CCL25-Eval 任务1系统报告：基于数据、训练、推理三阶协同增强的空间语义理解

华中天[1]，罗毅[1]，王梦园[1]，于美佳[2]，韩英杰[1]✉
[1]郑州大学　计算机与人工智能学院　郑州　450001
[2]河南科技大学　农业装备工程学院　洛阳　471023
{hzt1113,wangmengyuan}@gs.zzu.edu.cn
nancetide@stu.zzu.edu.cn
240320261531@stu.haust.edu.cn
ieyjhan@zzu.edu.cn

摘要

SpaCE2025以空间语义理解为核心，聚焦于具有较高难度的空间语义理解任务，旨在评估大语言模型（LLM）在空间语言能力和空间推理能力两方面的表现。面对空间语义复杂、训练数据缺失和模型参数限制等挑战，本文提出了一个基于数据、训练、推理三阶协同增强的模型优化框架，针对空间语言能力和空间推理能力两个子任务分别设计了两套不同的优化方案。对于空间语言能力任务，我们利用DeepSeek-R1结合空间词表对训练集进行了扩充，对Qwen系列LLM进行了LoRA微调，在推理过程中使用了测试时增强来进一步优化结果；对于空间推理能力任务，我们将空间语言能力数据集也纳入训练集，对DeepSeek-R1-Distill-Qwen-7B模型进行微调，并对模型预测结果进行了累计投票集成。最终，我们的方法排名第六，总体准确率得分为58.54%。此外，本文还报告了一些尝试过但未能提升模型表现的其他方法。

关键词： 空间语义；大语言模型；三阶协同增强

System Report for CCL25-Eval Task 1: Spatial Semantic Understanding based on the Triple-Stage Collaborative Augmentation of Data, Training and Inference

[1]Zhongtian Hua,[1]Yi Luo,[1]Mengyuan Wang,[2]Meijia Yu,[1]Yingjie Han✉
[1]Zhengzhou University,School of Computer and Artificial Intelligence,Zhengzhou　450001
[2]Henan University of Science and Technology,
College of Agricultural Equipment Engineering,Luoyang　471023
{hzt1113,wangmengyuan}@gs.zzu.edu.cn
nancetide@stu.zzu.edu.cn
240320261531@stu.haust.edu.cn
ieyjhan@zzu.edu.cn

Abstract

SpaCE2025 centers on spatial semantic understanding and focuses on spatial semantic understanding tasks of high difficulty, aiming to evaluate the performance of Large Language Models(LLM) in both spatial language ability and spatial reasoning ability.Facing challenges such as complex spatial semantics, missing training data, and model parameter constraints, we propose a model optimization framework based on the three-stage collaborative enhancement of data, training, and inference. Two different optimization schemes are respectively designed for the two sub-tasks of spatial language

第二十四届中国计算语言学大会论文集，第13页-第23页，济南，中国，2025年8月11日至14日。

and spatial reasoning.For spatial language tasks, we expanded the training set by combining DeepSeek-R1 with a spatial word list, and fine-tuned the Qwen series of LLM using LoRA. During the inference process, we used Test-Time Augmentation to further optimize the results. For spatial reasoning tasks, we included the spatial language dataset in the training set, fine-tuned the DeepSeek-R1-Distill-Qwen-7B model, and performed cumulative voting integration on the model's prediction results.Ultimately, our method ranked sixth and the overall accuracy score was 58.54%. Additionally, this paper also reports some other methods that were attempted but failed to improve the performance.

Keywords: Spatial semantic , Large Language Model , Triple-Stage Collaborative Augmentation

1 引言

近年来，空间语义理解在自然语言处理（Natural Language Processing, NLP）领域持续引发研究热潮。人类语言中蕴含着丰富的空间方位描述与关系表达，从日常对话中的方向指引到地理文本中的坐标定位，空间元素必不可少。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。随着人工智能逐步迈向通用领域，如何使模型拥有深度空间语义理解能力，能够解析并重构文本中的空间信息已成为NLP的关键问题之一。

大型语言模型（Large Language Models，LLM）是一种基于深度学习的大规模参数（通常为数十亿甚至数千亿）模型(Katikapalli, 2023)。它们在大规模文本数据上进行预训练，从而拥有能够捕获复杂的语言模式和语义信息的能力。LLM在第四届空间语义理解评测（SpaCE2024）的评测结果显示，在对空间认知能力要求较高的任务上其与人类平均水平相比仍存在较大差距。空间语义理解对大语言模型来说仍然是一项挑战性任务。因此，第五届空间语义理解评测（SpaCE2025）继续开展针对大语言模型的空间语义理解能力测试，关注大语言模型的空间语言能力和空间推理能力。

