AN ANALYSIS OF REASONING LENGTH SCALING AND POSITIONAL EFFECTS IN VISION LANGUAGE MODELS FOR SPATIAL REASONING

Hakan Muluk
Department of Computer Science
Bilkent University
Ankara, Türkiye
hakan.muluk@bilkent.edu.tr

ABSTRACT

Vision language models often produce step by step reasoning traces even in zero-shot settings, but it is unclear how this "reasoning length" scales with spatial problem complexity. We introduce the Largest Circle Puzzle, an easily scalable synthetic benchmark that requires connectivity-based relational filtering and relative size comparison under increasing visual clutter. By varying the number of circles, we control problem complexity, and by controlling answer location, we probe positional effects. Across several state of the art VLMs with explicit reasoning behavior, accuracy declines steadily as scenes become more crowded. On instances solved correctly, reasoning length usage typically grows with problem size and can follow an approximately linear trend, consistent with scan-like strategies. However, reasoning length alone does not predict robustness: some models exhibit strong location-dependent performance, with large accuracy gaps between corner placements even when successful reasoning length scaling remains stable.

1 INTRODUCTION

Vision language models perform strongly on many multimodal tasks and increasingly produce explicit step by step reasoning traces even in zero-shot settings, raising a key question for spatial reasoning: how does reasoning length scale with visual problem complexity, and does more reasoning yield more reliable decisions? This matters because spatial reasoning often requires combining visual perception with reasoning over relationships between multiple objects, rather than recognizing objects in isolation Liu et al. (2025). Moreover, spatial predictions can depend on where critical evidence appears in the image, making positional bias an essential factor to evaluate Zhu et al. (2025). Taken together, measuring accuracy and token usage across different complexity levels and target locations helps reveal whether VLM reasoning scales efficiently or fails under different settings.

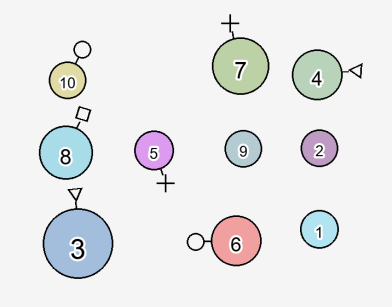


Figure 1: Example instance of the Largest Circle Puzzle with 10 labeled circles. The task is to output the label inside the largest circle among those connected to the specified shape type (here, the small hollow circle). The correct answer is 6.

In this work, we address these questions using the *Largest Circle Puzzle*[1], a scalable synthetic task where a model must output the label inside the largest circle among those connected to a specified shape type (for example, a small hollow circle). The task probes relative size comparison and connectivity-based relational filtering under increasing clutter. Because the scenes are synthetic, we can scale difficulty by varying the number of circles and test positional bias by placing the answer in different locations. Since humans can solve the puzzle in essentially linear time by scanning

[1]Reproducibility code: https://github.com/hakanmuluk/LargestCirclePuzzle

valid connections and tracking a maximum, we evaluate whether VLMs' reasoning scale similarly, reporting success rates, reasoning lengths and performance changes across placements.

2 Related Work

Reasoning traces in (V)LMs. Following Wei et al. (2022), chain of thought prompting emerged as a simple way to elicit multi step reasoning in LLMs, and it is now widely used in VLMs for multimodal reasoning, including spatial tasks (Liu et al., 2025). In parallel, recent work has begun to study reasoning length scaling by linking token usage to problem complexity (Estermann & Wattenhofer, 2025). We follow a similar analysis framework by extending scaling analyses to VLMs on a controlled spatial reasoning task, the Largest Circle Puzzle introduced in this work, and by incorporating spatial-specific factors such as answer location (positional) effects.

