Video Models Start to Solve Chess, Maze, Sudoku, Mental Rotation, and Raven' Matrices

Hokin Deng
Carnegie Mellon University
hokind@andrew.cmu.edu

Source Code: github.com/hokindeng/VMEvalKit **Results Page**

Abstract

We show that video generation models could reason now. Testing on tasks such as chess, maze, Sudoku, mental rotation, and Raven's Matrices, leading models such as Sora-2 achieve >60% success rates. We establish a robust experimental paradigm centered on the "Task Pair" design. We build a code framework, with 39 models available already, that supports this paradigm and allows for easy scaling—users can add models and tasks efficiently. We show our automated evaluation strongly correlates with human judgment, and therefore this paradigm is highly scalable. We see an opportunity, given the availability of our paradigm, to do reinforcement learning for improving reasoning in video models.

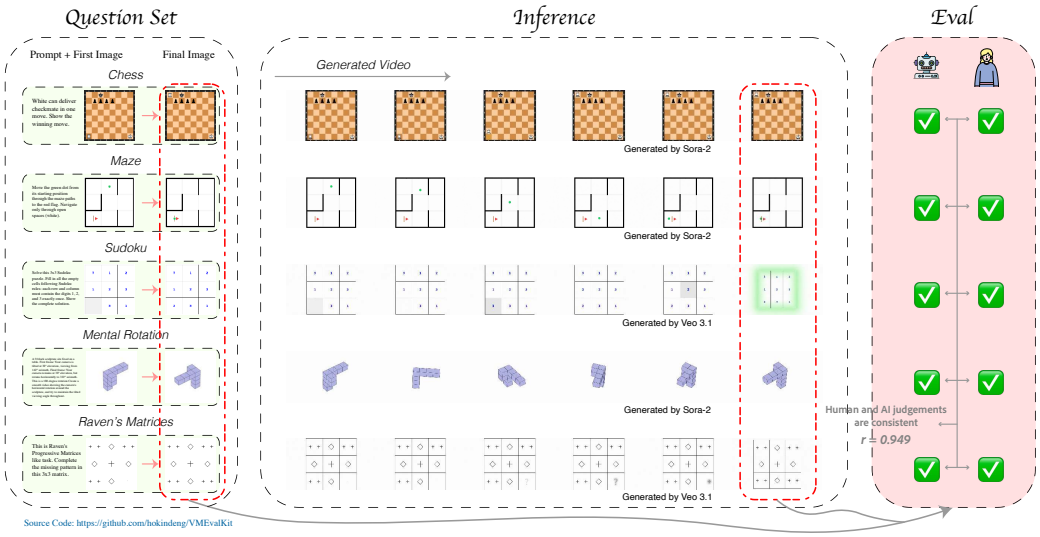

Figure 1: Representative examples of video reasoning across diverse cognitive tasks. We show 6-frame temporal sequences from videos. The robust correlation ($r = 0.949$, $p < 0.001$) between human and automated GPT-4o evaluations demonstrates the reliability of our evaluation framework for assessing reasoning in video generation models.

1 Introduction

The rapid evolution of video generation models has transformed our ability to synthesize realistic, high-fidelity videos from text descriptions. Systems like Sora-2 [OpenAI, 2025], Veo-3 [Google DeepMind, 2025], Runway Gen-3 [Runway Research, 2024], and Luma Dream Machine [Luma AI, 2024] can now generate videos that are visually indistinguishable from human-created content, depicting complex scenes with coherent motion, consistent objects, and plausible physics. These advances have been driven by scaling up training data, model parameters, and computational resources, coupled with architectural innovations. However, a fundamental question remains largely unexplored: **Can these models reason?**

Visual reasoning—the capacity to understand, manipulate, and solve problems through visual representations—represents a qualitatively different challenge from photorealistic synthesis. While generating a realistic video of a chess game or a person solving a Sudoku puzzle requires learning statistical patterns of motion and appearance, **actually solving** these problems requires understanding the underlying rules, constraints, and logical relationships that govern valid solutions [Wiedemer et al., 2025]. The distinction is critical: a model that has learned to mimic the visual appearance of problem-solving without understanding the problem itself cannot reliably generate correct solutions, nor can it adapt to novel problem instances requiring genuine reasoning [Deng et al., 2025].

Evaluating reasoning in video models isn't straightforward. Our paradigm provides (1) an initial image showing the unsolved problem, (2) a text instruction prompt explaining what to do, (3) a final image showing the correct solution (See Figure 1 and 2). In inference, we only show (1) and (2) to the video models and (3) is reserved for evaluation. We build VMEvalKit, a flexible open-source code framework for our experimental paradigm. Our codebase could (i.) create tasks for testing reasoning at scale, (ii.) run inferences of video models on the tasks at scale, and (iii.) do AI-powered evaluation at scale. It's also modularized which the users could add more tasks and models on their own easily. Now our tasks include chess, maze, Sudoku, mental rotation, and Raven's Matrices. The code framework now provides inferences for 39 video models spanning 11 model families.

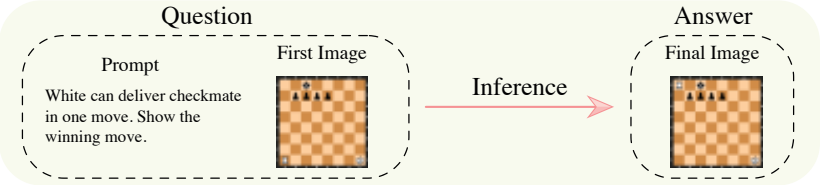

Figure 2: Each of our question unit includes three components: (1) an initial image showing the unsolved problem, (2) a text instruction describing the task, and (3) a final image showing the correct solution. During inference, models are only given the initial image and the instruction prompt. The final image is withheld and used solely for evaluation.

