北京大學
PEKING UNIVERSITY

SpaCE2025 第五届空间语义理解评测

秦宇航, 肖力铭, 胡楠, 邓思锐, 马敬原, 崔香, 张子涵, 蔡奇栩, 丁锦坤, 姜秀旻, 穗志方, 詹卫东

hezonglianheng@stu.pku.edu.cn

北京大学中国语言文学系
北京大学计算机学院
多媒体信息处理全国重点实验室
北京大学中国语言学研究中心

August 7, 2025

目录

任务背景介绍

空间表达是语言中的常见现象。理解语言中承载的空间信息依赖于上下文的语境和现实世界的情境，超越了“形式-意义”之间的简单对应，对机器的语言理解能力提出了更高的要求。

- 她缓缓地回过头，朝着**身后**带着潮气的泥土堆，深深地吸了一口气，慢慢闭上了眼睛。
- 她缓缓地回过头，朝着**面前**带着潮气的泥土堆，深深地吸了一口气，慢慢闭上了眼睛。

Figure: 回头时，“面前”即是“身后”

空间语义理解评测类别及任务

- 空间信息正误判断
- 异常空间信息识别
- 空间语义角色识别
- 空间参照实体判断
- 空间异形同义判别

} 主要关注空间语言理解能力 (语言能力类)

- 中文空间方位推理
- 英文空间方位推理

} 主要关注空间推理能力 (空间推理类)

SpaCE 评测子任务概览

SpaCE2025 在子任务的设置上主要做出了如下调整：

- 语言能力类任务上，不再设置较为依赖语言形式的“空间语义角色识别”“异常空间信息识别”任务；
- 将空间方位推理任务扩展至英文。

评测任务	SpaCE2021	SpaCE2022	SpaCE2023	SpaCE2024	**SpaCE2025**
空间信息正误判断	✓	✓	–	–	✓
异常空间信息识别	✓	✓	✓	✓	–
空间语义角色识别	–	✓	✓	✓	–
空间参照实体判断	–	–	–	✓	✓
空间异形同义判别	–	–	✓	✓	✓
中文空间方位推理	–	–	–	✓	✓
英文空间方位推理	–	–	–	–	✓

Table: 历届 SpaCE 评测任务概览

空间信息正误判断 (DSA)

模型需要判断给定句子中的空间信息是否正常，包括：

1. 空间信息是否符合常理；
2. 多个空间信息之间是否存在冲突。

Text：母企鹅张开嘴，让小企鹅把嘴伸到它嘴里，吃它**沿胃里呕出来**的食物。 Answer：错误 incorrect
Text：母企鹅张开嘴，让小企鹅把嘴伸到它嘴里，吃它**从胃里呕出来**的食物。 Answer：正确 correct

Figure: 空间信息正误判断任务示例

空间参照实体判断 (RSR)

现代汉语中，方位词前紧邻位置一般有一个作为参照物的物体。在一定的条件下，紧邻位置上的参照可以被省略。模型需要在这种情况下从上下文正确识别这个物体。

Text: 一栋四面透风的**土家吊脚楼**，楼上不到20平方米的面积被一道**木板**隔开，里面是老师的寝室，外面是学生的教室。
Interpretation 1: “里面是老师的寝室”是以“木板”为基准，确定“里面”所指的具体方位。 Answer: 正确
Interpretation 2: “里面是老师的寝室”是以“土家吊脚楼”为基准，确定“里面”所指的具体方位。 Answer: 错误

Figure: 空间参照实体判断任务示例

空间异形同义判别 (RSE)

现代汉语中，一般而言，不同形式的句子描述的是不同的空间场景。但在特定的条件下，两个句子可以描述同一个空间场景。
模型需要判断两个句子是否描述的是同一个空间场景。

例1 text1: 火车**上**没什么人。 text2: 火车**里**没什么人。 question: 判断text1和text2描述的空间场景是否相同。请只回答“相同”或“不同”。 answer: 相同
例 2 text1: 她在一只小盒子**里**，发现了一串项链。 text2: 她在一只小盒子**上**，发现了一串项链。 question: 判断text1和text2描述的空间场景是否相同。请只回答“相同”或“不同”。 answer: 不同

Figure: 空间异形同义判别任务示例

空间方位推理 (SPR)