Controlled spatial VQA benchmarks and size comparison. Spatial reasoning has long been studied through controlled VQA benchmarks that aim to separate compositional skills from dataset shortcuts, including classic testbeds such as CLEVR (Johnson et al., 2017) and GQA (Hudson & Manning, 2019), as well as newer spatial evaluations showing that even strong VLMs struggle with basic relations under controlled perturbations (Liu et al., 2023; Kamath et al., 2023; Wang et al., 2024; Stogiannidis et al., 2025). A closely related thread concerns relative attribute comparison, especially relative size comparison, where models must compare objects rather than merely recognize them. PhysBench highlights persistent failures on grounded comparative judgments (Chow et al., 2025), while stronger grounding signals, including 3D-aware supervision from structured cues such as scene graphs and depth, can improve spatial generalization (Cheng et al., 2024). Our work follows this diagnostic tradition but focuses on a compact primitive that combines relational filtering with comparative selection under increasing clutter.

Multi-object failure modes and improving spatial competence. Recent studies also highlight systematic biases and failure modes amplified in multi-object scenes. Predictions can be sensitive to where evidence appears in the image, motivating explicit tests of corner placements (Zhu et al., 2025). Multi-object settings increase confusion and hallucination, and analyses of CLIP-like representations suggest biases toward salient or larger objects, which can interfere with "largest under constraint" queries (Abbasi et al., 2024; Chen et al., 2024). Proposed improvements include inference time attention control and reinforcement learning based post-training for spatial competence (Chen et al., 2025; Pan & Liu, 2025). Finally, symbolic-graphics evaluations suggest a modality gap: when spatial structure is explicit and compositional, LLMs can be comparatively robust to translations and rotations, motivating measurement of both task accuracy and reasoning behavior in VLMs (Qiu et al., 2024).

3 Methods

3.1 The Largest Circle Problem

In this work, we use the *Largest Circle Puzzle*, a synthetic spatial reasoning task that is solvable by humans with sufficient effort, and scalable for evaluating test-time computation in vision language models. Each instance contains n labeled circles of varying radii (labels 1 to n), with large high-contrast text to avoid OCR confounds. Circles may be connected by a short line to a nearby small shape; the circles connected to a specified shape type (by default, a small hollow circle tag) form the candidate set $\mathcal{C}$, and the model must output the label of the largest candidate circle [2].

The puzzle requires two core skills: relational filtering, to identify which circles are connected to the specified shape, and relative size comparison, to select the largest among those candidates. To discourage shortcut strategies, each instance also includes at least one circle that is larger than the correct answer but connected to a different shape type, so choosing the globally largest circle produces an incorrect output. The correct answer circle is further guaranteed to be substantially larger than the other circles connected to the specified shape type, typically by about 40–60% in

[2] In our default data-generation setting, $|\mathcal{C}| = \frac{n}{5}$.

area relative to the largest competing candidate, with randomness in this margin to avoid near-ties. Because the data are synthetic, we can scale difficulty by varying n, which increases clutter and distractors in a controlled way. Finally, to study positional bias, we vary answer placement by placing the answer circle either randomly or fixing it to a corner (top-left, top-right, bottom-left, bottom-right). All models are evaluated zero-shot with a single prompt, and we record both accuracy and reasoning-token usage across problem sizes and placement modes.

