SCP-ISA: Spatial Cross-attention Projection for Textual Spatial Semantic Anomaly Detection and Attribution

Bei Xu **Mengqi Ding**
School of Computer Science, Nanjing University of Information Science and Technology
xubei@njupt.edu.cn dingmengqi720@gmail.com

Abstract

Textual spatial semantic anomaly detection and attribution requires deciding whether a sentence contains an implausible spatial relation and assigning the corresponding attribution label. We study controlled, sentence-level spatial consistency checking on mixed normal/anomalous inputs. We propose SCP-ISA, a Spatial Cross-attention Projection (SCP) method with an Inter-Slot Attention (ISA) layer that uses a frozen MacBERT-large encoder to build input-conditioned, role-oriented latent slots and injects the projected slots into Qwen2.5 as soft prefix embeddings. On CRSARC, SCP-ISA reaches 0.736 test cascade attribution F1 (Attr-F1), compared with 0.507 for parameter-matched Direct LoRA and 0.485 for a learned static prefix. The results support structured spatial conditioning, rather than an exact symbolic parser or complete slot-wise disentanglement.

1 Introduction

Textual spatial semantic anomaly detection and attribution asks a model to decide whether a sentence expresses an implausible spatial relation and, when an anomaly is present, identify its attribution label. Such errors arise from how entities, positional expressions, and events jointly describe a scene rather than from any single word. We focus on controlled sentence-level consistency checking in mixed normal/anomalous inputs. This setting is relevant to quality control for extracted, retrieved, or generated spatial claims, but does not by itself establish open-domain deployment or Web-scale prevalence. As a concrete example, consider *"There is a tall tree in front of the cottage, . . . dense branches extend behind the cottage."* The same tree is placed both *in front of* and *behind* the cottage—a direct spatial contradiction.

Existing methods thus have two gaps: prompt-only or LoRA-only approaches rely on token-level representations that do not explicitly compare structured spatial roles, and they lack an inductive bias for cross-triple conflicts.

We propose SCP-ISA, which derives six input-conditioned latent slots from a frozen MacBERT-large encoder, injects them into Qwen2.5 as soft prefix embeddings, and lets them interact through the ISA layer. The slots use role-oriented designations (S_i, P_i, E_i) as a structural inductive bias for relational conditioning; they are not explicitly supervised symbolic triples.

2 Related Work

Spatial cognition evaluation has developed through the SpaCE series (2030NLP, 2021; Xiao et al., 2023, 2024; Qin et al., 2025), revealing persistent LLM weaknesses in spatial relations and multi-hop constraints (Mirzaee and Kordjamshidi, 2023; Rizvi et al., 2024; Liu et al., 2023). Commonsense anomaly benchmarks further show that detecting physically implausible situations remains difficult (Ismayilzada et al., 2023; Bhagwatkar et al., 2025).

Prompting strategies such as Chain-of-Thought (Wei et al., 2022), Self-Consistency (Wang et al., 2023), and Least-to-Most (Zhou et al., 2023) elicit reasoning but do not impose spatial structural biases. Parameter-efficient fine-tuning (PEFT) methods (Hu et al., 2022; Li and Liang, 2021; Liu et al., 2022) efficiently condition large models. Q-Former-style designs (e.g., BLIP-2 (Li et al., 2023)) use learnable queries to bridge external representations and a frozen decoder; SCP-ISA adapts this idea to text and adds an inter-slot attention layer for relational conflicts.

3 Model

As shown in Figure 1, SCP-ISA contains three components. First, a frozen MacBERT-large (Cui et al., 2020) encodes the text. Second, six learnable queries summarize MacBERT states into

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33528–33534
October 24-29, 2026 ©2026 Association for Computational Linguistics