SpaCE2025以空间语义理解为核心，聚焦于具有较高难度的空间语义理解任务，要求模型具有较强的深层语义理解和空间认知能力，旨在评估大语言模型在空间语言能力和空间推理能力两方面的表现。该评测包含两大子任务共五个赛题，从不同维度考察模型处理空间信息的表现。针对本赛题的三大主要难点：空间语义复杂、训练数据缺失和严格资源限制，我们提出了一个基于数据、训练、推理三阶协同增强的模型优化框架，并根据两个子任务的不同特点在各个阶段分别采用了不同的增强方法来提升模型的性能。我们的贡献如下：

- 我们构建了一个基于数据、训练、推理三阶协同增强的模型优化框架，以提高LLM在空间语义理解领域中的性能表现。
- 我们根据空间语言能力任务和空间推理能力任务两大子任务的不同特点，在优化框架的基础上对其分别采用了包括多任务学习、测试时增强在内的不同增强策略，最终得到了58.54的准确率得分，在所有队伍中排行第六。
- 我们通过全面的对比实验验证了所采用的增强策略的有效性，同时也总结了一些我们在评测任务期间所尝试的一些其他方法，供后续的研究参考。

2 相关工作

2.1 空间语义理解研究

早期的空间语义信息提取主要采用基于机器学习的方法。Roberts等(Roberts and Ullman, 2012)基于支持向量机（SVM），使用多种方法联合的方式来识别和分类空间角色，首先使用CRF模型从数据中提取特征，捕捉词语之间的依赖关系，接着使用最大熵和朴素贝叶斯分

类器对词语之间的介词关系和歧义进行消除。Mazalov等(Alexey et al., 2015)提出了一个基于卷积神经网络的空间角色及关系标注系统。随着深度学习的不断发展，在NLP具有里程碑意义的Transformer模型被Vaswani等(Vaswani et al., 2017)提出，也有研究团队在基于Transformer的预训练语言模型上开展空间语义理解的相关研究。Shin等(Shin et al., 2020)提出了BERT空间模型，使用BERT从原始文本中提取空间元素，确定它们对应的空间角色，进一步使用R-BERT对空间角色的关系进行提取。而自从作为LLM代表的GPT-3系列模型的提出(Brown et al., 2020)，其强大的语义理解能力使其能够在各种下游任务中展现出良好的性能。本文基于LLM，开展了一系列空间语义理解领域的研究。

2.2 空间语义理解评测任务

随着人们对于空间语义理解任务的关注度不断增多，也有不少研究团队构建了空间语义理解相关的数据集并发布评测任务，以测试模型在空间语义理解领域中的表现。SemEval系列评测任务提出了面向空间语义理解的多个评测，关注模型的空间语义角色标注能力。具体来说，SemEval 2012(Kordjamshidi et al., 2012)引入了一个关注于静态空间关系的角色标注任务，SemEval 2013(Kolomiyets et al., 2013)则将空间关系扩展到动态，增加了评测任务的难度和复杂性，从而更加全面的考察模型的空间语义理解能力。

SpaCE系列评测任务作为中文空间语义理解评测的杰出代表，致力于考察模型在中文空间语义理解上的表现，近年来发布了一系列的评测任务。SpacE2021(詹卫东et al., 2022)提出了三个子任务，以考察模型能否正确区分正常与错误的空间语义表达和解释表达错误的原因。SpacE2022(Xiao et al., 2023a)扩大了任务类型，增加了信息标注任务，扩大了数据规模。SpacE2023(Xiao et al., 2023b)则在SpaCE2022基础上增加了对空间场景的关注，重点考察模型对于空间场景异同的判断。SpaCE2024(Xiao et al., 2024)更加注重针对LLM的空间语义理解能力的测试，将之前的测试题改为选择题，同时提高了专业领域的语料占比。本次SpaCE2025则在其基础上继续开展针对大语言模型的空间语义理解能力测试，关注大语言模型的空间语言能力和空间推理能力。