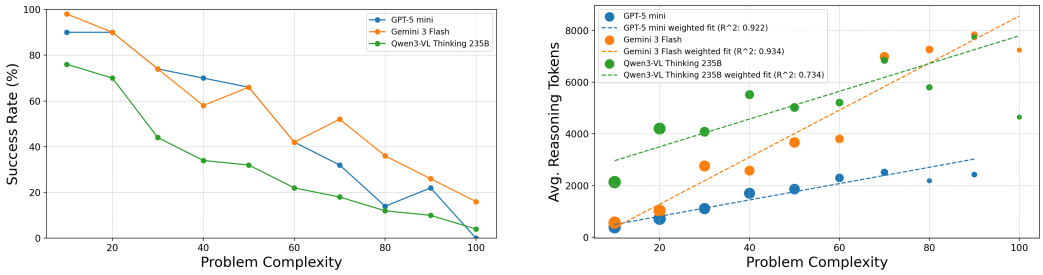
3.2 Evaluation Metrics

Our evaluation focuses on two metrics: success rate and reasoning length. Success is binary, indicating whether the model outputs the correct label for a Largest Circle Puzzle instance, which requires selecting the largest circle among those connected to the specified shape type. We measure problem complexity by the number of circles, n, which increases object count and distractors. As a proxy for the model's reasoning length, we measure total generated tokens (thinking + output), which we refer to as reasoning tokens. Because the output is constrained to a single number (typically 1–2 tokens) or "NONE," this total primarily reflects the length of the model's reasoning. We allow models to answer "NONE" (see Appendix A.1) in case they detect no circle connected to the target tag; however, in our dataset construction, an answer is always guaranteed to exist (at least one circle is connected to the specified shape type). We include the "NONE" option to avoid forcing a potentially arbitrary numeric guess when a model fails to identify any valid candidates. Also, since humans can solve the task in essentially linear time in n by scanning valid connections and tracking the maximum, we examine whether VLM computation scales similarly by fitting a linear model to average token usage versus n and reporting the R^2 goodness of fit. We also report average accuracy and average token usage, versus n. Finally, to evaluate positional bias, we compare accuracy across different answer location settings (random and fixed corners: top-left, top-right, bottom-left, bottom-right).

3.3 Considered Models

We evaluated the spatial reasoning performance of the following vision language models with explicit reasoning capabilities: GPT-5 mini and Gemini 3 Flash, both run with *medium* reasoning effort (Singh et al., 2025; Google DeepMind, 2025), and Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025).

4 Results



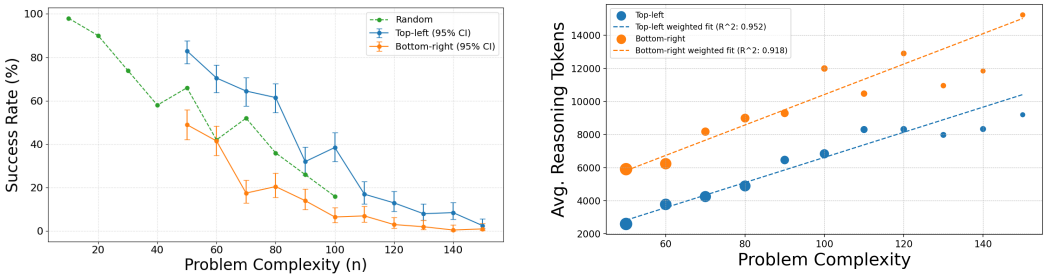
(a) (b)

Figure 2: Performance and reasoning length versus problem size n on the Largest Circle Puzzle for GPT-5 mini, Gemini 3 Flash, and Qwen3-VL-235B-A22B-Thinking with randomized answer locations. For each n, all models are evaluated on the same 50 instances. (a) Success rate versus n. (b) Average reasoning token count conditioned on successful attempts, after trimming the top/bottom 1% per model to reduce outliers; dashed lines show weighted linear fits ($R^2 = 0.922$ for GPT-5 mini, 0.934 for Gemini 3 Flash, 0.734 for Qwen3-VL Thinking 235B).

Figure 2a shows how accuracy changes with problem size n. All models degrade as n increases, indicating that adding more circles and distractors makes the relational filtering and size comparison steps substantially harder. Gemini 3 Flash is the most robust across most values of n, maintaining the highest success rates at larger problem sizes. GPT-5 mini performs strongly at small n but drops

sharply as complexity increases, reaching near-zero accuracy at the largest n. Qwen3-VL-235B-A22B-Thinking generally performs the worst across the range, with accuracy deteriorating quickly as the scene becomes more crowded.

