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

Anonymous ACL submission

Abstract

Textual spatial semantic anomaly detection and attribution requires deciding whether a sentence contains an implausible spatial relation and assigning the corresponding attribution label. Prior prompting and Low-Rank Adaptation (LoRA) fine-tuning methods rely on raw token-level representations, leaving spatial roles and cross-triple conflicts implicit. 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 spatial slots and injects the projected slots into Qwen2.5 as soft prefix embeddings. The ISA layer enables structured cross-slot interaction that collapses co-referring spatial mentions into a shared entity-anchored representation before generation. On the CRSARC benchmark, SCP-ISA improves test cascade attribution F1 (Attr-F1) by 22.7 percentage points (0.509 to 0.736) over Direct LoRA, and detection macro-F1 (Det-F1) from 0.846 to 0.881.

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. Moreover, most prior evaluation settings assume the input is already anomalous, leaving the mixed normal/anomalous scenario underexplored. 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 represents each spatial configuration as a triple (S_i, P_i, E_i) realized as a slot group, derives six such slots from the input via a frozen MacBERT-large encoder, injects them into Qwen2.5 as soft prefix embeddings, and lets the two slot groups interact via the ISA layer—providing a structured inductive bias for entity binding that enables downstream relational reasoning.

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 role-specific slots $(S_1, P_1, E_1, S_2, P_2, E_2)$, where S is

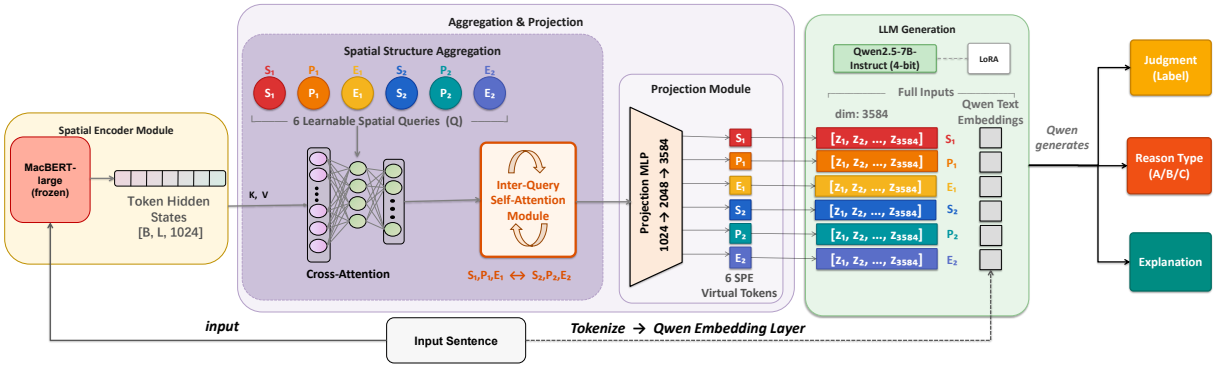

Figure 1: Overall architecture of SCP-ISA. MacBERT-large extracts six spatial slots, which are projected and prepended to Qwen2.5 text embeddings as soft prefix embeddings for generation.

the spatial entity, P the position, E the event trigger, and the subscript $i \in \{1, 2\}$ indexes two potentially conflicting spatial triples. The six-slot design provides a unified 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 enables explicit modeling of Type B conflicts; for Type A/C inputs, the second slot group (S_2, P_2, E_2) serves as a learnable auxiliary buffer whose contribution is down-weighted via attention. Third, slots are projected into Qwen2.5-7B-Instruct's (Yang et al., 2024) embedding space and prepended as soft prefix embeddings. Role assignment is induced by query position and the ISA layer (described below) rather than by explicit per-slot supervision.

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, structuring cross-slot dependencies before injection. As shown in Section 4.5, this manifests as an *entity-anchored attention sink* that binds co-referring mentions into a shared representation:

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

where SA is self-attention over the six slots (we refer to the variant omitting this ISA layer as SCP-Pfx). 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.

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.

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 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.

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 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 0.509 to 0.736 and Det-F1 from 0.846 to 0.881; the larger gain on attribution confirms that explicit cross-triple role comparison requires structural biases missed by token-level attention. 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. SCP-ISA thus offers an effective alternative for lightweight LLMs, injecting spatial structural biases directly. 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 ($p < 0.001$).

	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	0.509
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]

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

4.4 Ablation Study

Slot count and ISA layer. As shown in Table 3, 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

[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.

Backbone	Detection Val	Detection Test	Attribution Val	Attribution 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 4: Backbone ablation for SCP-ISA [Q=6]. Only the generation backbone changes; MacBERT-large is retained for slot encoding.

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 5: 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.

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 4, 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.

4.5 Case Study and ISA Inspection

Case study. Table 5 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.

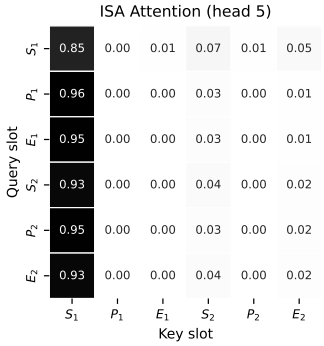

Figure 2: ISA layer attention (head 5) on a representative Type B test sample. Every query slot attends almost exclusively to S_1 (primary entity), forming an entity-anchored attention sink.

ISA inspection. We extract the ISA 6×6 attention matrix on all Type B test samples (8 heads each). A consistent pattern emerges: every query slot directs its maximum attention toward S_1 (the primary entity slot), forming an *entity-anchored attention sink* (Figure 2). The S_1 column accumulates more than 0.40 of the attention mass in over 95% of head-sample combinations (max 0.97); entity binding thus collapses co-referring mentions into a shared representation, enabling the decoder to detect positional contradictions via the residual pathway. Probing on 30 instances with gold spans confirms this: entity 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 slots stay near baseline (1.63–2.45).

5 Conclusion

We presented SCP-ISA, showing that MacBERT-large spatial slots with an inter-slot attention layer, injected as soft prefix embeddings into Qwen2.5, improve detection and attribution on CRSARC. Ablations also reveal that backbone choice affects attribution, with Qwen2.5 yielding the best Attr-F1 among comparable models. Together, these findings suggest that slot-level relational reasoning combined with a compatible backbone provides an effective inductive bias for spatial constraints that token-level representations alone cannot recover.

Limitations

Our work has two main limitations. First, the dual-group design models two spatial triples and does not natively extend to inputs with three or more conflicting triples; handling such cases would require either a hierarchical slot organization or a

dynamically-sized slot pool. Second, 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). Given the +22.7-point Attr-F1 improvement, we consider this a favorable accuracy–efficiency trade-off.

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. Enhancing spatial semantic understanding of large language models using chain-of-thought and voting ensemble. In *Proceedings of the 24th China National Conference on Computational Linguistics (Volume 2: Frontier Forum and Evaluation)*, pages 47–58.

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*.