3 方法

3.1 方法总览

我们基于提出的三阶协同增强优化框架为两个子任务分别设计了不同的优化方案。其中空间语言能力任务方法总览图如图1所示，空间推理能力任务方法总览图如图2所示。

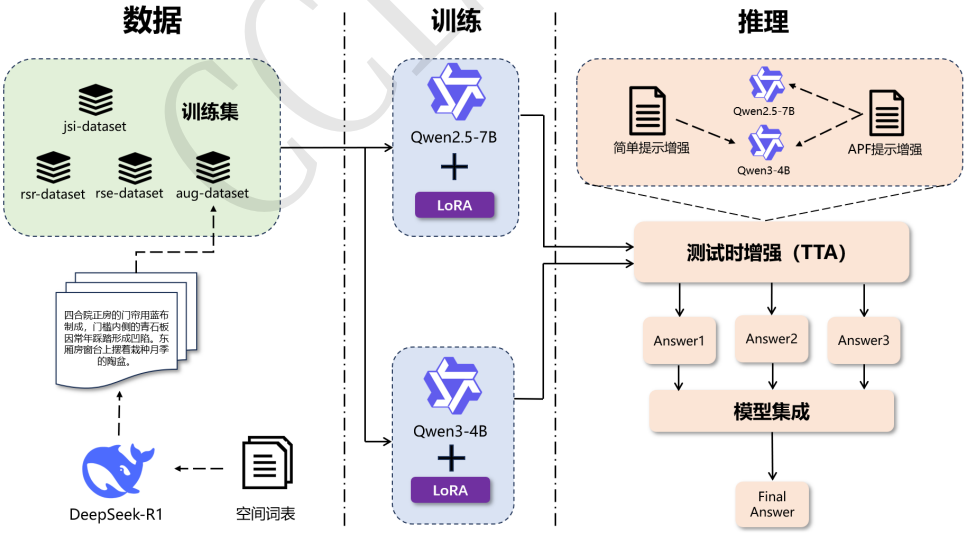

Figure 1: 空间语言能力任务方法总览图

对于空间语言能力任务，我们在数据部分利用了DeepSeek-R1模型(DeepSeek-AI et al., 2025)的强大文本生成能力，结合官方提供的空间词表和示例数据集来生成额外的训练数据。在训练部分，我们对符合要求的Qwen系列大语言模型（Qwen3_4B和Qwen2.5_8B）进行

了LoRA微调。在推理部分，我们使用了测试时增强（Test Time Augmentation，TTA）的方法，通过选用不同模型的检查点和以及我们自己所设计的一份混合增强提示框架（Augmented-PromptFusion Framework，APF）来生成几份不同的预测结果，最后对这些结果进行少数服从多数的投票集成以获取最佳的预测结果。

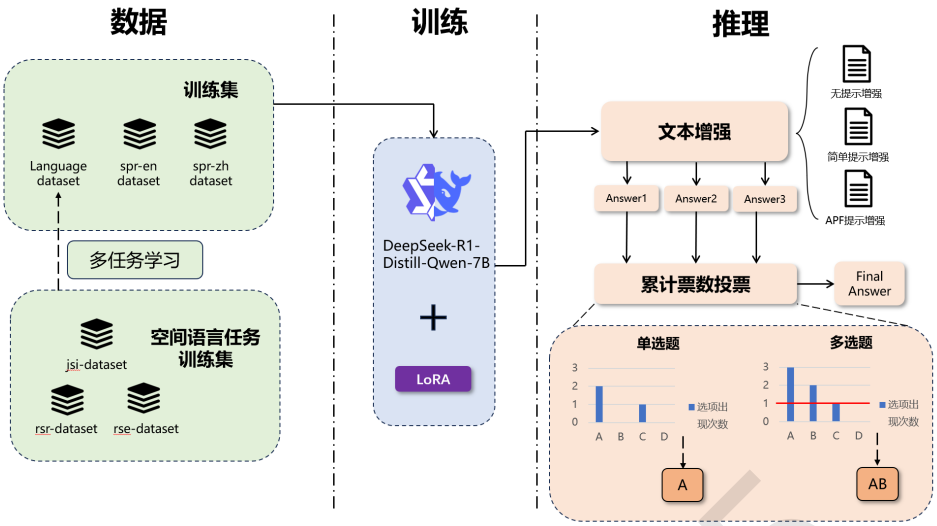

Figure 2: 空间推理能力任务方法总览图

对于空间推理能力任务，在数据部分，我们借鉴了多任务学习的思想，将空间语言能力数据集也纳入了原本的训练集来对模型进行协同训练。在训练过程中，对DeepSeek-R1-Distill-Qwen-7B模型采用了LoRA微调。在推理部分，我们对模型不同检查点的预测结果进行了适用于多项选择题的累计投票集成来提升最终的分数。