Figure 2b reports average reasoning-token counts conditioned on successful attempts at each n, after trimming the top and bottom 1% per model to reduce outlier effects and using weighted linear fits. Under this success-only view, GPT-5 mini and Gemini 3 Flash show strong linear scaling of reasoning tokens with problem size ($R^2 = 0.922$ and 0.934, respectively), consistent with a stable "scan-and-max" style strategy on instances they can solve. In contrast, Qwen3-VL-235B-A22B-Thinking exhibits weaker linearity ($R^2 = 0.734$), suggesting a less consistent scaling behavior. Overall, reasoning length increases predictably with n for the stronger models on solvable instances, yet this does not prevent the pronounced accuracy degradation with growing clutter in Figure 2a.



(a) (b)

Figure 3: Gemini 3 Flash accuracy and reasoning length versus problem size n under different answer locations. (a) Success rate for random, top-left, and bottom-right answer placement; 95% Wilson CIs are shown for top-left and bottom-right (shared 200 puzzles per n), while random uses the earlier 50-puzzle setting (no CIs). (b) Average reasoning tokens for top-left and bottom-right after trimming the top/bottom 1% reasoning lengths (per positional setting); dashed lines are weighted linear fits to successful attempts($R^2 = 0.952$ for top-left, 0.918 for bottom-right).

We conducted additional experiments to probe positional bias and analyzed the most notable cases (Appendix A.2). Among the evaluated models, Gemini 3 Flash exhibits the strongest location sensitivity. As shown in Figure 3a, top-left placement consistently improves success rates over bottom-right across problem sizes, with bottom-right remaining the weakest condition. This gap is striking because the task and nominal difficulty (as measured by n) are unchanged, suggesting that where critical evidence appears can materially affect spatial decision quality.

Figure 3b reports the corresponding average reasoning-token counts conditioned on successful attempts for the same placements, after trimming the top/bottom 1% token lengths and fitting weighted linear trends. Token usage scales highly linearly with problem size in both settings ($R^2 = 0.952$ top-left; $R^2 = 0.918$ bottom-right), indicating stable growth in reasoning length as n increases, for the successful attempts. Importantly, this argues against a simple "shortcut" explanation: even in the top-left setting, Gemini's reasoning length still increases predictably with n (also see Appendix A.4, Figure 9b), yet accuracy remains much higher than bottom-right. Also, the bottom-right setting clearly tends to use more tokens (also see Appendix A.4, Figure 9), yet achieves notably lower accuracy than top-left. Thus, answer location impacts success in a way that is not captured by problem complexity alone, consistent with positional sensitivity in visual processing or attention allocation.

5 Conclusion

This study analyzes reasoning length scaling in vision language models using the Largest Circle Puzzle, a scalable synthetic benchmark for connectivity-based filtering and size comparison under varying clutter. On correctly solved instances, models show roughly linear token growth with problem size, yet accuracy drops as scenes become more crowded. We also find positional effects: Gemini 3 Flash shows a persistent top-left vs. bottom-right accuracy gap, and bottom-right often uses more tokens despite lower accuracy. Overall, token growth can track complexity on solvable cases but is insufficient for robustness to clutter or positional shifts; future work should test additional primitives and interventions that improve reliability without simply increasing token budget.

REFERENCES

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Analyzing CLIP's performance limitations in multi–object scenarios: A controlled high–resolution study. In *Proceedings of the ECCV Workshop on Evaluation and Analysis for FoMo (EVAL-FoMo)*, 2024. URL https://arxiv.org/abs/2502.19828. arXiv preprint arXiv:2502.19828.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631. Technical report describing the Qwen3-VL model.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2503.01773. arXiv preprint arXiv:2503.01773.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F. Fouhey, and Joyce Chai. Multi–object hallucination in vision–language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2407.06192. Accepted to NeurIPS 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.01584. Project page available at https://www.anjiecheng.me.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision–language models for physical world understanding. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2501.16411. arXiv preprint arXiv:2501.16411.

