ASCII-Reasoning: A Structural Spatial Reasoning Benchmark for Large Language Models

Zhiwei Yu*[1], Xiangyu Wang†[1], Chengze Du†, Lin He*, Jiahai Yang†*
*Tsinghua University, †Quan Cheng Laboratory, ‡Beijing University of Posts and Telecommunications

***Abstract*—Spatial reasoning is a critical component of general intelligence, yet evaluating this capability in Large Language Models (LLMs) without visual encoders remains a significant challenge. Existing benchmarks either rely on multimodal inputs or lack structural complexity in pure text. To bridge this gap, we introduce the ASCII Reasoning Benchmark (ARB), a comprehensive framework comprising 2,400 samples across eight tasks, such as mechanical transmission and 3D folding. ARB utilizes discrete ASCII grids to represent topology, compelling models to construct internal mental models solely based on symbolic alignments. A systematic zero-shot evaluation of 16 LLMs reveals a significant capability gap, with the top-performing model achieving only 37.4% accuracy. Crucially, our analysis uncovers counter-intuitive insights: a *Parameter Efficiency Paradox*, where lightweight, code-optimized models frequently outperform larger counterparts, suggesting symbolic reasoning derives more from code synthesis than parameter scale; and distinct *Capability Specialization*, where models achieve state-of-the-art performance in specific niches like pathfinding despite lower overall rankings. Code and Data are available at https://anonymous.4open.science/r/ASCII-Based-Spatial-Reasoning-Benchmark-4C2B.**

***Index Terms*—Spatial Reasoning, Large Language Models, Benchmark, ASCII Art, Symbolic Reasoning**

I. Introduction

Large Language Models (LLMs) have achieved substantial progress in semantic modeling, natural language understanding, and code generation. However, as these models are increasingly integrated into embodied AI and the control of complex physical systems, their actual capabilities in the core dimensions of spatial cognition, reasoning, and planning remain insufficiently defined. A critical scientific question has emerged: can LLMs construct accurate spatial logical models through structured text alone, independent of visual sensory input? Answering this is essential for evaluating whether LLMs possess the attributes of a "world model" [1].

Existing evaluation frameworks exhibit two primary limitations. First, multimodal benchmarks rely heavily on the feature extraction of visual encoders [2, 3], making it difficult to decouple perceptual noise from the underlying logical reasoning capacity of the language model itself. Second, current text-based reasoning evaluations primarily focus on one-dimensional symbolic manipulation [4, 5]. When addressing two-dimensional or three-dimensional spatial tasks—such as mental rotation, topological connectivity, and physical constraint satisfaction—these benchmarks often lack controllability in data generation and interpretability in error analysis.

[1]Equal contribution.

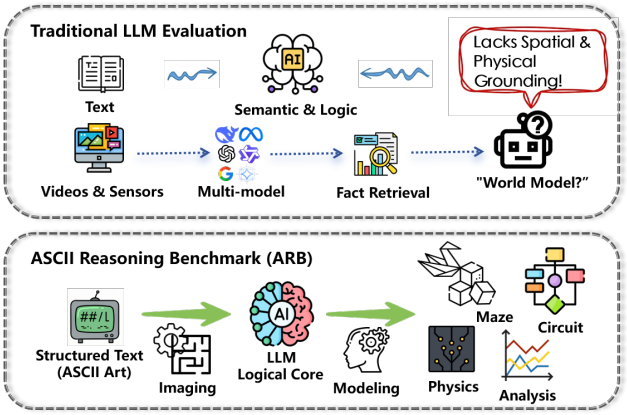

Fig. 1: A conceptual overview of the ARB. Unlike traditional LLM evaluations that focus on semantic logic and fact retrieval, ARB utilizes structured ASCII art to evaluate an LLM's spatial and physical grounding through complex tasks such as maze navigation, circuit analysis.

Consequently, the research community lacks a systematic understanding of the decision boundaries and robustness of LLMs when processing non-visual spatial structures.