We ran the VMEvalKit evaluation on 450 tasks across six cutting-edge video generation models: OpenAI Sora 2, Google Veo 3.0 and 3.1, Runway Gen4 Turbo, WaveSpeed WAN 2.2, and Luma Ray 2. Each model was tested on 75 tasks covering five reasoning tasks—chess, mazes, Raven's matrices, mental rotation, and Sudoku. Performance was measured using both human ratings and GPT-4o-based automated scoring, with strong agreement between the two (Pearson r = 0.949, Cohen's K = 0.867), confirming the reliability of automated evaluation. Sora 2 stood out with a 68% overall success rate, particularly excelling in maze navigation (87%), Raven's matrices (73%), and chess (73%). Veo 3.0 followed at 47%, achieving a perfect score on Sudoku tasks, while Veo 3.1 landed at 35%. The remaining models struggled: Gen4 Turbo reached 24%, WAN 2.2 hit 11%, and Luma Ray 2 scored just 1%. Among task types, Sudoku proved the easiest (57% average success), while mental rotation and chess were the most challenging (around 11%). We see all of our inference results from the video models, together with the prompt and question images, here at the **Results Page**

We argue that we have observed strong evidences that the video models have emerged reasoning abilities. There are important research opportunities on the horizon. First, it enables reinforcement learning: clear success signals (e.g., solved maze or legal Sudoku) make it possible to fine-tune models toward more consistent reasoning. Second, we need mechanistic interpretability to uncover

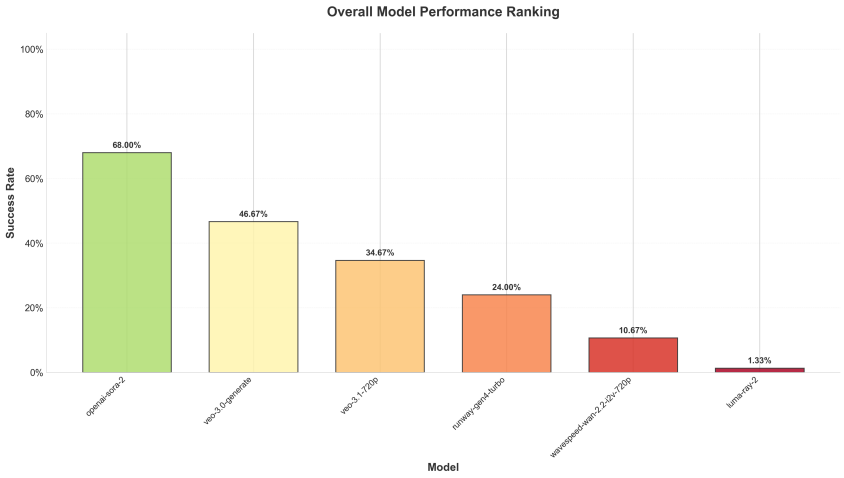

Figure 3: Success rates on all 5 tasks showing clear performance hierarchy across models.

how models represent rules, track state, and reason across steps. Third, functional decomposition—breaking tasks into sub-reasoning faculties—can help isolate where reasoning fails and how to improve it. Together, these works shall move us to a new paradigm-shift in AIs.

2 Related Works

2.1 Reasoning in Language Models

The emergence of reasoning capabilities in large language models has been extensively studied across multiple domains. Chain-of-thought prompting demonstrated that language models can solve complex reasoning tasks by generating intermediate reasoning steps, with performance scaling dramatically with model size [Wei et al., 2022]. This has been extended to multi-step reasoning in mathematics, logical inference, and strategic planning [Cobbe et al., 2021, Creswell et al., 2022, Yao et al., 2023].

Recent work has focused on evaluating and improving reasoning through specialized benchmarks. GSM8K evaluates mathematical problem-solving, while BIG-Bench assesses diverse reasoning capabilities including logical deduction, causal reasoning, and analogical thinking [Cobbe et al., 2021, Srivastava et al., 2023]. MMLU provides comprehensive evaluation across 57 subjects requiring factual knowledge and reasoning. Process reward models have shown that rewarding correct reasoning steps rather than just final answers substantially improves reliability [Hendrycks et al., 2021, Lightman et al., 2023]. However, these advances remain primarily text-based, with limited exploration of reasoning in visual and video domains.

2.2 Video Generation Evaluation

Community "arena" efforts aggregate large-scale pairwise preferences and formalize open evaluation with Elo-style ranking [Artificial Analysis, 2025, Jiang et al., 2024]. Sometime industries also published reflection memos [Google, 2025]. Dimensioned suites such as *VBench* decompose quality into text relevance, temporal coherence, style, and related facets [Huang et al., 2024], while *VideoBench* pushes closer agreement with human raters [Han et al., 2025]. Other measures include learning-based metrics, Fréchet Video distance, and Kernel Video distance. [He et al., 2024, Yan and Liu, 2024]. Benchmarks also focus on specific faculties such as temporal and metamorphic robustness [Yuan et al., 2024], physics/world-model fidelity [Li et al., 2024a, Motamed et al., 2024], and so on [Liu et al., 2024]. One recent paper argues that Veo 3 exhibits broad zero-shot perception and reasoning—spanning segmentation, edge detection, editing, physical/affordance understanding, tool-use simulation, and simple maze/symmetry reasoning—suggesting video models are trending toward unified, generalist vision foundation models [Wiedemer et al., 2025].