Benjamin Estermann and Roger Wattenhofer. Reasoning effort and problem complexity: A scaling analysis in llms. In *ICLR 2025 Workshop on Reasoning and Planning for Large Language Models*, 2025. URL https://openreview.net/forum?id=wfybjqEmx5. Published at ICLR 2025 Workshop on Reasoning and Planning for Large Language Models.

Google DeepMind. Gemini 3 flash model card. Model card published online, 2025. URL https://deepmind.google/models/gemini/flash/. Model card published December 2025.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real–world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, Long Beach, CA, 2019. URL https://arxiv.org/abs/1902.09506.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, Honolulu, HI, 2017. URL https://arxiv.org/abs/1612.06890.

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 (EMNLP)*, pp. 9161–9175, Singapore, 2023. URL https://arxiv.org/abs/2310.19785.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. URL https://arxiv.org/abs/2205.00363. TACL camera-ready version.

Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. *arXiv preprint arXiv:2511.15722*, 2025.

Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in vlms for the metaverse. *arXiv preprint arXiv:2503.18470*, 2025. URL https://arxiv.org/abs/2503.18470. Working paper.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Zhen Liu, Tim Z. Xiao, Katherine M. Collins, Joshua B. Tenenbaum, Adrian Weller, Michael J. Black, and Bernhard Schölkopf. Can large language models understand symbolic graphics programs? *arXiv preprint arXiv:2408.08313*, 2024. URL https://arxiv.org/abs/2408.08313. ICLR 2025 Spotlight.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank

Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision–language models. *arXiv preprint arXiv:2503.19707*, 2025. URL https://arxiv.org/abs/2503.19707. Submitted to arXiv on 25 March 2025.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision–language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.14852. Accepted to NeurIPS 2024.

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

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Youcheng Pan, Yongshuai Hou, Weili Guan, Jun Yu, and Min Zhang. Beyond the vision encoder: Identifying and mitigating spatial bias in large vision–language models. *arXiv preprint arXiv:2509.21984*, 2025. URL https://arxiv.org/abs/2509.21984. arXiv version dated 26 September 2025.

A Appendix

A.1 Prompt

The following prompt was used for all models.

```
Task:  Return the number inside the LARGEST circle among
the circles that are connected by a line to a small hollow
circle (tag).  Answer ONLY the number or NONE.
```

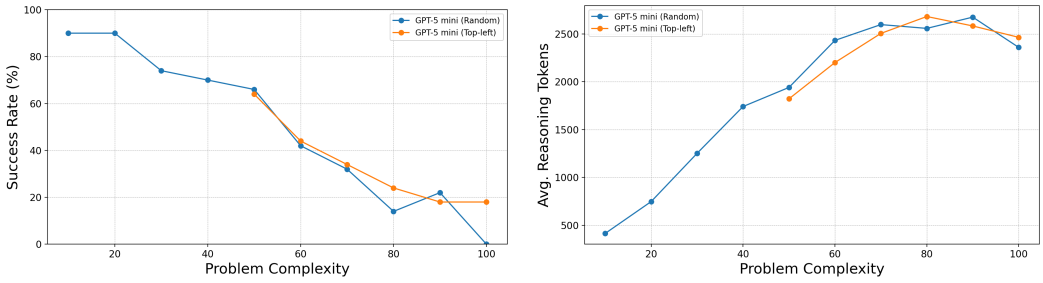
A.2 Additional positional-bias experiments

We report additional experiments on positional (locational) bias beyond those shown in the main paper. For an initial screening pass, we evaluated multiple answer location settings at a fixed problem complexity of $n = 75$ using 20 examples per setting (Tables 1, 2, and 3). Based on these preliminary results, we identified the settings with the largest deviations from the random-placement baseline and followed them up with higher-coverage evaluations using 50 examples per setting. In particular, we further analyzed the top-left placement for all models, and additionally the bottom-right placement for Gemini 3 Flash, extending the analysis to larger values of n. Except for this initial screening stage, all other experiments reported in this work use 50 examples per setting for each combination of problem complexity, model, and answer location setting; the only exception is Gemini 3 Flash in the top-left and bottom-right conditions, where we use 200 examples per setting.