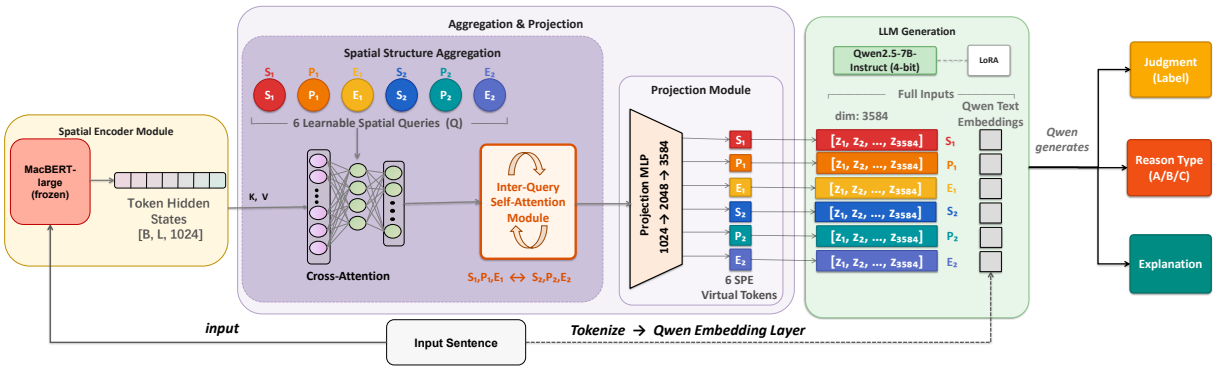


Figure 1: Overall architecture of SCP-ISA. MacBERT-large extracts six spatial slots, which are projected and prepended to Qwen2.5 text embeddings as soft prefix embeddings for generation.

role-oriented latent slots $(S_1, P_1, E_1, S_2, P_2, E_2)$, where the labels suggest spatial-entity, position, and event-trigger roles, and $i \in \{1, 2\}$ indexes two slot groups. The six-slot design provides a shared structural bias across all three SpaCE anomaly types: Type A (collocation anomaly), Type B (cross-triple conflict), and Type C (commonsense violation). The dual-group design is intended to facilitate comparison of potentially conflicting claims. Third, slots are projected into Qwen2.5-7B-Instruct's (Yang et al., 2024) embedding space and prepended as soft prefix embeddings. Role orientation is induced by query position and training rather than by per-slot supervision; we therefore do not claim an exact symbolic triple parser or complete slot-wise disentanglement.

For a concrete illustration, consider the Type B input in Section 4.5: a tree is described as *in front of* a cottage, while its branches extend *behind* the cottage. In an intuitive reading of the slot layout, (S_1, P_1, E_1) can summarize the tree–"in front of"–existence description, whereas (S_2, P_2, E_2) can summarize the branches–"behind"–extension description. The two groups need not correspond to exact text spans: the branches inherit the tree's identity through context, and each slot may attend to several words. Rather, the layout gives the model two compact places in which to preserve the candidate descriptions before comparing them. For a local Type A or Type C anomaly, the first group can carry the primary entity–position–event evidence and the second can retain contextual evidence; Type B is the setting in which comparison between both groups is especially useful.

The MacBERT-large branch outputs token-level hidden states $\mathbf{H} \in \mathbb{R}^{B \times L_b \times 1024}$ and a padding mask A_b. The aggregation module converts this into six fixed-size slots using query embeddings $\mathbf{q} \in \mathbb{R}^{6 \times 1024}$:

$$\mathbf{R} = \mathrm{LN}(\mathrm{CA}(\mathbf{q}, \mathbf{H}, \mathbf{H}; A_b) + \mathbf{q}) \quad (1)$$

where LN denotes layer normalization and CA denotes 8-head cross-attention masked by A_b. The ISA layer then applies self-attention over the six slots, enabling input-dependent cross-slot interaction before injection. The inspection in Section 4.5 provides qualitative evidence of an entity-centered attention pattern, rather than deterministic symbolic binding:

$$\hat{\mathbf{R}} = \mathrm{LN}(\mathrm{SA}(\mathbf{R}) + \mathbf{R}) \quad (2)$$

where SA is self-attention over the six slots (we refer to the variant omitting this ISA layer as SCP-Pfx). In the example above, ISA gives the "in front of" and "behind" summaries an opportunity to exchange evidence before Qwen generates a label. It is not programmed with a geometric contradiction rule: it may also use the entity and event summaries to decide whether the two phrases concern the same situation. The residual connection retains the sentence-conditioned evidence in each slot, while the attention update can add relational context from the other five slots. Finally, a two-layer MLP maps the slots to Qwen2.5's hidden dimension ($\mathbf{Z} \in \mathbb{R}^{B \times 6 \times 3584}$) for use as soft prefix embeddings. Qwen still receives the original text, so the prefix is a compact structural cue rather than the sole information source for the decision.