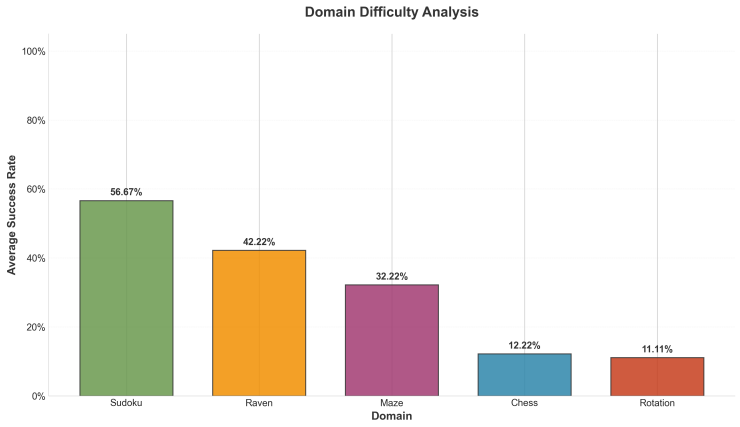

Figure 4: Average success rates by reasoning domain reveal distinct level of challenges.

2.3 Visual Cognition Evaluation

Most of the visual cognition evaluation for AI systems are done in MLLMs. One recent project used human developmental trajectory to assess visual cognition in MLLMs [Li et al., 2024b, Luo et al., 2025c]. Other works focus on particular visual cognitive faculties such perspective taking and theory of mind [Gao et al., 2024, Wang et al., 2025, Abrini et al., 2025], mechanical reasoning [Sun et al., 2024], perceptual constancy [Sun et al., 2025], gaze understanding [Zhang et al., 2025], physical conservation [Luo et al., 2024], cognitive control [Luo et al., 2025d], and so on [Li et al., 2025, Luo et al., 2025a, Deng, 2025].

3 Methods

3.1 Video Inference

Our VMEvalKit provides to evaluate 39 text-to-video models through a unified inference framework supporting both commercial APIs and open-source implementations. Our system dynamically loads model wrappers from a centralized catalog containing detailed configurations for 9 model families:

- Commercial API Models: OpenAI Sora (sora-2, sora-2-pro), Google Veo (3.0-generate, 3.1 via WaveSpeed proxy), Runway ML (gen4-turbo, gen4-aleph, gen3a-turbo), Luma Dream Machine (ray-2, ray-flash-2), WaveSpeed WAN (2.1/2.2 variants with 480p/720p/5B configurations).
- Open-Source Models: LTX-Video (13B-distilled, 13B-dev, 2B-distilled), HunyuanVideo-I2V, VideoCrafter2-512, DynamiCrafter (256p/512p/1024p variants).

Each model implements sophisticated image preprocessing to ensure integrity and homogeneity:

- Resolution Standardization: Input images undergo automatic padding/resizing to model-specific requirements (Sora: 1280×720/720×1280, VEO: 16:9/9:16 with padding, Runway: 1280×768/768×1280 letter boxing)
- Color Space Conversion: Automatic RGB conversion with neutral gray padding (128,128,128) to prevent harsh borders
- Aspect Ratio Management: Scale-preserving letter boxing where images fit within target dimensions maintaining aspect ratio, then center-padded to exact specifications
- Format Standardization: Base64 encoding for API models, tensor conversion for local models, MIME type validation (image/jpeg, image/png, image/webp)
- Inference Parameters: During inference, all models are run using standardized settings: each video is generated with an 8-second duration, a temperature of 0.7 is applied where the model allows for sampling control, and a fixed random seed of -1 is used to promote variability across generations. Resolution handling is streamlined through automatic padding to meet each model's input requirements.

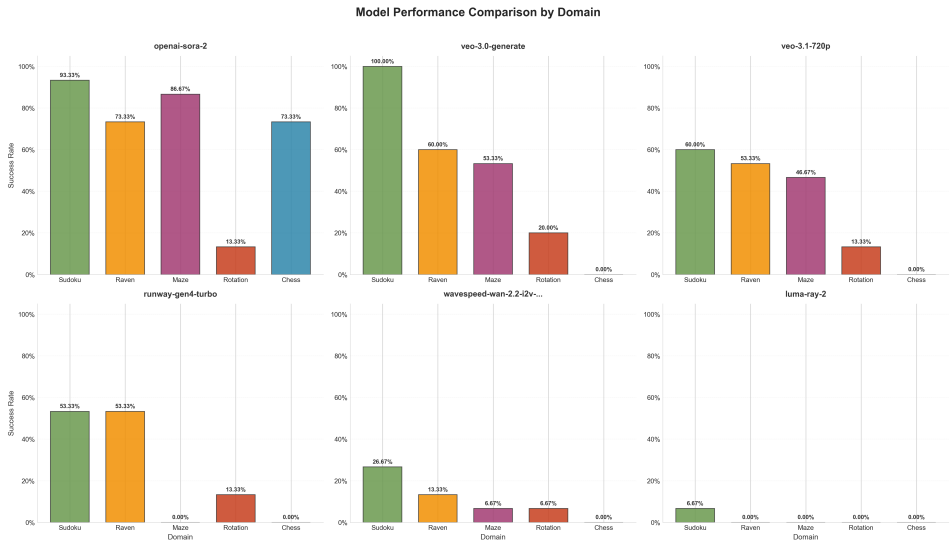

Figure 5: Individual model performance across all reasoning domains.