As illustrated in Figure 1, we introduce and open-source the ASCII Reasoning Benchmark (ARB). This benchmark utilizes two-dimensional ASCII character arrangements to represent complex geometric and physical structures. It requires models to construct internal "mental models" solely through the row-column alignment of characters under zero-shot, text-only conditions. ARB encompasses eight representative scenarios, including mechanical transmission, 3D folding, logic circuits, and pathfinding, totaling 2,400 samples. Using this framework, we conducted an empirical evaluation of 16 frontier models, including the GPT [6], Gemini [7], Claude [8], and Qwen [9] series. Our experimental results indicate that current models face significant bottlenecks in symbolic spatial reasoning, with average accuracies generally falling below 40%.

Our empirical analysis reveals two noteworthy phenomena. First, we identify a "parameter efficiency paradox": in structured symbolic reasoning tasks, lightweight models such as gpt-4.1-mini frequently outperform their larger counterparts. This suggests that such capabilities may stem more from specific data fine-tuning or architectural optimizations than from simple parameter scaling [10]. Second, we observe "ca-

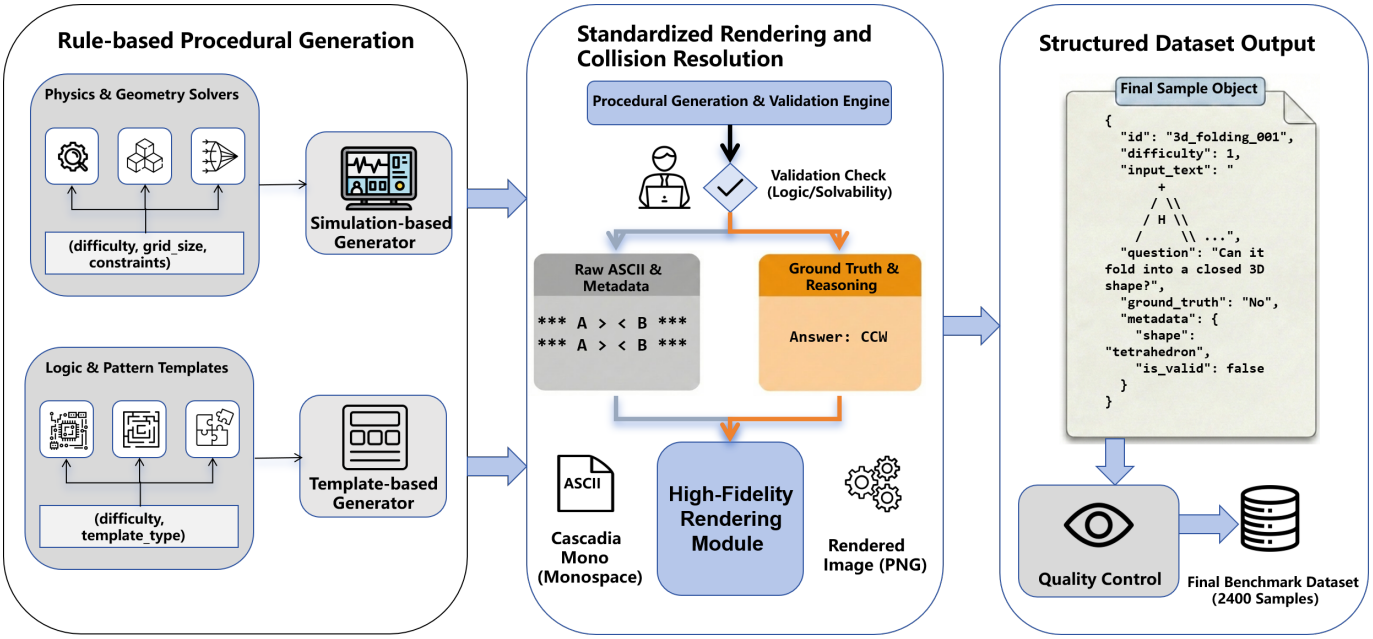

Fig. 2: Overview of the three-stage ARB framework. The pipeline transitions from simulation-based configuration and logical constraints to high-fidelity ASCII rendering, culminating in rigorous automated and human verification.

pability specialization," where different models exhibit distinct skill preferences. For instance, Grok-4.1 [11] demonstrates superior performance in long-distance pathfinding, while GPT-4o shows a relative advantage in physical dynamics reasoning. These findings highlight a significant gap in the generalization from linguistic representations to spatial logic.

