

Learning to Draw ASCII Improves Spatial Reasoning in Language Models

Shiyuan Huang*, Li Liu*, Jincheng He, Leilani H. Gilpin[†]

University of California, Santa Cruz

{shuan101, lliu112, jhe516, lgilpin}@ucsc.edu

Abstract

When faced with complex spatial problems, humans naturally sketch layouts to organize their thinking, and the act of drawing further sharpens their understanding. We ask whether the same principle holds for Large Language Models (LLMs): can learning to construct explicit visual layouts from spatial descriptions instill generalizable spatial understanding? We introduce TEXT2SPACE¹, a dataset that pairs natural language descriptions with ground-truth ASCII grid layouts and spatial QA pairs, enabling us to separate failures in constructing spatial representations from failures in reasoning over them. We adopt ASCII because it is human-readable, operates entirely within the token space of language models, and encodes spatial relations in a structurally verifiable form. Our evaluation reveals a pronounced “Read-Write Asymmetry”: LLMs interpret ASCII representations effectively but struggle to produce them from text, and these construction errors propagate to incorrect answers downstream. We therefore train models on layout construction (Desc→ASCII) and find that it substantially improves spatial reasoning from text alone, even without producing any ASCII at inference time. Combining construction with comprehension training further amplifies these gains on several benchmarks. Crucially, these improvements hold across model families and scales and transfer to three external benchmarks, demonstrating that, much as sketching sharpens human spatial thinking, learning to construct explicit layouts instills spatial understanding that generalizes beyond the training format.

1 Introduction

Spatial reasoning (Byrne & Johnson-Laird, 1989; Newcombe & Huttenlocher, 2000) is fundamental to human cognition (Malanchini et al., 2020). As large language models (LLMs) are increasingly applied to spatial reasoning tasks, a growing body of work has begun evaluating their capabilities in this domain (Li et al., 2024; Cohn & Hernandez-Orallo, 2023; Wang et al., 2024; Liu et al., 2023; Cheng et al., 2024).

However, existing benchmarks are limited to evaluating final answers (Yamada et al., 2024; Liu et al., 2023). When a model fails, it is unclear whether the error stems from a flawed spatial representation or from faulty reasoning over a correct one. More critically, even correct answers may reflect surface-level pattern matching rather than robust spatial understanding. A model can exploit lexical cues in simple configurations yet fail unpredictably as spatial complexity grows. Prior work on faithful reasoning has shown that correct answers produced through unreliable intermediate steps offer no guarantee of consistent behavior on novel problems (Jacovi & Goldberg, 2020; Creswell & Shanahan, 2022; Parcalabescu & Frank, 2024; Xiong et al., 2025). Without a way to inspect and verify the model’s intermediate spatial representation, these failure modes remain invisible.

*Equal Contribution

[†]Corresponding Author: lgilpin@ucsc.edu

¹Code and Dataset: <https://github.com/aiea-lab/Text2Space>

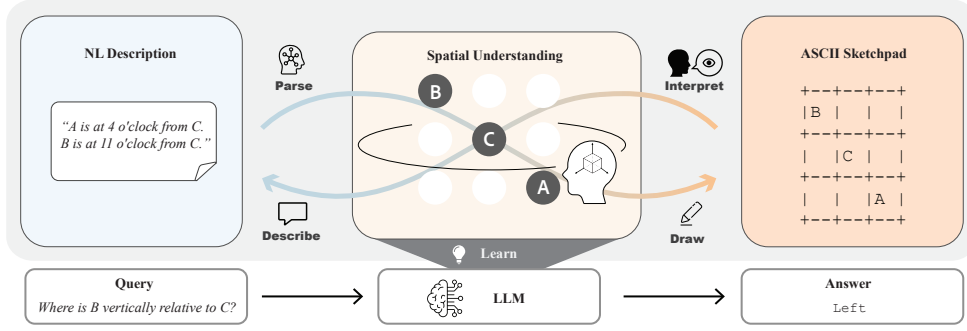


Figure 1: Overview of the study design. Inspired by empirical cognitive human strategies, we utilize Natural Language (NL) and ASCII representations as independent reasoning sources. This bidirectional cycle connects the two modalities via Parsing, Drawing, Interpreting, and Describing steps. We find that training LLMs to mimic this human-like, sketch-based reasoning process substantially improves accuracy on spatial reasoning tasks.

To open this black box, we use ASCII grid layouts as an explicit, structured text that makes the reasoning process itself observable. ASCII is well suited as this diagnostic lens because it lives within the model’s own token space, is immediately interpretable by humans, and can be formally verified against ground truth. We introduce TEXT2SPACE, a benchmark of diverse spatial configurations paired with natural language descriptions, images, and three styles of ASCII grids. Each instance includes a ground-truth layout that enables direct evaluation of the model’s intermediate spatial representation.

Figure 1 outlines our study design, which explores the bidirectional cycle between natural language and ASCII representations. Using TEXT2SPACE, we contribute three investigations. First, we **benchmark** language models on ASCII comprehension, generation, and spatial reasoning. Second, we examine the **reasoning process** by assessing whether intermediate ASCII representations faithfully support model reasoning. Most importantly, we investigate whether the spatial understanding acquired through ASCII-based learning **generalizes** to purely linguistic spatial reasoning tasks.

Our evaluation yields three main findings that progressively identify a representation bottleneck (what we call the *construction gap*) and show how to mitigate it. **Finding 1: Models can read spatial layouts but struggle to construct them.** Across all models, interpreting ASCII diagrams is substantially easier than producing them from text, identifying construction as the primary constraint. **Finding 2: This construction gap directly harms reasoning.** When models sketch layouts as an intermediate step, their construction errors propagate to the final answer. Yet when accurate ASCII is provided instead, reasoning improves substantially over text alone, showing that the bottleneck is not reasoning itself but the quality of the representation it operates over. **Finding 3: Training on construction reduces the asymmetry and the benefits generalize.** Training models on layout construction improves spatial reasoning from text alone, even without producing any ASCII at inference time. Comprehension training alone does not consistently help, isolating construction as the active ingredient. Adding comprehension on top of construction yields the strongest results on several benchmarks. These gains replicate across model families and scales and transfer to three external benchmarks, confirming that construction training builds transferable spatial understanding rather than format-specific heuristics.

We present three contributions: (1) TEXT2SPACE, a benchmark pairing natural-language spatial descriptions with ground-truth ASCII layouts, rendered images, and spatial QA pairs; (2) a diagnostic protocol that separates representation failures from reasoning failures by measuring the construction, interpretation, and utilization of ASCII layouts; and (3) empirical evidence that learning to construct spatial layouts builds transferable spatial understanding, validated on three external benchmarks.

2 Related Work

Linguistic Spatial Reasoning Benchmarks. Early work on spatial language processing focused on extracting formal representations from natural language text (Kordjamshidi et al., 2011). Subsequent benchmarks like bAbI (Weston et al., 2016) introduced positional reasoning, while STEPGame (Shi et al., 2022) and SPARTQA (Mirzaee et al., 2021) advanced the field with more complicated reasoning and broader linguistic coverage. However, a critical limitation remains across these benchmarks: they rely exclusively on final-answer accuracy. This "black-box" evaluation obscures the underlying inference process, making it impossible to distinguish between failures in representation and failures in reasoning. Our work addresses this methodological gap by introducing ground-truth intermediate representations, enabling precise diagnosis of spatial reasoning errors.

While integrating visual modalities (VLMs) offers a potential remedy for spatial understanding, recent studies reveal that VLMs face their own challenges in visual perception (Liu et al., 2023; 2024; Song et al., 2025; Kamath et al., 2023) and spatial reasoning (Chen et al., 2024; Zhang et al., 2025a; Ogezi & Shi, 2025). Rather than relying on a separate visual encoder, we use ASCII grids to represent spatial layouts entirely within the language model’s native token space. These text-native grids serve a dual purpose: they act not only as a reasoning source within the context but also as a verifiable representation of all spatial relations.

ASCII and Spatial Representation. The utility of intermediate reasoning steps is well-established through Chain-of-Thought (CoT) prompting (Wei et al., 2022) and its extensions, such as Tree-of-Thoughts (Yao et al., 2023) and latent reasoning (Hao et al., 2025). Beyond purely linguistic traces, prior work has shown that symbolic, text-native spatial representations can effectively support spatial reasoning in language models, including grid-based worlds and discrete relational layouts (Côté et al., 2018). Visualization-of-Thought (Wu et al., 2024; Li et al., 2025) draws on human mental imagery to generate symbolic or pixel-level visualizations for spatial reasoning. While we share the intuition that explicit spatial construction aids reasoning, we focus on a different question: when and why does this process fail, and can training on it improve performance? We use ASCII as a diagnostic medium because, unlike natural language, its spatial outputs are structurally verifiable against objective ground truth.

ASCII Environments and Game-Based Grounding. ASCII has long served as a productive spatial medium: agents in the NetHack Learning Environment (Küttler et al., 2020) and MiniHack (Samvelyan et al., 2021) learn complex spatial behavior directly from ASCII observations, and ASCII-Bench (Luo et al., 2025) evaluates language models on understanding ASCII art. Our work instead treats ASCII as a structured encoding of spatial layouts and uses it as a training signal for downstream reasoning. A related line grounds evaluation in game play: Clembench (Chalamalasetti et al., 2023) assesses chat-optimized models through dialogue games, Hakimov et al. (2025) extend this paradigm to multimodal models, and Suglia et al. (2024) survey visually grounded language learning more broadly. We share the motivation of structurally verifiable interaction but isolate the spatial-reasoning axis, making layout construction the training objective itself, a text-native form of grounding free of perceptual confounds.

