Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.083755

ARTICLE

Can Causal Circuits Enable Efficient Spatial Reasoning for Geoinformatics in Large Language Models?

Sahil Tripathi[1], **Manaswi Kulahara**[2], **Abdul Khader Jilani Saudagar**[3] **and Hatoon S. AlSagri**[3,*]

[1]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India
[2]Department of Geoinformatics, TERI School of Advanced Studies, Delhi, India
[3]Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

*Corresponding Author: Hatoon S. AlSagri. Email: hsagri@imamu.edu.sa

Received: 09 April 2026; Accepted: 15 June 2026; Published: 27 July 2026

ABSTRACT: Spatial reasoning, defined as the ability to infer and compose relations among entities—is fundamental to geoinformatics applications such as spatial querying and map understanding. However, existing works such as Bidirectional Encoder Representations from Transformers (BERT)-based spatial Question Answering (QA) models and neuro-symbolic models rely on dataset-specific patterns, leading to *shortcut learning*, where reliance on superficial lexical cues rather than true relational understanding. Recent Large Language Models (LLMs)-based works, including fine-tuning and Chain-of-Thought (CoT) prompting, partially alleviate *shortcut learning* but remain limited by *non-causal reasoning*, where predictions depend on spurious correlations rather than stable relational structure. To address these limitations, we propose `Causal Inference and Reasoning via Compact sUbnetwork IdenTification (CIRCUIT-X)`, motivated by the hypothesis that spatial reasoning in LLMs is governed by compact causal parameter subsets (a.k.a *causal circuits*). `CIRCUIT-X` operates in two stages: (i) Causal Importance Estimation (Stage I) via structured interventions to mitigate *shortcut learning*, and (ii) Minimal Circuit Discovery (Stage II) via structured pruning to mitigate *non-causal reasoning*. Empirically, `CIRCUIT-X` achieves up to 91% accuracy on SPAtial Reasoning on Textual Question Answering (SPARTQA) and 87% on StepGame, outperforming State-of-the-Art (SOTA) methods while improving intervention robustness by up to +11% and causal consistency by up to +16%. Therefore, it retains up to 96% of full-model performance using only 3%–6% of active parameters, while demonstrating strong robustness under cross-domain transfer with improvements of up to +10% in accuracy and substantially higher intervention stability under distribution shifts.

KEYWORDS: Causal circuits; geoinformatics; large language models; parameter-efficient learning; spatial reasoning; structured pruning

1 Introduction

Spatial reasoning, defined as the ability to infer and compose relational structures among entities [1,2], is a fundamental capability for geoinformatics applications such as spatial querying [3,4], map interpretation [5], and geographic knowledge reasoning [6,7]. For example, answering queries like "What lies north of the river and east of the city center?" requires composing multiple spatial relations across entities. In natural language settings, this involves multi-hop relational inference, where models must integrate intermediate constraints while remaining robust to paraphrasing, ambiguity, and incomplete spatial descriptions (see Fig. 1).

Limitations of Existing Spatial Reasoning vs. CIRCUIT-X

Input Spatial Query	Existing LLM Behavior	CIRCUIT-X Behavior
Case 1: Multi-Hop Composition What lies north of the river and east of the city center?	**Shortcut Learning** Predicts: east of city center ✗ Uses dominant lexical cue only	**Causal Relational Reasoning** Predicts: northeast of city center ✓ Correctly composes north + east relations
Case 2: Paraphrased Spatial Relations Which area is south of the park and west of the lake?	**Non-Causal Reasoning** Predicts: west of lake ✗ Relies on dominant relation	**Intervention-Stable Reasoning** Predicts: southwest of lake ✓ Preserves relational structure under paraphrasing
Case 3: Missing Spatial Cues Village north of forest. Forest north of river. (River cue partially removed)	**Shortcut-Based Inference** Predicts: village north of river ✗ Assumes unsupported direct relation	**Robust Multi-Hop Inference** Predicts: village north of forest (indirectly north of river) ✓ Infers stable relational dependency

Key Insight: Existing LLMs often rely on superficial spatial cues or dominant lexical patterns, leading to shortcut learning and non-causal reasoning. In contrast, CIRCUIT-X identifies compact intervention-stable parameter subsets that preserve robust relational reasoning under paraphrasing, missing cues, and multi-hop composition.

Figure 1: Illustration of spatial reasoning failures in existing LLMs and the motivation behind CIRCUIT-X. Existing models frequently rely on dominant lexical cues or shortcut patterns, leading to incorrect predictions under multi-hop composition, paraphrasing, and incomplete spatial information. In contrast, CIRCUIT-X identifies compact intervention-stable subnetworks that preserve relational structure and improve robustness under spatial interventions.

However, existing works such as Bidirectional Encoder Representations from Transformers (BERT)-based spatial Question Answering (QA) models [8] and neuro-symbolic models [9], despite achieving strong in-distribution accuracy, are largely driven by dataset-specific patterns. This leads to *shortcut learning*, where models rely on superficial lexical cues (e.g., "left of", "above") or rigid templates instead of reasoning over underlying relational structure. Furthermore, these models often fail under distribution shifts [10], paraphrased inputs [11], or when spatial cues are partially removed [12], indicating limited generalization beyond benchmark settings. Recent Large Language Model (LLM)-based works, including fine-tuning [13] and Chain-of-Thought (CoT) prompting [14], have shown improvements in spatial reasoning tasks by encouraging intermediate reasoning steps. However, these works remain fundamentally limited by *non-causal reasoning*, where predictions are driven by spurious correlations rather than stable relational dependencies. As a result, such works show poor cross-dataset transfer, limited interpretability [15], and high computational overhead [16], making them less reliable for real-world geoinformatics applications.

To address these limitations, we propose `Causal Inference and Reasoning via Compact sUbnetwork IdenTification (CIRCUIT-X)`, motivated by the hypothesis that spatial reasoning in LLMs is governed by compact, causally grounded parameter subsets rather than the full model (a.k.a *causal circuits*). This perspective shifts the focus from full-model optimization to identifying the minimal components responsible for reasoning behavior. The `CIRCUIT-X` framework operates in two stages. In Stage I (Causal Importance Estimation), we use structured interventions to identify parameter groups that contribute to mitigate *shortcut learning*. In Stage II (Minimal Circuit Discovery), we perform iterative structured pruning to isolate a compact subnetwork that captures stable relational dependencies to mitigate *non-causal reasoning*. In summary, the main contributions of this work are summarized as follows:

- We propose `CIRCUIT-X`, a two stage causal circuit-based framework that mitigates *shortcut learning* and *non-causal reasoning* by identifying and isolating compact parameter subsets responsible for multi-hop spatial reasoning.
- Empirically, we demonstrate that `CIRCUIT-X` consistently improves performance across SPAtial Reasoning on Textual Question Answering (SPARTQA) and StepGame, achieving up to 91% and 87%

accuracy, respectively, while substantially improving intervention robustness and causal consistency compared to State-of-the-Art (SOTA) methods. Despite activating only 3%–6% of model parameters, CIRCUIT-X retains up to 96% of full-model performance and shows strong generalization under cross-domain transfer, improving robustness and reasoning stability across distribution shifts.

The work is organized as follows: Section 2 presents related work, Section 3 describes the proposed methodology, Section 4 outlines the experimental setup, Section 5 discusses results, Section 6 provides ablation studies, Section 7 offers additional analysis, and Section 8 concludes the work.

2 Related Works

To summarize the limitations of prior works and position our work-Table 1 compares existing works with CIRCUIT-X across key properties.

Table 1: Comprehensive comparison of general spatial reasoning and geoinformatics-oriented approaches with CIRCUIT-X across key reasoning and robustness properties. "Multi-Hop Reasoning" denotes the ability to compose multiple spatial relations, "Cross-Dataset Generalization" measures robustness across different benchmarks, "Interpretability" indicates whether the model provides transparent reasoning behavior, "Parameter Efficiency" refers to achieving strong performance using compact parameter subsets, "Reduced Shortcut Reliance" measures resistance to superficial lexical cues, and "Intervention Stability" measures robustness under relational perturbations and paraphrased spatial interventions. The abbreviated column headers correspond to: Multi-Hop. = Multi-Hop Reasoning, Cross-Dataset. = Cross-Dataset Generalization, Interp. = Interpretability, Param. Eff. = Parameter Efficiency, Reduced Shortcut. = Reduced Shortcut Reliance, and Intervention Stable. = Intervention Stability.

Method	Multi-Hop.	Cross-Dataset.	Interp.	Param. Eff.	Reduced Shortcut.	Intervention Stable.
General Spatial Reasoning						
BERT-based Models [8]	✗	✗	✗	✓	✗	✗
BiLSTM-based Models [17]	✗	✗	✗	✓	✗	✗
Relation Classification Models [18]	✗	✗	✗	✓	✗	✗
PISTAQ [9]	✓	✗	✓	✗	✗	✗
SREQA [9]	✓	✗	✓	✗	✗	✗
NSM-based Models [19]	✓	✗	✓	✗	✗	✗
GPT-4 [20]	✓	✗	✗	✗	✗	✗
LLaMA-based Models [21]	✓	✗	✗	✗	✗	✗
Instruction-Tuned Models [22]	✓	✗	✗	✗	✗	✗
Fine-Tuned LLMs [13]	✓	✗	✗	✗	✗	✗
CoT [14]	✓	✗	✗	✗	✗	✗
Self-Consistency [23]	✓	✗	✗	✗	✗	✗
Geoinformatics-Oriented Spatial Reasoning						
GIS/Spatial Databases (PostGIS) [24]	✓	✗	✓	✓	✓	✓
GeoQA Models [25]	✓	✗	✗	✗	✗	✗
Geospatial Knowledge Graph Models [26]	✓	✗	✗	✗	✗	✗
GNN Geospatial Models [27]	✓	✗	✗	✗	✗	✗
Remote Sensing Reasoning Models [28]	✗	✗	✗	✗	✗	✗
Multimodal Geospatial Models [29]	✗	✗	✗	✗	✗	✗
LLM-based Geospatial QA [30]	✓	✗	✗	✗	✗	✗
Multimodal LLMs (Geospatial) [31]	✓	✗	✗	✗	✗	✗
CIRCUIT-X	✓	✓	✓	✓	✓	✓

2.1 General Spatial Reasoning

Spatial reasoning in natural language has been extensively studied through QA benchmarks such as SPARTQA [32] and StepGame [33], which require multi-hop relational inference and compositional reasoning across entities. These benchmarks have served as standard testbeds for evaluating a model's ability to infer implicit spatial relations under varying levels of complexity. Early neural works, including BERT-based models [8], Bidirectional Long Short-Term Memory (BiLSTM)-based reasoning models [17], and relation classification models [18], achieved strong in-distribution performance by learning statistical patterns from training data. However, these models primarily rely on surface-level correlations and struggle to generalize beyond seen patterns. To address this, neuro-symbolic works introduced structured reasoning pipelines that model spatial relations. Models such as PIpeline model for SpaTiAl Question answering (PISTAQ) [9], Spatial Relation Extraction for QA (SREQA) [9], and Neural Symbolic Machine (NSM)-based models [19] leverage symbolic rules or intermediate representations to improve compositional reasoning. While these works achieve higher accuracy under controlled settings, they depend on task-specific designs and lack scalability to open-domain scenarios. More recently, LLMs, including GPT-4 [20], LLaMA-based models [21], and instruction-tuned variants [22], have been explored for spatial reasoning tasks. Furthermore, works such as fine-tuning [13], CoT prompting [14], and self-consistency decoding [23] further enhance performance by encouraging intermediate reasoning steps. Yet, despite these improvements, the LLMs are brittle when exposed to paraphrased input, missing spatial cues, or cross-dataset transfer. Across all of these paradigms, a common limitation remains as discussed in Section 1-models relieve heavily on data set specific patterns, allowing *shortcut learning*, and even when intermediate reasoning is encouraged, models show non-causal reasoning. All these shortcomings highlight the need for causally grounded and parameter efficient mechanisms for robust spatial reasoning, as addressed by `CIRCUIT-X`.