3.2 数据阶段的数据扩充方法

在本小节中，我们将介绍在数据阶段我们对于两个子任务分别使用的不同的数据增强方法。在3.2.1节中，我们将介绍如何利用DeepSeek-R1模型来生成新的训练集以应对空间语言能力任务数据集缺失的问题。在3.2.2节中，我们将介绍如何将多任务学习的思想应用到空间推理能力任务当中。

3.2.1 面向语言任务的基于DeepSeek-R1的数据集扩充方法

针对比赛方未提供空间语言能力任务数据集的问题，我们利用官方提供的空间实体词表和空间方位词表加上数据样例，精心设计了一份提示词，将其送给DeepSeek-R1来让其仿造示例来生成新的数据以构建训练集。具体来说，我们首先对官方提供的空间实体词表和空间方位词表进行随机采样，每次从中随机提取出50个方位词和100个实体添加到提示词中。再将官方提供的空间语言能力任务示例集进行正负样本配对，每次提取一个正负样本对添加到提示词中。最后再设计额外的提示词来提示DeepSeek-R1从提供的方位词和实体词中选择合适的词语，仿照样本来生成新的数据。最终，我们针对空间语言能力任务中空间信息正误判断、空间参照实体判断和空间异形通义判断三个赛题分别生成了2388、2114、1664条训练集。

3.2.2 面向推理任务的多任务学习方法

多任务学习是一种机器学习方法，其基本原理是利用不同任务之间的相关性，通过共享知识来提高模型的训练效果。这种方法的优势在于可以提高模型的泛化能力，通过学习多个任务，模型能够捕捉到更广泛的特征和模式。同时，当数据量有限时，多任务学习可以利用不同任务的数据来丰富模型的训练数据，从而减少过拟合的风险。

在本次任务中，我们发现空间语言能力任务数据中包含了丰富的空间语义特征，通过这些特征，模型能够学习到更加丰富和细致的空间语义知识，从而提升其在空间推理能力任务上的表现。因此，我们将空间语言能力任务数据集也纳入到了空间推理能力任务的训练集当中，以对原有的空间推理能力训练集进行扩充。

3.3 训练阶段的微调方法

在本小节中，我们将介绍在训练阶段所采用的策略。在3.3.1中，我们将介绍在本次评测任务中所选用的基座LLM并阐述为何选择了这些模型。在3.3.2中，我们将介绍如何对这些模型进行高效参数微调。

3.3.1 模型选择

出于对比赛公平性以及后续评估研究的考量，本次评测任务的主办方限定了空间语言能力任务所使用的模型大小不能超过7B；空间推理能力任务所使用的模型只能是DeepSeek-R1-Distill-Qwen-7B(DeepSeek-AI et al., 2025)。因此，我们的模型选择主要针对于空间语言能力任务。

Qwen 系列大模型是由通义实验室开发的一系列先进开源大语言模型，涵盖从1.8B 到72B等不同参数规模版本。得益于在海量中文语料上的深度训练，其在处理中文语料的任务中拥有着出色的性能。因此，我们在本次评测任务中选择了Qwen系列最新发布且符合参数规模的Qwen2.5_7B(Yang et al., 2024)以及Qwen3_4B(Yang et al., 2025)模型来参加本次评测。

3.3.2 参数高效微调

参数高效微调（Parameter-Efficient Fine-Tuning，PEFT）是一种在LLM上进行微调的新技术，旨在降低LLM微调的计算和存储成本，使得微调过程更为高效。低秩适应（Low-Rank Adaptation，LoRA）是PEFT 支持的多种微调方法之一。在本次比赛中，我们采用了LoRA方法进行指令微调，并通过多种不同策略进一步优化微调效果、显存利用和微调效率。

具体来说，为了实现在训练过程中实时计算每个检查点的准确率得分，我们手动实现了一套简易的训练框架来进行微调。同时还使用了余弦退火学习率调度方法来优化微调的效果。而在显存利用的方面，采用了8bit AdamW优化器、梯度检查点等策略。针对Qwen2.5模型，我们额外采用了liger-kernel的线性融合交叉熵损失来替代原本的LLM前向传播方法中的损失函数，这一改进进一步降低了显存占用。而为了加快收敛，我们采用了掩码策略，将模型输入序列中的instruction部分全部设为-100，仅计算output部分的交叉熵损失，从而提升了微调的效率。

3.4 推理阶段的推理增强方法

在本小节中，我们将会阐述在推理阶段时我们采取的增强策略。在3.4.1中介绍我们所提出的混合增强提示框架APF，在3.4.2中介绍在空间语言能力任务当中所采用的测试时增强方法。在3.4.3中介绍在空间推理能力任务当中我们所采用的累计票数投票法。