Generalization of Spatial Reasoning. Recent research demonstrates that enhancing specific capabilities in LLMs often yields transfer benefits to other domains. For instance, Shao et al. (2024) indicate that training on mathematics improves general reasoning benchmarks, not just quantitative tasks. Prior work shows that spatial training benefits visual reasoning tasks (Wei et al., 2025; Han et al., 2025; Wang et al., 2025; Hu et al., 2025), typically through training on perception or comprehension objectives. We study a complementary mechanism: whether learning to construct spatial layouts can strengthen spatial understanding in language models, yielding transferable gains in spatial reasoning.

3 TEXT2SPACE Dataset

We introduce TEXT2SPACE, a spatial reasoning QA dataset with text-based graph (ASCII) representations designed to validate the reasoning process in LLMs. This dataset evaluates how well LLMs interpret, align, and reason about spatial relationships. It presents diverse spatial configurations across four formats: ASCII art, natural language, rendered images, and Query-Answer pairs. An example is shown in Figure 2.

3.1 Data Generation

The dataset is constructed from connected spatial graphs where entities are linked via eight directional relationships: *above*, *below*, *left*, *right*, *upper-left*, *upper-right*, *lower-left*, *lower-right*. All relations adhere to a unified coordinate system, serving as the ground truth for both ASCII diagrams and rendered images. Instances are generated using a breadth-first search (BFS) procedure to guarantee connectivity. We enforce strict validation rules: self-loops and duplicate edges are prohibited, and any configuration conflicting with existing spatial constraints is rejected. These safeguards ensure that every instance represents a valid, internally consistent topology. A detailed data generation process is presented in the Appendix G.

We vary complexity from simple two-component structures to intricate graphs with up to eight components and twelve relations, capturing configurations that require transitive reasoning across multiple inference steps. Each instance includes natural language descriptions synthesized from the spatial graph, using diverse synonyms to ensure linguistic variation. Corresponding queries are designed so that target object pairs admit unique inferability, yielding a single unambiguous spatial relation.

3.2 Multi-Format Representations

We derive both ASCII and image representations from the same underlying structure to facilitate multimodal evaluation and cross-format consistency.

ASCII Visualization: We include three rendering styles to test robustness against formatting variations. The *simple format* uses minimal spacing without borders for clarity. The *grid format* introduces structure using +, -, and | characters. The *panel format* employs box-drawing characters for a visually distinct layout.

Image Representation: The spatial arrangement is also rendered in a standardized image format with configurable grid lines and directional arrows, serving as both a multimodal evaluation target and a human-readable reference. Further dataset details and data examples are provided in Appendix A.

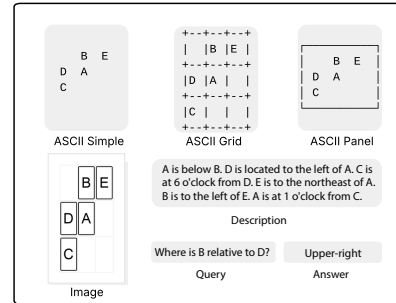


Figure 2: An example in our TEXT2SPACE dataset: we have natural language description, Q-A pairs, three types of ASCII, and an image.

4 Experiments

We split TEXT2SPACE into training and test sets, with the test set consisting of 1,000 samples used in the following experiments. We choose the ASCII *grid format* throughout our experiments; details of this format are provided in Appendix B. Our experiments address three questions in sequence. **Benchmarking:** Can models both read and draw spatial layouts, and where does the gap lie? **Reasoning process:** When models sketch a layout as an intermediate step, does the quality of that sketch help or hurt downstream reasoning? **Generalization:** Does training on layout construction instill spatial understanding that transfers to text-only reasoning and external benchmarks, even when no layout is produced at inference time?

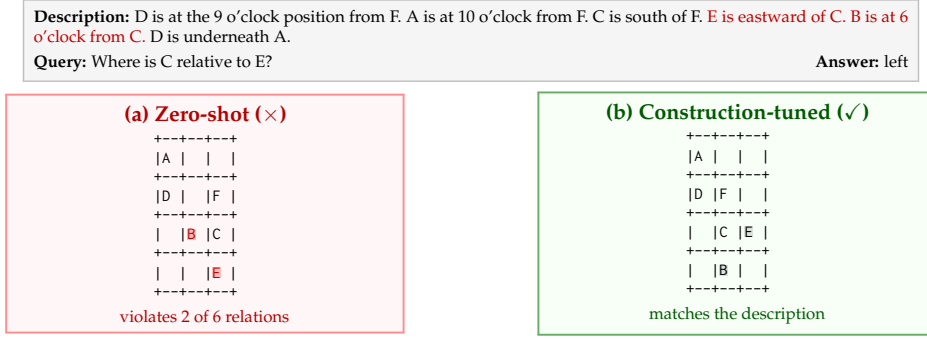


Figure 3: An example of construction task ($Desc \rightarrow ASCII$) with our TEXT2SPACE. Zero-shot, QWEN3-30B-A3B misplaces **B** and **E** (a): the layout is locally plausible but breaks the two statements shown in red in the description. After construction training ($FT_{D \rightarrow A}$), the model places both entities correctly (**B**, **E**) and reproduces the ground truth exactly (b). Such near-miss sketches typify the failures analyzed in Section 5.1.

4.1 Models

We evaluate two model categories. **Instruct models:** including LLAMA-3-70B-INSTRUCT (AI@Meta, 2024), QWEN2.5-7B-INSTRUCT (Yang et al., 2024; Team, 2024), QWEN3-30B-A3B-INSTRUCT-2507, QWEN3-235B-A22B-INSTRUCT-2507 (Team, 2025), GPT-4.1 (OpenAI, 2025). **Thinking model:** GPT-OSS-20B (OpenAI, 2025). Unless otherwise specified, we use default settings and apply 5-shot prompting (Brown et al., 2020) to ensure consistency.

4.2 Task Setups

Since ASCII is a unique representation that encodes two-dimensional spatial structure within a one-dimensional token sequence, we first evaluate language models’ ability to understand and generate ASCII representations. Specifically, we assess two complementary capabilities: generating an ASCII representation from a textual description ($Desc \rightarrow ASCII$) and generating a natural-language description from an ASCII representation ($ASCII \rightarrow Desc$). Figure 3 shows a worked example of the construction task, contrasting a failed zero-shot generation with a successful one after construction training.

We then evaluate spatial reasoning tasks, denoted $\mathcal{R}$, in which the model is given a reasoning source and asked to answer queries about spatial relationships. $\mathcal{R}_{Desc}$ refers to reasoning from textual descriptions alone, while $\mathcal{R}_{ASCII}$ refers to reasoning from ASCII representations alone. Beyond reasoning from a single source, we investigate the role of ASCII as an intermediate step alongside descriptions through two generation orders. In $\mathcal{R}_{Desc(Ans, ASCII)}$, the model first produces the answer and then generates an ASCII representation. In $\mathcal{R}_{Desc(ASCII, Ans)}$, the model first generates an ASCII representation and then uses it to answer the query. For both orders, we consider *one-turn generation*, where the model produces both outputs within a single response, and *two-turn generation*, where the second output is elicited via a follow-up prompt. The two generation orders mirror established self-explanation paradigms (Huang et al., 2023; Madsen et al., 2024): we compare two generation orders: predict-then-explain versus explain-then-predict (Camburu et al., 2018).

Finally, we evaluate $\mathcal{R}_{Desc+ASCII}$. In this phase, we substitute the model-generated ASCII with the ground-truth ASCII before the model answers the query. This replacement strategy simulates an ideal reasoning process, allowing us to assess the upper-bound performance of the model when guided by an ideal spatial representation.

4.3 Evaluation Methods

Our evaluation protocol targets three key aspects of the model’s performance: **QA Correctness:** For the spatial reasoning QA tasks, we validate the prediction against a predefined

set of relation aliases. **Representation Alignment:** For the $ASCII \leftrightarrow Desc$ conversion tasks, we measure the correspondence between converted representations. For $ASCII \rightarrow Desc$, the ASCII diagram is treated as the reference. We evaluate, across all generated descriptions, the percentage that can be correctly inferred from the given ASCII representation. For $Desc \rightarrow ASCII$, we evaluate the generated ASCII diagrams, $\mathcal{R}_{Desc(Ans, ASCII)}$ and $\mathcal{R}_{Desc(ASCII, Ans)}$, by taking the description as the reference and measuring the proportion of the description that is correctly reflected in the ASCII output. **Reasoning Consistency:** For the joint generation settings, we introduce two complementary metrics to analyze internal coherence. **Consistency (Cons):** Measures the alignment between the generated ASCII and the generated Answer. This determines if the model’s intermediate output supports its final prediction. **Consistency Accuracy (C_Acc):** Measures answer accuracy strictly on instances where the ASCII and Answer are consistent. By controlling for internal coherence, this isolates the model’s reasoning capability, helping distinguish whether errors arise from failures in constructing spatial representations or from limitations in reasoning over them.