3.2 Task Generation

The VMEvalKit now provides parameterized ways to generating the following 5 tasks. For our experiments, we first have used VMEvalKit task sets to generate several batches of tasks. Then, we use human labor to manually go through each task to ensure the task correctness and integrity. Upon our inspection, all the tasks generated by the VMEvalKit parameterized methods are correct and good-to-go for testing the models.

Chess Reasoning: Mate-in-1 puzzles using a self-contained generator producing 100+ verified positions through back-rank mates, queen corner patterns, and tactical combinations. Each task shows an initial board position and requires demonstrating the winning move sequence.

Maze Navigation: 3×3 grid mazes generated via Kruskal's algorithm with professional rendering. Tasks display a start position (green dot) and goal (red flag), requiring path-finding demonstrations from start to finish.

Raven Progressive Matrices: 3×3 pattern completion puzzles testing abstract visual reasoning. Our RPM generator creates rule-based patterns (shape progression, rotation, color sequences) with systematic difficulty control.

3D Mental Rotation: Voxel-based structures (8–9 cubes) rendered from different camera viewpoints with 180° horizontal rotations and 20–40° elevation angles. Tasks require demonstrating spatial understanding through smooth camera transitions.

Sudoku Solving: Simplified 3×3 grids with exactly one missing number, testing logical deduction capabilities. Each puzzle has a unique solution following standard Sudoku constraints.

All tasks follow a consistent (first_frame, final_frame, prompt) format enabling standardized evaluation across domains. See Appendix A for details.

4 Evaluation

Our VMEvalKit provides infrastructure for both human and AI automated evaluation of the results.

- Automated Evaluation: Vision-language model (GPT-4o) assessment comparing generated video final frames against ground truth solutions. The evaluator receives first frames, expected final frames, and actual video outputs, providing 1–5 correctness scores with explanations. Evaluation prompts are task-specific with detailed scoring criteria.

- Human Evaluation: Expert annotators using a Gradio interface assess video generation quality and reasoning correctness. The system supports session resumption and consistent scoring across multiple evaluators.

Both methods use identical 5-point scales where scores 4–5 indicate successful reasoning demonstration, enabling statistical comparison of automated vs. human assessment reliability. This has been stated clearly in the prompts for both humans and AIs.

5 Experiments

5.1 Overall Results

We conduct comprehensive evaluation on 6 representative models: OpenAI Sora-2, Google Veo 3.0/3.1, Runway Gen4-Turbo, WaveSpeed WAN 2.2, and Luma Ray-2. Each model generates videos for 75 reasoning tasks (15 per domain) totaling 450 evaluations. All models use 8-second generation duration with task-appropriate prompts. OpenAI Sora-2 achieves the highest success rate at 68.0% (3.853/5 average score), followed by Google Veo 3.0 (46.7%), Veo 3.1 (34.7%), Runway Gen4 (24.0%), WaveSpeed (10.7%), and Luma (1.3%). This establishes a clear performance hierarchy across different model architectures and training approaches (see Figure 3).

Reasoning domains exhibit distinct difficulty profiles. Sudoku emerges as most tractable (56.7% average success), followed by Raven matrices (42.2%), Maze navigation (32.2%), Chess solving (12.2%), and 3D Mental Rotation (11.1%). This ranking reflects the varying cognitive reasoning demands of each reasoning type for video models to solve (see Figure 4). Superior models show consistent performance across domains, while weaker models exhibit domain-specific failures. This reveals both general reasoning capabilities and domain-specific strengths and weaknesses. Notably, Veo 3.0 achieves very high Sudoku performance but fails Chess reasoning, indicating specialized rather than general reasoning capabilities (see Figure 5).

The failure cases exhibit clear and often idiosyncratic patterns, but we currently lack an idea of principled framework to systematically analyze or interpret them. For now, we present some examples in the Appendix B and all examples in Results Page. We strongly encourage readers to explore them, as they may reveal valuable insights for future investigation.

5.2 Solving Chess

Only OpenAI Sora-2 showed strong chess reasoning (see Figure 6), succeeding on 73% of tasks (11/15), especially with back-rank mates (100%) and queen-corner mates (85%). It failed on all knight-fork problems and multi-piece tactics, revealing limits in complex patterns. All other models—Veo 3.0/3.1, Runway Gen4-Turbo, WaveSpeed WAN 2.2, and Luma Ray-2—scored 0%, either making legal but tactically useless moves or showing basic rule confusion. Overall, chess remain extremely challenging, with a total success rate of just 12.22% across 90 evaluations (see Figure 5).

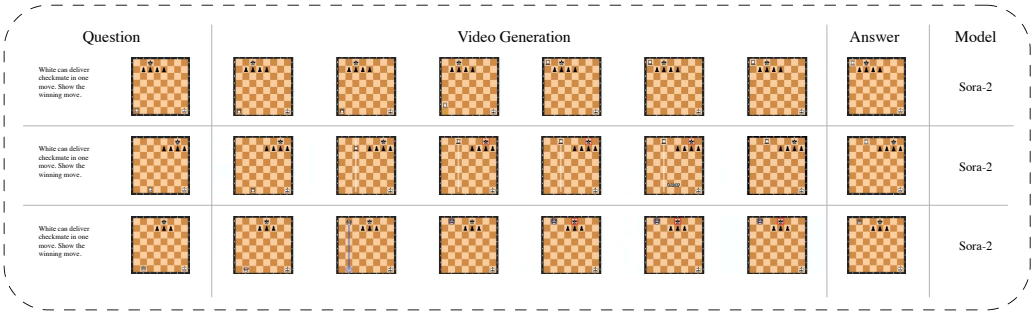

Figure 6: Example runs from video models trying to solve our chess problems. Results Page

5.3 Solving Maze

Overall success was low, while video models do solve mazes (see Figure 5 and Figure 7). Sora-2 was the only consistently strong model, showing reliable backtracking and stable green-dot tracking. Veo 3.0/3.1 were middling (half of tasks), better on simple, linear layouts and weaker on branching or backtracking. WaveSpeed WAN 2.2 managed a rare win on the easiest maze, while Runway Gen4-Turbo and Luma Ray-2 failed across the board.