3.4.1 混合增强提示框架

提示工程是一种为LLM设计更加工程化、更加符合任务需求的提示词设计技术。通过精心设计输入提示，引导模型按照使用者的需求来响应。有效的提示工程可以大大提高大型语言模型在特定任务中的性能。在本次评测中，我们通过集成角色扮演、问题分解等提示工程技术，设计了一份针对于空间评测任务的混合增强提示框架（Augmented-Prompt Fusion framework，APF），并根据两大子任务中五个赛题的不同特点，在该框架的基础上精心设计了五份提示模板。同时，为了获取多样化的答案输出以用于最后的模型集成，我们还设计了五份只使用了角色扮演技术的简单提示模板。我们所使用的提示模板的样例在文章的附录中给出。

3.4.2 面向语言任务的测试时增强方法

测试时增强（Test-Time Augmentation, TTA）是一种在模型推理阶段进行数据增强的方法。TTA在模型测试时通过对输入数据进行多次变换（如使用不同的提示增强方法构建的提示词），然后对这些增强后的数据进行预测，最终通过对多个预测结果的平均或投票等策略来得到最终的预测输出。这种方法能够有效提高模型的泛化能力，是一种简单且有效的提升模型性能的技术。

在空间语言能力任务中，我们在推理时应用了TTA方法进一步提升模型表现。具体来说，我们在简单增强数据集和APF数据集两组数据集上对Qwen2.5模型和Qwen3模型分别进行了微调。在推理阶段同样应用了不同的提示模板，从而使两个模型分别产生了两组多样化的输出。然后对这四个结果中得分相近的三个进行少数服从多数的模型投票硬集成以得到最终结果。

3.4.3 面向推理任务的累计票数投票法

在空间推理能力任务中，由于模型的种类被限定了，因此我们额外使用了一份不包含提示增强的提示模板来训练模型，从而同样得到了三份答案输出。同时，由于空间推理能力任务的题目均为单项选择题或多项选择题，这就导致虽然投票集成的方式可以应用于单项选择题，却无法直接应用于多项选择题中。因此，我们引入了同时适用于单项与多项选择题的累计票数投票法。对于单项选择题，我们沿用了少数服从多数的投票机制，票数最多的选项会被选为最终答案选项。若出现平票的极端情况，则会选择得分最高的检查点生成的答案；对于多项选择题，我们统计了每个选项的得票数，最终将所有票数大于等于2的选项作为最终答案选项。

3.5 其他尝试方法

3.5.1 数据阶段所尝试的其他方法

在数据阶段，我们曾经尝试过对空间推理能力任务也进行数据集扩增的方法。然而由于空间推理能力任务的复杂性，由DeepSeek-R1生成的新数据质量较为一般，无法对模型起到正面影响。我们也尝试过模型蒸馏的方法，由DeepSeek-R1产生对训练集每个答案生成过程的思维链，让小型模型去学习大型模型的思考过程。由于小型模型的模型参数过小，无法吸收大型模型的思考过程，这种方法甚至会降低小型模型的回答准确率。

3.5.2 训练阶段所尝试的其他方法

在训练阶段中，我们在模型选择方面和微调方法选择方面都尝试过一些别的方法。在模型选择方面，由于Qwen3_8B所表现出的出色性能，我们也尝试在不影响其性能的情况下对其进行剪枝操作以符合参数要求。我们统计了训练集和测试集所有可能出现的token索引，裁剪掉除此之外的嵌入层token行权重，然而裁剪后的模型权重大小为7.6B，仍然超过了比赛的限制，而进一步的剪枝又可能会影响模型的性能，因此我们最终放弃了这一方案。而在训练方法方面，受强化学习算法GRPO在DeepSeek-R1模型上取得的巨大成功的影响，我们也尝试对三个模型分别进行GRPO微调。然而，由于GRPO算法所需的算力资源和微调时间过多，即使我们使用了unsloth框架来降低其显存需求并加快训练，还是没有足够的时间来对三个模型进行完整的GRPO训练。最终，我们对DeepSeek-R1-Distill-Qwen-7B进行了一轮GRPO微调，发现GRPO微调的效果也远不如预期。这可能是由训练轮次不足和GRPO算法只能强化模型现有能力而无法增加其他能力的特性共同导致的。