A.2.1 GPT-5 mini

Table 1: Success rates (%) for GPT-5 mini on the Largest Circle Puzzle under different answer location settings. "Random" places the answer circle uniformly at random (reported at $n \in \{70, 80\}$), while the corner settings fix the answer circle to the specified corner (reported at $n = 75$).

Setting (n)	Success rate (%)
Random ($n = 70$)	32.00
Random ($n = 80$)	14.00
Top-left ($n = 75$)	53.33
Top-right ($n = 75$)	50.00
Bottom-left ($n = 75$)	16.67
Bottom-right ($n = 75$)	20.00



(a) (b)

Figure 4: (a) Mean success rate versus problem complexity n for GPT-5 mini, comparing random placement and top-left placement. (b) Average reasoning token usage versus n for the same two settings.

In the initial screening (Table 1), the top-left setting showed the largest deviation from the random-placement baseline for GPT-5 mini, so we selected it for a higher coverage follow-up. When we reran random and top-left with 50 examples per setting and swept n from 50 to 100, the two curves were largely similar across most complexities (Figure 4). A clearer separation only emerges at $n =$

100, where the random setting falls to near-zero accuracy while top-left remains higher, indicating that any location effect for GPT-5 mini is limited in this range and most visible at the highest tested complexity.

A.2.2 GEMINI 3 FLASH

Table 2: Success rates (%) for Gemini 3 Flash on the Largest Circle Puzzle under different answer location settings. "Random" places the answer circle uniformly at random (reported at $n \in \{70, 80\}$), while the corner settings fix the answer circle to the specified corner (reported at $n = 75$).

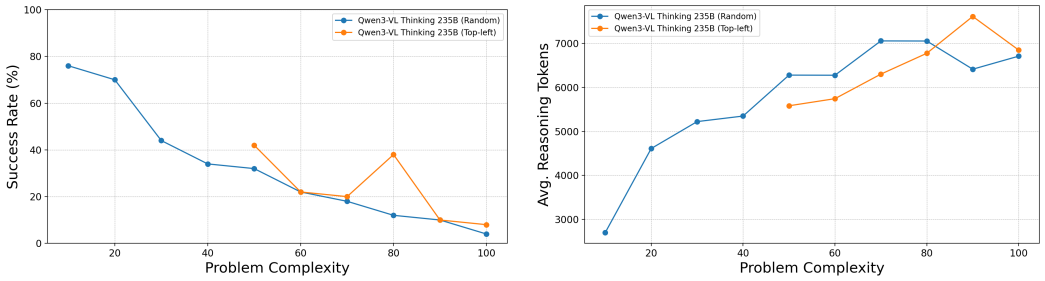
Setting (n)	Success rate (%)
Random ($n = 70$)	52.00
Random ($n = 80$)	36.00
Top-left ($n = 75$)	66.67
Top-right ($n = 75$)	56.67
Bottom-left ($n = 75$)	56.67
Bottom-right ($n = 75$)	13.33

In the initial screening for Gemini 3 Flash (Table 2), the top-left position achieved the highest success rate, while bottom-right showed a sharp drop relative to both random placement and the other corners. Given this strong spread, we selected top-left (best performing) and bottom-right (worst performing) for a higher-coverage follow-up. Figure 3 compares these two positions against the random baseline across a wider sweep of problem sizes, showing that the top-left advantage persists over increasing complexity, whereas bottom-right remains consistently weaker and degrades more rapidly as n grows.