在一个情境中，物体之间具有一系列位置关系。模型需要在给定的情境中，根据文本中的已知条件，推理出物体的位置，以及它们之间的位置关系。

Figure: 四人卡座布局示意图

吕洞宾、铁拐李、姜子牙、张果老四人来到火锅店吃火锅，选了四人卡座坐下。卡座分列一张长方形桌子长边两侧，每排卡座上坐两人。面对面而坐。 已知：张果老在吕洞宾同侧右边，坐在铁拐李右手边的是姜子牙。 问题：___的正对面是吕洞宾的右邻。 A.姜子牙 B.张果老 C.铁拐李 D.以上选项都不是 答案：C
Robert, James, Jason, Mary, - Four people went to a tea restaurant to eat and chose a four-person booth. The booth is arranged along the long sides of a rectangular table, with two people sitting on each side, facing each other. It is known that: Mary is on Robert's right on the same side; Jason is the one sitting on the right-hand side of James. Question:___ is across from the right neighbor of Robert. A. Jason B. Mary C. James D. None of the above Answer：C

Figure: 四人卡座布局空间方位推理试题示例

SpaCE2025 数据集规模统计

对于语言能力类任务，数据集分为用于设计 prompt 的示例集[1]和测试集两个部分。
对于空间推理类任务，数据集分为训练集、验证集和测试集三个部分，可以用于模型的微调。

子任务	示例集	训练集	验证集	测试集	合计
DSA	20	0	0	3,500	3,520
RSR	20	0	0	1,763	1,783
RSE	20	0	0	1,100	1,120
SPR(Chinese)	0	2,000	500	3,500	6,000
SPR(English)	0	2,000	500	3,500	6,000
Total	60	4,000	1,000	13,363	18,423

Table: SpaCE2025 总体及各个子任务的数据集规模

[1]示例集提供了任务数据的少量示例，参赛者可以用这些示例构造 prompt。这个数据集数据过少，因此难以用于训练。

SpaCE2025 数据集的特点

- **推理双语数据集：**今年首次在空间推理类任务上将数据集扩展至英文。相比已有的空间语言能力评测数据集，SpaCE2025 可以同时考察模型在多语言上的空间语义理解能力。
- **专注于高认知难度任务评测：**先前工作发现，模型在高度依赖语言形式的任务上已经接近或达到人类水平。因此，本次评测聚焦于超越语言形式-意义配对的任务，在高认知难度的试题上评测模型的空间语义理解能力。
- **数据量和平衡性：**增加了测试数据量，并保证各子任务上试题标签类型的平衡性。

参赛队伍情况

- 报名队伍：38 支
- 最终提交队伍：12 支
- 总奖金池：3.6 万元（华为技术有限公司赞助）

名次	队伍名称	所属单位
1	outofmemory	中国石油大学（华东）
2	zyy	上海大学
3	一二三四 1234	个人
4	PassionAI	中国平安
5	多宝儿	郑州大学
6	ZZUNLP_Han	郑州大学

Table: SpaCE2025 获奖队伍名称及单位

语言能力类任务上的方法

- 任务要求：可通过设计提示词或微调的方式，使用参数量不大于 7B 的大语言模型参赛。
- 参赛队伍普遍基于提示工程 (prompt engineering) 的方式参赛。

名次	模型	方法
1	DeepSeek-R1-7B	为每个子任务设计逐步分析 prompt
2	Qwen2.5-7B-instruct	使用 In-Context Learning(ICL) 方法获得 prompt，微调模型
3	Qwen3-4B	基于提示工程 (prompt engineering) 的方式
4	Qwen3-4B	使用 DeepSeek-Chat 获得思考规则
5	DeepSeek-R1-Distill-Qwen-7B、Qwen2.5-7B、Qwen3-4B	使用含有 CoT 的 prompt 进行投票
6	Qwen2.5-7B 和 Qwen3-4B	使用提供的空间词表和示例集，合成了一批语言能力类任务数据微调

Table: 获奖队伍在语言能力类任务上使用的方法

空间推理类任务上的方法

- 任务要求：必须通过微调DeepSeek-R1-Distill-Qwen-7B的方式参赛。
- 参赛队伍普遍使用了 LoRA(Low-Rank Adaption) 方法对模型进行微调。