2.2 Geoinformatics-Oriented Spatial Reasoning

In geoinformatics, spatial reasoning is fundamental for tasks and has been extensively studied through traditional models, including Geographic Information Systems (GIS) [34] and spatial database engines [35], which rely on rule-based reasoning and spatial query languages such as Structured Query Language (SQL) extensions (e.g., PostGIS [24]). These models encode spatial relations (e.g., topological and directional constraints) and support precise querying, but lack flexibility in handling natural language inputs and complex multi-hop reasoning. To improve flexibility, recent works have integrated machine learning with geospatial data. Works such as GeoQA models [25], spatial knowledge graph reasoning models [2], and Graph Neural Network (GNN)-based geospatial models [27] attempt to learn spatial dependencies from structured data. Similarly, remote sensing-based reasoning models [28] and multimodal geospatial learning models [29] incorporate visual and geographic features to enhance spatial understanding. However, these models typically rely on structured inputs, predefined ontologies, or modality-specific representations, limiting their ability to generalize to open-ended natural language reasoning tasks. With the emergence of LLMs, there has been growing interest in applying language models to geospatial reasoning. Recent efforts include LLM-based geographic QA [31], map-based reasoning with language models [36], and multimodal LLMs for geospatial understanding [37]. While these works improve flexibility in handling natural language queries, they inherit limitations from general spatial reasoning systems. In particular as discussed in Section 1, they show *shortcut learning*, relying on dominant spatial cues, and *non-causal reasoning*, where predictions are driven by spurious correlations rather than stable relational dependencies. These limitations highlight the need for models that can capture causally grounded spatial dependencies while remaining efficient and interpretable, as achieved by `CIRCUIT-X`.

3 Methodology

This section presents the proposed CIRCUIT-X framework to address *shortcut learning* and *non-causal reasoning* (see Fig. 2). We begin by formalizing the spatial reasoning problem to support our hypothesis of compact causal circuits.

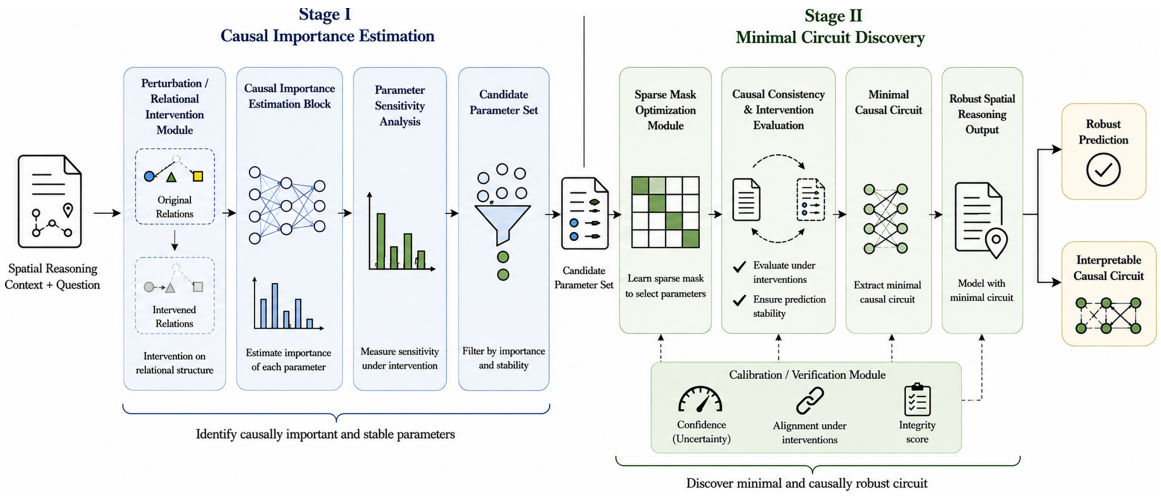

Figure 2: Overview of the proposed CIRCUIT-X framework for causal circuit-based spatial reasoning. Given an input (C, Q), the full model M_θ is first analyzed in Stage I, where structured interventions on the relational structure $\mathcal{R}$ are used to estimate parameter importance. Sensitivity (Δ_i) and causal stability $(\Delta_i^{\text{causal}})$ are computed to identify a subset θ_s that mitigates *shortcut learning*. In Stage II, a sparsity-constrained optimization is applied over θ_s to obtain a minimal subnetwork θ_c, implementing robustness under interventions and mitigating *non-causal reasoning*. The resulting causal circuit θ_c for efficient, interpretable, and generalizable multi-hop spatial reasoning.

3.1 Problem Formulation

We consider the task of text-based spatial reasoning (see Fig. 1), as it provides a natural and controlled setting to study multi-hop relational inference in LLMs, which is central to geoinformatics applications. Intuitively, the task requires a model to understand how multiple spatial relations connect together (e.g., if object A is left of B, and B is left of C, then A is left of C), similar to how humans reason about maps, locations, and directions. In this task, a model must infer spatial relations by composing multiple relational statements expressed in natural language.

Formally, each instance consists of a context, question pair (C, Q) and a target answer A. The context C describes a set of entities $E = \{e_1, \ldots, e_n\}$ and a set of spatial relations $\mathcal{R}$ among them, while the question Q queries the relation between a target pair $(e_i, e_j) \in E$, often requiring multi-hop relational composition. For example, the context may describe the relative positions of landmarks or objects, while the question asks the model to infer a new relation that is not stated but must be derived through reasoning. Let M_θ denote a pretrained LLM with parameters $\theta \in \mathbb{R}^d$. Given an input $x = [C; Q]$, the model produces Eq. (1).

$$P(A \mid C, Q; \theta) = M_\theta(x) \tag{1}$$

Existing works (e.g., [9,24–27,31]) improve performance either by optimizing the full parameter set θ (e.g., fine-tuning, denoted as $\theta \to \theta'$) or by modifying inference via prompting strategies π (e.g., CoT), yielding Eq. (2). In simple terms, these approaches attempt to improve reasoning either by updating the entire model through training or by guiding the model using carefully designed prompts and reasoning steps.

$$P(A \mid C, Q; \theta', \pi) \tag{2}$$

However, such works often rely on *shortcut learning*, where the model exploits spurious features $\phi(x)$ (e.g., lexical cues) instead of true relational structure as shown in Eq. (3), and show *non-causal reasoning*, where predictions depend on correlations that are not invariant under interventions on the relational structure. For instance, a model may incorrectly associate words such as "left of" or "above" with a specific answer pattern without actually understanding the underlying spatial arrangement. As a result, small changes in wording or relation order may cause incorrect predictions even when the spatial meaning remains unchanged. Formally, for an intervention $\text{do}(\mathcal{R})$ on the underlying relational structure, a causally grounded model should satisfy as per Eq. (4), whereas non-causal models as per Eq. (5).

$$P(A \mid x; \theta) \approx P(A \mid \phi(x); \theta) \tag{3}$$

$$P(A \mid \text{do}(\mathcal{R}); \theta) = P(A \mid \mathcal{R}; \theta) \tag{4}$$

$$P(A \mid \text{do}(\mathcal{R}); \theta) \neq P(A \mid \mathcal{R}; \theta) \tag{5}$$

In contrast, we hypothesize that spatial reasoning in LLMs is governed by a compact subset of parameters $\theta_c \subset \theta$, referred to as a *causal circuit*, which captures stable relational dependencies. Intuitively, instead of relying on the entire model, we hypothesize that only a small subset of parameters is primarily responsible for robust spatial reasoning, and identifying these parameters can improve both interpretability and robustness. Our objective is to identify a minimal subset θ_c as per Eq. (6).

$$P(A \mid C, Q; \theta_c) \approx P(A \mid C, Q; \theta) \tag{6}$$

3.2 Proposed Framework

We propose `CIRCUIT-X`, a two-stage framework that identifies a compact causal circuit $\theta_c \subset \theta$ responsible for multi-hop spatial reasoning. The framework builds upon the hypothesis formulation in Section 3, where the objective is to approximate the full model behavior using a minimal subset of parameters (see Eq. (6)).

3.2.1 Stage I: Causal Importance Estimation

The goal of Stage I is to identify a subset of parameters $\theta_s \subset \theta$ that are causally relevant for spatial reasoning while being robust to shortcut features. Intuitively, this stage attempts to determine which parts of the model genuinely contribute to spatial reasoning, rather than simply memorizing common wording patterns or dataset-specific cues. We begin by noting that model predictions can be implicitly influenced by both relational structure $\mathcal{R}$ and spurious features $\phi(x)$ as shown in Eq. (7).

$$P(A \mid x; \theta) = P(A \mid \mathcal{R}, \phi(x); \theta) \tag{7}$$

A causally grounded model should depend primarily on $\mathcal{R}$, whereas shortcut-driven behavior arises when predictions are dominated by $\phi(x)$. To distinguish these effects, we construct perturbed inputs $\tilde{x}$ by intervening on the relational structure $\mathcal{R}$ (e.g., removing or altering spatial cues), while keeping other components unchanged. For example, if a sentence contains "left of" or "above", we modify or remove these cues to test whether the model still understands the underlying spatial relationship rather than relying on keywords alone. This approximates an intervention $\text{do}(\mathcal{R})$ as introduced in Eqs. (4) and (5). For each parameter (or parameter group) θ_i, we measure its causal importance via sensitivity under intervention as per Eq. (8), where $\theta^{(i \to 0)}$ denotes the model with θ_i ablated.

$$\Delta_i = \mathbb{E}_x\Big[\big|P(A \mid x;\theta) - P(A \mid \tilde{x};\theta^{(i\to 0)})\big|\Big] \tag{8}$$

To further isolate causal contributions, we compare sensitivity under original and intervened inputs as per Eq. (9).

$$\Delta_i^{\text{causal}} = \mathbb{E}_x\Big[\big|P(A \mid x;\theta^{(i\to 0)}) - P(A \mid \tilde{x};\theta^{(i\to 0)})\big|\Big] \tag{9}$$

Parameters that show high Δ_i and maintain stability under Δ_i^{causal} are considered causally important. In simple terms, we retain parameters that strongly influence reasoning performance while remaining stable under relational changes, since such behavior suggests more reliable spatial understanding. We therefore define the candidate set as per Eq. (10), where τ_1 and τ_2 control importance and stability, respectively.

$$\theta_s = \Big\{\theta_i \in \theta \;\Big|\; \Delta_i > \tau_1 \;\wedge\; \Delta_i^{\text{causal}} < \tau_2\Big\} \tag{10}$$

This formulation filters out parameters that are primarily sensitive to spurious features $\phi(x)$, while retaining those that capture stable relational dependencies. The resulting subset θ_s serves as the input to Stage II.