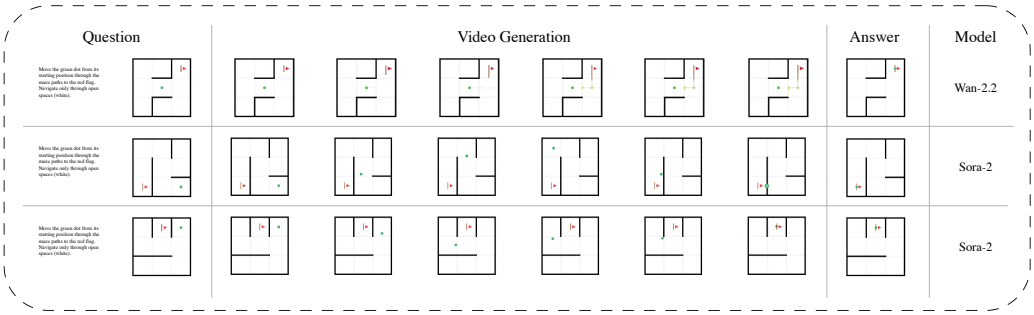

Figure 7: Example runs from video models trying to solve our maze problems. Results Page

5.4 Solving Sudoku

It turns out that Sudoku might be the most tractable domain overall, as most of the models are able to solve some (see Figure 8). Veo 3.0 was almost flawless, and the same for Sora-2, with one isolated miss due to a final-frame inconsistency. Veo 3.1 and Runway Gen4-Turbo were middling—good on simpler fills, weaker when constraints interacted. WaveSpeed WAN 2.2 was weak, and Luma Ray-2 just solved one (see Figure 5).

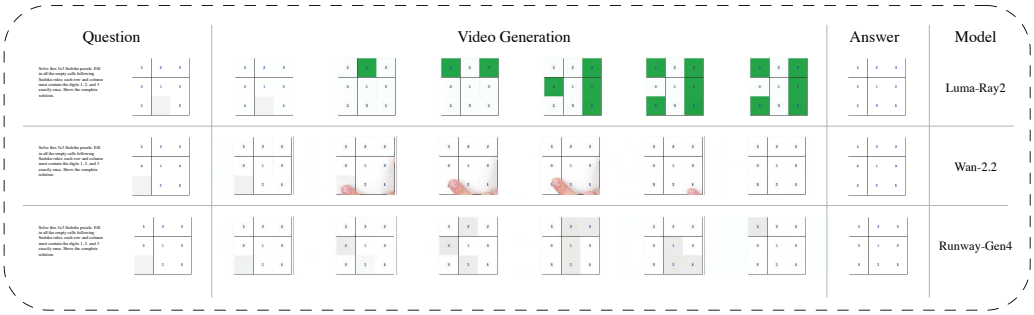

Figure 8: Example runs from video models trying to solve our Sudoku problems. Results Page

5.5 Solving Mental Rotation

Mental rotation seems to be a very hard problem, but weirdly video models sometimes do maintain consistency and are able to solve the rotation problems. Some failures are due to de-generation of the transformation, while other failures include not rotating to the right angle that's required in the prompt (see Figure 9). Veo 3.0 led but only on the simplest shapes and basic 180° turns; Sora-2, Veo 3.1, and Runway Gen4-Turbo managed occasional wins on easy layouts; WAN 2.2 rarely succeeded, and Luma Ray-2 failed entirely (see Figure 5).

5.6 Solving Raven's Matrices

The models do exhibit a level of competency in solving our Raven's Matrices problems (see Figure 10). Sora-2 excelled at shape and number progressions but stumbling on rotations and multi-rule combos.

Veo 3.0 was strong on color sequences and simple rotations; Veo 3.1 and Runway Gen4-Turbo were middling and sensitive to visual clarity. WaveSpeed WAN 2.2 was weak, and Luma Ray-2 failed entirely (see Figure 5).

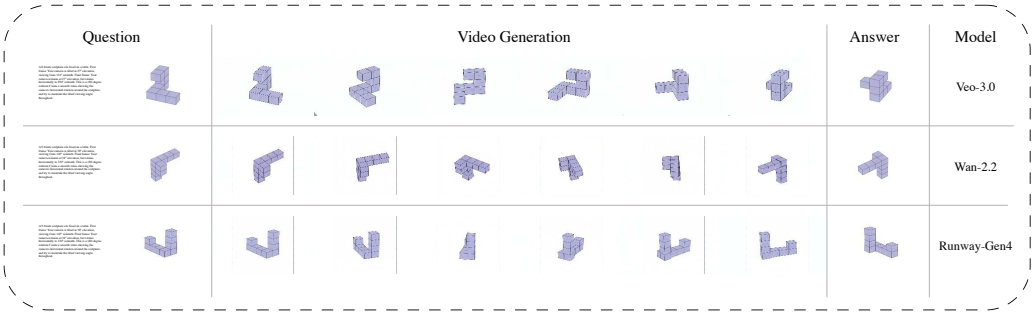

Figure 9: Example runs from video models trying to solve our Sudoku problems. Results Page