4 Experiments

4.1 Dataset and Evaluation

We evaluate on CRSARC, a unified real-text spatial anomaly benchmark that we construct by cleaning and re-integrating annotations from the SpaCE

Split	Anomalous	Normal	Total
Train	4,966	3,710	8,676
Validation	313	462	775
Test	315	467	782

Table 1: CRSARC split statistics after integration.

2021–2025 series (2030NLP, 2021; Xiao et al., 2023, 2024; Qin et al., 2025); the split statistics are summarized in Table 1. We report detection macro-F1 (Det-F1) over normal vs. anomalous. For attribution, we compute cascade attribution F1 (Attr-F1) over all gold anomalous instances with $\{A, B, C\}$ labels (Section 3); if a method predicts a gold anomaly as normal, it counts as an attribution error, so detection failures are directly penalized.

4.2 Experimental Setups

Trainable models use Qwen2.5-7B-Instruct with 4-bit `NF4` quantization via bitsandbytes (Dettmers et al., 2023). LoRA is applied to projection matrices with rank 16, alpha 32, and dropout 0.05. The learning rate is 1×10^{-4} with a batch size of 1 and gradient accumulation of 8, training for 3 epochs. The SCP-ISA modules add approximately 18.3M trainable parameters on top of the LoRA baseline configuration. All models optimize a token-level cross-entropy loss, selecting the best checkpoint via validation Attr-F1. For the controlled comparison below, Direct LoRA and the learned static-prefix baseline are parameter-matched to SCP-ISA's trainable budget.

Prompt-only baselines apply zero-shot (ZS) and in-context learning (ICL) variants, each with chain-of-thought (Wei et al., 2022) (CoT) or majority-vote self-consistency (Wang et al., 2023) (Vote) decoding. ICL uses 4 demonstrations curated from the training set to cover all three anomaly types (A/B/C) with balanced representation; each demonstration includes a complete spatial reasoning chain that identifies conflicting elements and explains the anomaly mechanism. Vote samples 5 outputs at T=0.7. This represents the strongest prompting-only upper bound without parameter updates; its failure (Section 4.3) shows that spatial attribution fundamentally requires structural inductive biases beyond in-context demonstrations. The year suffix (2024/2025) in Table 2 indicates which SpaCE edition's protocol was followed (Xiao et al., 2024; Qin et al., 2025). The top fine-tuned systems of SpaCE2024/2025 report numbers on their own data splits and release

	Detection		Attribution	
Method	Val	Test	Val	Test
ZS-CoT (2024)	0.573	0.591	0.198	0.275
ZS-Vote (2025)	0.573	0.591	0.198	0.241
ICL-CoT (2024)	0.564	0.549	0.198	0.236
ICL-Vote (2025)	0.564	0.549	0.215	0.217
Direct LoRA	0.835	0.846	0.458	0.507
Direct LoRA + CoT	0.790	0.744	0.405	0.403
SCP-ISA (Ours)	**0.878**	**0.881**	**0.688**	**0.736**

Table 2: Main validation/test results. Detection is macro-F1; Attribution is cascade Attr-F1. **Bold** marks the column maximum.[1]

neither training code nor checkpoints; since CRSARC is a re-integrated split (Table 1), their leaderboard numbers are not directly comparable. We therefore reproduce their published prompt configurations under the unified CRSARC protocol and compare against Direct LoRA (LoRA on Qwen2.5 without spatial slot projection) and Direct LoRA + CoT, which augments the LoRA training data with spatial chain-of-thought annotations following Liu et al. (2025); Wu et al. (2024); Cho et al. (2025)—each training target is prefixed with a structured reasoning process that identifies spatial entities, analyzes relations, and checks cross-triple consistency before the final judgment.