A.2.3 QWEN3-VL-235B-A22B-THINKING

Table 3: Success rates (%) for Qwen3-VL-235B-A22B-Thinking on the Largest Circle Puzzle under different answer location settings. "Random" places the answer circle uniformly at random (reported at $n \in \{70, 80\}$), while the corner settings fix the answer circle to the specified corner (reported at $n = 75$).

Setting (n)	Success rate (%)
Random ($n = 70$)	18.00
Random ($n = 80$)	12.00
Top-left ($n = 75$)	30.00
Top-right ($n = 75$)	30.00
Bottom-left ($n = 75$)	20.00
Bottom-right ($n = 75$)	16.67



(a) (b)

Figure 5: (a) Mean success rate versus problem complexity n for Qwen3-VL-235B-A22B-Thinking, comparing random placement and top-left placement. (b) Average reasoning token usage versus n for the same two settings.

In the initial screening (Table 3), Qwen3-VL-235B-A22B-Thinking showed its strongest corner performance in the top-left setting, so we selected top-left for follow-up comparisons against the random baseline. In the higher coverage sweep across $n \in [50, 100]$ (Figure 5), the random and top-left curves are broadly similar, indicating no consistent location effect for Qwen3-VL in this range. The main exception occurs at $n = 80$, where top-left performs noticeably better than random, while at other complexities the differences are small and negligible.

A.3 Example Puzzles

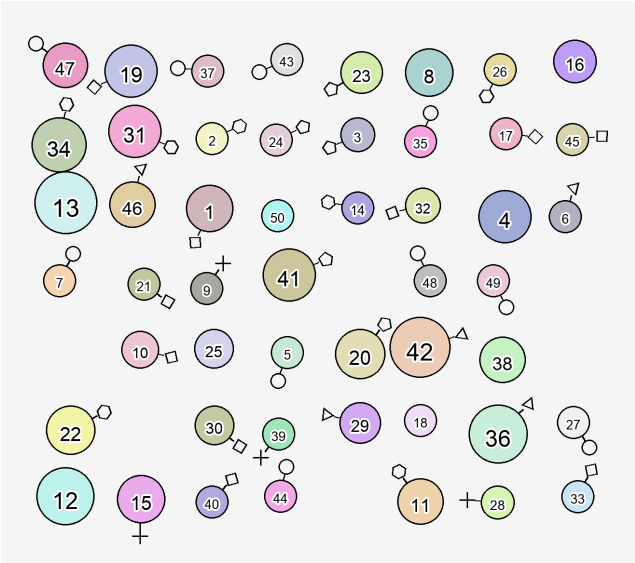


Figure 6: Example Largest Circle Puzzle instance from the top-left answer location setting, with n = 50 and 47 as the answer.

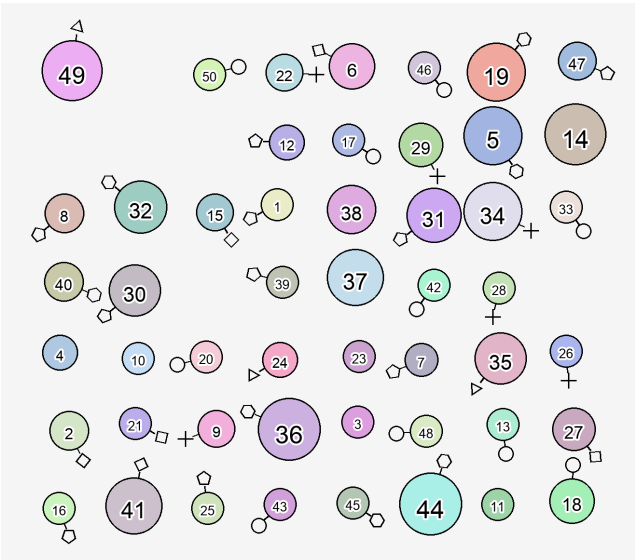
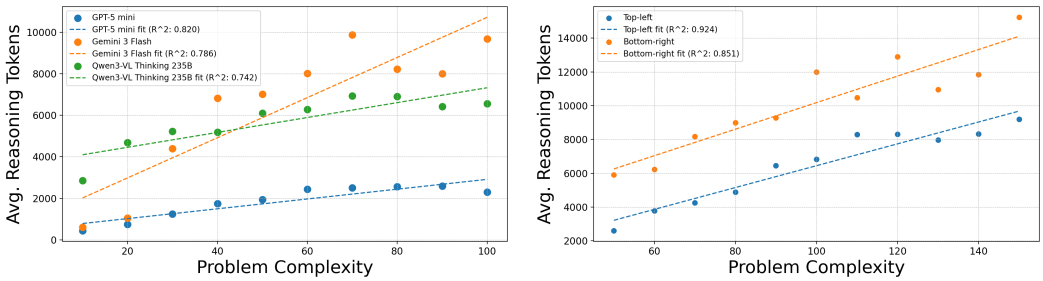