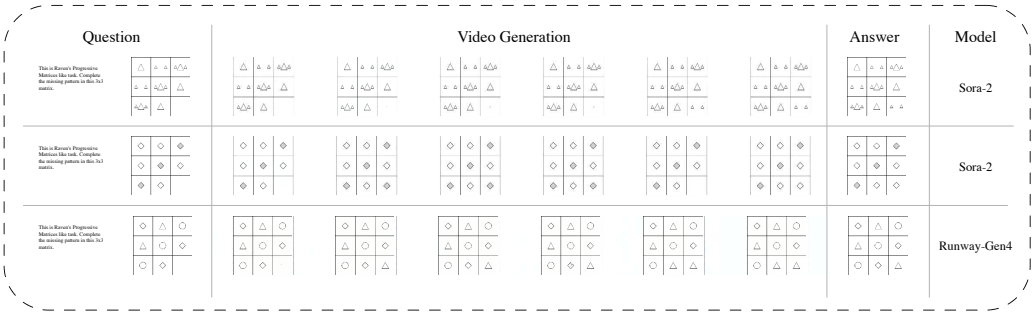

Figure 10: Example runs trying to solve our Raven's Matrices problems. Results Page

5.7 Robust Evaluation

To validate our evaluation methodology, we conducted statistical comparison between GPT-4O assessments and human expert annotations. Statistical analysis reveals exceptionally strong agreement between evaluation methods with Pearson correlation coefficient r = 0.949, indicating near-perfect concordance in scoring patterns. The scatter plot analysis shows consistent linear relationship between automated and human assessments across all score levels (1-5), with point clustering demonstrating frequent agreement on identical score assignments (see Figure 16).

Figure 11: AI Automated versus Human Evaluation

6 Discussion

Our work adds evidences that scaling laws and the philosophy of emergence start to work in the pixel domain, together with another recent work which reports similar findings in Veo-3 [Wiedemer et al., 2025]. Our work extends their efforts by buiding an integrated modular code framework VMEvalKit which allow users to add models and tasks at scale. The emergence of reasoning in video models is another victory of the scaling hypothesis. It is believed that intelligence comes from compression and therefore in order to achieve intelligence is the path is to scale data, and useful data [Sutton, 2019, Sutskever, 2023]. This belief worked in language space, and will probably work in video space.

However, several interesting yet hard problems are lying in front of us. What is the best way to evaluate reasoning, and particularly language-free reasoning, in video models? Analogical questions have been asked in humans. Could humans without language ability reason? The answer is yes: aphasia patients could reason and solve difficult problems perfectly [Fedorenko et al., 2024]. However, counter-arguments exist as most of the aphasia patients lose the abilities of language later in life and a human brain which does not acquire language throughout its entire life is particular rare [Penfield and Roberts, 1959]. Therefore, it seems to be the case that there exists language-free reasoning, but how to find it, as function as such, seems to be particularly difficult both in the human and the machine minds. In current video models, it's obvious that their training data are contaminated with language.

Not only hard is its ontology, but also its metrology. For human scientists, how to prompt a patient without language ability to understand the task that you want the patient to play. This could be put as the "rule-following" problem for language-free thought [Wittgenstein, 1958, Kripke, 1982]. In layman's terms, how do you know both of you are playing the same game when two of you could not communicate via language. On the opposite side, it's could also be put that if the aphasia patient is able to win a competent in a game of chess, it's very unlikely if the patient is not playing chess at all, and not understanding the rule of chess [Cappelen and Dever, 2024, Cao and Warren, 2025, Cappelen and Dever, 2025]. Nevertheless, assessing reasoning and thought without language at all, both in humans and machines, still remain a very hard problem. In our paradigm, we still take the advantage of text prompts to import "rule-following".

Let's say we are able to find existence of language-free reasoning and evaluate it fine. What is its eidos? One interesting paper about hippocampus has shown that human brains use the same substrate for retrieving the past and imagining the future [Hassabis et al., 2007], while other recent papers argue probably it's also how human conceptual reasoning are done [Xiao et al., 2025, Veselic et al., 2025]. The central belief behind this line of ideas is that language is not the substrate of the reasoning and thought but very rich experiential content [Luo et al., 2025b], with some empirical support for and against it [Balaban and Ullman, 2025a,b]. In other words, we could believe that there is such thing that one could reason with only world content, the very rich experiences of senses and actions, which essentially comes from the compression of world itself.

There are also some engineering opportunities laying in front of us. First, one should be able to do supervised finetuning on the reasoning perceptual flow of the video models. Just like how language models are finetuned on chain-of-thought data, we should be able to display correct reasoning flow, and use that to finetune the reasoning ability in video models. Second, given that we have correctness signal on each task, we should also be able to use reinforcement learning to improve reasoning in video models. Third, understanding the functional decompositions of each ability in the video models offer another opportunity to use video models as experiential intelligence in silico as such, just like how we use language models as models of language [Millière, 2024]. Fourth, it also offers tremendous mechanistic interpretability opportunities in understanding what's happening in video models that enable reasoning, just like of language reasoning in language models [Wu et al., 2025a,b, Sheta et al., 2025].