The primary contributions of this paper are threefold:

- **ARB Benchmark:** We propose the first comprehensive benchmark for text-based spatial reasoning, featuring eight tasks across geometric, physical, and logical dimensions with controlled difficulty levels.
- **Empirical Findings:** We evaluate 16 mainstream LLMs, quantifying their performance boundaries and identifying key phenomena such as the "parameter efficiency paradox" and task-specific specialization.
- **Open-source Pipeline:** We provide a standardized "generation-rendering-evaluation" framework and open-source the dataset and generator to support reproducible research in symbolic spatial reasoning.

II. Related Work

LLM Reasoning Capabilities Evaluation. Existing benchmarks generally fall into linguistic or visual-spatial categories. Linguistic evaluations focus on symbolic logic and knowledge retrieval rather than spatial perception [4, 5, 12]. Visual benchmarks depend on encoders to extract features, which complicates the isolation of reasoning capacity from perceptual noise [2, 3]. Current methods lack a systematic approach to assessing physical constraints and topological relationships based strictly on character alignment without visual aids.

Structured Reasoning Tasks. Structured reasoning typically prioritizes algorithmic logic generation, formal proofs, or image-based prediction [13–15]. In contrast, our work requires models to perform mental rotation and causal propagation within pure text grids. This approach shifts the analytical focus from one-dimensional deduction to two-dimensional structural interpretation based on character alignment.

Applications of ASCII Art in AI. Research on ASCII art has shifted from generation to object recognition [16, 17]. Current studies primarily assess identification rather than logical inference based on structural rules. While benchmarks like ArtPrompt and ASCIIBench focus on visual perception or alignment challenges [16–18], ARB utilizes ASCII as a medium for logical reasoning. This framework quantitatively evaluates model performance in processing structured symbolic information across physical and geometric dimensions.

III. ASCII Reasoning Benchmark

A. Overview

We present the ARB, a comprehensive evaluation framework comprising 2,400 samples designed to assess the intrinsic spatial reasoning capabilities of Large Language Models without reliance on visual encoders. Unlike traditional multimodal benchmarks, ARB requires models to construct internal spatial mental models solely through the row-column alignment of discrete ASCII character grids. The benchmark encompasses eight representative tasks across three core domains: *Physical Dynamics*, *Geometry and Topology*, and *Logic and Structured Data*. To rigorously quantify performance boundaries, each task is further stratified into Easy, Medium, and Hard difficulty levels based on topological complexity and component density. This hierarchical design facilitates a systematic analysis of model robustness and reveals performance degradation patterns across varying degrees of structural depth.

Task	Grid Size	Symbols	Output Format	$\lvert\mathcal{Y}\rvert$	Chance
Mechanical Transmission	10×12 (5×5–30×33)	A–Z, *, ¡, ¿, v, caret, =	clockwise / counterclockwise / same / opposite	4	25.00%
Maze Pathfinding	11×11 (5×5–21×21)	#, S, E, space	Yes / No	2	50.00%
Logic Circuit Analysis	8×23 (1×8–15×54)	digits, letters, #, ¡, ¿, =, ?, +, -, *, /, backslash, —, [ ], arrows	integer / Yes–No / short string	54	1.85%
3D Geometric Folding	12×25 (5×9–18×33)	letters, +, -, /, backslash, —, *, #, $, %, &, @, ?, !, =, tilde	Yes–No / face labels / relations	130	0.77%
Geometric Ray Tracing	15×30 (fixed)	#, S, targets (letters/digits), /, backslash	visible / blocked / target sets	18	5.56%
Physical Equilibrium	3×29 (2×19–5×47)	letters/digits, -, :	stable / unstable / true / false	4	25.00%
Structural Pattern Completion	5×7 (5×3–8×13)	block/box glyphs, *, letters/digits	A / B / C / D	4	25.00%
Structured Data Analysis	12×49 (6×36–14×102)	digits, letters, *, —, +, -, /, #, apostrophe	peak label set (comma-separated)	189	0.53%

TABLE I: Detailed specifications of ARB tasks. Grid size is reported as height×width. Output space $\lvert\mathcal{Y}\rvert$ is the number of distinct ground-truth answers in the dataset (first question per sample). Chance level is $1/\lvert\mathcal{Y}\rvert$.

System Prompt: You are an expert in spatial reasoning and pathfinding algorithms. You will analyze ASCII mazes to determine reachability.