4.4 Learning Spatial Relations with ASCII

We propose that ASCII layouts function as a structured reasoning intermediate: a verifiable “cognitive map” that makes spatial relationships explicit. We hypothesize that the learning process to construct such maps (Lightman et al., 2024; Zhang et al., 2025b) consolidates spatial logic during training. If so, models trained on layout construction should improve on spatial reasoning tasks even without producing any layout at inference time, and the benefit should transfer beyond the training format. To test and disentangle the understanding and generation of ASCII representations, we design three configurations. **Learning to Construct** ($Desc \rightarrow ASCII$) forces the model to actively construct spatial layouts from text (performing “mental mapping”). **Learning to Comprehend** ($ASCII \rightarrow Desc$) focuses on parsing and verbalizing ASCII representations. **Bidirectional Alignment** ($Desc \leftrightarrow ASCII$) combines both tasks simultaneously.

We fine-tune QWEN3-30B-A3B on 4x NVIDIA A40 GPUs using the SFTTrainer from the HuggingFace TRL library (von Werra et al., 2020). We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022; Mangrulkar et al., 2022) for parameter-efficient training, utilizing a split of 4,000 training and 500 validation examples. We adhere to the default library configurations for all other settings; full training details are provided in Appendix 7. To assess generality across model families and scales, we additionally fine-tune LLAMA-3-8B-INSTRUCT and LLAMA-3-70B-INSTRUCT under the same procedure.

We evaluate performance on the Spatial Reasoning Task ($\mathcal{R}_{Desc}$), where models answer spatial relations solely from natural language descriptions without explicit ASCII input. To assess whether the benefits extend beyond our benchmark, we further evaluate on three external tasks that each target a different facet of spatial cognition: STEP GAME (Shi et al., 2022; Rizvi et al., 2024) examines compositional directional navigation, BABI task 19 (Weston et al., 2016) tests route planning through connected locations, and SPARTQA (Mirzaee et al., 2021) challenges models with complex spatial questions in rich natural language. None of these tasks involve ASCII, and we evaluate the model’s performance with zero-shot prompting setting. If construction training instills generalizable spatial understanding, gains should appear on these tasks despite the format mismatch.

5 Results and Discussion

5.1 Benchmarking

ASCII Understanding and Generation Table 1 presents the results for the $Desc \leftrightarrow ASCII$ conversion tasks. Across all models, we observe a clear and consistent asymmetry: performance on the interpretation task ($ASCII \rightarrow Desc$) consistently exceeds that of the generation task ($Desc \rightarrow ASCII$). This pattern suggests that while models are relatively proficient at “reading” or interpreting patterns in ASCII, they struggle considerably more with “drawing” it correctly. Even the reasoning model GPT-OSS-20B renders only 44.0% of the described relations faithfully, well below the 58.1% it achieves on the ASCII interpretation task.

	Base models						Qwen3-30B Fine-tuned		
	QWEN2.5-7B	LLAMA-3-70B	GPT-4.1	GPT-OSS-20B	QWEN3-235B	QWEN3-30B	FT _{D→A}	FT _{D←A}	FT _{D↔A}
$D \rightarrow A$	16.1	43.9	53.6	44.0	46.0	29.2	67.9	24.1	64.6
$A \rightarrow D$	38.4	57.4	72.5	58.1	59.7	55.8	53.5	66.5	68.4

Table 1: Bidirectional mapping results for generating ASCII from textual descriptions ($\text{Desc} \rightarrow \text{ASCII}$) and identifying descriptions from ASCII ($\text{ASCII} \rightarrow \text{Desc}$). Each score is the average proportion of relations correctly aligned between the two representations.

Model	$\mathcal{R}_{\text{Desc}}$	$\mathcal{R}_{\text{ASCII}}$	$\mathcal{R}_{\text{Desc}(\text{Ans}, \text{ASCII})}$		$\mathcal{R}_{\text{Desc}(\text{ASCII}, \text{Ans})}$		$\mathcal{R}_{\text{Desc}+\text{ASCII}}$
	Ans	Ans	Ans	ASCII	Ans	ASCII	Ans
<i>One-turn generation</i>							
QWEN2.5-7B	42.0	39.9	<u>43.6</u>	12.5	36.7	12.0	55.5
LLAMA-3-70B	71.7	<u>75.3</u>	71.1	42.2	66.9	41.7	78.4
GPT-4.1	85.8	<u>90.1</u>	86.5	50.6	83.8	50.5	92.8
GPT-OSS-20B	93.3	<u>92.7</u>	94.3	38.9	91.9	42.8	97.0
QWEN3-235B-A22B	84.8	<u>89.0</u>	84.2	45.3	82.3	45.4	90.1
QWEN3-30B-A3B	75.0	<u>83.2</u>	71.1	27.9	66.6	27.5	83.7
QWEN3-30B-A3B FT _{D→A}	82.2	78.6	<u>82.5</u>	64.2	79.1	64.7	84.9
QWEN3-30B-A3B FT _{D←A}	73.7	<u>87.4</u>	68.0	22.0	70.2	20.0	87.2
QWEN3-30B-A3B FT _{D↔A}	82.7	81.3	82.9	60.2	<u>83.1</u>	61.0	86.2
<i>Two-turn generation</i>							
QWEN2.5-7B			<u>43.6</u>	13.5	42.1	13.7	52.3
LLAMA-3-70B			<u>70.9</u>	42.8	69.0	42.3	78.8
GPT-4.1			<u>85.9</u>	50.8	81.0	50.6	93.0
GPT-OSS-20B			<u>93.1</u>	45.0	87.7	47.9	94.3
QWEN3-235B-A22B			<u>85.6</u>	45.1	77.8	45.3	89.0
QWEN3-30B-A3B			<u>74.3</u>	27.4	69.3	26.4	82.6
QWEN3-30B-A3B FT _{D→A}			80.7	65.1	<u>81.3</u>	64.7	83.7
QWEN3-30B-A3B FT _{D←A}			<u>70.1</u>	22.1	65.2	20.2	87.7
QWEN3-30B-A3B FT _{D↔A}			<u>83.1</u>	61.4	82.7	61.8	87.5

Table 2: Spatial reasoning accuracy across reasoning sources and settings. $\mathcal{R}_{\text{Desc}(\text{Ans}, \text{ASCII})}$ and $\mathcal{R}_{\text{Desc}(\text{ASCII}, \text{Ans})}$ denote whether the answer precedes or follows the ASCII output, generated jointly (*one-turn*) or in separate calls (*two-turn*); the two-turn $\mathcal{R}_{\text{Desc}+\text{ASCII}}$ provides gold-standard ASCII. Answer accuracy: label matching; ASCII accuracy: average proportion of NL relations correctly depicted. **Bold**: best reasoning source; underlined: second best.

The disparity spans 13 to 27 percentage points across all six base models (Table 1), pointing to a systematic bottleneck rather than an incidental failure mode. Interpretation relies on local pattern recognition within a provided grid, whereas generation demands global constraint satisfaction, jointly positioning all entities so that every described relation is upheld. As we show next, this construction deficit propagates directly to reasoning failures, and mitigating it through targeted training yields spatial understanding that transfers beyond the ASCII format (Section 5.3).

Spatial Reasoning Performance Table 2 reports accuracy on spatial reasoning tasks ($\mathcal{R}$) under different settings. On the text-only baseline ($\mathcal{R}_{\text{Desc}}$), accuracy ranges from 42.0% (QWEN2.5-7B) to 93.3% (GPT-OSS-20B). This spread confirms that our benchmark meaningfully differentiates spatial reasoning ability, yet most models have substantial room to improve, motivating our examination of whether ASCII can narrow this gap.

Providing ASCII as input substantially refines performance. Most models reason more accurately from ASCII alone ($\mathcal{R}_{\text{ASCII}}$) than from text descriptions ($\mathcal{R}_{\text{Desc}}$), and the combination of both ($\mathcal{R}_{\text{Desc}+\text{ASCII}}$) yields the highest accuracy. This shows that as an input, ASCII provides the layout directly, without the transitive composition that natural language requires.

In contrast, model-generated ASCII as an intermediate step does not improve reasoning. Answering first and then drawing ($\mathcal{R}_{\text{Desc}(\text{Ans}, \text{ASCII})}$) roughly matches the text-only baseline ($\mathcal{R}_{\text{Desc}}$), while drawing first and then answering ($\mathcal{R}_{\text{Desc}(\text{ASCII}, \text{Ans})}$) drops below it across almost all base models in both *one-turn* and *two-turn* settings. With generated ASCII accuracy at only 12-50% across models, the sketches are not accurate enough to serve as a promising reasoning scaffold. Since provided ASCII yields the highest accuracy while

model-generated ASCII degrades it, the bottleneck is sketch quality, not the format itself. This echoes prior findings that generating intermediate artifacts can reduce task accuracy (Camburu et al., 2018; Huang et al., 2023). However, unlike unstructured self-explanations or textual CoT (Wei et al., 2022), where errors in intermediate steps are difficult to detect, ASCII is structurally verifiable. This allows us to directly attribute the degradation to **flawed construction** rather than treating it as a black-box effect.

This diagnosis points to a concrete path forward: since the deficiency lies in representation construction, the next step is to examine whether training on this capability can alleviate this deficiency and strengthen the model’s spatial understanding. We first analyze whether the model’s generated representations are internally coherent with its answers, using consistency as an unsupervised signal of reasoning reliability (Section 5.2). We then investigate whether training on layout construction can instill the spatial understanding needed to bridge this discrepancy (Section 5.3).