7 Conclusion

We present the evidence that video generation models can reason across hard tasks, enabled by our scalable VMEvalKit framework, which reveals strong performance in top models and paves the way for future advances in evaluation, fine-tuning, and interpretability.

References

Mouad Abrini, Omri Abend, Dina Acklin, Henny Admoni, Gregor Aichinger, Nitay Alon, Zahra Ashktorab, Ashish Atreja, Moises Auron, Alexander Aufreiter, et al. Proceedings of 1st workshop on advancing artificial intelligence through theory of mind. *arXiv preprint arXiv:2505.03770*, 2025.

Artificial Analysis. Video-generation-arena leaderboard. https://huggingface.co/spaces/ArtificialAnalysis/Video-Generation-Arena-Leaderboard, 2025. Community leaderboard hosted on Hugging Face.

Halely Balaban and Tomer D Ullman. The capacity limits of moving objects in the imagination. *Nature Communications*, 16(1):5899, 2025a.

Halely Balaban and Tomer D Ullman. Physics versus graphics as an organizing dichotomy in cognition. *Trends in Cognitive Sciences*, 2025b.

Rosa Cao and Jared Warren. Mental representation,"standing-in-for", and internal models. *Philosophical Psychology*, 38(2):379–396, 2025.

Herman Cappelen and Josh Dever. Introspective machines: Are llms better at self-reflection than humans? *Philosophical Perspectives*, 38(1):189–196, 2024.

Herman Cappelen and Josh Dever. Going whole hog: A philosophical defense of ai cognition. *arXiv preprint arXiv:2504.13988*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.

Hokin Deng. Reinforcement learning versus natural language programs: Where is flexible planning and problem solving in natural intelligence coming from, 2025.

Hokin Deng, Raphaël Millière, Alessandro Palmarini, Jeff Goldsmith, Melanie Mitchell, and John Krakauer. Knowing how versus knowing that: Competence without comprehension in language models, 2025. Manuscript under review.

Evelina Fedorenko, Steven T Piantadosi, and Edward AF Gibson. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, 2024.

Qingying Gao, Yijiang Li, Haiyun Lyu, Haoran Sun, Dezhi Luo, and Hokin Deng. Vision language models see what you want but not what you see. *arXiv preprint arXiv:2410.00324*, 2024.

Google. Gemini share: Reflections on video generation evaluation. https://gemini.google.com/share/aadd18fe8234, 2025. URL https://gemini.google.com/share/aadd18fe8234. Shared post via Google Gemini.

Google DeepMind. Fuel your creativity with new generative media models and tools, May 2025. URL https://deepmind.google/discover/blog/fuel-your-creativity-with-new-generative-media-models-and-tools/. Announcing Veo 3. Accessed: 2025-10-30.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li jia Li, and Yongxin Ni. Video-bench: Human-aligned video generation benchmark, 2025. Benchmark emphasizing alignment with human preferences.

Demis Hassabis, Dharshan Kumaran, Seralynne D Vann, and Eleanor A Maguire. Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, 104(5):1726–1731, 2007.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Mantisscore: Building automatic metrics to simulate fine-grained human feedback for video generation, 2024. Introduces a learned metric that correlates with human ratings.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2024. doi: 10.48550/arXiv.2311.17982. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.pdf.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models, 2024. URL https://arxiv.org/abs/2406.04485. NeurIPS 2024; v4 (2024-11-11).

Saul A. Kripke. *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Harvard University Press, Cambridge, MA, USA, 1982. ISBN 0-674-95401-7.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models, 2024a. Benchmark emphasising instruction following and physics adherence.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2024b.

Yijiang Li, Bingyang Wang, Tianwei Zhao, Qingying Gao, Hokin Deng, and Dezhi Luo. Evaluating multi-modal language models through concept hacking. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Weiqi Ye, and Jiawei Zhang. A survey of ai-generated video evaluation: Toward unified, multi-facet assessment, 2024.

Luma AI. Dream machine, 2024. URL https://lumalabs.ai/dream-machine. Initial launch June 2024; Accessed: 2025-10-30.

Dezhi Luo, Haiyun Lyu, Qingying Gao, Haoran Sun, Yijiang Li, and Hokin Deng. Vision language models know law of conservation without understanding more-or-less. *arXiv preprint arXiv:2410.00332*, 2024.

Dezhi Luo, Qingying Gao, and Hokin Deng. Rethinking the simulation vs. rendering dichotomy: No free lunch in spatial world modelling. *arXiv preprint arXiv:2510.20835*, 2025a.

Dezhi Luo, Qingying Gao, and Hokin Deng. Rethinking the simulation vs. rendering dichotomy: No free lunch in spatial world modelling, 2025b. Discusses simulation vs rendering trade-offs in spatial world models:contentReference[oaicite:4]index=4.

Dezhi Luo, Yijiang Li, and Hokin Deng. The philosophical foundations of growing ai like a child. *arXiv preprint arXiv:2502.10742*, 2025c.

Dezhi Luo, Maijunxian Wang, Bingyang Wang, Tianwei Zhao, Yijiang Li, and Hokin Deng. Machine psychophysics: Cognitive control in vision-language models. *arXiv preprint arXiv:2505.18969*, 2025d.

Raphaël Millière. Language models as models of language. *arXiv preprint arXiv:2408.07144*, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2024. Introduces Physics-IQ benchmark assessing physical understanding.

OpenAI. Sora 2 is here, September 2025. URL https://openai.com/index/sora-2/. Accessed: 2025-10-30.

Wilder Penfield and Lamar Roberts. *Speech and brain mechanisms*. Princeton University Press, 1959.