3.2.2 Stage II: Minimal Circuit Discovery

The goal of Stage II is to identify a minimal subset $\theta_c \subset \theta_s$ that preserves the predictive behavior of the full model while ensuring causal robustness. Intuitively, this stage attempts to compress the reasoning process into a much smaller subset of parameters without significantly reducing performance. Starting from the candidate set θ_s obtained in Stage I, we introduce a binary mask $m \in \{0,1\}^{|\theta_s|}$ over parameters, where $m_i = 1$ indicates that θ_i is retained. The pruned model is defined as per Eq. (11), where $\odot$ denotes element-wise masking.

$$M_{\theta\odot m}(x) \tag{11}$$

We formulate circuit discovery as a constrained optimization problem as per Eq. (12), where $\|m\|_0$ for sparsity and ϵ controls allowable performance degradation.

$$\min_m \|m\|_0 \quad \text{s.t.} \quad \mathbb{E}_x\big[\ell(M_{\theta\odot m}(x), A)\big] \le \epsilon \tag{12}$$

Since the ℓ_0 constraint is combinatorial, we relax it using a weighted objective as per Eq. (13), where $\|m\|_1$ promotes sparsity and λ_1 controls the trade-off.

$$\mathcal{L}(m) = \mathbb{E}_x\big[\ell(M_{\theta\odot m}(x), A)\big] + \lambda_1\|m\|_1 \tag{13}$$

To enforce causal consistency, we further incorporate robustness under relational interventions as per Eq. (14), where $\tilde{x}$ denotes inputs under interventions on $\mathcal{R}$, and λ_2 for invariance as defined in Eqs. (4) and (5). This encourages the model to produce stable predictions even when the wording or relational expressions are modified, thereby reducing sensitivity to superficial correlations.

$$\mathcal{L}_{\text{total}}(m) = \mathbb{E}_x\big[\ell(M_{\theta\odot m}(x), A)\big] + \lambda_2\,\mathbb{E}_{\tilde{x}}\big[\ell(M_{\theta\odot m}(\tilde{x}), A)\big] + \lambda_1\|m\|_1 \tag{14}$$

Finally, the causal circuit is obtained as per Eq. (15).

$$\theta_c = \theta \odot m^*, \quad \text{where} \quad m^* = \arg\min_m \mathcal{L}_{\text{total}}(m) \tag{15}$$

This formulation ensures that the resulting subnetwork θ_c is both minimal and causally robust, capturing stable relational dependencies while discarding parameters associated with *shortcut learning* and *non-causal reasoning*.

4 Experimental Setup

4.1 Datasets

We evaluate CIRCUIT-X on two widely used benchmarks for text-based spatial reasoning, **SPARTQA** (https://huggingface.co/datasets/tasksource/spartqa-mchoice) [32] and **StepGame** (https://huggingface.co/datasets/tasksource/stepgame) [33], chosen to systematically directly align with our objective of mitigating *shortcut learning* and *non-causal reasoning*. We use the official train, validation, and test splits for both datasets, following prior works [32,33], which approximately correspond to a 70%/15%/15% split for training, validation, and testing, respectively. We do not introduce any additional annotations or modify the original data partitions. Table 2 summarizes key dataset statistics. **SPARTQA** [32] is a synthetic spatial QA benchmark consisting of natural language descriptions of scenes involving multiple entities and spatial relations (e.g., directional and topological). It is designed to evaluate compositional reasoning under incomplete or implicit spatial cues, making it suitable for assessing robustness against shortcut features $\phi(x)$. **StepGame** [33] focuses on multi-hop spatial reasoning over grid-based relational descriptions, where each instance requires composing multiple directional relations across entities. Compared to **SPARTQA** [32], **StepGame** [33] involves longer reasoning chains and stricter compositional constraints, making it particularly suitable for evaluating causal consistency under relational perturbations.

Table 2: Statistics of the spatial reasoning benchmarks used in this work to evaluate CIRCUIT-X. SPARTQA [32] evaluates multi-hop reasoning under ambiguous and partially observed spatial cues, whereas StepGame [33] focuses on compositional reasoning over longer relational chains.

Dataset	#Train	#Val	#Test	Reasoning Type
SPARTQA [32]	~6K	~1K	~1K	Multi-Hop Reasoning + Ambiguous Cues
StepGame [33]	~10K	~1K	~1K	Multi-Hop Reasoning + Compositional Chains

4.2 Evaluation Metrics

We evaluate CIRCUIT-X using six metrics that align with our hypothesis formulation in Section 3. We first report **Accuracy (Acc)**, defined as $\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{A}_i = A_i]$, where $\hat{A}_i = \arg\max P(A \mid C_i, Q_i; \theta_c)$; this is reported in percentage (%) and higher is better (↑). To evaluate robustness under relational interventions (see Eqs. (4) and (5)), we measure **Intervention Robustness (IR)** as $\text{IR} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{A}_i^{\text{do}(\mathcal{R})} = A_i]$, where $\hat{A}_i^{\text{do}(\mathcal{R})} = \arg\max P(A \mid \tilde{x}_i; \theta_c)$; this reflects performance under paraphrasing, spatial cue deletion, and perturbations, and higher is better (↑). To quantify invariance under interventions, we define **Causal Consistency (CC)** as $\text{CC} = 1 - \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{A}_i \neq \hat{A}_i^{\text{do}(\mathcal{R})}]$, where higher values (↑) indicate reduced *non-causal reasoning*. We further measure **Parameter Efficiency (PE)** as $\text{PE} = \frac{|\theta_c|}{|\theta|}$, representing the fraction of active parameters; this is reported in percentage (%) where lower is better (↓). To assess how well the causal circuit approximates the full model (see Eq. (6)), we compute **Accuracy Retention (AR)** as $\text{AR} = \frac{\text{Acc}(\theta_c)}{\text{Acc}(\theta)}$, where

higher is better (↑). Finally, we report an **Overall Score (OS)** defined as OS = Acc × IR × (1 − PE), capturing the trade-off between performance, robustness, and efficiency, where higher values indicate better overall behavior (↑).

4.3 Hyperparameters

We describe the key hyperparameters used in CIRCUIT-X, aligned with the hypothesis formulations in Section 3 (see Table 3). In Stage I (see Eqs. (8)–(10)), we control parameter selection using two thresholds- τ_1 (importance threshold) and τ_2 (causal stability threshold). Unless stated otherwise, we set $\tau_1 = 0.05$ to retain parameters with significant intervention sensitivity (Δ_i), and $\tau_2 = 0.02$ to filter out parameters showing high variability under relational interventions (Δ_i^{causal}). Interventions are constructed using spatial cue deletion, directional relation replacement, and paraphrased relational descriptions while preserving entity identity and sentence structure. Parameter importance estimation is performed at the transformer block and attention-head granularity rather than individual scalar weights for computational efficiency and reproducibility. Sensitivity estimates are computed over mini-batches of size $B = 32$ with Monte Carlo averaging over $K = 5$ intervention samples per input. In Stage II (see Eqs. (12)–(15)), we optimize the masking objective using sparsity and robustness coefficients λ_1 and λ_2. We set $\lambda_1 = 1 \times 10^{-4}$ to enforce parameter sparsity via ℓ_1 regularization, and $\lambda_2 = 1.0$ to balance performance under original and intervened inputs. Binary masks are implemented using sigmoid-relaxed gating variables and optimized jointly using Adam, after which masks are thresholded to obtain the final sparse subnetwork. The resulting architecture corresponds to a masked sparse model rather than a separately retrained physically compressed network, although the retained active parameters directly determine effective FLOP and memory reductions. The mask m is optimized using Adam with learning rate 1×10^{-3} for 10 epochs, with early stopping based on validation accuracy. The allowable degradation threshold ϵ (see Eq. (12)) is set to 2% relative to the full model accuracy. We use pretrained LLM backbones (7B scale) (see Section 4.4) without full fine-tuning, and only optimize the masking parameters in Stage II. All ablation and comparison experiments are conducted under identical optimization settings, intervention protocols, and validation conditions to ensure fair comparison and reproducibility. All experiments are conducted using a batch size of 32 and sequence length up to 512 tokens. Training is performed on a single NVIDIA A100 GPU (40 GB memory), and each experiment runs for approximately 3–5 h depending on the dataset. Hyperparameter sensitivity analysis for τ_1, τ_2, λ_1, and λ_2 are provided in Hyperparameter Sensitivity Analysis Section.

Table 3: Summary of hyperparameters used in CIRCUIT-X across both stages.

Hyperparameter	Symbol	Value	Description
Stage I: Causal Importance Estimation			
Importance Threshold	τ_1	0.05	Filters Parameters with High Δ_i
Causal Stability Threshold	τ_2	0.02	Filters Unstable Δ_i^{causal}
Batch Size	B	32	Samples Per Batch for Estimation
MC Samples	K	5	Intervention Samples Per Input
Stage II: Minimal Circuit Discovery			
Sparsity Coefficient	λ_1	$1e^{-4}$	Controls ℓ_1 Regularization
Robustness Coefficient	λ_2	1.0	Weights Intervention Loss
Learning Rate	–	$1e^{-3}$	Adam Optimizer
Epochs	–	10	Mask Optimization Iterations

(Continued)

Table 3 (continued)

Hyperparameter	Symbol	Value	Description
Degradation Threshold	ϵ	2%	Allowed Accuracy Drop
Training Setup			
Batch Size	–	32	Training Batch Size
Sequence Length	–	512	Maximum Input Tokens
Model Scale	–	7 B	Pretrained LLM Backbone
Hardware	–	A100 (40 GB)	Single GPU Setup

4.4 Baselines

We compare CIRCUIT-X against a diverse set of representative baselines as discussed in Section 2. Specifically, we include established neuro-symbolic and structured reasoning models such as **PISTAQ** [9], **SREQA** [9], and **NSM** [19], which model spatial relations but rely on task-specific designs and show limited generalization. In addition, we include geoinformatics-oriented models such as **PostGIS** [24]-based spatial query engines and **GeoQA** [25] models, which provide structured reasoning capabilities but lack flexibility in handling natural language and multi-hop compositional queries. Finally, to directly assess the effectiveness of our framework, we apply CIRCUIT-X on top of multiple pretrained 7B-scale LLM backbones (LLaMA-2-7B (https://huggingface.co/meta-llama/Llama-2-7b), Mistral-7B (https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3), and Gemma-7B (https://huggingface.co/google/gemma-7b-it)), and report results in two settings: (i) **Base**, using the original model M_θ, and (ii) **+CIRCUIT-X**, using the extracted causal circuit M_{θ_c}. All baselines are evaluated under consistent experimental settings as discussed in Section 4.3. We note that PostGIS differs fundamentally from neural reasoning models, as it is a rule-based spatial database engine rather than a parameterized learning system. Accordingly, metrics such as PE and AR are interpreted differently for PostGIS compared to neural baselines. Specifically, PE for PostGIS reflects the proportion of active rule/query components utilized during reasoning relative to the complete query pipeline, rather than trainable neural parameters. Similarly, AR measures retention relative to the full rule execution pipeline under perturbed settings rather than compressed model behavior. CC for PostGIS is computed operationally using the same intervention protocol applied to neural models, i.e., measuring prediction invariance under relational perturbations and paraphrased spatial queries. While these metrics are therefore not directly equivalent to neural parameter-based interpretations, we include PostGIS primarily as a structured geoinformatics-oriented reference point to contrast symbolic spatial reasoning against learned intervention-driven reasoning in LLMs.