4.3 Main Results

Table 2 shows that SCP-ISA achieves the best Attr-F1 and Det-F1. Compared with Direct LoRA, SCP-ISA improves test Attr-F1 from 0.507 to 0.736 and Det-F1 from 0.846 to 0.881. The controlled comparison in Table 3 further distinguishes structured, input-conditioned slot construction from added capacity and a generic learned prefix. Direct LoRA + CoT *degrades* performance relative to vanilla Direct LoRA (test Det-F1 0.744, Attr-F1 0.403), consistent with CoT harming sub-10B models (Liu et al., 2025) by increasing sequence complexity without providing relational structure. Prompt-only baselines mostly fall below the chance-level Attr-F1 of 0.33 (uniform over A/B/C). On the test set (782 instances), paired bootstrap resampling (N=1,000) confirms that SCP-ISA's gains over Direct LoRA are significant on both metrics ($p < 0.001$).

Controlled comparisons. Table 3 tests whether the attribution gain reflects structured conditioning

[1]Prompt-only baselines (ZS/ICL) are our reproductions under the unified CRSARC split, following published SpaCE 2024/2025 protocols; they are not directly comparable to original leaderboard entries.

Condition	Test Attr-F1
Parameter-matched Direct LoRA	0.507
Learned static prefix	0.485
MacBERT pooled prefix	0.545
SCP-ISA	**0.736**

Table 3: Controlled test attribution comparison.

Variant	Detection	Attribution
Direct LoRA	0.835	0.458
SCP-Pfx [Q=1]	0.863	0.572
SCP-Pfx [Q=3]	**0.884**	0.655
SCP-Pfx [Q=6]	0.876	0.524
SCP-Pfx [Q=12]	0.876	0.503
SCP-ISA [Q=6, Qwen-hs]	0.838	0.432
SCP-ISA [Q=6] (Ours)	0.878	**0.688**

Table 4: Validation ablation results (macro-F1 / cascade Attr-F1).

rather than additional parameters or a task-level prefix. Parameter-matched Direct LoRA and a learned static prefix reach 0.507 and 0.485 test Attr-F1, respectively; a MacBERT pooled prefix reaches 0.545. In separate validation controls, role-oriented slots without ISA reach 0.578, ISA raises this to 0.688, and substituting Qwen hidden states for MacBERT states yields 0.562. These controlled variants are distinct from the fixed-Q slot-count ablation in Table 4; together, they support input-conditioned slot construction and inter-slot interaction without establishing exact symbolic roles for individual slots.

4.4 Ablation Study

Slot count and ISA layer. As shown in Table 4, SCP-Pfx denotes our model without the ISA layer. Here Q denotes the number of learnable queries (slots): Q=3 corresponds to one spatial triple (S, P, E), Q=6 to two triples (our default); Q=1 collapses all information into a single slot, and Q=12 doubles each role. Increasing Q without ISA yields non-monotonic attribution gains. Although SCP-Pfx [Q=3] reaches a marginally higher Val Det-F1 (0.884 vs. 0.878), its Attr-F1 is substantially lower (0.655 vs. 0.688); we adopt Q=6 with ISA as our default. Adding ISA raises validation Attr-F1 from 0.524 to 0.688, confirming that role-specific slots must communicate before injection.

Encoder choice and backbone. Critically, replacing MacBERT-large with Qwen2.5's own hidden states as the slot source (denoted Qwen-hs) drops Attr-F1 from 0.688 to 0.432—below Direct LoRA (0.458)—while Det-F1 also falls to 0.838. This isolates the necessity of an independent bidirectional encoder for spatial slot construction; the ISA layer alone cannot compensate for the lack of fine-grained role information in the causal hidden states of the generation backbone. As shown in Table 5, Qwen2.5 achieves the best Attr-F1 among backbones, suggesting better compatibility between its embedding space and our projected slot vectors compared to LLaMA 3.1 or Mistral; InternLM2.5 attains a slightly higher Test Det-F1 (0.884 vs. 0.881), but we keep Qwen2.5 as the default backbone for its stronger attribution performance.