5.2 Consistency Analysis in Spatial Reasoning

Table 3 shows the alignment between ASCII representations and textual answers across both output orders and turn settings. We report three key metrics: (1) Answer Accuracy measures overall task performance; (2) Consistency quantifies the logical alignment between the generated ASCII representation and the textual answer; and (3) Consistency Accuracy evaluates correctness specifically on instances where the representation and answer correspond.

In the *one-turn* generation setting, producing the ASCII representation before the answer does not improve consistency. Similar consistency across output orders suggests that the model does not clearly use the ASCII as a reasoning scaffold, with both outputs likely reflecting the same underlying representation.

In the *two-turn* generation setting, we observed an improved consistency rate. The consistency gap between $\mathcal{R}_{\text{Desc}(\text{Ans}, \text{ASCII})}$ and $\mathcal{R}_{\text{Desc}(\text{ASCII}, \text{Ans})}$ indicates that the response to the follow-up query is strongly influenced by the previously generated output. Despite this notable improvement in consistency, answer accuracy increases only marginally. This limited gain may be partly due to the suboptimal quality of the generated ASCII representations, which correlates with lower answer accuracy (Table 2).

Nevertheless, consistency remains a valuable indicator, particularly given that it operates without requiring ground-truth annotations for either the spatial layout or the final answer. While prior work has leveraged Natural Language Inference (NLI) to assess reasoning via consistency (Wang et al., 2023; Lyu et al., 2023; Kumar & Talukdar, 2020), our approach benefits from the explicit verifiability of the generated artifact. Unlike unstructured text, ASCII representations enforce strict spatial constraints and encode the full spatial configuration, not just the queried pair, allowing us to objectively assess the alignment between the representation and the answer. When the model’s ASCII representation and its textual answer agree, accuracy is substantially higher than on disagreeing instances, so agreement serves as an unsupervised reliability signal; conversely, accuracy on inconsistent instances falls well below average (Appendix F).

From a forward-looking perspective, our results show “ASCII as a Verifier” as a promising direction for verifiable reasoning: agreement between two representations of the same content offers a label-free signal of which outputs to trust, both for error detection and as a training objective for representation construction.

5.3 Learning Spatial Relations with ASCII

Table 1 details the performance on conversion tasks after fine-tuning. We note that construction training is an auxiliary task: the model learns to produce ASCII layouts, not to answer spatial questions. Construction training ($\text{FT}_{\text{D} \rightarrow \text{A}}$) raises generation accuracy ($\text{Desc} \rightarrow \text{ASCII}$) from 29.2% to 67.9%, a much larger gain than comprehension training achieves ($\text{Desc} \leftarrow \text{ASCII}$: 55.8% to 66.5%). The disparity confirms that ASCII generation is the primary bottleneck, with far more room for improvement. Unidirectional training improves the targeted direc-

Model	One-turn generation						Two-turn generation					
	$\mathcal{R}_{\text{Desc(Ans, ASCII)}}$			$\mathcal{R}_{\text{Desc(ASCII, Ans)}}$			$\mathcal{R}_{\text{Desc(Ans, ASCII)}}$			$\mathcal{R}_{\text{Desc(ASCII, Ans)}}$		
	Ans	Cons	C_Acc	Ans	Cons	C_Acc	Ans	Cons	C_Acc	Ans	Cons	C_Acc
QWEN2.5-7B	43.6	23.0	65.2	36.7	20.5	64.4	43.6	24.4	69.3	42.1	31.1	56.9
LLAMA-3-70B	71.1	65.2	87.7	66.9	65.4	83.8	70.9	63.9	89.9	69.0	71.3	81.9
GPT-4.1	86.5	74.2	94.9	83.8	78.7	90.2	85.9	77.3	93.9	81.0	81.3	86.0
GPT-OSS-20B	94.3	75.9	97.0	91.9	75.5	96.9	93.1	71.8	96.7	87.7	77.2	91.6
QWEN3-235B	84.2	67.8	95.0	82.3	72.5	89.9	85.6	70.6	93.2	77.8	78.4	84.1
QWEN3-30B	71.1	46.1	88.5	66.6	46.3	82.5	74.3	50.8	88.8	69.3	58.6	75.7
QWEN3-30B FT _{D→A}	82.5	73.2	97.5	79.1	71.9	96.5	80.7	74.1	98.1	81.3	74.9	96.7
QWEN3-30B FT _{D←A}	68.0	39.2	84.9	70.2	46.4	81.2	70.1	42.8	85.7	65.2	55.7	69.7
QWEN3-30B FT _{D↔A}	82.9	73.7	97.4	83.1	72.9	96.6	83.1	73.8	96.9	82.7	76.1	94.2

Table 3: Answer accuracy (Ans), consistency between the generated ASCII and the generated answer (Cons), and accuracy on consistent instances (C_Acc) across generation settings. For each model and setting, the higher Cons value is in **bold**.

tion but slightly degrades the other, while joint training ($\text{Desc} \leftrightarrow \text{ASCII}$) improves both tasks (64.6% generation, 68.4% understanding) without this trade-off.

We now evaluate whether these improvements transfer to the downstream spatial reasoning task ($\mathcal{R}_{\text{Desc}}$), which does not generate ASCII at inference time but requires answering spatial queries from text alone, a task format entirely absent from fine-tuning. Table 2 shows that construction training ($\text{FT}_{\text{D} \rightarrow \text{A}}$) improves accuracy by over 7 percentage points, indicating that learning to construct spatial layouts strengthens spatial reasoning from text alone. In contrast, comprehension training ($\text{FT}_{\text{D} \leftarrow \text{A}}$) does not reliably improve performance despite equal data exposure, and even fall slightly below the baseline, suggesting that better comprehension alone may not translate into better reasoning.

Fine-tuning on layout construction ($\text{FT}_{\text{D} \rightarrow \text{A}}$) substantially improves ASCII accuracy for $\mathcal{R}_{\text{Desc(Ans, ASCII)}}$ and $\mathcal{R}_{\text{Desc(ASCII, Ans)}}$, reversing the pattern from Section 5.1: accurate sketches now scaffold reasoning rather than mislead. The bottleneck was not the representation format but **sketch quality**: after fine-tuning, ground-truth ASCII ($\mathcal{R}_{\text{Desc+ASCII}}$) still outperforms model-generated sketches, a gap that further improvements in construction may close.

The persistent advantage of externally provided layouts, even after construction training, parallels human spatial cognition (Larkin & Simon, 1987). In a preliminary human validation (Appendix J, $n = 11$), participants who drew the layout themselves answered less accurately than when a correct layout was supplied (With accuracy 0.87 vs. 0.98). Reconstructing spatial structure is harder than reading a provided layout, for humans and models alike.

The consistency analysis (Table 3) tells a complementary story. Construction training improves consistency between the model’s sketch and its answer far more than comprehension training does. The mechanism is intuitive: when the sketch is accurate, it reinforces the textual answer; when it remains flawed, better comprehension only makes the model more aware of the conflict between its noisy sketch and the description. The confusion matrices (Appendix, Figure 5) corroborate this: construction training reduces confusions between spatially adjacent classes, sharpening directional boundaries.

Table 4 tests whether these gains extend beyond TEXT2SPACE by evaluating all three base models (QWEN3-30B, LLAMA-3-8B, and LLAMA-3-70B) on three external benchmarks under zero-shot prompting, without ASCII at inference time. Construction training ($\text{FT}_{\text{D} \rightarrow \text{A}}$) is the only one of the three configurations with generally non-negative deltas on all benchmark cells. Comprehension training alone ($\text{FT}_{\text{D} \leftarrow \text{A}}$) is inconsistent: it degrades STEPGAME, which tests precise directional navigation, for all three models, while its large gains on SPARTQA, which emphasizes parsing complex spatial language, appear only for QWEN3-30B and LLAMA-3-70B, indicating that comprehension alone cannot substitute for construction. Bidirectional training ($\text{FT}_{\text{D} \leftrightarrow \text{A}}$) further improves over construction alone on several model-benchmark pairs (Table 4), showing that comprehension reinforces construction.

Model	STEPGAME	BABI Task 19	SPARTQA
QWEN3-30B (base)	54.9	17.8	26.5
W/ Construction ($FT_{D \rightarrow A}$)	+9.9	+17.0	+43.3
W/ Comprehension ($FT_{D \leftarrow A}$)	-6.1	+10.7	+44.2
W/ Bidirectional ($FT_{D \leftrightarrow A}$)	+8.4	+20.6	+49.0
LLAMA-3-8B (base)	44.8	17.7	51.3
W/ Construction ($FT_{D \rightarrow A}$)	+1.5	+2.6	+9.9
W/ Comprehension ($FT_{D \leftarrow A}$)	-0.2	+4.7	-3.4
W/ Bidirectional ($FT_{D \leftrightarrow A}$)	+7.0	+10.2	+5.5
LLAMA-3-70B (base)	55.7	56.3	59.2
W/ Construction ($FT_{D \rightarrow A}$)	+10.7	+4.4	+13.5
W/ Comprehension ($FT_{D \leftarrow A}$)	-4.3	-12.1	+13.6
W/ Bidirectional ($FT_{D \leftrightarrow A}$)	+2.3	-0.5	+21.3

Table 4: Accuracy on STEPGAME, BABI Task 19, and SPARTQA under a zero-shot setting without ASCII at inference time. STEPGAME evaluation uses the SpaRP PS2 1k test split. Fine-tuned rows report deltas relative to each base model, with the best delta per benchmark and model in **bold**.