名次	方法
1	提出了一套数学化的框架以精确表示实体的空间关系
2	使用 GPT-o3-mini-high 提取试题中物体位置的约束关系
3	使用 DeepSeek-R1 生成推理链数据进行训练
4	使用 DeepSeek-Chat 生成推理链数据进行训练
5	将中英文数据集合并进行多任务学习
6	考虑到两类任务数据的相似性，在空间推理类任务上微调时也使用语言能力类任务的数据

Table: 获奖队伍在空间推理类任务上使用的方法

评测指标

本次工作使用准确率 (accuracy) 作为评测指标。
总体准确率是语言能力类任务和空间推理类任务准确率的平均。
对于每类任务，其准确率是其全部子任务准确率的平均。

$$S = \frac{S_{language} + S_{reasoning}}{2}$$

$$S_{language} = \sum_{i=1}^{3} Acc_i$$

$$S_{reasoning} = \sum_{i=1}^{2} Acc_i$$

$$Acc = \frac{\#correct}{\#total}$$

参赛队伍的整体表现情况

队伍排名	语言能力类任务				常识推理类任务			合计
	DSA	RSR	RSE	合计	SPR(Chinese)	SPR(English)	合计	
1	**0.7089**	**0.8491**	0.6600	**0.7393**	**0.8686**	**0.8251**	**0.8469**	**0.7931**
2	0.6454	0.7720	0.7082	0.7085	0.6254	0.5997	0.6126	0.6606
3	0.6911	0.8259	0.6827	0.7332	0.6914	0.3766	0.5340	0.6336
4	0.6766	0.7992	0.6527	0.7095	0.5106	0.5263	0.5184	0.6140
5	0.6889	0.7856	**0.7200**	0.7315	0.4694	0.4609	0.4651	0.5983
6	0.6626	0.8100	0.6973	0.7233	0.4446	0.4503	0.4474	0.5854
7	0.6637	0.8332	0.6355	0.7108	0.4431	0.4703	0.4567	0.5838
8	0.6994	0.8168	0.6636	0.7266	0.4349	0.4283	0.4316	0.5791
9	0.6829	0.8168	0.6336	0.7111	0.4329	0.4374	0.4351	0.5731
10	0.6017	0.7079	0.6173	0.6423	0.3151	0.2714	0.2933	0.4678
11	0.5177	0.5394	0.5309	0.5293	0.3734	0.3986	0.3860	0.4577
12	0.6003	0.7317	0.6336	0.6552	0.2426	0.2034	0.2230	0.4391
Baseline	0.6274	0.6375	0.5809	0.6153	0.2266	0.2977	0.2622	0.4388

Table: 12 支参赛队伍的表现情况[1]

[1] 排行榜：https://pku-space.github.io/SpaCE2025/leaderboard.html

参赛队伍的整体表现分析

12 支队伍的表现均超过了我们的基线系统。此外：

- 任务所关注的空间表达的数量和复杂度决定了任务的难度。任务的难度大小：RSR < RSE < DSA < SPR，模型在 RSR 等较为简单的任务上表现较好，但在 DSA、SPR 等复杂任务上表现较差。
- 在语言能力类任务上，各队伍表现差别较小，单纯依赖提示工程难以激发模型的空间语义理解能力。
- 在空间推理类任务上，各队伍表现差别较大，能激发模型推理能力的微调方法和 prompt 是关键。

人机表现对比

为了对比人类和模型在空间语义理解上的表现，我们从各子任务的测试集中抽取一部分，组织了人类测试，并与模型在对应试题上的表现做对比。

- 模型表现并未超过人类。模型在较简单的任务，如 RSR 上接近人类水平，但在较复杂任务上与人类表现仍有较大差异，复杂语义理解仍有待增强；
- 人类一致性总体较高，这说明本次评测的数据具有较高的质量。

队伍排名	语言能力类任务			常识推理类任务	
	DSA	RSR	RSE	SPR(Chinese)	SPR(English)
1	0.62	0.85	0.70	0.87	0.93
2	0.51	0.72	0.72	0.63	0.53
3	0.63	0.83	0.75	0.73	0.43
4	0.59	0.80	0.71	0.47	0.43
5	0.64	0.81	0.74	0.43	0.43
6	0.60	0.79	0.73	0.47	0.43
队伍平均	0.60	0.80	0.73	0.60	0.53
人类平均	**0.82**	**0.85**	**0.89**	**0.97**	**0.93**