Task Prompt: Analyze the following ASCII maze:
`{ascii_maze}`
`{question}`

Instructions:
- '#' represents walls; ' ' (space) represents open paths.
- 'S' is start; 'E' is end.
- Move up, down, left, right only.
- Answer exactly "Yes" or "No".

Your answer:

Fig. 3: The standardized prompt structure employed in ARB, using the Maze Pathfinding task as an example.

B. Task Taxonomy

Mechanical Transmission. Gears are represented using asterisks, with connections denoted by angle brackets or equality signs. The model propagates motion from a driver wheel through the chain to predict the rotation direction of a target gear or detect jamming from conflicting forces.

3D Geometric Folding. This task uses two-dimensional nets built from characters such as plus signs, hyphens, and pipes. The model must verify whether the net forms a valid closed polyhedron such as a cube or tetrahedron. The task requires the model to deduce the spatial orientation of specific faces within the reconstructed three-dimensional shape.

Logic Circuit Analysis. We depict control flow graphs using diamond and rectangle characters to indicate logic branches and operational nodes. The model traces the execution path based on specific input variables. It must parse the directed acyclic graph structure to compute the final output value.

Geometric Ray Tracing. The grid contains obstacles marked by hash symbols and mirrors represented by slashes. We define a light source and a target point within this space. The model applies geometric optical rules to determine if a reflected ray can reach the target. This process requires precise coordinate updates and collision detection.

Physical Equilibrium. This task visualizes vertical stacks of character blocks with varying widths under gravity. The model analyzes the distribution of torque to assess stability. It must also identify any critical support blocks that would cause the structure to collapse if removed.

Structured Data Analysis. Unformatted ASCII bar charts and tables encode magnitude via character repetition. Models resolve row and column alignment to extract values and identify extremes or trends.

Structural Pattern Completion. The input displays a two-dimensional character pattern with a masked section. The model selects the correct ASCII fragment from a set of candidates to fill the void. Success depends on restoring local texture continuity and global structural symmetry.

Maze Pathfinding. This task uses hash symbols for walls and dots for paths to define a navigation problem between a start and end point. The model verifies reachability by executing search algorithms internally. It must traverse the grid while strictly adhering to boundary constraints.

Task Specifications and Complexity Baselines. We detail the structural parameters of each task in Table I, including grid dimensions and symbol vocabularies. Crucially, we define the **Chance Level** for each category to contextualize model performance. Tasks involving boolean verifications have a chance baseline of 50%, whereas generative tasks like *Maze Pathfinding* and *Data Analysis* possess an open-ended output space where random guessing yields near 0% accuracy. This distinction highlights that while an average accuracy of 37.4% may appear low in absolute terms, it represents a capability leap over random baselines in complex generative scenarios.

C. Data Construction Pipeline

Rule-based Procedural Generation. As illustrated in Figure 2, data construction begins with deterministic scripts to preclude hallucinations common in model-synthesized data. We implemented dedicated rule-based algorithms for each task. Taking *3D Folding* as an instance, the algorithm initializes the topological structure of a cube and employs spatial transformation logic to derive the 2D net and its corresponding 3D face mappings as ground truth. To ensure a rational difficulty gradient, we did not rely solely on automated parameter scaling. Instead, we employed **Manual Annotation** for difficulty stratification. Experts classified the generated samples into Easy, Medium, and Hard levels based on visual complexity and reasoning depth, ensuring the validity of the evaluation hierarchy.

Standardized Rendering and Collision Resolution. The logical topology is mapped onto a 2D character grid. The primary objective of this stage is to ensure the **Unique Determinism** and **Consistency** of the visual representation, eliminating ambiguities or "garbled" artifacts caused by misalignment. To achieve this, we enforce a strict rendering protocol: (i) **Grid**