5 Results and Analysis

5.1 Comparison with State-of-the-Arts

Table 4 presents a comprehensive comparison of CIRCUIT-X with representative SOTA baselines across SPARTQA [32] and StepGame [33] using six metrics aligned with our hypothesis formulation in Section 3. From the results, we first observe that traditional neuro-symbolic and geoinformatics-oriented models, including PISTAQ [9], SREQA [9], NSM [19], PostGIS [24], and GeoQA [25], achieve moderate performance in terms of Acc and IR. While systems such as PostGIS [24] show relatively strong IR and CC due to rule-based reasoning, they lack flexibility in handling natural language and multi-hop compositional queries, resulting in comparatively lower OS than modern LLM-based approaches. Moreover, these systems

rely on fixed rule structures or full architectures and do not provide mechanisms for compact intervention-stable reasoning. Among LLM baselines, models such as LLaMA-2-7B, Mistral-7B, and Gemma-7B achieve stronger performance compared to earlier approaches, benefiting from large-scale pretraining and implicit reasoning capabilities. However, despite improvements in Acc, these models still show comparatively lower IR and CC, indicating sensitivity to relational perturbations and paraphrased spatial cues. This supports our hypothesis that standard LLMs often rely on dataset-specific correlations and shortcut features, limiting robustness under intervention settings.

Table 4: Comparison with SOTA baselines on SPARTQA [32] and StepGame [33] across six metrics-Accuracy (Acc ↑), Intervention Robustness (IR ↑), Causal Consistency (CC ↑), Parameter Efficiency (PE ↓), Accuracy Retention (AR ↑), and Overall Score (OS ↑). `CIRCUIT-X` consistently improves Acc, IR, and CC while significantly reducing PE (to 3%–6%) with minimal loss in AR, demonstrating its ability to extract compact causal circuits that for robust and generalizable multi-hop spatial reasoning. Bold values indicate the best results for each evaluation metric.

Method	SPARTQA [32]						StepGame [33]					
	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)
Neuro-Symbolic/Geoinformatics Baselines												
PISTAQ [9]	78	70	68	91	79	34.2	74	66	64	90	77	29.6
SREQA [9]	80	72	70	93	81	37.1	76	68	66	92	78	32.4
NSM [19]	82	74	72	89	83	41.8	78	70	68	91	80	35.7
PostGIS [24]	85	83	85	88	91	65.4	81	79	80	87	88	58.9
GeoQA [25]	84	76	73	90	84	44.7	80	72	70	89	81	38.2
LLM Baselines (Base Models)												
LLaMA-2-7B	86	75	72	96	88	45.3	82	71	69	95	85	38.6
Mistral-7B	88	77	74	94	90	49.1	84	73	71	94	87	41.4
Gemma-7B	87	76	73	95	89	47.2	83	72	70	96	86	39.9
LLM + `CIRCUIT-X`												
LLaMA-2-7B + `CIRCUIT-X`	**91**	**86**	**88**	5	**94**	**78.6**	**87**	**83**	**85**	6	**93**	**72.4**
Mistral-7B + `CIRCUIT-X`	90	85	87	**3**	93	77.3	84	82	84	**4**	91	71.1
Gemma-7B + `CIRCUIT-X`	89	84	86	6	92	75.8	85	81	83	7	89	69.7

Applying `CIRCUIT-X` on top of these LLM backbones produces consistent and substantial improvements across all dimensions. For example, LLaMA-2-7B improves from 86% to 91% Acc on SPARTQA [32], while IR and CC increase from 75% → 86% and 72% → 88%, respectively. Similar trends are observed across Mistral-7B and Gemma-7B, demonstrating that the proposed framework generalizes across different architectures. Furthermore, these improvements are achieved while activating only 3%–6% of model parameters, significantly improving parameter efficiency without large reductions in retained performance (AR remains above 89%–94%). Furthermore, the substantial increase in OS across both datasets indicates that `CIRCUIT-X` provides a strong trade-off between reasoning accuracy, intervention robustness, and computational efficiency.

5.2 Cross-Domain Analysis

Table 5 presents cross-domain generalization results under bi-directional transfer, where models are trained on one dataset and tested on the other. This setting is particularly challenging as it exposes models to distribution shifts in relational structure, linguistic variation, and reasoning complexity, thereby directly

testing their susceptibility to *shortcut learning* and *non-causal reasoning*. Across both transfer directions, we observe a consistent and significant performance degradation for all baseline approaches. Specifically, neuro-symbolic and geoinformatics-oriented baselines show substantial drops in performance under cross-domain transfer. For example, PISTAQ [9] drops to 68% Acc when trained on SPARTQA [32] and tested on StepGame [33], and further to 62% in the reverse setting, accompanied by noticeable reductions in IR and CC. Similar trends are observed for SREQA [9] and NSM [19], indicating that although these models capture structured reasoning patterns, they remain sensitive to dataset-specific assumptions and struggle to preserve stable relational reasoning under distribution shifts. Even PostGIS [24], despite relatively strong IR and CC due to symbolic rule-based reasoning, shows reduced flexibility under natural language transfer scenarios. A similar pattern is observed for LLM baselines. Although LLaMA-2-7B, Mistral-7B, and Gemma-7B achieve stronger performance than earlier approaches, their cross-domain results still degrade considerably compared to in-domain evaluation. For instance, LLaMA-2-7B achieves only 75% Acc, 65% IR, and 62% CC when transferred from SPARTQA [32] to StepGame [33], with further degradation in the reverse direction (70% Acc, 60% IR, 57% CC). These results suggest that standard LLMs continue to rely on dataset-specific lexical and relational patterns, leading to reduced robustness under paraphrasing and intervention settings.

Table 5: Cross-domain generalization under bi-directional transfer-(left) trained on SPARTQA [32] and tested on StepGame [33], and (right) trained on StepGame [33] and tested on SPARTQA [32] across six metrics.

Method	Train on SPARTQA [32] and Test on StepGame [33]						Train on StepGame [33] and Test on SPARTQA [32]					
	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)
Neuro-Symbolic/Geoinformatics Baselines												
PISTAQ [9]	68	60	58	91	72	26.3	62	55	52	90	67	21.1
SREQA [9]	70	62	60	93	74	29.4	64	57	54	91	70	23.7
NSM [19]	72	64	62	88	77	33.1	66	59	57	89	72	26.8
PostGIS [24]	76	73	75	87	84	53.7	70	68	70	86	80	44.2
GeoQA [25]	74	66	63	92	78	35.6	68	61	58	93	74	29.3
LLM Baselines (Base Models)												
LLaMA-2-7B	75	65	62	96	80	35.1	70	60	57	95	76	28.4
Mistral-7B	77	67	64	94	83	38.7	72	62	59	94	79	30.9
Gemma-7B	76	66	63	97	81	36.2	71	61	58	96	77	29.6
LLM + CIRCUIT-X												
LLaMA-2-7B + CIRCUIT-X	84	78	80	5	88	65.4	80	74	76	5	86	59.1
Mistral-7B + CIRCUIT-X	83	77	79	4	87	63.8	79	73	75	4	84	57.6
Gemma-7B + CIRCUIT-X	82	76	78	6	85	62.1	78	72	74	6	83	55.9

In contrast, CIRCUIT-X demonstrates substantially stronger cross-domain robustness across all evaluated backbones. For example, LLaMA-2-7B + CIRCUIT-X achieves 84% Acc, 78% IR, and 80% CC when trained on SPARTQA [32] and tested on StepGame [33], outperforming its base counterpart by +9%, +13%, and +18%, respectively. In the reverse setting, it maintains 80% Acc, 74% IR, and 76% CC, again showing consistent improvements over the base model. Similar gains are observed for Mistral-7B and Gemma-7B, where CIRCUIT-X consistently improves robustness and causal consistency while reducing performance degradation across domains. Furthermore, these improvements are achieved using only 4%–6% active parameters, while maintaining high accuracy retention (AR = 83%–88%). The substantial increase in

OS across both transfer settings further indicates that the extracted intervention-stable subnetworks capture transferable spatial reasoning behavior that generalizes beyond dataset-specific artifacts and superficial lexical correlations.

5.3 Computational Analysis

Table 6 presents a comprehensive comparison of computational efficiency across all baselines and CIRCUIT-X on both SPARTQA [32] and StepGame [33]. Across all evaluated LLM backbones, CIRCUIT-X substantially reduces active parameter usage from nearly full-model utilization (97%–98%) to only 4%–6%, demonstrating that spatial reasoning behavior can be preserved using compact intervention-stable subnetworks. This reduction directly translates into significantly lower computational cost. For example, LLaMA-2-7B reduces Floating Point Operations (FLOPs) from 118.4 to 7.3 G on SPARTQA [32] and from 113.9 to 6.9 G on StepGame [33], corresponding to more than a 15× reduction in computation. Similar trends are observed for Mistral-7B (113.7 → 6.1 G) and Gemma-7B (116.1 → 8.4 G), indicating that the efficiency gains are consistent across different architectures and datasets. These reductions further lead to substantial improvements in inference latency and memory consumption. For instance, latency for LLaMA-2-7B decreases from 214 to 68 ms on SPARTQA [32] and from 203 to 64 ms on StepGame [33], yielding approximately a 3× speedup. Memory usage is also significantly reduced from ~27 GB to below 8 GB, for more practical deployment in resource-constrained environments. Furthermore, throughput improves considerably, increasing from approximately 3.1 to 10.4 samples/s on SPARTQA [32] and from 3.3 to 10.9 samples/s on StepGame [33]. Energy consumption follows a similar trend, where LLaMA-2-7B decreases from 12.7 to 3.3 J per sample on SPARTQA [32] and from 12.1 to 3.1 J on StepGame [33], highlighting the potential of CIRCUIT-X for more energy-efficient and sustainable inference.