Figure 7: Example Largest Circle Puzzle instance from the bottom-right answer location setting, with n = 50 and 18 as the answer.

A.4 ADDITIONAL RESULTS



(a) (b)

Figure 8: Average reasoning tokens versus problem size n, computed over all attempts (not only successful ones). In all plots, we trim the top and bottom 1% of reasoning token lengths (per model) to reduce outlier effects, and show weighted linear fits (dashed). (a) Trimmed averages over 50 instances per model–n: GPT-5 mini ($R^2 = 0.820$), Gemini 3 Flash ($R^2 = 0.786$), and Qwen3-VL-235B-A22B-Thinking ($R^2 = 0.742$). (b) Trimmed averages for Gemini 3 Flash with 200 instances per n under top-left ($R^2 = 0.924$) and bottom-right ($R^2 = 0.851$) answer locations.

Figure 8 reports average reasoning token usage as a function of problem size n when all attempts are included, rather than conditioning on success. This view captures both solved and failed runs, so it reflects how each model allocates test-time computation under increasing clutter even when it does not reach the correct answer. In Figure 8(a), all three models show an overall upward trend in trimmed average tokens with n, and the weighted linear fits indicate that token usage remains approximately linear in this aggregate sense (GPT-5 mini $R^2 = 0.820$, Gemini 3 Flash $R^2 = 0.786$, Qwen3-VL-235B-A22B-Thinking $R^2 = 0.742$). Compared to the success-only analyses in the main text, these fits are generally weaker, which is expected because failures introduce additional variability in both stopping behavior and reasoning length. Figure 8(b) performs the same analysis for Gemini under higher coverage (200 instances per n) and shows that linear scaling persists across placements, with stronger linearity for top-left than bottom-right ($R^2 = 0.924$ vs. 0.851). Together, these results suggest that average token consumption still grows with problem size even when unsuccessful attempts are included, while positional changes can measurably affect the stability of this scaling.

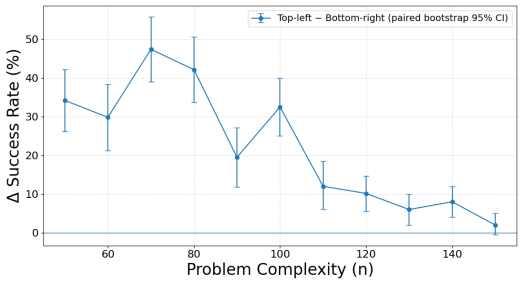


Figure 9: Accuracy gap between top-left and bottom-right answer locations for Gemini 3 Flash as a function of problem size n. Points show the paired difference in success rate, Δ = top-left − bottom-right, evaluated on the same 200 puzzle instances per n; error bars indicate 95% paired bootstrap confidence intervals. The horizontal line at $\Delta = 0$ marks no positional effect.