为1。而“正下方”则是一个弱约束，仅表示Y轴坐标更小，但距离不确定（可能隔层）。我们的系统在约束提取时，未能区分这两种表述在约束强度上的差异，将其泛化为相同的相对位置关系。这种混淆导致在排列遍历时引入了错误的假设，最终推理失败。这暴露了当前格式化约束定义在粒度上的不足以及解析器在处理语义模糊性上的挑战。

7 结语

在第五届中文空间语义理解评测（SpaCE2025）中，我们的系统展示了在空间语言理解与推理任务中的稳健表现，在测试集上，模型在信息正误判断、异形同义判断、参照实体判断、中文方位推理和英文方位推理任务中的准确率分别为0.6454、0.7082、0.7720、0.6254和0.5997，总排名第二。实验结果表明，基于上下文的有监督微调方法有效提升了模型在语言能力任务中的表现，特别是在参照实体判断任务中表现突出；而基于格式化约束的逻辑推理框架则在处理复杂空间推理任务时展现了较高的鲁棒性，尤其在方桌和架子约束子任务中准确率分别达到了0.8360与0.6884。同时，环形约束任务的较低准确率（约0.53）暴露了系统在处理循环排列歧义和多解场景时的局限性。

未来，为进一步提升模型的空间语义理解能力，我们计划从以下方向优化：首先，针对环形约束等高歧义场景，改进约束提取的精准性和解析器的鲁棒性，探索更精细的提示词设计或引入外部知识库以增强模型对复杂空间结构的理解；其次，优化备用求解器的推理能力，提升其在约束提取失败场景下的容错表现；最后，结合跨语言数据增强，强化模型在中英文推理任务中的泛化能力。这些改进将为构建更智能、更通用的空间语义理解系统奠定基础。

参考文献

DeepSeek-AI. 2025. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* arXiv:2501.12948.

Mark Greatrix. 2024. *Can Large Language Models Create New Knowledge for Spatial Reasoning Tasks?* arXiv:2405.14379.

Daniel Hershcovich et al. 2022. *Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings.* arXiv:2309.08591.

Yifan Hu et al. 2024. *Large Language Models Are Cross-Lingual Knowledge-Free Reasoners.* arXiv:2406.16655.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. *SemEval-2015 Task 8: SpaceEval. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894.

Qi Su and Weimin Zhan. 2019. *From Minimal Contrast to Meaning Construct: Corpus-based, Near Synonym Driven Approaches to Chinese Lexical Semantics.* Springer.

Yuki Yamada, Yutong Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. 2023. *Evaluating Spatial Understanding of Large Language Models.* arXiv:2310.14540.

An Yang, Baosong Yang, et al. 2024. *Qwen2.5 Technical Report.* arXiv:2412.15115.

Peng Yue, Chen Wang, Chao Sun, Weimin Zhan, and Zhenzhen Sui. 2023. 中文空间语义理解评测数据集质量评估研究. 语言文字应用, (01):101–113.

Yifan Zhao and Jie Wei. 2023. *Holistic Spatial Reasoning for Chinese Spatial Language Understanding. Applied Sciences*, 13(21):11712.

Appendix A 求解器路径示例

为展示第5节中描述的推理路径，本节以一个具体的“环形约束”推理任务为例，完整展示从原始问题到最终求解器调用的全过程。

A.1 原始问题

系统接收的原始输入包含题干、问题和选项，格式如下。

原始问题示例

题干(Text): 何仙姑、吕洞宾、张天师、张果老、铁拐李、韩湘子六座神像在神坛中围成一个圆圈，每座神像都正对神坛中心。六个神像的位置恰好落在正六边形的六个顶点上。任意相邻两个神像之间的距离相等，大约为一米。已知：(1)铁拐李的右边紧接着就是韩湘子；(2)吕洞宾在张果老逆时针方向的第5个位置；(3)张天师在何仙姑顺时针方向数第1个位置；(4)韩湘子在吕洞宾逆时针方向的第5个位置；(5)铁拐李在何仙姑逆时针方向的第1个位置；(6)张果老在张天师左边数起第1个位置。

问题(Question): 何仙姑在()顺时针方向数第2个位置。

选项(Options): A: 吕洞宾　B: 张天师　C: 张果老　D: 以上选项都不是

A.2　步骤一：约束提取

解析器$\mathcal{P}$ 首先将非结构化的题干文本转换为格式化的约束。对于此例，提取结果如下：

提取约束结果

```
entity_num=6

entity=['何仙姑', '吕洞宾', '张天师', '张果老', '铁拐李', '韩湘子']

restraint_list=[

Relative(entity_1='韩湘子', entity_2='铁拐李', relative_pos=-1),

Relative(entity_1='吕洞宾', entity_2='张果老', relative_pos=-5)

Relative(entity_1='张天师', entity_2='何仙姑', relative_pos=1),

Relative(entity_1='韩湘子', entity_2='吕洞宾', relative_pos=-5)

Relative(entity_1='铁拐李', entity_2='何仙姑', relative_pos=-1)

Relative(entity_1='张果老', entity_2='张天师', relative_pos=1)

]
```

A.3　步骤二：排列求解

接下来，遍历程序根据提取的约束条件，对所有可能的实体排列（$6! = 720$种）进行检验。在本例中，程序找到了唯一的合法排列。

- **唯一解(Solution):** [何仙姑, 张天师, 张果老, 吕洞宾, 韩湘子, 铁拐李]

该解表示从何仙姑开始，按顺时针方向排列的实体顺序。

A.4　步骤三：求解器调用

由于找到了唯一解，系统将采用“解析器-求解器路径”。它会将原始问题与找到的解一起组合成新的提示词，并将其输入求解器模型$\mathcal{S}_k$ 以获得最终答案。

求解器输入提示词(有解路径)

题目是单选题，有一个正确答案。答案选项必须与标准答案完全一致才能得分。请逐步思考，并最终输出答案选项。

<题干>

现在已求得俯视按顺时针排列如下：何仙姑-张天师-张果老-吕洞宾-韩湘子-铁拐李

问题:何仙姑在()顺时针方向数第2个位置。
选项:
A:吕洞宾
B:张天师
C:张果老
D:以上选项都不是

答案:

A.5 备用路径（对比）

作为对比，如果在步骤二中，遍历程序未能找到唯一解（例如，找到多个解或没有解），系统将触发“备用求解器路径”。此时，输入给备用求解器模型$\mathcal{S}_{\text{aux}}$ 的提示词将不包含求解部分，仅依赖模型自身的端到端推理能力。

备用求解器输入提示词(无解/多解路径)

题目是单选题，有一个正确答案。答案选项必须与标准答案完全一致才能得分。请逐步思考，并最终输出答案选项。

<题干>

问题:何仙姑在()顺时针方向数第2个位置。
选项:
A:吕洞宾
B:张天师
C:张果老
D:以上选项都不是

答案:

Appendix B 微调超参数

本节列出我们在微调过程中采用的主要超参数配置。

参数名称	值
LoRA秩(LoRA Rank)	128
LoRA作用层(LoRA Target)	q_proj, v_proj, k_proj, o_proj
学习率(Learning Rate)	1.0×10^{-4}
训练轮数(Num Train Epochs)	3.0
单设备批量大小(Per Device Train Batch Size)	1
梯度累积步数(Gradient Accumulation Steps)	2
优化器调度策略(LR Scheduler Type)	cosine

Table 7: LoRA微调主要超参数设置

第二十四届中国计算语言学大会论文集，第1页-第12页，济南，中国，2025年8月11日至14日。