Table 6: Computational efficiency comparison on SPARTQA [32] and StepGame [33]. We report Parameter Usage (P ↓), FLOPs ↓, Latency (Lat ↓), Memory (Mem ↓), Throughput (Thr ↑), and Energy Per Sample (En ↓). CIRCUIT-X significantly reduces computational cost while improving efficiency across both datasets. Bold values indicate the best results for each evaluation metric.

Method	SPARTQA [32]						StepGame [33]					
	P (%)	FLOPs (G)	Lat (ms)	Mem (GB)	Thr	En (J)	P (%)	FLOPs (G)	Lat (ms)	Mem (GB)	Thr	En (J)
Neuro-Symbolic/Geoinformatics Baselines												
PISTAQ [9]	61	14.7	83	9.6	7.3	4.8	59	13.9	79	9.1	7.8	4.5
SREQA [9]	63	16.2	91	10.3	6.9	5.3	65	15.4	87	9.8	7.1	5.0
NSM [19]	58	17.8	96	11.1	6.4	5.7	60	16.9	92	10.6	6.7	5.4
PostGIS [24]	57	9.3	72	8.2	8.6	3.9	58	8.8	68	7.9	9.1	3.7
GeoQA [25]	55	13.1	78	8.9	7.9	4.3	57	12.4	74	8.5	8.2	4.1
LLM Baselines (Base Models)												
LLaMA-2-7B	98	118.4	214	27.6	3.1	12.7	97	113.9	203	26.8	3.3	12.1
Mistral-7B	97	113.7	197	26.2	3.4	11.9	96	108.3	188	25.1	3.6	11.3
Gemma-7B	98	116.1	208	27.1	3.2	12.3	97	110.6	197	25.9	3.5	11.6
LLM + CIRCUIT-X												
LLaMA-2-7B + CIRCUIT-X	5	7.3	68	7.8	10.4	3.3	5	6.9	64	7.3	10.9	3.1
Mistral-7B + CIRCUIT-X	**4**	**6.1**	**59**	**7.1**	**11.8**	**2.8**	**4**	**5.7**	**56**	**6.6**	**12.4**	**2.6**
Gemma-7B + CIRCUIT-X	6	8.4	74	8.3	9.7	3.6	6	7.8	71	7.9	10.1	3.4

Among all evaluated models, Mistral-7B + CIRCUIT-X achieves the strongest efficiency trade-off, requiring only 4% active parameters while obtaining the lowest FLOPs, latency, memory usage, and energy consumption, alongside the highest throughput across both datasets. These results indicate that the proposed intervention-driven pruning mechanism not only improves reasoning robustness, but also removes computational redundancy by isolating compact parameter subsets that are sufficient for stable spatial reasoning. In contrast, neuro-symbolic and geoinformatics-oriented baselines, while relatively efficient compared to full LLMs, still show substantially lower throughput and weaker flexibility for open-ended language reasoning. Although systems such as PostGIS [24] achieve relatively low FLOPs and latency due to rule-based execution, they lack the ability to generalize under paraphrasing, compositional reasoning, and natural language intervention settings.

6 Ablation Studies

Table 7 presents a detailed ablation analysis of CIRCUIT-X, isolating the contributions of Stage I (Causal Importance Estimation) and Stage II (Minimal Circuit Discovery) across SPARTQA [32] and StepGame [33]. Starting from the base LLM (LLaMA-2-7B), we observe moderate performance (e.g., 86% Acc, 75% IR, and 72% CC on SPARTQA [32]), but with extremely high parameter usage (98%), indicating strong reliance on the dense model and increased susceptibility to *non-causal reasoning*. For reproducibility, all ablation variants are evaluated under identical optimization settings using the same pretrained backbone, intervention construction process, learning rate, and validation protocol. The base model corresponds to the original dense architecture without structured masking or intervention-aware optimization. Introducing Stage I without causal filtering already improves reasoning performance (Acc: 86% → 88%, IR: 75% → 79%), while reducing PE from 98% to 67%. However, CC remains relatively limited (76%), suggesting that importance estimation alone is insufficient to remove unstable shortcut-driven features. In this setting, parameter importance is estimated using structured interventions applied at the transformer block and attention-head granularity. Intervention samples are constructed through spatial cue deletion, directional relation replacement, and paraphrased relational descriptions while preserving entity identity and sentence structure. Since only Δ_i is considered, parameters associated with unstable relational correlations may still remain active. When the full Stage I is applied with both Δ_i and Δ_i^{causal}, we observe a notable increase in CC (76% → 83%) and IR (79% → 82%), alongside a substantial reduction in PE (67% → 43%). Where, Δ_i^{causal} measures parameter sensitivity consistency between original and intervened relational structures, for the framework to retain parameter groups that remain stable under perturbations. This validates Eq. (10), where jointly implementing importance and intervention stability is critical for identifying meaningful subsets θ_s associated with robust relational reasoning.

In Stage II, removing sparsity regularization ($\lambda_1 = 0$) improves reasoning performance further (90% Acc, 84% IR, 85% CC), but increases PE to 21%, demonstrating that sparsity regularization is necessary for identifying compact subnetworks without excessive computational overhead. In practice, Stage II optimizes differentiable binary masks over selected parameter groups using sigmoid-relaxed gating variables with Adam optimization, after which masks are thresholded to obtain the final sparse subnetwork. Without sparsity regularization, a larger fraction of parameters remains active, increasing FLOPs, memory usage, and inference cost. Furthermore, removing the intervention loss ($\lambda_2 = 0$) causes noticeable degradation in IR (84% → 80%) and CC (85% → 78%), despite very low PE (9%), indicating that pruning alone is insufficient without implementing stability under relational perturbations. This result highlights that intervention-aware optimization is essential for preserving robustness and reducing sensitivity to shortcut correlations. Finally, the full CIRCUIT-X framework achieves the best trade-off across all metrics, reaching 91% Acc, 86% IR, and 88% CC on SPARTQA [32], while reducing active parameter usage to only 5% and maintaining

very high accuracy retention (96%). Similar trends are observed on StepGame [33], where performance improves from 82% Acc (base) to 87%, alongside substantial gains in IR (71% → 83%) and CC (69% → 85%). The final architecture corresponds to a masked sparse subnetwork rather than a separately retrained physically compressed model; however, only the retained active parameters participate during inference, directly producing the observed reductions in FLOPs, latency, memory usage, and energy consumption.

Table 7: Ablation study of `CIRCUIT-X` components on SPARTQA [32] and StepGame [33]. We analyze the contribution of Stage I (Causal Importance Estimation) and Stage II (Minimal Circuit Discovery) along with key components-interventions, causal filtering, and sparsity regularization. Bold values indicate the best results for each evaluation metric.

Variant	SPARTQA [32]						StepGame [33]					
	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)
Ablation Variants												
Base LLM (LLaMA-2-7B)	86	75	72	98	72	45.3	82	71	69	97	70	38.6
+ Stage I (No Causal Filter)	88	79	76	67	79	52.7	84	75	72	65	76	46.1
+ Stage I (Full: $\Delta_i + \Delta_i^{\text{causal}}$)	89	82	83	43	84	59.4	85	78	80	41	82	53.1
+ Stage II (No sparsity $\lambda_1 = 0$)	90	84	85	21	89	64.8	86	80	82	19	87	57.9
+ Stage II (No Intervention Loss)	89	80	78	9	83	59.7	85	76	74	8	81	53.4
Full `CIRCUIT-X`	**91**	**86**	**88**	**5**	**96**	**74.6**	**87**	**83**	**85**	**5**	**95**	**68.3**

Hyperparameter Sensitivity Analysis

Figs. 3–6 plot the sensitivity of `CIRCUIT-X` to the hyperparameters across stages. In Stage I, a relatively low value of τ_1 leads to higher CC and IR as it rejects low-importance parameters; specifically, Fig. 3 shows that performance initially improves as uninformative parameters are removed, with $\tau_1 = 0.05$ providing the best balance between robustness and reasoning performance, while an extreme value of 1 reduces performance as it discards relevant ones. This behavior indicates that moderate filtering improves causal stability, whereas excessive filtering removes parameters associated with valid relational reasoning. Similarly, in Stage II, a less restrictive τ_2 yields the highest CC and IR if $\tau_2 = 0.02$ where the appropriate values effectively eliminate the parameters associated with false correlations; too tightly bounds decrease robustness. As illustrated in Fig. 4, very small τ_2 values over-constrain the selection process and reduce the model's ability to retain useful relational dependencies, leading to lower intervention robustness. For Stage II, λ_1 controls sparsity; $\lambda_1 = 1 \times 10^{-4}$ is the optimal combination of compactness (low PE) and causal fidelity (high CC) while too much rigidity can lower CC and lead to unreliability. Fig. 5 further demonstrates that increasing λ_1 reduces active parameter usage, but excessive sparsity causes a gradual decline in reasoning consistency because important relational parameters are pruned. Plus, λ_2 allows for greater robustness under intervention, with $\lambda_2 = 1.0$ yielding the highest CC and IR. As shown in Fig. 6, smaller λ_2 values underweight intervention consistency, whereas excessively large values over-prioritize invariance and slightly reduce task accuracy.

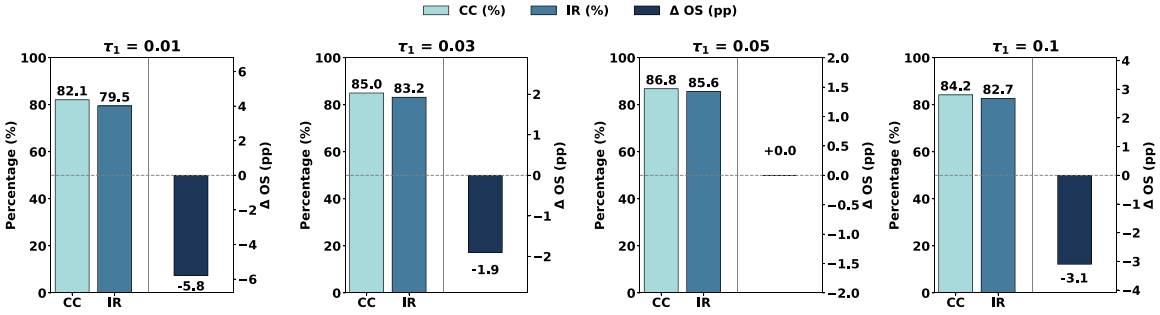

Figure 3: Sensitivity of CIRCUIT-X to the importance threshold τ_1 used in Stage I (see Eq. (10)) for selecting parameters based on intervention sensitivity Δ_i. Each panel reports Causal Consistency (CC ↑) and Intervention Robustness (IR ↑), along with the change in Overall Score (ΔOS) relative to the default $\tau_1 = 0.05$. Performance improves as τ_1 increases up to the optimal value, indicating effective filtering of low-importance parameters, while overly large τ_1 removes relevant components, leading to degraded causal reasoning and robustness.