	Detection		Attribution	
Backbone	Val	Test	Val	Test
LLaMA 3.1-8B-Instruct	0.813	0.829	0.587	0.615
Mistral-7B-Instruct-v0.3	0.798	0.803	0.572	0.598
InternLM2.5-7B-Chat	0.872	**0.884**	0.683	0.729
Qwen2.5-7B-Instruct (Default)	**0.878**	0.881	**0.688**	**0.736**

Table 5: Backbone ablation for SCP-ISA [Q=6]. Only the generation backbone changes; MacBERT-large is retained for slot encoding.

4.5 Case Study and ISA Inspection

Coverage across anomaly types. The scope is not limited to Type B cross-triple conflicts: test type-wise F1 is 0.577, 0.791, and 0.839 for Type A, Type B, and Type C, respectively. These results support coverage across the three CRSARC attribution types, while remaining within the controlled sentence-level setting.

Case study. Table 6 contrasts model predictions on a real Type B example from CRSARC:

> *"There is a tall tree in front of the cottage, dense branches extend behind the cottage, and the thick shade envelops the cottage."*

The same tree is described as being both *in front of* and *behind* the cottage—two spatially incompatible positions, constituting a Type B cross-triple conflict. SCP-Pfx misses the anomaly entirely; Direct LoRA detects it but attributes to Type C, treating "behind" as a commonsense violation in isolation; SCP-ISA correctly identifies the cross-triple conflict and assigns Type B.

Interpreting the prediction difference. The three explanations suggest a plausible division of labor. Direct LoRA sees the full sentence but must recover entity binding and the relation between the

Method	Prediction	Explanation (excerpt)
Direct LoRA	Anomaly, Type C	Branches cannot extend "behind" the cottage; violates commonsense.
SCP-Pfx [Q=6]	Normal	— (no anomaly produced)
SCP-ISA (Ours)	Anomaly, **Type B**	"in front of" vs. "behind" contradict each other for the same tree.

Table 6: Case study on a Type B input from CRSARC (id `2022-task2-2-train-366`). The input describes the same tree as both *in front of* and *behind* the cottage, two spatially incompatible positions.

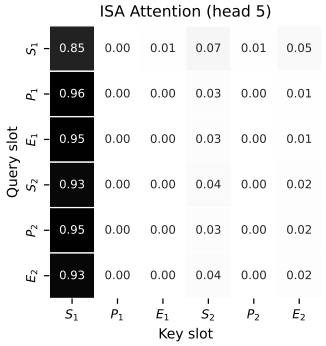


Figure 2: ISA layer attention (head 5) on a representative Type B test sample. The displayed head shows a strong entity-centered attention pattern toward S_1.

two descriptions only within the decoder; its rationale instead isolates "behind" and treats it as a local Type C issue. SCP-Pfx receives input-conditioned slot summaries, yet without ISA it may preserve the two descriptions without explicitly reconciling them, which is consistent with its normal prediction in this example. SCP-ISA names both phrases and their shared entity in its explanation, as expected when relational information has been made available to the decoder through interacting slots. This single example is not causal evidence that a particular head performs symbolic comparison; rather, it motivates the hypothesis that cross-slot communication helps the model shift from judging an individual phrase to judging the compatibility of two descriptions.

ISA inspection. We extract the ISA 6×6 attention matrix on all Type B test samples (8 heads each). A recurring pattern is that query slots often direct their maximum attention toward S_1 (Figure 2), forming an *entity-centered attention pattern*. The S_1 column accumulates more than 0.40 of the attention mass in over 95% of head-sample combinations (max 0.97). Probing on 30 instances with gold spans finds that the entity-oriented slots align with their gold spans at $7.24\times$ (S_1) and $4.14\times$ (S_2) the off-diagonal baseline (2.69), while position- and event-oriented slots stay near baseline (1.63–2.45). This is compatible with structured spatial conditioning, but does not prove a causal mechanism or complete slot-wise semantic disentanglement.