6 Conclusion

We investigate whether training language models to construct explicit spatial layouts can instill generalizable spatial understanding. Using ASCII grids as a diagnostic instrument that is text-native, human-readable, and structurally verifiable, we introduce TEXT2SPACE, a benchmark to separate construction failures from reasoning failures.

Our results reveal a clear read-write asymmetry: models interpret spatial layouts effectively but struggle to construct them from text, and these construction errors propagate to downstream reasoning. Consistency analysis further shows that agreement between a model’s ASCII and its textual answer is an unsupervised reliability signal: accuracy is far higher when they agree. Training models to construct layouts addresses this bottleneck directly: spatial reasoning from text improves even without sketching at inference time, and the benefit transfers to three external benchmarks. Only construction training generalizes across all benchmarks and models, and adding comprehension on top generally yields the largest gains on linguistically complex tasks. The spatial understanding gained through layout construction thus transfers beyond the training format.

Limitations and Future Work

Every boundary of this study follows from one requirement: the intermediate layout must be verifiable relation by relation, since that is what separates construction failures from reasoning failures. ASCII grids meet it, and TEXT2SPACE uses a categorical vocabulary of fixed-frame 2D relations, which places projective relations, free-form spatial language, and 3D beyond what we measure. The study is text-only, and visual settings remain untested. Configurations are capped at eight components and twelve relations, a scale at which the strongest tested model still draws only about half of the relations correctly (Table 1).

Each boundary is also a direction. Replacing the ASCII target with coordinate lists or scene graphs while holding data and training fixed would test whether the medium matters or only the act of construction. Projective relations (“behind”) and full 3D descriptions first require establishing unambiguous ground truth, since “above” may mean directly overhead or merely higher. The human preference for 2D projections (Tversky, 2011) suggests stacked 2D slices as a first step. A multimodal model could likewise construct layouts grounded in images, and we expect the same read-write asymmetry on visual benchmarks such as VSI-Bench (Yang et al., 2025) and OmniSpatial (Jia et al., 2026). Because TEXT2SPACE is constructed procedurally, scaling to more components and relations is a direct extension.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-24-1-0149.

This work used Delta GPU at NCSA through allocation CIS251224 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program (Boerner et al., 2023), which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. This work used resources available through the National Research Platform (NRP) (Weitzel et al., 2025) at the University of California, San Diego. NRP has been developed, and is supported in part, by funding from National Science Foundation, from awards 1730158, 1540112, 1541349, 1826967, 2112167, 2100237, and 2120019, as well as additional funding from community partners.

References

- AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. ACCESS: Advancing innovation: NSF’s advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing (PEARC ’23)*, pp. 173–176, Portland, OR, USA, 2023. Association for Computing Machinery. doi: 10.1145/3569951.3597559.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Ruth M.J Byrne and P.N Johnson-Laird. Spatial reasoning. *Journal of Memory and Language*, 28(5):564–575, 1989. ISSN 0749-596X. doi: [https://doi.org/10.1016/0749-596X\(89\)90013-2](https://doi.org/10.1016/0749-596X(89)90013-2). URL <https://www.sciencedirect.com/science/article/pii/0749596X89900132>.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.
- Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11174–11219, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.689. URL <https://aclanthology.org/2023.emnlp-main.689/>.
- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.
- An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.
- Anthony G Cohn and Jose Hernandez-Orallo. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of llms. *arXiv preprint arXiv:2304.11164*, 2023.

- Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Anderson Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew J. Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. In *Workshop on Computer Games*, pp. 41–75. Springer, 2018.
- Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*, 2022.
- Sherzod Hakimov, Yerkezhan Abdullayeva, Kushal Koshti, Antonia Schmidt, Yan Weiser, Anne Beyer, and David Schlangen. Using game play to investigate multimodal and conversational grounding in large multimodal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5686–5718, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.coling-main.381/>.
- Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying llm visual priors from language pre-training. *arXiv preprint arXiv:2509.26625*, 2025.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025. URL <https://openreview.net/forum?id=Itxz7S4Ip3>.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- Zhe Hu, Jing Li, Zhongzhu Pu, Hou Pong Chan, and Yu Yin. Praxis-VLM: Vision-grounded decision making via text-driven reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=U806q3iILo>.
- Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*, 2023.
- Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386.
- Mengdi Jia, Zekun Qi, Shaochen Zhang, Wen Yao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, 2026. URL <https://arxiv.org/abs/2506.03135>.
- Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568.
- Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):1–36, 2011.
- Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8730–8742, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.771. URL <https://aclanthology.org/2020.acl-main.771/>.
- Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The NetHack learning environment. *Advances in neural information processing systems*, 33, 2020. URL <https://papers.nips.cc/paper/2020/hash/569ff987c643b4bedf504efda8f786c2-Abstract.html>.

- Jill H Larkin and Herbert A Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100, 1987.
- Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2025.
- Fangjun Li, David C Hogg, and Anthony G Cohn. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18500–18507, 2024.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *International Conference on Learning Representations*, 2024.
- Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.
- Li Liu, Diji Yang, Sijia Zhong, Kalyana Suma Sree Tholeti, Lei Ding, Yi Zhang, and Leilani Gilpin. Right this way: Can vlms guide us to see more to answer questions? *Advances in Neural Information Processing Systems*, 37:132946–132976, 2024.
- Kerry Luo, Michael Fu, Joshua Peguero, Husnain Malik, Anvay Patil, Joyce Lin, Megan Van Overborg, Ryan Sarmiento, and Kevin Zhu. ASCIIBench: Evaluating language-model-based understanding of visually-oriented text. *arXiv preprint arXiv:2512.04125*, 2025. URL <https://arxiv.org/abs/2512.04125>.
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AAACL 2023)*, 2023.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 295–337, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.19.
- Margherita Malanchini, Kaili Rimfeld, Nicholas G Shakeshaft, Andrew McMillan, Kerry L Schofield, Maja Rodic, Valerio Rossi, Yulia Kovas, Philip S Dale, Elliot M Tucker-Drob, and Robert Plomin. Evidence for a unitary structure of spatial cognition beyond general intelligence. *npj Science of Learning*, 5(1):9, 2020.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022.
- Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 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*, pp. 4582–4598, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.364.
- Nora Newcombe and Janellen Huttenlocher. *Making space: The development of spatial representation and reasoning*. MIT press, 2000.
- Michael Ogezi and Freda Shi. SpaRE: Enhancing spatial reasoning in vision-language models with synthetic data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7855–7875, Vienna, Austria, 2025. Association for Computational Linguistics.

- OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL <https://arxiv.org/abs/2508.10925>.
- OpenAI. Gpt-4.1 model documentation. <https://platform.openai.com/docs/models/gpt-4.1>, 2025. Accessed: 2025-12-20.
- Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, 2024.
- Md Imbesat Rizvi, Xiaodan Zhu, and Iryna Gurevych. 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)*, pp. 4750–4767, 2024.
- Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, Jack Parker-Holder, Minqi Jiang, Eric Hambro, Fabio Petroni, Heinrich Küttler, Edward Grefenstette, and Tim Rocktäschel. MiniHack the planet: A sandbox for open-ended reinforcement learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL <https://openreview.net/forum?id=skFwlyefkWJ>.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL <https://api.semanticscholar.org/CorpusID:267412607>.
- Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11321–11329, Jun. 2022. doi: 10.1609/aaai.v36i10.21383. URL <https://ojs.aaai.org/index.php/AAAI/article/view/21383>.
- Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *International Conference on Learning Representations*, 2025.
- Alessandro Suglia, Ioannis Konostas, and Oliver Lemon. Visually grounded language learning: A review of language games, datasets, tasks, and models. *Journal of Artificial Intelligence Research*, 79:173–239, 2024. doi: 10.1613/jair.1.15185. URL <https://www.jair.org/index.php/jair/article/view/15185>.
- Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL <https://qwenlm.github.io/blog/qwen2.5/>.
- Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Barbara Tversky. Visualizing thought. *Topics in cognitive science*, 3(3):499–535, 2011.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>, 2020.
- Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024.
- Sophie L Wang, Phillip Isola, and Brian Cheung. Words that make language models perceive. *arXiv preprint arXiv:2510.02425*, 2025.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, Jia Wang, Chunrui Han, Yuang Peng, Qi Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, and Vishal M. Patel. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=oEgybA04dY>.
- Derek Weitzel, Ashton Graves, Sam Albin, Huijun Zhu, Frank Wuerthwein, Mahidhar Tatineni, Dmitry Mishin, Elham Khoda, Mohammad Sada, Larry Smarr, Thomas DeFanti, and John Graham. The national research platform: Stretched, multi-tenant, scientific kubernetes cluster. In *Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration*, PEARC '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713989. doi: 10.1145/3708035.3736060. URL <https://doi.org/10.1145/3708035.3736060>.
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations*, 2016. URL <https://arxiv.org/abs/1502.05698>.
- Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind’s eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2024.
- Zidi Xiong, Shan Chen, Zhenting Qi, and Himabindu Lakkaraju. Measuring the faithfulness of thinking drafts in large reasoning models. In *Advances in Neural Information Processing Systems*, 2025.
- Yutaro Yamada, Yihan Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating spatial understanding of large language models. *Transactions on Machine Learning Research*, 2024.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL <https://arxiv.org/abs/2412.14171>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.
- Hongjie Zhang, Hourui Deng, Jie Ou, and Chaosheng Feng. Mitigating spatial hallucination in large language models for path planning via prompt engineering. *Scientific Reports*, 15(1):8881, 2025a.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL*

2025, pp. 10495–10516, Vienna, Austria, 2025b. Association for Computational Linguistics.
doi: 10.18653/v1/2025.findings-acl.547.