3.5.3 推理阶段所尝试的其他方法

在推理阶段，我们曾尝试过将思维链技术（Chain-of-Thought）应用到APF模板当中。但是效果也同样不佳。这与Nayab等(Nayab et al., 2024)的结论是一致的：CoT技术只适用于参数规模大于10B的大型LLM上，将其应用在小型LLM上不仅不能提升模型性能反而会影响其表现。

4 实验与结果

4.1 实验设置

4.1.1 数据集

SpaCE2025数据集总数据量为18,423 题。其中空间语言能力任务的三大赛题：空间信息正误判断、空间参照实体判断、空间异形同义判断任务在多种不同类型的真实语料上进行改写工作，包括：报刊语料、文学作品语料、中小学课本语料、交通事故描述文本、人体动作文本、地理百科文本。空间推理能力任务则是运用基于知识库的数据合成方法生成的高质量合成数据。数据分布如表1所示：

4.1.2 评价指标

SpaCE2025使用的评价标准为两大类任务的综合得分S，$S1$代表空间语言能力类评测任务的得分，$S2$ 代表空间推理能力类评测任务的得分，Acc_i 代表各子任务的准确率（Accuracy，Acc）。公式如下：

赛题	示例集	训练集	验证集	测试集	数据总量
空间信息正误判断	20	0	0	3500	3520
空间参照实体判断	20	0	0	1763	1783
空间异形同义判断	20	0	0	1100	1120
中文空间方位关系推理	0	2000	500	3500	6000
英文空间方位关系推理	0	2000	500	3500	6000
合计	60	4000	1000	13363	18423

Table 1: 数据集分布

$$S = 0.5 \cdot S1 + 0.5 \cdot S2\ ,\ S1 = \frac{1}{3}\sum_{i=1}^{3} Acc_i\ ,\ \ S2 = \frac{1}{2}\sum_{i=1}^{2} Acc_i\ ,\ \ Acc_i = \frac{\#correct}{\#total} \quad (1)$$

4.1.3 基线与细节

对于空间语言能力任务，选用了未经微调的Qwen2.5_7B模型使用未增强的提示词直接预测作为基线模型；对于空间推理能力任务，也同样使用未经微调不使用提示增强的DeepSeek-R1-Distill-Qwen-7B模型作为基线模型。实验过程中与模型训练相关的所有超参数设置如表2所示:

参数名称	参数值
epoch	5
batch size	1
gradient_accumulation	4
learning rate	1.8e-5
max grad norm	1.0
lora_r	64
lora_alpha	512

Table 2: 超参数设置

4.2 实验结果与分析

空间语言能力任务实验结果如表3所示，其中Acc_{jsi}、Acc_{rse}、Acc_{rsr} 分别代表了空间信息正误判断、空间参照实体判断、空间异形通义判断的正确率得分。具体来看，不论是数据阶段的训练集扩增方法还是推理阶段的APF增强方法，即便是单独使用都可以给模型带来不小的提升，综合分数分别相对于基线提升了6.53和3.93个百分点。而在3.5小节中阐述的对DeepSeek-R1的思维链进行蒸馏的方法并没有奏效，这可能是因为模型参数带来的限制。需要特别说明的是，模型推理时所使用的提示模板要和微调时的提示模板匹配，因此若在推理时想使用思维链技术让模型生成带思维链的输出，在训练阶段也必须用带思维链的数据集去微调。最终我们选择将有效的增强方法组合，对模型起到一个协同增强的作用，并将Qwen3_4B和Qwen2.5_8B的最好答案集成投票，取得了我们在本次评测中空间语言能力任务的最佳准确率得分72.33。

空间推理能力任务实验结果如表4所示,其中Acc_{spr_zh}、Acc_{spr_en}分别代表了中文空间方位关系推理和英文空间方位关系推理的准确率得分。具体来看，首先比较GRPO和LoRA两种不同的微调方式，结果证明LoRA微调的效果远超GRPO，这可能是强化学习本身特性和训练轮次不够两个因素共同导致的结果。此外，出于对DeepSeek-R1-Distill-Qwen-7B模型本身就是带思维链的推理模型的考量，我们也尝试了对其使用带思维链的训练数据来微调，然而实验结果表明其效果仍然不如直接让其输出答案，这进一步证明了我们先前的结论。与之相反的是，实验结果证明了多任务学习的方法和文本增强的有效性，相对于基线带来了29.43%的提升。即使与仅使用LoRA微调的结果来比较也有4.79%的提升。最终我们将推理阶段无增强、简单增强