Table: 在人类测试子集上模型和人类的表现对比

模型选项的偏向性

任务名称	正确	错误
DSA	0.88	0.33

Table: 获奖队伍在 DSA 任务中不同答案试题上的平均准确率

任务名称	正确	错误
RSR	0.90	0.79

Table: 获奖队伍在 RSR 任务中不同答案试题上的平均准确率

任务名称	相同	不同
RSE	0.55	0.82

Table: 获奖队伍在 RSE 任务中不同答案试题上的平均准确率

模型选项的偏向性

统计前 6 名队伍在 DSA、RSR、RSE 任务中不同答案试题上的准确率可以发现：

- 在 DSA 和 RSR 任务中，模型在答案为“正确”的试题上准确率高于答案为“错误”的试题；
- 在 RSE 任务中，模型在答案为“相同”的试题上准确率低于答案为“不同”的试题。

这可能与模型训练语料的性质有关，其中的空间表达多数都是正确的，且不同形式的空间表达大部分意义不同。

对同源题的分析

DSA 和 RSE、RSR 和 RSE 的数据集当中，有的试题共享相同的 text，对这些文本的理解应当具有相似性，我们将这些试题称为“同源题”。此外，在 SPR 任务中，中文和英文试题的文本具有一一对应的关系，也称为“同源题”。

<table>
<tr><td>Text: 列宾生命的最后几年，右手已不能握画笔，他就改用左手画；调色板则借助皮带挂在脖子上面。
Answer: 正确</td><td rowspan="2">Text1:列宾生命的最后几年，右手已不能握画笔，他就改用左手画；调色板则借助皮带挂在脖子上面。
Text2: 列宾生命的最后几年，右手已不能握画笔，他就改用左手画；调色板则借助皮带挂在脖子下面。
Answer: 相同</td></tr>
<tr><td>Text:列宾生命的最后几年，右手已不能握画笔，他就改用左手画；调色板则借助皮带挂在脖子下面。
Answer: 正确</td></tr>
</table>

Figure: DSA 和 RSE 的同源题示例，2 道 DSA 试题和 1 道 RSE 试题共享相同的 text。

同源题的数量情况

语言能力类任务的子任务中同源题的数量如下表所示。在空间推理类任务中，中文和英文试题的文本具有一一对应的关系，因此同源题的对数和单个子任务的试题数量相同。

相关联的子任务	DSA	RSR	RSE
DSA-RSE	1779	–	973
RSR-RSE	–	452	112

Table: 在空间语言能力评估的子任务中，共享相同文本的试题数量。DSA-RSE 表示 DSA 问题的文本也出现在 RSE 中，RSR-RSE 表示 RSR 问题的文本也出现在 RSE 中。

模型在同源题上的表现

统计发现，模型在同源题上的表现相关性较差。这说明模型在完成不同任务时使用的是不同的能力，模型并不存在整体性的空间语义理解能力。

队伍排名	DSA-RSE		RSR-RSE		SPR(Chinese)-SPR(English)	
	ρ	p	ρ	p	ρ	p
1	0.046	0.193	-0.009	0.922	0.467	*
2	0.037	0.291	-0.077	0.423	0.505	*
3	0.045	0.203	-0.029	0.760	0.257	*
4	0.013	0.716	0.180	0.058	0.301	*
5	0.009	0.804	0.143	0.133	0.641	*
6	0.127	*	0.089	0.352	0.577	*
Average	0.090	0.011	0.132	0.164	0.480	*

Table: 各类同源题上的 Spearman 相关系数 (ρ) 及 p-value。* 表示 ρ 的绝对值在 0.001 和-0.001 之间或者 p-value 小于 0.001。

语言能力类任务上表现的特点

在语言能力类任务上，模型受文本结构的影响较大，难以捕获较复杂的语言成分之间的关系。

- 在空间信息正误判断任务中，模型更难识别由于句子中多个空间表达之间存在冲突导致的错误；
- 在空间异形同义判别任务中，模型在需要模型进行参照实体判断后才能解答的试题上表现较差。