Appendix

A Dataset Details

Table 5 summarizes the possible values for each field in TEXT2SPACE. The dataset is in English, covering various terminologies including cardinal directions (e.g., ‘North’), clock-face references (e.g., ‘3 o’clock’), and standard spatial prepositions (e.g., ‘left of’). Figure 4 and Table 8 provides an overview of statistics for the full dataset and test set.

Field	Possible Values
query_type	full, vertical, horizontal
label (full)	above, below, left, right, upper-left, upper-right, lower-left, lower-right
label (vertical)	above, below, same level
label (horizontal)	left, right, same column
terminology_used	spatial, cardinal, clock, spatial+cardinal, spatial+clock, cardinal+clock, spatial+cardinal+clock
ascii	simple, grid, panel
num_components	2–8
num_relations	1–12
is_directly_stated	true, false
has_unique_layout	true, false
ambiguous_stages	0–11

Table 5: TEXT2SPACE Dataset Schema, including spatial query types, terminology variations, and structural constraints (components and relations).

B ASCII Format Choice

Using three different ASCII representations as input, we conduct two experiments to determine the most suitable format. The first experiment evaluates the interpretation of natural language descriptions from ASCII ($ASCII \rightarrow Desc$), while the second assesses spatial reasoning directly from ASCII ($\mathcal{R}_{ASCII}$). As shown in Table 6, performance on $ASCII \rightarrow Desc$ is comparable across all three formats. In contrast, for $\mathcal{R}_{ASCII}$, the grid-based representation yields substantially higher performance. Accordingly, we adopt the grid format in all subsequent experiments.

ASCII Format	QWEN3-30B-A3B
$ASCII \rightarrow Desc$	
simple	57.2
grid	55.8
panel	54.5
$\mathcal{R}_{ASCII}$	
simple	76.8
grid	83.2
panel	79.2

Table 6: Accuracy of QWEN3-30B-A3B across the three ASCII formats. The grid format offers the strongest reasoning performance and is adopted in all subsequent experiments.

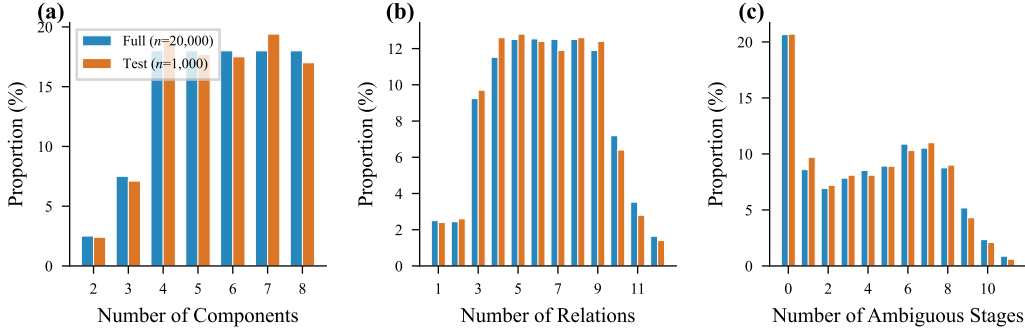


Figure 4: Distribution of (a) number of components, (b) number of relations, and (c) number of ambiguous stages. An ambiguous stage occurs when a newly introduced relation does not uniquely determine the position of an entity relative to those already placed; higher counts indicate configurations requiring more transitive inference steps. The dataset contains 20,000 instances with an average of 5.7 components and 6.4 relations.

C Training configurations

Configuration	Value
Base Model	QWEN3-30B-A3B
Frameworks	HuggingFace Transformers, TRL, PEFT
Fine-tuning Method	LoRA (Rank $r = 16$, $\alpha = 32$)
Target Modules	$q, k, v, o, gate, up, down$ projections
Precision	bfloat16
Optimizer	AdamW
Learning Rate	2×10^{-4} (Cosine Schedule)
Batch Size	4 (per device)
Gradient Accumulation	4 steps
Epochs	1
Dataset Split	4,000 (Train) / 500 (Validation)
Hardware	4 $\times$ NVIDIA A40

Table 7: Training configuration details.

For LLAMA-3-8B-INSTRUCT and LLAMA-3-70B-INSTRUCT (Table 4), we keep the same data split, objectives, and LoRA rank, apply the adapters to the attention projections (q, k, v, o), and train for one epoch with a learning rate of 1×10^{-4} and an effective batch size of 16; LLAMA-3-70B is loaded in 8-bit precision (QLoRA) to fit GPU memory.

D Confusion matrices for spatial reasoning predictions

Figure 5 presents the confusion matrices for spatial reasoning predictions before and after fine-tuning on the ASCII-description conversion tasks. Fine-tuning results in stronger concentration along the diagonal and reduces systematic confusions between spatially related classes, indicating that the residual errors concentrate less on spatially adjacent relations after fine-tuning.

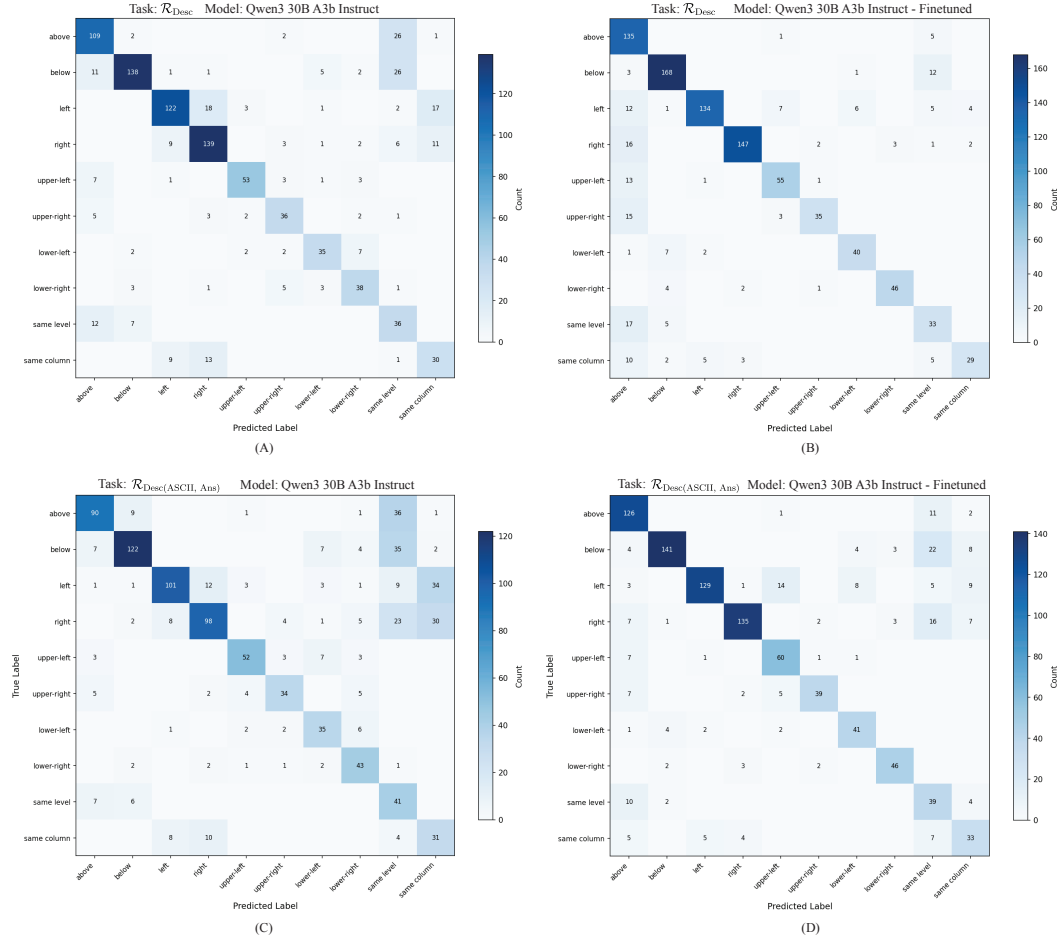


Figure 5: Confusion matrices for spatial reasoning predictions by Qwen3-30B-A3B before and after fine-tuning (Qwen3-30B-A3B $FT_{D \rightarrow A}$). The top row ((A)-(B)) shows results on the text-only spatial reasoning task ($\mathcal{R}_{Desc}$), while the bottom row ((C)-(D)) shows results when ASCII is generated as an intermediate step ($\mathcal{R}_{Desc(Ans, ASCII)}$). Left panels ((A),(C)) correspond to the base model, and right panels ((B),(D)) to the fine-tuned model. Rows indicate ground-truth spatial relations and columns indicate predicted relations. Darker diagonal cells indicate correct predictions.

Property	Full	Test
Instances	20,000	1,000
<i>Structural properties</i>		
Components (range)	2–8	2–8
Components (μ)	5.67	5.67
Relations (range)	1–12	1–12
Relations (μ)	6.44	6.33
Direct relations (%)	52.4	53.2
Unique layouts (%)	41.9	42.4
Ambiguous stages (μ)	4.13	4.02
<i>Query type (%)</i>		
Full	33.3	33.7
Vertical	33.3	32.6
Horizontal	33.3	33.7
<i>Terminology (%)</i>		
Spatial	14.3	14.9
Cardinal	14.3	14.3
Clock	14.3	14.7
Spatial + Cardinal	14.3	15.4
Spatial + Clock	14.3	12.2
Cardinal + Clock	14.3	13.5
All three	14.3	15.0

Table 8: Dataset statistics for the full dataset and the evaluation subset.