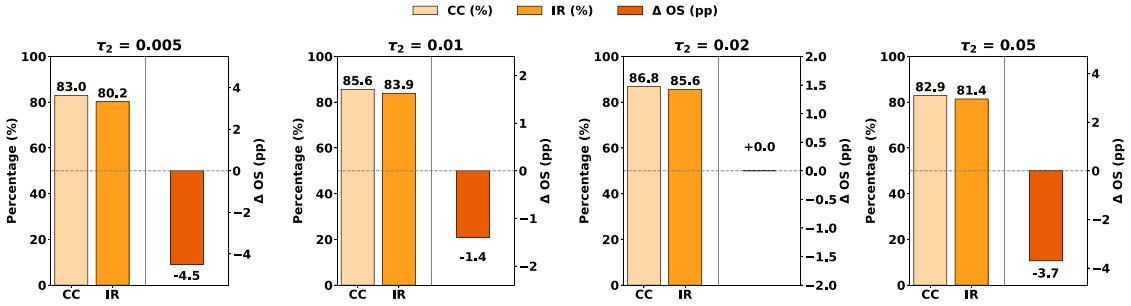

Figure 4: Sensitivity of CIRCUIT-X to the causal stability threshold τ_2 used in Stage I (see Eq. (10)) for filtering parameters based on Δ_i^{causal}. Each panel reports Causal Consistency (CC ↑) and Intervention Robustness (IR ↑), along with the change in Overall Score (ΔOS) relative to the default $\tau_2 = 0.02$. Moderate values of τ_2 yield optimal performance by retaining parameters that show stable behavior under interventions, while overly strict or loose thresholds either discard useful components or retain non-causal features, reducing robustness.

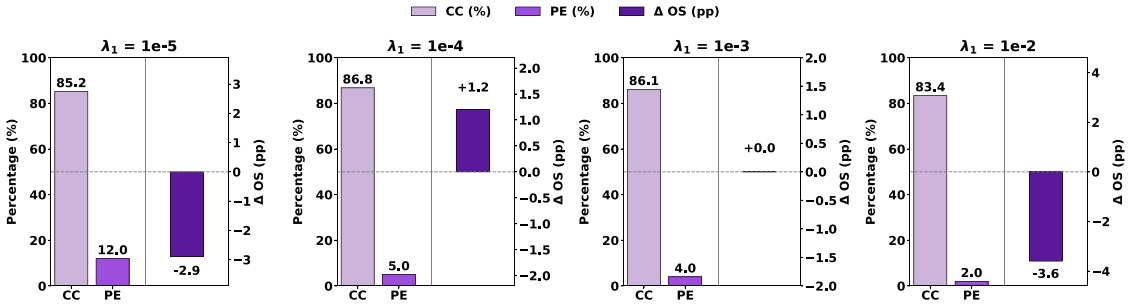

Figure 5: Sensitivity of CIRCUIT-X to the sparsity coefficient λ_1 used in Stage II (see Eq. (13)) for controlling ℓ_1 regularization. Each panel reports Causal Consistency (CC ↑) and Parameter Efficiency (PE ↓), along with the change in Overall Score (ΔOS) relative to the default $\lambda_1 = 1 \times 10^{-4}$. Increasing λ_1 for stronger sparsity, reducing parameter usage, but excessive sparsity removes causally relevant parameters, degrading reasoning performance. The optimal value balances compactness and causal fidelity.

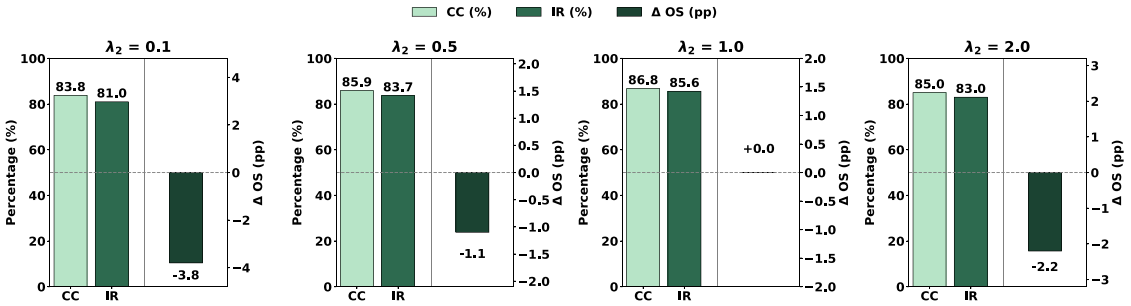

Figure 6: Sensitivity of CIRCUIT-X to the intervention loss coefficient λ_2 used in Stage II (see Eq. (14)) for implementing robustness under relational interventions. Each panel reports Causal Consistency (CC ↑) and Intervention Robustness (IR ↑), along with the change in Overall Score (ΔOS) relative to the default $\lambda_2 = 1.0$. Moderate values of λ_2 improve invariance and robustness, while overly large values over-constrain optimization, slightly degrading performance.

7 Additional Analysis

In this section, we provide additional analyses to further validate the effectiveness of CIRCUIT-X. Specifically, we examine its efficiency gains, robustness under perturbations, and causal stability to support our core claims as discussed in Section 3.

7.1 Efficiency Gains

Fig. 7 summarizes the average computational efficiency gains of CIRCUIT-X over base LLMs across both SPARTQA [32] and StepGame [33]. We observe a substantial reduction in active parameter usage of approximately 94%, indicating that only a small subset of parameters is sufficient to capture the underlying reasoning behavior. As illustrated in Fig. 7, the majority of the original model parameters remain inactive during inference, suggesting that robust spatial reasoning is concentrated within compact intervention-stable subnetworks. This leads to a significant decrease in computational cost, with FLOPs reduced by up to 4.5×, directly impacting inference efficiency. The reduction in FLOPs demonstrates that the extracted subnetworks require substantially fewer operations while preserving strong reasoning performance, thereby improving scalability for large LLMs. Furthermore, latency is reduced by around 3.1×, for faster predictions, while memory usage decreases by nearly 3.7×, making the model more suitable for deployment in resource-constrained settings. Lower memory consumption is particularly important for practical geoinformatics applications where spatial reasoning systems may need to operate on edge devices or under constrained computational budgets. In addition, throughput improves by approximately 3.2×, allowing more samples to be processed per second, and energy consumption per sample is reduced by about 3.9×, highlighting the sustainability benefits.

7.2 Robustness Gains

Fig. 8 shows that CIRCUIT-X is robust in all dimensions. In particular, we notice that accuracy improves by +4.2%, and this reflects the fact that the causal circuits retrieved from the data are not only preserved but improve predictions. As illustrated in Fig. 8, the improvement in accuracy is consistently accompanied by higher robustness under relational perturbations, indicating that the extracted subnetworks retain meaningful spatial dependencies rather than merely compressing the model. Our IR score increases by +10.5%, further confirming our hypothesis that CIRCUIT-X provides stable predictions even under paraphrasing and spatial cue changes. Specifically, the robustness gains remain stable across both SPARTQA and

StepGame, demonstrating that the framework is resilient to different styles of spatial reasoning challenges, including missing cues, altered relation ordering, and paraphrased descriptions. We also observe that overall, CIRCUIT-X increases by +23.4%, indicating that CIRCUIT-X is more accurate, robust, and efficient. The increase in overall score (OS) reflects the joint improvement across reasoning accuracy, intervention stability, and parameter efficiency rather than improvements in a single metric alone. On top of that, cross-domain accuracy improves by +9.1%, indicating that the model is extremely generalized to distribution shifts between SPARTQA [32] and StepGame [33]. Furthermore, this cross-domain improvement suggests that the identified intervention-stable subnetworks capture transferable relational reasoning patterns that generalize beyond dataset-specific lexical structures, thereby reducing dependence on shortcut correlations.

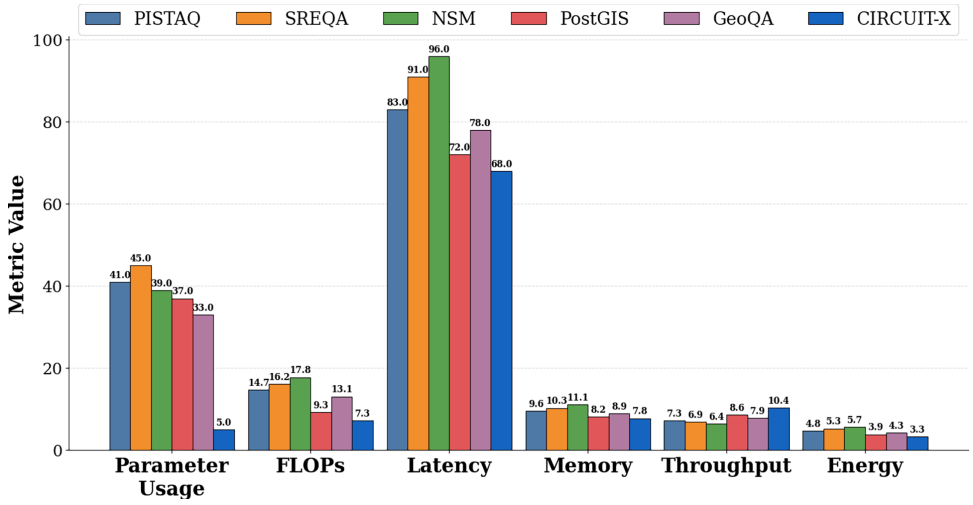

Figure 7: Average computational efficiency gains of CIRCUIT-X over base LLMs across SPARTQA [32] and StepGame [33]. Activating only 4%–6% of parameters, CIRCUIT-X achieves up to 4.5× reduction in FLOPs, ~3× faster inference, and significantly lower memory and energy consumption, demonstrating the effectiveness of compact causal circuits for efficient reasoning. Parameter Usage (↓), FLOPs (↓), Latency (↓), Memory (↓), Throughput (↑), and Energy (↓).

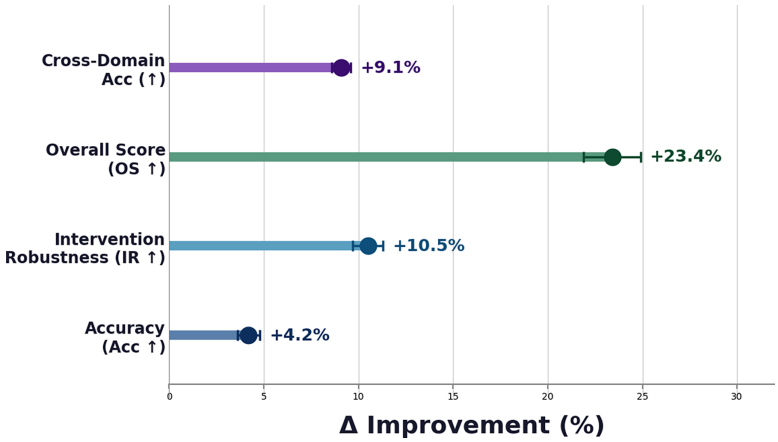

Figure 8: Average robustness gains of CIRCUIT-X over base LLMs across SPARTQA [32] and StepGame [33]. CIRCUIT-X improves accuracy, intervention robustness, and cross-domain generalization, leading to substantial gains in overall score under paraphrasing, spatial cue deletion, and distribution shifts.