模型	数据/训练/推理增强方法	Acc_{jsi}	Acc_{rse}	Acc_{rsr}	$S1$
Qwen2.5_7B	-/-/-	59.20	62.81	66.70	62.90
	-/-/APF增强	65.37	64.90	70.22	66.83
	扩增训练集/LoRA/-	66.11	65.63	76.57	69.43
	扩增训练集（带CoT）/LoRA/CoT	60.42	57.90	72.20	63.51
	扩增训练集/LoRA/APF增强	66.25	68.36	77.76	70.80
Qwen3_4B	-/-/-	62.45	65.72	63.30	63.82
	扩增训练集/LoRA/-	62.82	67.27	78.50	69.53
	扩增训练集/LoRA/APF增强	61.51	68.18	80.54	70.08
模型集成	扩增训练集/LoRA/集成投票	**66.26**	**69.73**	**81.00**	**72.33**

Table 3: 空间语言能力任务实验结果

和APF增强的三种文本增强方法产生的结果进行累计投票集成，得到了我们本次评测中的最好成绩44.74。

模型	数据/训练/推理增强方法	Acc_{spr_zh}	Acc_{spr_en}	$S2$
DeepSeek-R1-Distill-Qwen-7B	-/-/-	16.24	14.37	15.31
	-/LoRA/-	40.23	39.64	39.95
	-/GRPO/-	17.10	15.23	16.17
	带CoT训练集/LoRA/CoT	38.42	38.76	38.59
	多任务学习/LoRA/-	41.26	41.06	41.16
	多任务学习/LoRA/简单增强	42.43	44.86	43.64
	多任务学习/LoRA/APF增强	43.39	43.71	43.70
	多任务学习/LoRA/累计投票集成	**44.46**	**45.03**	**44.74**

Table 4: 空间推理能力任务实验结果

4.3 总结

在本次CCL2025空间语义理解评测任务（SpaCE2025）中，我们针对空间语义复杂、数据集缺失和模型限制等问题，构建了一套基于数据、训练和推理三阶协同增强的模型优化框架，并针对空间语言能力任务和空间推理能力任务两个子任务具体构建了两种不同的优化方案。我们进行了充分的对比实验来验证方法的有效性并取得了模型的最佳预测结果，最终，我们取得了58.54的综合分数，在所有参数队伍中排行第六。

参考文献

Alexey Mazalov, Bruno Martins, and David Martins de Matos. 2015. Spatial role labeling with convolutional neural networks. In Ross S. Purves and Christopher B. Jones, editors,*Proceedings of the 9th Workshop on Geographic Information Retrieval, GIR 2015, Paris, France, November 26-27, 2015,*,pages 12:1–12:7. ACM.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu,Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671.*

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao

Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Llion, Aiden N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. Attention is All you Need. In*Advances in Neural Information Processing Systems (NIPS),*,pages 5998-6008.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J.L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R.J. Chen, R.L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S.S. Li et al. (100 additional authors not shown). 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLM via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Hyeong Jin Shin, Jeong Yeon Park, Dae Bum Yuk, and Jae Sung Lee. 2020. Bert-based spatial information extraction. In*Proceedings of the Third International Workshop on Spatial Language Understanding,*,pages 10–17.

Katikapalli Subramanyam Kalyan. 2023. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. *arXiv preprint arXiv:2310.12321*.

Kirk Roberts and Sanda Harabagiu. 2012. UTD-SpRL: A joint approach to spatial role labeling. In**SEM2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*,pages 419-424.

Liming Xiao, Weidong Zhan, Zhifang Sui, Yuhang Qin, Chunhui Sun, Dan Xing, Nan Li, Fangwei Zhu, and Peiyi Wang. 2023. ccl23-eval 任务4 总结报告: 第三届中文空间语义理解评测(overview of ccl23-eval task 4: The 3rd chinese spatial cognition evaluation). In*Proceedings of the 22nd Chinese National Conference on Computational Linguistics,* pages 150–158.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens and Steven Bethard. 2013. Semeval-2013 task 3: Spatial role labeling. In*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*,pages 255-262.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. SemEval-2012 task 3: Spatial role labeling. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors,**SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*,pages 365–373, Montr´eal, Canada, 7-8 June. Association for Computational Linguistics.

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, Fabrizio Giacomelli. 2024. Concise Thoughts: Impact of Output Length on LLM Reasoning and Cost. *arXiv preprint arXiv:2407.19825*.

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, Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Xiao Liming, Sun Chunhui, Zhan Weidong, Xing Dan, Li Nan, Wang Chengwen, and Zhu Fangwei. 2023. Space2022 中文空间语义理解评测任务数据集分析报告(a quality assessment report of the chinese spatial cognition evaluation benchmark). In*Proceedings of the 22nd Chinese National Conference on Computational Linguistics,*pages 547–558.