E F1 Scores

Table 9 reports the F1 scores for spatial reasoning tasks. In line with Table 2, reasoning from ASCII alone generally outperforms reasoning from descriptions alone. While generating ASCII before answering hurts performance without fine-tuning, fine-tuning with explicit ASCII generation, either unidirectional (Desc $\rightarrow$ ASCII) or bidirectional (Desc $\leftrightarrow$ ASCII), consistently improves performance when only descriptions are available at inference time.

F Accuracy on Inconsistent Instances

Table 10 reports accuracy on inconsistent instances (Inc_Acc) alongside C_Acc from Table 3. Across all nine models, Inc_Acc lies well below the corresponding overall accuracy: disagreement between the generated ASCII and the stated answer signals a probable error, without reference to ground truth. The effect is most pronounced after construction training, where Inc_Acc (30.9–41.4%) falls to approximately half of the overall accuracy (79.1–82.5%), establishing inconsistency as a reliable, label-free indicator of error.

Model	One-turn generation				Two-turn generation			
	$\mathcal{R}_{\text{Desc}(\text{Ans}, \text{ASCII})}$		$\mathcal{R}_{\text{Desc}(\text{ASCII}, \text{Ans})}$		$\mathcal{R}_{\text{Desc}(\text{Ans}, \text{ASCII})}$		$\mathcal{R}_{\text{Desc}(\text{ASCII}, \text{Ans})}$	
	C_Acc	Inc_Acc	C_Acc	Inc_Acc	C_Acc	Inc_Acc	C_Acc	Inc_Acc
QWEN2.5-7B	65.2	37.1	64.4	29.6	69.3	35.3	56.9	35.4
LLAMA-3-70B	87.7	39.9	83.8	35.0	89.9	37.2	81.9	36.9
GPT-4.1	94.9	62.4	90.2	60.1	93.9	58.6	86.0	59.4
GPT-OSS-20B	97.0	85.9	96.9	76.3	96.7	84.0	91.6	74.3
QWEN3-235B	95.0	61.5	89.9	62.2	93.2	67.3	84.1	55.1
QWEN3-30B	88.5	56.2	82.5	52.8	88.8	59.3	75.7	60.3
QWEN3-30B FT _{D$\rightarrow$A}	97.5	41.4	96.5	34.6	98.1	30.9	96.7	35.5
QWEN3-30B FT _{D$\leftarrow$A}	84.9	57.1	81.2	60.7	85.7	58.4	69.7	59.6
QWEN3-30B FT _{D$\leftrightarrow$A}	97.4	42.2	96.6	46.7	96.9	44.3	94.2	46.0

Table 10: Accuracy on consistent (C_Acc, from Table 3) and inconsistent (Inc_Acc) instances across generation settings. Instances without parsable ASCII are excluded (at most 2%).

Model	$\mathcal{R}_{\text{Desc}}$	$\mathcal{R}_{\text{ASCII}}$	$\mathcal{R}_{\text{Desc(Ans, ASCII)}}$	$\mathcal{R}_{\text{Desc(ASCII, Ans)}}$	$\mathcal{R}_{\text{Desc+ASCII}}$
<i>One-turn generation</i>					
QWEN2.5-7B	37.7	33.4	<u>38.6</u>	33.9	48.3
LLAMA-3-70B	<u>63.3</u>	62.2	<u>62.3</u>	52.2	68.8
GPT-4.1	81.2	<u>85.8</u>	82.2	77.5	89.6
GPT-OSS-20B	91.9	92.0	<u>93.2</u>	90.5	96.8
QWEN3-235B-A22B	81.9	<u>86.4</u>	81.5	79.6	88.0
QWEN3-30B-A3B	72.3	<u>80.3</u>	69.6	66.2	81.3
QWEN3-30B-A3B FT _{D→A}	79.4	72.7	79.5	<u>76.9</u>	81.9
QWEN3-30B-A3B FT _{D←A}	71.4	<u>84.1</u>	67.4	69.4	84.7
QWEN3-30B-A3B FT _{D↔A}	78.2	73.3	<u>80.0</u>	79.3	82.6
<i>Two-turn generation</i>					
QWEN2.5-7B			<u>38.6</u>	<u>38.6</u>	46.9
LLAMA-3-70B			<u>61.0</u>	56.2	67.2
GPT-4.1			<u>81.0</u>	76.1	90.0
GPT-OSS-20B			<u>92.3</u>	86.8	94.0
QWEN3-235B-A22B			<u>83.0</u>	73.1	85.0
QWEN3-30B-A3B			<u>71.8</u>	64.1	78.8
QWEN3-30B-A3B FT _{D→A}			<u>76.2</u>	74.9	78.3
QWEN3-30B-A3B FT _{D←A}			<u>68.2</u>	57.8	82.6
QWEN3-30B-A3B FT _{D↔A}			<u>77.3</u>	74.0	80.1

Table 9: F1-score results. The best result is shown in **bold**, and the second-best is underlined.

G Data Generation Pipeline

Pseudo-code for the spatial dataset generation pipeline with multi-dimensional balancing.

Algorithm 1 Text2Space Dataset Generation

Require:

- 1: N_{target} : Target count
- 2: R_{comp}, R_{rel} : Complexity ranges (components, relations)

Ensure:

- 3: $\mathcal{D}$: Final spatial dataset
 - 4: $\mathcal{D} \leftarrow \emptyset$
 - 5: Initialize *Tracker* to monitor distribution
 - 6: **while** $|\mathcal{D}| < N_{target}$ **do**
 1. **Select Parameters**
 - 7: Identify imbalanced categories using *Tracker*
 - 8: $P \leftarrow$ select parameters (e.g., node count) to balance
 2. **Build Connected Structure**
 - 9: $V \leftarrow \{A, B, C, \dots\}$
 - 10: $G \leftarrow (V, E = \emptyset)$
 - 11: $C \leftarrow \{\text{Random}(V)\}$
 - 12: **while** $V \setminus C \neq \emptyset$ **do**
 - 13: $u \leftarrow \text{Random}(V \setminus C)$
 - 14: $v \leftarrow \text{Random}(C)$
 - 15: $dir \leftarrow \text{Random}(8 \text{ Compass Directions})$
 - 16: **if** edge (u, v, dir) is valid **then**
 - 17: $E \leftarrow E \cup \{(u, v, dir)\}$
 - 18: $C \leftarrow C \cup \{u\}$
 - 19: **end if**
 - 20: **end while**
 - 21: Add extra relations to E to reach target complexity
 3. **Generate Description**
 - 22: $S \leftarrow \text{Random}(\{\text{Spatial}, \text{Cardinal}, \text{Clock}\})$
 - 23: $Desc \leftarrow \text{GRAPHTONATURALLANGUAGE}(G, S)$
 4. **Create Query**
 - 24: $\mathcal{P}_{infer} \leftarrow$ Identify uniquely inferable pairs in G
 - 25: $(Q, A) \leftarrow$ Sample pair and Type from $\mathcal{P}_{infer}$
 - 26: Determine if relationship is Direct vs. Inferred
 5. **Render Visualizations**
 - 27: $Vis \leftarrow \text{RENDERASCII}(G, \{\text{Simple}, \text{Grid}, \text{Panel}\})$
 6. **Apply Rejection Sampling**
 - 28: $I \leftarrow \{G, Desc, Q, A, Vis, S\}$
 - 29: $\Delta \leftarrow \text{CALCBALANCEDEVIATION}(I, \text{Tracker})$
 - 30: **if** I worsens balance **then**
 - 31: **Reject** with probability $P \propto \Delta^2$
 - 32: **else**
 - 33: $\mathcal{D} \leftarrow \mathcal{D} \cup \{I\}$
 - 34: Update *Tracker* with instance I
 - 35: **end if**
 - 36: **end while**
 - 37: **return** $\mathcal{D}$
-

H Evaluation Algorithm for ASCII and Descriptions

Pseudo-code for evaluating generated descriptions and ASCII layouts.