Model	Mech. Trans.	Maze Path.	Logic Circ.	3D Fold.	Geom. Ray.	Phys. Equil.	Struct. Pat.	Struct. Data	Avg
GPT-5.1	53.33	55.00	50.00	18.67	**24.00**	5.00	27.67	19.00	31.58
GPT-4o	**70.33**	51.67	29.33	27.33	17.33	21.67	20.67	3.67	26.89
GPT-4.1-mini	67.33	53.33	53.33	25.00	21.33	40.00	20.33	18.67	**37.42**
Gemini-2.5-pro	57.33	9.00	60.00	6.67	4.33	0.33	0.33	**47.00**	23.12
Gemini-2.5-flash	57.33	52.67	**60.67**	24.67	23.67	10.67	17.33	40.00	35.87
Gemini-2.0-flash	48.33	49.67	38.67	**28.00**	22.67	9.33	20.33	22.33	29.92
Claude-sonnet-4	43.00	54.00	48.33	25.00	19.00	23.67	2.00	6.33	27.67
Qwen-max	67.00	53.00	35.67	23.00	21.00	3.67	28.67	6.00	29.75
Qwen2.5-72B	65.67	54.33	42.67	23.00	20.00	3.33	**30.67**	2.67	30.29
Qwen2.5-32B	62.67	51.33	39.33	24.00	19.33	12.33	27.00	15.00	31.37
Qwen2.5-14B	59.33	52.00	29.67	23.00	14.33	12.33	0.67	5.67	24.62
Qwen2.5-7B	60.33	47.33	27.67	21.67	19.33	36.33	28.00	4.33	30.62
DeepSeek-V3	46.67	42.67	31.00	23.33	17.00	16.67	9.00	15.00	25.17
DeepSeek-V3.2	49.67	54.00	32.67	23.00	10.33	23.00	6.33	11.00	26.25
Kimi-K2	58.33	49.67	29.67	**28.00**	15.00	**45.67**	4.67	14.00	30.63
Grok-4.1	37.33	**73.67**	49.00	3.00	20.34	0.67	1.33	4.33	23.71

TABLE II: Quantitative comparison of various LLMs on eight ASCII-based visual reasoning tasks. We report the **Mean Accuracy (%)** $\pm$ **Standard Deviation** over $N = 3$ independent runs. The chance level varies by task.

Discretization: All geometric elements are aligned to integer coordinates under a monospaced font constraint, ensuring a strict correspondence between a character's physical position and its logical index. (ii) **Collision Resolution**: For complex intersections, the renderer applies layer prioritization rules to automatically detect and correct illegal character overlaps, ensuring the generated ASCII artifacts remain topologically coherent and unambiguous.

Dual-Layer Verification. To maintain a noise-free benchmark, we implement a "Automated Filter + Human Audit" protocol. Initially, scripts automatically discard rendered outputs containing character collisions or topological discontinuities. Subsequently, a Human-in-the-loop verification is conducted, where three engineering experts independently audit a 10% stratified sample of the dataset. The focus is to identify potential visual ambiguities or "optical illusions" where character proximity might mislead reasoning. Only batches achieving 100% Inter-Annotator Agreement between human labels and solver-generated truths are retained, resulting in a finalized set of 2,400 high-quality samples.

As illustrated in Figure 3, our prompt establishes strict boundaries. We include a clear role definition (e.g., *"You will analyze ASCII mazes..."*) and enforce a specific output format (e.g., *"Answer with exactly [Format]"*). These constraints prevent irrelevant text generation and ensure the output directly matches our ground-truth labels, which facilitates the regex extraction described below.

IV. Experiments

A. Experimental Setup

Models Evaluated. We systematically evaluate 16 mainstream LLMs, spanning both proprietary and open-weight architectures. The evaluation suite includes proprietary families such as **GPT** [6], **Gemini** [19], **Claude**, and **Grok**, alongside open-source series like **Qwen2.5** ranging from 7B to 72B and **DeepSeek** [20]. This selection covers a diverse spectrum of parameter scales and architectures such as dense, MoE.

Implementation Details. To decouple the inductive bias introduced by in-context learning, all experiments are executed under a **Zero-Shot** setting [21]. We adhere to a strict standardized prompt template and parsing protocol (see Sec. III) to ensure evaluation fairness. To guarantee statistical robustness and mitigate the randomness inherent in LLM generation ($T = 0.7$), we conducted $N = 3$ independent runs for each model-task pair using different random seeds. The results reported in this paper represent the mean accuracy $\pm$ standard deviation. This protocol ensures that the observed "Parameter Efficiency Paradox" is a consistent behavioral pattern rather than a stochastic anomaly. **Implementation Details & Protocol.** All experiments are conducted in a **Zero-Shot** setting to assess intrinsic spatial grounding capabilities [21]. We employ a standardized prompt structure to minimize prompt engineering sensitivity:

To distinguish reasoning failures from formatting errors, we implement a Robust Parsing Protocol: (i) **Regex Extraction**: We use task-specific regular expressions to isolate the answer key from verbose responses (e.g., extracting '$\dot{+}$' for numeric answers). (ii) **Self-Correction Retry**: If the output fails regex matching, we trigger a single automatic retry with the prompt *"Invalid format. Please output ONLY the final answer in [Format]."*. Samples are marked as incorrect only if parsing fails after this correction step. We report metrics averaged over $N = 3$ runs with temperature $T = 0.7$. **Evaluation Metrics.** We adopt **Accuracy** as the primary metric, defined as the ratio of correctly answered samples $N_{correct}$ to the total number of

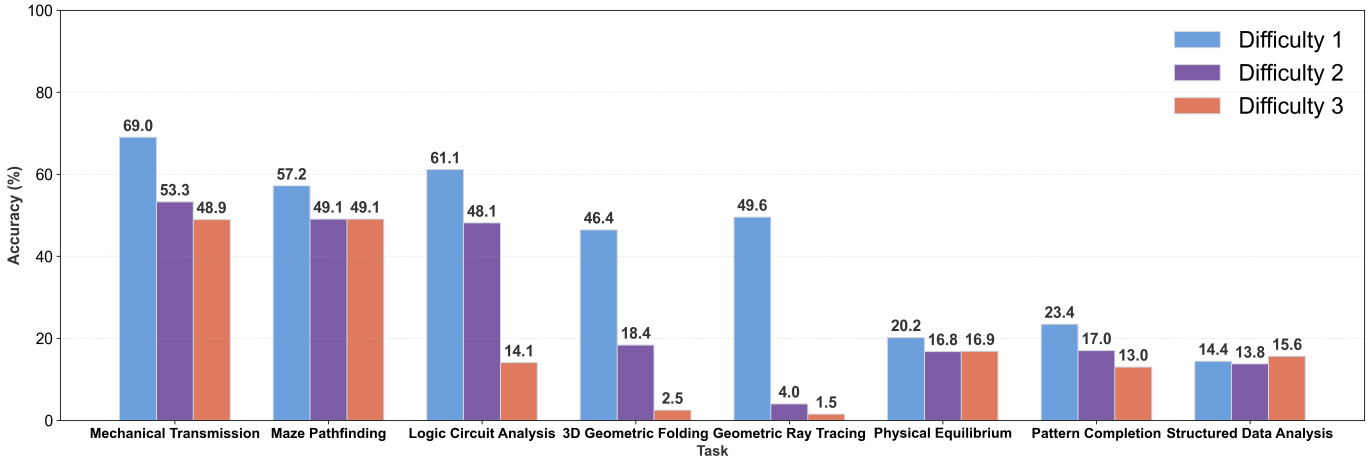

Fig. 4: **Performance degradation across difficulty levels.** The bar chart illustrates the average accuracy of evaluated models across three escalating difficulty tiers. While linear tasks like Maze Pathfinding show resilience, complex topological tasks like Ray Tracing exhibit a "complexity cliff," dropping to near-zero accuracy at Level 3.

samples N_{total}:

$$Accuracy = \frac{N_{correct}}{N_{total}} \tag{1}$$

To address each task's distinct physical and logical characteristics, we categorize correctness criteria into three dimensions: (i) **Directional Accuracy**, evaluating causal chain derivation correctness in *Mechanism* and *Balance* tasks; (ii) **Topological Validity**, judging structural closure and logical flow connectivity in *3D Folding* and *Circuit* tasks; and (iii) **Reachability Score**, measuring path planning compliance with geometric constraints in *Maze* and *Ray Tracing* tasks.

B. Main Result

Table II presents the comprehensive performance of all evaluated models on the ARB.