成因类型	准确率
空间语言表达错误	0.71
空间逻辑推理错误	**0.29**

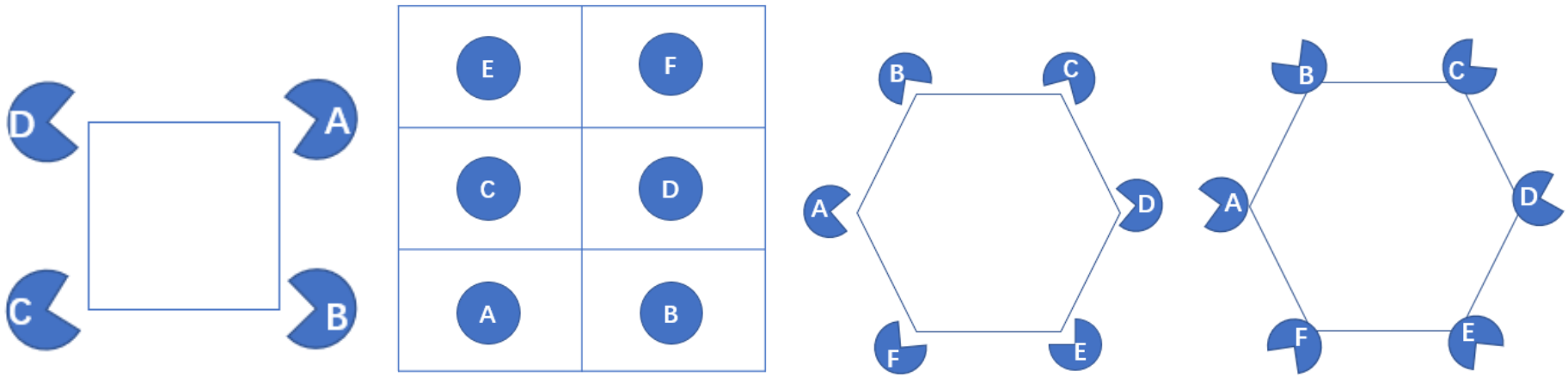
Table: DSA 任务上不同成因类型试题获奖队伍的平均准确率

成因类型	准确率
空间图式交集	0.70
涉及趋向动词	0.75
涉及多个参照	**0.41**
实体投影关系	0.73
语义包含关系	0.50

Table: RSE 任务上不同成因类型试题获奖队伍的平均准确率

空间推理不同布局类型上的表现

空间推理任务的布局类型主要分为四种：四人卡座布局、三层两列布局、向心六角布局和离心六角布局。



(a) 四人卡座布局　(b) 三层两列布局　(c) 向心六角布局　(d) 离心六角布局

Figure: 空间方位推理任务的 4 种布局类型

空间推理不同布局类型上的表现

对不同布局类型的试题统计分析的结果表明，不同的空间情景难度不同，一般而言，实体数量越多、实体间关系越复杂，试题的难度越高，模型的表现越差。

排名	四人卡座	三层两列	向心六角	离心六角
1	0.99	0.98	0.71	0.91
2	0.84	0.46	0.64	0.57
3	0.73	0.80	0.53	0.77
4	0.70	0.67	0.35	0.48
5	0.83	0.45	0.41	0.32
6	0.84	0.40	0.37	0.31
平均	0.82	0.63	0.50	0.56

Table: 各布局类型上的模型平均准确率

评测总结

- 各子领域的准确率情况：RSR > RSE > DSA > SPR. 模型在较简单的任务 (如 RSR) 上接近人类水平，但在较复杂的任务 (如 DSA、SPR) 上表现较差，说明模型在空间语义理解方面仍有待提升。
- 模型不具有完整的空间语义理解能力。一方面，模型在同源题上的表现相关性较差，说明模型在完成不同任务时使用的是不同的能力；另一方面，模型的空间语义理解受到上下文距离和文本复杂性等的影响，说明模型的空间语义理解能力仍然有限。
- 基于 prompt 的方法对模型空间语义理解能力的提升较为有限。微调在空间推理类任务上取得了良好的效果，将结果从 0.2622 提升到 0.8469。这说明无训练的方法难以提升模型表现，需要做进一步的微调。这再次强调了高质量、大规模数据的必要性。

评测展望

- 对于语言能力类任务：
  - 增加数据量，供测试和微调使用；
  - 增加同源题的比例，以测试模型不同难度层次的空间语义理解能力。
- 对于空间推理类任务：
  - 增加新的关系类型和布局类型。

此外，与多模态任务的结合也是未来的一个重要方向。

谢谢！

评测官网：https://pku-space.github.io/SpaCE2025
数据集：
https://github.com/PKU-SpaCE/SpaCE2025/tree/main/data