5 Conclusion

We presented SCP-ISA, which injects input-conditioned, role-oriented latent slots from MacBERT-large into Qwen2.5 and applies inter-slot attention before generation. Controlled comparisons support the value of structured spatial conditioning over matched capacity and generic prefixes on CRSARC. These findings concern controlled sentence-level anomaly checking; they do not establish complete symbolic parsing, full slot-wise disentanglement, or open-domain spatial-claim validation.

Limitations

Our work has three main limitations. First, the dual-group design does not natively extend to inputs with three or more potentially conflicting claims; longer inputs and complex reference chains remain challenging. Dynamic slots, retrieval of supporting context, and explicit coreference handling are promising directions. Second, CRSARC supports a controlled sentence-level evaluation, not a measurement of Web-scale prevalence or an open-domain deployment claim. Third, the SCP-ISA modules add 18.3M trainable parameters (0.24% of Qwen2.5-7B) and require an additional frozen MacBERT-large forward pass at inference time. We benchmark on a single T4 GPU (batch size 1, sequence length 256): the MacBERT encoding adds 57 ms per sample, and end-to-end generation latency increases by 163 ms (4.6% relative to Direct LoRA), while peak GPU memory remains effectively unchanged (+0.01 GB). This added cost should be weighed against the attribution improvement for the intended controlled quality-control setting.

Acknowledgements

This research was supported by the National Key Research and Development Program of China (Grant No. 2024YFC3307502) and the Excellent Postdoctoral Program of Nanjing.

References

2030NLP. 2021. SpaCE2021: The first chinese spatial cognition evaluation. https://github.com/2030NLP/SpaCE2021.

Rishika Bhagwatkar, Syrielle Montariol, Angelika Romanou, Beatriz Borges, Irina Rish, and Antoine Bosselut. 2025. CAVE: Detecting and explaining commonsense anomalies in visual environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27110–27151, Suzhou, China. Association for Computational Linguistics.

Ikhyun Cho, Changyeon Park, and Julia Hockenmaier. 2025. The power of bullet lists: A simple yet effective prompting approach to enhancing spatial reasoning in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3047–3057.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. CRoW: Benchmarking commonsense reasoning in real-world tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9785–9821.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4582–4597.

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

Haixin Liu, Hongying Zan, Jinwang Song, Yifan Li, and Lulu Kong. 2025. System report for CCL25-Eval task 1: Spatial semantic understanding using chain-of-thought and voting ensemble enhanced large language models. In *Proceedings of the 24th China National Conference on Computational Linguistics (CCL 2025)*, pages 24–32.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 61–68.

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

Yuhang Qin, Liming Xiao, Nan Hu, Sirui Deng, Jingyuan Ma, Hyang Cui, Zihan Zhang, Chi Hsu Tsai, Jinkun Ding, Sumin Kang, Zhifang Sui, and Weidong Zhan. 2025. Overview of CCL25-Eval task 1: The fifth spatial cognition evaluation (SpaCE2025). In *Proceedings of the 24th China National Conference on Computational Linguistics*, pages 33–46, Jinan, China. Chinese Information Processing Society of China.

Md Imbesat Rizvi, Xiaodan Zhu, and Iryna Gurevych. 2024. SpaRC and SpaRP: Spatial reasoning characterization and path generation for understanding spatial reasoning capability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4750–4767. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 37.

Liming Xiao, Nan Hu, Weidong Zhan, Yuhang Qin, Sirui Deng, Chunhui Sun, Qixu Cai, and Nan Li. 2024. The fourth evaluation on chinese spatial cognition. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 122–134, Taiyuan, China. Chinese Information Processing Society of China.

Liming Xiao, Chunhui Sun, Weidong Zhan, Dan Xing, Nan Li, Chengwen Wang, and Fangwei Zhu. 2023. 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, Harbin, China. Chinese Information Processing Society of China.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 2 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*.