Overall Performance Bottleneck and Task Stratification. The experimental data indicates that current LLMs exhibit limited proficiency in text-based spatial reasoning. The top-performing model, `gpt-4.1-mini`, attained an average accuracy of 37.42%. We observe that performance varies based on the structural requirements of the task. Models achieve relatively higher scores on tasks characterized by linear causal structures. Specifically, the *Mechanical Transmission* task yielded an average accuracy of approximately 57%, while *Maze Pathfinding* averaged around 50%. In contrast, accuracy decreases in tasks that demand global pattern matching or precise row-column alignment. For the *ASCII Structural Pattern Completion* and *Data Analyst* tasks, most model scores approach the level of random guessing. For instance, `Claude-Sonnet-4` recorded an accuracy of 2.00% in the Structural Pattern Completion task. These results suggest that models struggle to process unstructured visual alignments without visual sensory input.

The Parameter Efficiency Paradox. The experiments reveal an inverse scaling pattern, where lightweight models outperform larger ones in structured symbolic reasoning. As shown in Table II, `gpt-4.1-mini` achieves 37.42% vs. 26.89% for `gpt-4o`, and `gemini-2.5-flash` reaches 35.87% vs. 23.12% for `gemini-2.5-pro`. Among open-weight models, `Qwen2.5-32B-Instruct` (31.37%) exceeds `Qwen-max` (29.75%), suggesting that performance does not strictly scale with parameters and may depend on factors such as instruction tuning or pre-training data.

Capability Specialization. We observe performance variations across model architectures, where specific models excel in particular domains. While `gpt-4o` had a lower mean accuracy, it achieved 70.33% in the *Mechanical Transmission* task. In comparison, `grok-4.1` recorded an overall average of 21.17% yet attained 73.67% in *Maze Pathfinding*. Additionally, `gemini-2.5-flash` showed consistent results, with 60.67% in *ASCII Circuit* and 40.00% in *Data Analyst*, suggesting reliance on task-specific patterns rather than generalized spatial reasoning.

C. Analysis of Reasoning Robustness

To investigate the boundaries of LLM spatial reasoning, we further analyze the performance trajectory across the three difficulty strata defined in our data construction pipeline. Figure 4 illustrates the aggregated accuracy of all models across distinct difficulty levels.

Validation of Difficulty Gradient. The experimental results show a monotonic decrease in accuracy across most tasks as the complexity level increases. This trend indicates that procedural generation parameters such as grid density and constraint chains effectively differentiated the reasoning difficulty. In the Physical Equilibrium and Structured Data Analysis tasks, models demonstrated consistently low performance across all three difficulty levels. These data points suggest that current architectures struggle with these specific domains independent of the complexity of the individual samples.

The Performance Drop in Geometric Recursion. Accuracy in recursive state tracking and high-precision spatial manip-

ulation decreased significantly. In Geometric Ray Tracing, performance plummeted from 49.6% at Level 1 to 1.5% at Level 3; similarly, 3D Geometric Folding scores fell from 46.4% to 2.5%. These drops suggest that while models handle simple Level 1 patterns, they fail to maintain coherence during the long-horizon transformations required for Level 3 [22].

Resilience in Linear Causal Chains. In contrast, tasks with linear sequential logic demonstrate greater robustness, with *Mechanical Transmission* and *Maze Pathfinding* showing smoother decay. Notably, Maze Pathfinding stabilizes around **49.1%** at both Level 2 and Level 3, suggesting that current LLMs are more adept at sequential search and linear propagation than at processing 2D/3D topological structures or resolving collision constraints.

Logic Circuit Sensitivity. In the *Logic Circuit Analysis* task, accuracy decreased from 61.1% at Level 1 to 48.1% at Level 2. Subsequently, the performance fell to 14.1% at Level 3. This pattern indicates that models maintained performance on simpler branching logic but accuracy declined as the circuit depth and node density increased.

V. Conclusion

We presented **ARB**, a framework designed to assess LLMs' intrinsic spatial reasoning by decoupling perceptual recognition from logical modeling. Our evaluation of 2,400 samples reveals that current models lack robust spatial grounding and exhibit a *Parameter Efficiency Paradox*, suggesting that code-centric priors may contribute more to symbolic logic than sheer parameter scale. However, a limitation of our work lies in the discrete nature of ASCII representations, which serve as an abstraction and may not fully capture the continuous dynamics of real-world physics. Future research should address the observed "Complexity Cliff" through strategies like spatial Chain-of-Thought or topology-aware instruction tuning, effectively bridging the gap between semantic fluency and structural logic for embodied intelligence.