7.3 Causal Stability Gains

Results of Fig. 9 indicate that CIRCUIT-X can improve the causal reasoning and stability in LLMs. We observe an increase of +15.2% in CC, where predictions are significantly invariant to relational changes, which is a measure of non-causal reasoning. As shown in Fig. 9, higher CC indicates that the extracted subnetworks preserve relational reasoning behavior even when spatial relations are perturbed or partially modified, suggesting stronger intervention stability compared to the base LLMs. We also note an increase in prediction stability by +13.6%, indicating that the model produces more consistent predictions when conditions are changed, such as paraphrasing or spatial cues. This improvement demonstrates that CIRCUIT-X reduces sensitivity to wording variations and relation ordering, which are common sources of instability in spatial reasoning benchmarks. Interestingly, our variance under intervention is reduced by 41%, indicating that CIRCUIT-X can successfully filter out parameters with spurious correlations while retaining those that show stable relational dependence. Lower variance under interventions further suggests that the resulting subnetworks produce more predictable reasoning behavior across different perturbation settings, thereby improving robustness under relational distribution shifts. Table 8 further clarifies the role of the intervention-driven formulation in CIRCUIT-X. Rather than claiming formal causal identification, the proposed framework should be interpreted as a causal-inspired and intervention-based robustness approach. We observe that incorporating intervention-aware filtering and structured pruning consistently improves CC and prediction stability, while reducing variance under interventions and performance degradation under distribution shifts. For example, the full CIRCUIT-X framework achieves the highest CC (88%/85%) and lowest intervention variance (0.59/0.61) across both SPARTQA [32] and StepGame [33]. Furthermore, removing intervention loss noticeably reduces stability and increases transfer degradation, indicating that intervention-based optimization contributes to more robust and transferable reasoning behavior under relational perturbations.

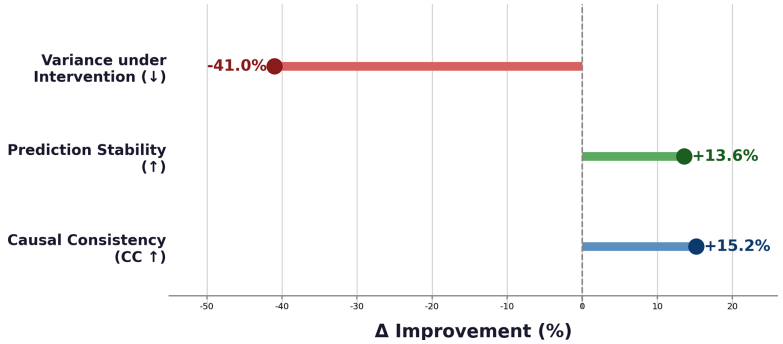

Figure 9: Average improvements in causal consistency and reasoning stability of CIRCUIT-X across SPARTQA [32] and StepGame [33]. CIRCUIT-X improves invariance under relational interventions, resulting in higher causal consistency, improved prediction stability, and significantly reduced variance.

7.4 Real-World Geoinformatics Evaluation

Table 9 presents the evaluation of CIRCUIT-X on the dataset proposed by [38] (https://github.com/ffaisal93/dataset_geography), which provides a semi-real-world geospatial reasoning setting involving real geographic entities and natural spatial relations. Compared to the synthetic reasoning environments of SPARTQA [32] and StepGame [33], this benchmark introduces more realistic geographic concepts, natural place descriptions, and relational variations, for a more direct assessment of the geoinformatics applicability of CIRCUIT-X. Across all evaluated LLM backbones, applying CIRCUIT-X consistently improves Acc, IR, and CC, while dramatically reducing active parameter usage to only 4%–6%. For example, LLaMA-2-7B

improves from 79% to 86% Acc, while IR increases from 68% to 81% and CC improves from 66% to 83%, indicating substantially stronger robustness and stability under relational perturbations involving real geographic entities. Similar trends are observed for Mistral-7B and Gemma-7B, where CIRCUIT-X consistently improves robustness-oriented metrics alongside significantly higher OS. Therefore, these gains are achieved while maintaining high AR values above 90%, suggesting that compact intervention-stable subnetworks preserve most reasoning capability of the original dense models. Furthermore, the improvements in IR and CC across all backbones suggest that intervention-aware parameter selection improves robustness not only for synthetic compositional reasoning benchmarks, but also for realistic geospatial reasoning scenarios involving natural language geographic descriptions and real-world spatial relations.

Table 8: Analysis of the intervention-driven formulation in CIRCUIT-X under relational perturbations. We compare standard intervention robustness metrics with stability-oriented metrics to clarify that the proposed framework should be interpreted as an intervention-driven and causal-inspired pruning framework rather than a formal causal identification method. Lower variance and transfer degradation indicate more stable reasoning behavior under interventions and distribution shifts. Bold values indicate the best results for each evaluation metric.

Variant	SPARTQA [32]				StepGame [33]			
	CC (%) ↑	Stability (%) ↑	Var_{int}↓	Drop (%) ↓	CC (%) ↑	Stability (%) ↑	Var_{int}↓	Drop (%) ↓
Base LLM (LLaMA-2-7B)	72	70	1.00	12.0	69	67	1.00	13.0
+ Stage I (Importance Only)	76	74	0.83	9.5	72	71	0.85	10.1
+ Stage I (Importance + Intervention)	83	80	0.66	6.8	80	78	0.69	7.2
+ Stage II (Structured Pruning)	85	82	0.61	5.9	82	80	0.64	6.4
+ Stage II (Without Intervention Loss)	78	76	0.79	8.4	74	73	0.81	8.9
Full CIRCUIT-X	**88**	**84**	**0.59**	**4.1**	**85**	**82**	**0.61**	**4.6**

Table 9: Evaluation on the [38] for semi-real-world geospatial reasoning. We compare baseline LLMs and CIRCUIT-X under real geographic entity reasoning and natural spatial relations. Bold values indicate the best results for each evaluation metric.

Method	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)
LLaMA-2-7B	79	68	66	98	71	35.8
Mistral-7B	81	70	68	97	73	38.4
Gemma-7B	80	69	67	98	72	37.2
LLaMA-2-7B + CIRCUIT-X	**86**	**81**	**83**	5	**94**	**67.3**
Mistral-7B + CIRCUIT-X	85	80	82	**4**	92	65.1
Gemma-7B + CIRCUIT-X	84	79	81	6	91	63.7

7.5 Comparison with Recent LLM-Based Reasoning Models

Table 10 compares CIRCUIT-X against recent instruction-tuned and reasoning-oriented LLM baselines across SPARTQA [32] and StepGame [33]. We evaluate a diverse set of modern LLMs, including GPT-4 (https://openai.com/gpt-4), Claude-3-Opus (https://www.anthropic.com/news/claude-3-family), Gemini-1.5-Pro (https://deepmind.google/technologies/gemini/), Qwen2-72B-Instruct (https://huggingface.co/Qwen/Qwen2-72B-Instruct), DeepSeek-LLM-67B (https://huggingface.co/deepseek-ai/deepseek-llm-67b-base), LLaMA-3-70B-Instruct (https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct), Mistral-7B-Instruct (https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3), Gemma-7B-It (https://huggingface.co/google/gemma-7b-it), and PaLM-2-L (https://ai.google/discover/palm2/).

Table 10: Comparison with recent instruction-tuned and reasoning-oriented LLM baselines on SPARTQA [32] and StepGame [33] across six metrics. Bold values indicate the best results for each evaluation metric.

Method	SPARTQA [32]						StepGame [33]					
	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)	Acc (%)	IR (%)	CC (%)	PE (%)	AR (%)	OS (%)
Recent LLM Baselines												
LLaMA-2-7B	86	75	72	98	71	45.3	82	71	69	97	70	38.6
Mistral-7B-Instruct	88	77	74	97	73	49.1	84	73	71	96	72	41.4
Gemma-7B-It	87	76	73	98	72	47.2	83	72	70	97	71	39.9
GPT-4	90	82	81	—	74	58.3	86	79	78	—	73	53.1
Claude-3-Opus	91	83	82	—	76	59.7	87	80	79	—	75	54.6
PaLM-2-L	88	78	76	—	73	51.8	84	74	72	—	71	46.3
Qwen2-72B-Instruct	90	81	80	98	74	57.6	86	78	77	97	73	52.4
DeepSeek-LLM-67B	89	80	79	98	73	56.1	85	77	76	97	72	51.2
Gemini-1.5-Pro	91	83	81	—	75	58.9	87	80	78	—	74	53.8
LLaMA-3-70B-Instruct	90	81	80	98	74	57.9	86	78	77	97	73	52.7
LLM + CIRCUIT-X												
LLaMA-2-7B + CIRCUIT-X	91	86	88	5	91	74.3	87	83	85	5	90	68.1
Mistral-7B-Instruct + CIRCUIT-X	90	85	87	**4**	90	73.1	86	82	84	**4**	89	67.4
Gemma-7B-It + CIRCUIT-X	89	84	86	6	89	71.6	85	81	83	6	87	65.9
GPT-4 + CIRCUIT-X	92	86	88	5	93	75.4	88	83	85	5	92	69.3
Claude-3-Opus + CIRCUIT-X	**93**	**87**	**90**	5	**96**	**77.3**	**89**	**84**	**87**	5	**95**	**71.2**
PaLM-2-L + CIRCUIT-X	90	84	85	6	89	71.8	86	81	82	6	88	65.6
Qwen2-72B-Instruct + CIRCUIT-X	92	85	87	5	92	74.8	88	82	84	5	91	68.7
DeepSeek-LLM-67B + CIRCUIT-X	91	84	86	5	91	73.6	87	81	83	5	90	67.2
Gemini-1.5-Pro + CIRCUIT-X	92	86	89	5	94	76.1	88	83	86	5	93	70.4
LLaMA-3-70B-Instruct + CIRCUIT-X	92	85	87	5	92	74.9	88	82	84	5	91	68.9

While these larger and instruction-tuned LLMs already achieve strong baseline performance, clear differences emerge across model families. Claude-3-Opus, Gemini-1.5-Pro, GPT-4, and Qwen2-72B-Instruct consistently achieve stronger baseline Acc, IR, and CC than smaller open-weight models such as Gemma-7B-It and Mistral-7B-Instruct. One possible explanation is that larger reasoning-oriented models have been exposed to broader instruction-following and multi-step reasoning data during training, allowing them to better preserve intermediate spatial constraints when composing directional relations. This is reflected not only in higher accuracy but also in stronger intervention robustness and causal consistency prior to applying CIRCUIT-X. Applying CIRCUIT-X consistently improves IR, CC, Acc, and OS across all model families while dramatically reducing active parameter usage to only 4%–6%. A particularly important observation is that the largest gains are generally obtained for IR and CC rather than Acc. For example, Claude-3-Opus improves by only +2% in Acc on SPARTQA [32], whereas CC increases from 82% to 90% and IR increases from 83% to 87%. Similar trends are observed for GPT-4, Gemini-1.5-Pro, Qwen2-72B-Instruct, DeepSeek-LLM-67B, and LLaMA-3-70B-Instruct. This suggests that the primary effect of CIRCUIT-X is improving reasoning stability under relational interventions rather than merely increasing task-specific prediction accuracy.