Xiao Liming, Hu Nan, Zhan Weidong, Qin Yuhang, Deng Sirui, Sun Chunhui, Cai Qixu, Li Nan. 2024. The Fourth Evaluation on Chinese Spatial Cognition In*Proceedings of the 23rd Chinese National Conference on Computational Linguistics,* pages 122–134.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐, and 秦梓巍. 2022. 空间语义理解能力评测任务设计的新思路—space2021 数据集的研制. 语言文字应用, 2:99–110.

附录A.提示模板样例

```
### **空间信息正误判断**
#### **1. 核心任务定义**
**目标**：空间信息正确的文本可以构造出合乎常理的空间场景，而错误的文本不能，本任务要求机器判断文本的空间信息是否正确,请基于给定的文本材料（text）判断其中各个实体的空间信息表达是否正确，输出“正确”或“错误”。
---
#### **2. 输入数据示例**
text: 碰撞发生后，“VANMANILA”轮船长迅速跑去驾驶台左侧外部的桥翼查看情况，当班三副杰哈尔跟在后面看见“XIANGZHOU”轮甲板以下已没入水上，就迅速跑回驾驶台发布全船广播，船长回驾驶室见其在进行减速操作，下令恢复原速继续航行。
---
#### **3. 模型推理逻辑指令**
**分阶段分析**：
1. **关键信息提取**
识别文本中的实体，包括但不限于：- 各种生物与非生物实体（如“母鸡”、“轮船”、“窗户”、“肩膀”等）- 指代实体的代词（如“我”、“你”、“这家伙”等）;识别文本中的空间表达，包括但不限于：- 方向性描述（如“左侧”、“右侧”、“前方”）- 空间相对关系（如“在……之上/下”、“内部/外部”）- 运动轨迹（如“进入”、“离开”、“没入”- 位置状态（如“悬浮于”、“沉入”）
2. **一致性与逻辑性检查**
检查细粒度的实体空间表达是否互相矛盾，如：“没入水上”、“沉入空中”均属于矛盾;评估实体的空间关系是否符合现实物理规律，如：“影子在灯的上方”、“后视镜中前方的车子”均属于有误;确认实体的运动路径是否合乎逻辑，如：“从房间内跳出窗外，然后打开门进入房间”、“由南向北行驶，右转驶向西侧”均属于有误
3. **最终判定**
- 若文本中的空间信息完全符合现实逻辑，输出“正确”,若存在任何空间矛盾或不合理表达，输出“错误”
---
#### **4. 特殊场景处理规则**
1. **隐含空间关系推理**
需要推断未明确说明但隐含在文本中的空间逻辑。如“我们沿着蜿蜒崎岖而略微有些陡峭的山路上行”意味着“我们”原本处于山脚，正在往山顶前进，因此若后文中若出现表达“我们”正接近山脚远离山顶的内容则属于表达矛盾。
2. **上下文一致性**
判断同一实体在上下文的空间表达是否保持一致，如“列车向北行驶，然后左转驶入南站”存在矛盾，应为“右转驶入南站”。
3. **实体隐含空间关系**
请注意推理某些特殊实体自身所隐含的空间信息。如在“后视镜”中所看到的东西一定位于“后视镜”的后面，而不会在其前面。
4. **修辞手法说明**
请注意判别文本中使用的修辞手法，如夸张、拟人等修辞虽然会夸大空间关系，但本质上不属于错误的空间表达
---
#### **5. 最终输出格式要求**
请只输出“正确”或“错误”，禁止在最终输出中包含额外解释或推理过程**。
---
```

Figure 3: APF模板使用样例

任务名称：空间信息正误判断

任务背景：空间信息正确的文本可以构造出合乎常理的空间场景，而错误的文本不能，本任务要求机器判断文本的空间信息是否正确

任务详情：假设你是一个拥有强大空间语义理解能力的专家，现在请你判断text的空间信息表达是否正确。

任务规则：
1.请简单思考，text所描述的空间场景不会太复杂
2.判断时既要判断材料所描述的空间场景是否正确，也要根据基本空间常识判断文本的空间逻辑表达是否正确
3.在给出最终答案时只需输出“正确”或“错误”。

Figure 4: 简单增强模板使用样例

假设你是一个拥有强大空间语义理解能力的专家，现在请你判断text的空间信息表达是否正确。

Figure 5: 无增强模板使用样例