Algorithm 2 Evaluate Description vs. ASCII Layouts

Require:

- 1: G : ASCII Grid (Model Output)
- 2: D : Natural Language Description
- 3: $Mode$: "verify_ascii" or "verify_desc"

Ensure:

- 4: R : Result tuple (Pass, Accuracy, Breakdown)
 - 1. Parse Description**
 - 5: $Segs \leftarrow \text{Split}(D, \{.,, \})$
 - 6: $Claims \leftarrow \emptyset, \quad Bad \leftarrow \emptyset$
 - 7: **for** $s \in Segs$ **do**
 - 8: Match s with patterns (e.g., "A is left of B")
 - 9: **if** match found **then**
 - 10: $(u, v, dir) \leftarrow \text{Extract}(s)$
 - 11: $Claims.add((u, v, dir))$
 - 12: **else**
 - 13: $Bad.add(s)$
 - 14: **end if**
 - 15: **end for**
 - 16: $Objs_T \leftarrow \text{UniqueObjects}(Claims)$
 - 2. Build Grid Graph**
 - 17: $Pos \leftarrow \text{ScanCoordinates}(G)$
 - 18: Normalize Pos (top-left at 0,0)
 - 19: $Truth \leftarrow \emptyset$
 - 20: **for** $(u, v) \in Pos$ **do**
 - 21: $\vec{d} \leftarrow Pos[u] - Pos[v]$
 - 22: $Truth[u, v] \leftarrow \text{VectorToDir}(\vec{d})$
 - 23: **end for**
 - 3. Validate Relations**
 - 24: $Corr \leftarrow \emptyset, \quad Err \leftarrow \emptyset$
 - 25: **for** $(u, v, dir) \in Claims$ **do**
 - 26: **if** $u \notin Pos$ **or** $v \notin Pos$ **then**
 - 27: $Err.add((u, v, "Missing"))$
 - 28: **else**
 - 29: **if** $dir == Truth[u, v]$ **then**
 - 30: $Corr.add((u, v, dir))$
 - 31: **else**
 - 32: $Err.add((u, v, Truth[u, v]))$
 - 33: **end if**
 - 34: **end if**
 - 35: **end for**
 - 4. Final Scoring**
 - 36: **if** $Mode == \text{"verify_desc"}$ **then** ▷ ASCII is Truth
 - 37: $Pass \leftarrow (Err = \emptyset)$
 - 38: **else** ▷ Text is Truth
 - 39: $Extra \leftarrow Pos.keys \setminus Objs_T$
 - 40: $Pass \leftarrow (Err = \emptyset \wedge Bad = \emptyset \wedge Extra = \emptyset)$
 - 41: **end if**
 - 42: $Acc \leftarrow |Corr| / |Claims|$
 - 43: **return** $\{Pass, Acc, Corr, Err, Bad\}$

I Evaluation Algorithm for Consistency

Pseudo-code for evaluating the consistency between generated ASCII layouts and query relations.

Algorithm 3 Evaluate Consistency Between ASCII Grid and Reference relations

Require:

- 1: G : ASCII representation (Model Grid)
- 2: Q : Natural language query
- 3: L : Reference label (relation)

Ensure:

- 4: R : Evaluation result (correctness, objects, relations)
 - 1. **Parse the Query**
 - 5: Match Q vs. patterns: "Vertical", "Horizontal", or "Full"
 - 6: Extract target objects X and Y from Q
 - 7: **if** no valid pattern found **then**
 - 8: **return** ERROR("Query parsing failed")
 - 9: **end if**
 - 2. **Parse the ASCII Grid**
 - 10: Scan G to find coordinates for all characters
 - 11: $Pos \leftarrow \{(obj, x, y) \mid obj \text{ found in } G\}$
 - 12: **if** $X \notin Pos$ or $Y \notin Pos$ **then**
 - 13: **return** ERROR("Queried object missing in grid")
 - 14: **end if**
 - 15: Normalize Pos so top-left object is at $(0, 0)$
 - 3. **Infer Relation from Grid**
 - 16: $(x_1, y_1) \leftarrow Pos[X], (x_2, y_2) \leftarrow Pos[Y]$
 - 17: $Rel_{inf} \leftarrow \emptyset$
 - 18: **if** Type is "Vertical" **then**
 - 19: **if** $y_1 < y_2$ **then** $Rel_{inf} \leftarrow \text{"above"}$
 - 20: **else if** $y_1 > y_2$ **then** $Rel_{inf} \leftarrow \text{"below"}$
 - 21: **else** $Rel_{inf} \leftarrow \text{"same level"}$
 - 22: **end if**
 - 23: **else if** Type is "Horizontal" **then**
 - 24: **if** $x_1 < x_2$ **then** $Rel_{inf} \leftarrow \text{"left"}$
 - 25: **else if** $x_1 > x_2$ **then** $Rel_{inf} \leftarrow \text{"right"}$
 - 26: **else** $Rel_{inf} \leftarrow \text{"same column"}$
 - 27: **end if**
 - 28: **else** ▷ Full Spatial Query
 - 29: Compute vector $\vec{v} = (x_1 - x_2, y_1 - y_2)$
 - 30: $Rel_{inf} \leftarrow \text{Map } \vec{v} \text{ to } \{\text{upper-left, below, right, ...}\}$
 - 31: **end if**
 - 4. **Compare with Reference Label**
 - 32: $Match \leftarrow (\text{LowerCase}(Rel_{inf}) == \text{LowerCase}(L))$
 - 33: $R \leftarrow \{\text{Pass} : Match, \text{Exp} : L, \text{Act} : Rel_{inf}\}$
 - 34: **return** R
-

J Human Validation

Human Validation Setup. To quantify the impact of different input formats on spatial reasoning accuracy, we conducted a controlled human validation study with 11 evaluators divided into two groups: Team A ($n = 7$) and Team B ($n = 4$). The study used a within-subjects design in which question sets were counterbalanced across formats, disentangling the effect of input format from question difficulty. We curated three problem sets (Set X, Set Y, and Set Z), each containing 10 spatially complex examples, and assigned them to three sessions using the Description-only, ASCII-only, and Description+ASCII formats (Table 11). In the Description-only session, participants were given a blank grid worksheet and asked to draw the described layout before selecting an answer, mirroring the *Desc*→*ASCII* construction task we pose to models. No drawing was required in the other two sessions, where a correct layout was already supplied.

- **Sessions 1 and 3 (Crossover):** We swapped the assignment of problem sets and formats between teams. Team A solved Set X using **Description-only** (Session 1) and Set Y using **Description+ASCII** (Session 3). Conversely, Team B solved Set Y using **Description-only** (Session 1) and Set X using **Description+ASCII** (Session 3).
- **Session 2 (ASCII Visualization):** Both teams evaluated Set Z using the **ASCII-only** format. Performance was similar across groups (Team A: 0.986, Team B: 0.950).

This cross-over design enables that Sets X and Y were evaluated in both text-only and combined formats, allowing for a direct comparison on semantically identical questions.

Task Sessions	Team A ($n = 7$)	Team B ($n = 4$)
Session 1 (<i>Draw-then-Answer</i>)	Description Only (Set X)	Description Only (Set Y)
Session 2 (<i>ASCII Visualization</i>)	ASCII Only (Set Z)	ASCII Only (Set Z)
Session 3 (<i>Description+ASCII</i>)	Desc + ASCII (Set Y)	Desc + ASCII (Set X)

Table 11: Human Validation Study Design. The experiment followed a three-session protocol with counterbalanced question sets.

Validation Results Table 12 shows the substantial improvement achieved when reasoning from an ASCII-augmented description compared to reasoning from the description alone.

Question Set	Desc Only (NL Format)	Desc + ASCII (Combined)	Improvement
Set X	0.857 (Team A)	0.975 (Team B)	+0.118
Set Y	0.875 (Team B)	0.986 (Team A)	+0.111
Average	0.866	0.981	+0.115

Table 12: Human performance comparison across different input formats, reported in percentage points. Counterbalancing Sets X and Y controls for question difficulty.

Participants and Ethics We assessed the need for ethical review: the study consists of benign behavioral tasks with no collection of personally identifiable information (PII), and participants responded through a form without recorded identifiers.

Please answer by choosing the most appropriate option from the drop-down menu. Specifically, if A is upper-right of B, then "upper-right" is a more accurate answer than simply "above" or "right."

Important: Each question is independent of the others.

You will be given a text description, an ASCII representation, and a query.

- In the first part, you will see only the text description and the query.
- In the second part, you will see only the ASCII representation and the query.
- In the third part, you will see all three components together.

Task format:

Your question zone

Example 1:
Text Description:
 E is at the 9:00 position from A. B is at the 6 o'clock position from A. C is at 12 o'clock from A. F is at the 1 o'clock position from B. D is at 4:30 from B. A is at the 9:00 position from F. F is at 12 o'clock from D.

ASCII Representation:

```

  C
  |
E A F
  |
  B
  |
  D
  
```

Query:
 Where is E relative to B? (E->B)

Your answer zone

Choose

Choose

Above

Below

Right

Left

Upper-Right

Upper-Left

Lower-Right

Lower-Left

Task format:

Horizontal directions

Choose

Choose

Right

Left

Same Column (Horizontal-only relations)

Vertical directions

Choose

Choose

Above

Below

Same Level (Vertical-only relations)

Full directions

Choose

Choose

Above

Below

Right

Left

Upper-Right

Upper-Left

Lower-Right

Lower-Left

Text Description + Query

Session 1

Example 1:
Text Description:
 E is at the 9:00 position from A. B is at the 6 o'clock position from A. C is at 12 o'clock from A. F is at the 1 o'clock position from B. D is at 4:30 from B. A is at the 9:00 position from F. F is at 12 o'clock from D.

Query:
 Where is B relative to C? (B->C)

Choose

ASCII visualization + Query

Session 2

Example 1:

```

  C
  |
E A F
  |
  B
  |
  D
  
```

Query:
 Where is B relative to C? (B->C)

Choose

Text + ASCII visualization + Query

Session 3

Example 1:
Text Description:
 E is at the 9:00 position from A. B is at the 6 o'clock position from A. C is at 12 o'clock from A. F is at the 1 o'clock position from B. D is at 4:30 from B. A is at the 9:00 position from F. F is at 12 o'clock from D.

```

  C
  |
E A F
  |
  B
  |
  D
  
```

Query:
 Where is B relative to C? (B->C)

Choose

Figure 6: Overview of the human validation interface and instruction protocol. Participants solved spatial reasoning tasks across three counterbalanced sessions: (1) Text Description Only, (2) ASCII Visualization Only, and (3) Combined Text + ASCII. We use Google Form to collect the response from the participants. The interface requires participants to select the most precise directional term from a standardized drop-down menu.