Acknowledgments

This work is supported by the Research Project of Quan Cheng Laboratory, China (Grant No. QCL20250108) and the Research Project of Provincial Laboratory of Shandong, China (Grant No. SYS202201). Lin He and Jiahai Yang are the corresponding authors.

References

[1] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *NeurIPS*, 2018, pp. 2455–2467.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: visual question answering," in *ICCV*. IEEE Computer Society, 2015, pp. 2425–2433.

[3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[4] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[5] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[6] J. Zhao, T. Wang, W. Abid, G. Angus, A. Garg, J. Kinnison, A. Sherstinsky, P. Molino, T. Addair, and D. Rishi, "Lora land: 310 fine-tuned llms that rival gpt-4, A technical report," *CoRR*, vol. abs/2405.00732, 2024.

[7] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. P. Lillicrap, J. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, I. Antonoglou, R. Anil, S. Borgeaud, A. M. Dai, K. Millican, E. Dyer, M. Glaese, T. Sottiaux, B. Lee, F. Viola, M. Reynolds, Y. Xu, J. Molloy, J. Chen, M. Isard, P. Barham, T. Hennigan, R. McIlroy, M. Johnson, J. Schalkwyk, E. Collins, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, C. Meyer, G. Thornton, Z. Yang, H. Michalewski, Z. Abbas, N. Schucher, A. Anand, R. Ives, J. Keeling, K. Lenc, S. Haykal, S. Shakeri, P. Shyam, A. Chowdhery, R. Ring, S. Spencer, E. Sezener, and et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *CoRR*, vol. abs/2403.05530, 2024.

[8] Anthropic, "The claude 3 model family: Opus, sonnet, haiku." [Online]. Available: https://api.semanticscholar.org/CorpusID:268232499

[9] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," *CoRR*, vol. abs/2412.15115, 2024.

[10] I. R. McKenzie, A. Lyzhov, M. Pieler, A. Parrish, A. Mueller, A. Prabhu, E. McLean, A. Kirtland, A. Ross, A. Liu, A. Gritsevskiy, D. Wurgaft, D. Kauffman, G. Recchia, J. Liu, J. Cavanagh, M. Weiss, S. Huang, T. F. Droid, T. Tseng, T. Korbak, X. Shen, Y. Zhang, Z. Zhou, N. Kim, S. R. Bowman, and E. Perez, "Inverse scaling: When bigger isn't better," *Trans. Mach. Learn. Res.*, vol. 2023, 2023.

[11] xAI, "Grok-1 model card," 2024. [Online]. Available: https://x.ai/blog/grok-os

[12] I. Comsa and S. Narayanan, "A benchmark for reasoning with spatial prepositions," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 16 328–16 335. [Online]. Available: https://aclanthology.org/2023.emnlp-main.1015/

[13] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, "Solving olympiad geometry without human demonstrations," *Nat.*, vol. 625, no. 7995, pp. 476–482, 2024.

[14] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, and R. Girshick, "PHYRE: A new benchmark for physical reasoning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019. [Online]. Available: https://player.phyre.ai/

[15] W. Wu, S. Mao, Y. Zhang, Y. Xia, L. Dong, L. Cui, and F. Wei, "Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models," in *NeurIPS*, 2024.

[16] Q. Jia, X. Yue, S. Huang, Z. Qin, Y. Liu, B. Y. Lin, Y. You, and G. Zhai, "Asciieval: Benchmarking models' visual perception in text strings via ascii art," 2025.

[17] K. Luo, M. Fu, J. Peguero, H. Malik, A. Patil, J. Lin, M. V. Overborg, R. Sarmiento, and K. Zhu, "Asciibench: Evaluating language-model-based understanding of visually-oriented text," 2025. [Online]. Available: https://arxiv.org/abs/2512.04125

[18] F. Jiang, Z. Xu, L. Niu, Z. Xiang, B. Ramasubramanian, B. Li, and R. Poovendran, "ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 157–15 173. [Online]. Available: https://aclanthology.org/2024.acl-long.809/

[19] G. Team, "Gemini: A family of highly capable multimodal models," *CoRR*, vol. abs/2312.11805, 2023.

[20] DeepSeek-AI, "Deepseek-v3 technical report," *CoRR*, vol. abs/2412.19437, 2024.

[21] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *NeurIPS*, 2022.

[22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.