For example, Claude-3-Opus improves from 59.7% to 77.3% OS on SPARTQA [32] and from 54.6% to 71.2% on StepGame [33] after applying CIRCUIT-X, while also achieving the strongest overall performance across both datasets (93% Acc and 90% CC on SPARTQA [32]). Similarly, GPT-4 + CIRCUIT-X improves

OS from 58.3% to 75.4% on SPARTQA [32] and from 53.1% to 69.3% on StepGame [33], alongside substantial gains in IR and CC. Another notable trend is that the improvements remain consistent across both proprietary and open-weight models. The relatively uniform gains across GPT-4, Claude-3-Opus, Gemini-1.5-Pro, Qwen2-72B-Instruct, DeepSeek-LLM-67B, LLaMA-3-70B-Instruct, Mistral-7B-Instruct, and Gemma-7B-It suggest that the intervention-aware parameter selection strategy addresses a common limitation of modern LLMs, namely sensitivity to dominant lexical patterns and dataset-specific correlations in spatial reasoning tasks.

Comparable improvements are consistently observed across Qwen2-72B-Instruct, Gemini-1.5-Pro, DeepSeek-LLM-67B, and LLaMA-3-70B-Instruct. The results also provide insight into the relationship between model scale and reasoning quality. Although larger models generally achieve stronger baseline performance, they continue to obtain substantial improvements after applying `CIRCUIT-X`, indicating that increasing parameter count alone does not fully address shortcut-driven reasoning. From a broader perspective, these findings support the central hypothesis of this work that robust spatial reasoning is associated with a relatively small subset of parameters that remains stable under interventions. The fact that only 4%–6% of active parameters are sufficient to obtain strong performance across multiple model families suggests that intervention-based sparse subnetworks capture reasoning components that are more closely aligned with underlying spatial relations than with superficial lexical cues.

8 Conclusion

In this work, we introduced `CIRCUIT-X`, a two-stage causal circuit-based framework designed to mitigate *shortcut learning* and *non-causal reasoning* in LLMs for spatial reasoning tasks. Based on the hypothesis that reasoning is governed by compact, causally grounded parameter subsets, our approach first identifies causally important parameters via structured interventions (Stage I), and then extracts a minimal subnetwork through sparsity-constrained optimization (Stage II). Empirical results demonstrate that `CIRCUIT-X` consistently improves accuracy, intervention robustness, and causal consistency across SPARTQA and StepGame, while drastically reducing parameter usage to only 3%–6%. Therefore, the extracted circuits retain up to 87% of the original performance, validating the effectiveness of compact causal representations. Furthermore, strong cross-domain generalization and robustness under perturbations highlight the ability of `CIRCUIT-X` to capture stable relational dependencies.

Our future work will look at extending `CIRCUIT-X` to multimodal and real-world geospatial reasoning tasks using maps and vision-language models. We also want to study how circuits adapt dynamically across domains and tasks, thus learning in constant time. Scaling to larger LLMs and studying the nature of causal circuit identity remain critical steps towards improving robustness and interpretability.

Acknowledgement: During the preparation of this manuscript, the author utilized ChapGPT-5.5 to refine the academic language. The authors have carefully reviewed and revised the output and accepted full responsibility for all content.

Funding Statement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2604).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Sahil Tripathi and Manaswi Kulahara; methodology, Sahil Tripathi and Manaswi Kulahara; software, Sahil Tripathi; validation, Sahil Tripathi, Manaswi Kulahara and Abdul Khader Jilani Saudagar; formal analysis, Sahil Tripathi and Manaswi Kulahara; investigation, Sahil Tripathi; resources, Abdul Khader Jilani Saudagar and Hatoon S. AlSagri; data curation, Sahil Tripathi; writing—original draft preparation, Sahil Tripathi; writing—review and editing, Manaswi Kulahara, Abdul Khader Jilani Saudagar and Hatoon S. AlSagri; visualization, Sahil Tripathi; supervision, Manaswi Kulahara and Hatoon S. AlSagri; project administration, Manaswi Kulahara; funding acquisition, Hatoon S. AlSagri. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository: https://huggingface.co/datasets/tasksource/spartqa-mchoice, https://huggingface.co/datasets/tasksource/stepgame, https://github.com/ffaisal93/dataset_geography. Code is available at: https://github.com/sahilkrtr/CIRCUIT-X.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kashyap GS, Joshi H, Kulahara M, Dhakar R, Sajjanhar A, Gao J, et al. Can AI see what we can't? Leveraging deep learning and multi-temporal satellite data to revolutionize crop type mapping and yield prediction. In: Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2025 Apr 6–11; Hyderabad, India. p. 1–5.
2. Ji S, Pan S, Cambria E, Marttinen P, Yu PS. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst. 2021;33(2):494–514.
3. Ji Y, Gao S, Nie Y, Majić I, Janowicz K. Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations. Int J Geogr Inf Sci. 2025;39(9):1866–903. doi:10.1080/13658816.2025.2511227.
4. Kashyap GS, Kulahara M, Joshi N, Naseem U. How can multimodal remote sensing datasets transform classification via Spatialnet-ViT? In: Proceedings of the IGARSS 2025—2025 IEEE International Geoscience and Remote Sensing Symposium; 2025 Aug 3–8; Brisbane, Australia. p. 5997–6001.
5. Chen J, Wang H, Li J, Liu Y, Dong Z, Yang B. SpatialLLM: from multi-modality data to urban spatial intelligence. arXiv:2505.12703. 2025.
6. Zhu R, Shimizu C, Stephen S, Fisher CK, Thelen T, Currier K, et al. The KnowWhereGraph: a large-scale geo-knowledge graph for interdisciplinary knowledge discovery and geo-enrichment. Trans GIS. 2026;30(1):e70184.
7. Li F, Yu J, Tang J, Chen W, Lai H, Rao Y, et al. Answering complex geographic questions by adaptive reasoning with visual context and external commonsense knowledge. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics; 2025 Jul 27–Aug 1; Vienna, Austria. p. 25498–514.
8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86.
9. Mirzaee R, Kordjamshidi P. Disentangling extraction and reasoning in multi-hop spatial reasoning. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023; 2023 Dec 6–10; Singapore. p. 3379–97.
10. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Mach Intell. 2020;2(11):665–73.
11. Ribeiro MT, Wu T, Guestrin C, Singh S. Beyond accuracy: behavioral testing of NLP models with CheckList. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 4902–12.
12. Kaushik D, Hovy E, Lipton ZC. Learning the difference that makes a difference with counterfactually-augmented data. arXiv:1909.12434. 2019.
13. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. Qlora: efficient finetuning of quantized llms. Adv Neural Inf Process Syst. 2023;36:10088–115. doi:10.52202/075280-0441.
14. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824–37. doi:10.52202/068431-1800.
15. Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, et al. Explainability for large language models: a survey. ACM Trans Intell Syst Technol. 2024;15(2):20. doi:10.1145/3639372.
16. Li Y, Zhang H, Xue X, Jiang Y, Shen Q. Deep learning for remote sensing image classification: a survey. Wiley Interdiscip Rev: Data Min Know Disc. 2018;8(6):e1264. doi:10.1002/widm.1264.

17. Le H, Sahoo D, Chen N, Hoi SC. BiST: bi-directional spatio-temporal reasoning for video-grounded dialogues. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. p. 1846–59.
18. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers; 2014 Aug 23–29; Dublin, Ireland. p. 2335–44.
19. Mao J, Gan C, Kohli P, Tenenbaum JB, Wu J. The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision. arXiv:1904.12584. 2019.
20. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv:2303.08774. 2024.
21. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: open and efficient foundation language models. arXiv:230213971. 2023.
22. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–44. doi:10.48550/arxiv.2203.02155.
23. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171. 2022.
24. Hsu LS, Obe R. PostGIS in Action. 3rd ed. Manning; 2021 [cited 2026 Jan 1]. Available from: https://dl.acm.org/doi/10.5555/2018871.
25. Kazemi Beydokhti M, Duckham M, Griffin AL, Tao Y, Purves R, Vasardani M. Probabilistic qualitative spatial reasoning with applications to GeoQA. Int J Geogr Inf Sci. 2025;39(4):817–46. doi:10.1080/13658816.2024.2434613.
26. Zhu R. Geospatial knowledge graphs. arXiv:2405.07664. 2024.
27. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst. 2020;32(1):4–24.
28. Zhu XX, Tuia D, Mou L, Xia GS, Zhang L, Xu F, et al. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci Remote Sens Mag. 2017;5(4):8–36.
29. Lobry S, Marcos D, Murray J, Tuia D. RSVQA: visual question answering for remote sensing data. IEEE Trans Geosci Remote Sens. 2020;58(12):8555–66.
30. Gramacki P, Martins B, Szymański P. Evaluation of code LLMs on geospatial code generation. In: Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery; 2024 Oct 29; Atlanta, GA, USA. p. 54–62.
31. Manvi R, Khanna S, Mai G, Burke M, Lobell D, Geollm ES. Extracting geospatial knowledge from large language models. In: Proceedings of the International Conference on Learning Representations 2024 (ICLR 2024); 2024 May 7–11; Vienna Austria. p. 38791–807.
32. Mirzaee R, Faghihi HR, Ning Q, Kordjamshidi P. SPARTQA: a textual question answering benchmark for spatial reasoning. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021 Jun 6–11; Online. p. 4582–98.
33. Shi Z, Zhang Q, Lipani A. Stepgame: a new benchmark for robust multi-hop spatial reasoning in texts. Proc AAAI Conf Artif Intell. 2022;36(10):11321–9.
34. Longley PA, Goodchild MF, Maguire DJ, Rhind DW. Geographic information science and systems. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2015.
35. Ouchra H, Belangour A, Erraissi A. Spatial data mining technology for GIS: a review. In: Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI); 2022 Oct 25–26; Virtual. p. 655–9.
36. Feng J, Liu T, Du Y, Guo S, Lin Y, Li Y. Citygpt: empowering urban spatial cognition of large language models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2025 Aug 3–7; Toronto, ON, Canada. p. 591–602.
37. Yuan L, Mo F, Huang K, Wang W, Zhai W, Zhu X, et al. Omnigeo: towards a multimodal large language models for geospatial artificial intelligence. arXiv:2503.16326. 2025.
38. Faisal F, Wang Y, Anastasopoulos A. Dataset geography: mapping language data to language users. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022 May 22–27; Dublin, Ireland. p. 3381–411.