Runway Research. Introducing gen-3 alpha: A new generation of foundation models, June 2024. URL https://runwayml.com/research/introducing-gen-3-alpha. Accessed: 2025-10-30.

Hala Sheta, Eric Huang, Shuyu Wu, Ilia Alenabi, Jiajun Hong, Ryker Lin, Ruoxi Ning, Daniel Wei, Jialin Yang, Jiawei Zhou, et al. From behavioral performance to internal competence: Interpreting vision-language models with vlm-lens. *arXiv preprint arXiv:2510.02292*, 2025.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. OpenReview ID: uyTL5Bvosj.

Haoran Sun, Qingying Gao, Haiyun Lyu, Dezhi Luo, Yijiang Li, and Hokin Deng. Probing mechanical reasoning in large vision language models. *arXiv preprint arXiv:2410.00318*, 2024.

Haoran Sun, Suyang Yu, Yijiang Li, Qingying Gao, Haiyun Lyu, Hokin Deng, and Dezhi Luo. Probing perceptual constancy in large vision language models. *arXiv preprint arXiv:2502.10273*, 2025.

Ilya Sutskever. An observation on generalization. https://www.youtube.com/watch?v=AKMuA_TVz3A, August 2023. Talk at the Simons Institute workshop "Large Language Models and Transformers".

Richard S. Sutton. The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html, March 2019. Essay.

Sebastijan Veselic, Nour Mohsen, Lennart Luettgau, Elena Gutierrez, Maria K Eckstein, Steven W Kennerley, Timothy EJ Behrens, Timothy H Muller, and Zeb Kurth-Nelson. Reasoning with programs in replay. *bioRxiv*, pages 2025–10, 2025.

Bingyang Wang, Yijiang Li, Qingyang Zhou, Hui Yi Leong, Tianwei Zhao, Letian Ye, Hokin Deng, Dezhi Luo, and Nuno Vasconcelos. Do vision-language models infer human intention without visual perspective-taking? towards a scalable "one-image-probe-all" dataset. In *Proceedings of the ICML 2025 Workshop on Assessing World Models*, 2025. ICML 2025 World Models Workshop Paper.

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. *arXiv preprint arXiv:2201.11903*, 2022.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328. v2, 29 Sep 2025.

Ludwig Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 1958.

Shuyu Wu, Ziqiao Ma, Xiaoxi Luo, Yidong Huang, Josue Torres-Fonseca, Freda Shi, and Joyce Chai. The mechanistic emergence of symbol grounding in language models. *arXiv preprint arXiv:2510.13796*, 2025a.

Yiwei Wu, Atticus Geiger, and Raphaël Millière. How do transformers learn variable binding in symbolic programs? *arXiv preprint arXiv:2505.20896*, 2025b.

Zhibing Xiao, Xiongfei Wang, Jinbo Zhang, Jianxin Ou, Li He, Yukun Qu, Xiangyu Hu, Timothy EJ Behrens, and Yunzhe Liu. Human hippocampal ripples align new experiences with a grid-like schema. *Neuron*, 2025.

Qi Yan and Jiahe Liu. A review of video evaluation metrics. https://qiyan98.github.io/blog/2024/fvmd-1/, 2024. Blog post reviewing FVD, KVD and alternative metrics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation, 2024. Evaluates time-lapse and metamorphic robustness of T2V models.

Zory Zhang, Pinyuan Feng, Bingyang Wang, Tianwei Zhao, Suyang Yu, Qingying Gao, Hokin Deng, Ziqiao Ma, Yijiang Li, and Dezhi Luo. Can vision language models infer human gaze direction? a controlled study. *arXiv preprint arXiv:2506.05412*, 2025.

A Detailed Task Generation

Chess Reasoning - Tactical Pattern Generation

Our chess module implements a self-contained mate-in-1 generator producing 150+ verified positions without external dependencies. The system employs multiple strategic templates:

Back-Rank Mate Generation: Systematic enumeration of king positions (8 files), pawn structures (20+ configurations), and attacking piece placements. Algorithm generates all combinations of king positions k7, 1k6, 2k5... with corresponding pawn barriers ppp5, 1ppp4, 2ppp3... and attacking pieces R6K, Q6K, 1R5K producing 50+ unique back-rank scenarios.

Queen Corner Patterns: Algorithmic generation of Queen+King endgames with enemy king in corners. System tests 10 queen positions × 4 corner king placements × 5 white king support positions, validating mate moves Qa8#, Qb8#, Qc8# through chess engine verification.

Tactical Combination Templates: Knight mates, rook endgames, and piece coordination patterns with automated validation. Each position undergoes legal move verification and mate confirmation through the python-chess library.

Position Validation Pipeline: All generated positions pass through rigorous validation: (1) FEN string correctness, (2) legal position verification, (3) mate-in-1 confirmation, (4) uniqueness checking via position hashing, (5) visual rendering validation.

Maze Navigation - Graph-Theoretic Generation

Maze creation employs Kruskal's minimum spanning tree algorithm ensuring unique solution paths:

Grid Generation: 3×3 lattice graphs with systematic edge enumeration and random weight assignment. Kruskal's algorithm produces maze connectivity guaranteeing single solution paths between arbitrary start/end positions.

Professional Rendering: Custom matplotlib-based visualization with precise coordinate systems, consistent styling, and high-resolution output (832×480 pixels, 100 DPI). Green circle markers indicate current position, red flag markers show goals.

Difficulty Scaling: Grid size limitations (3×3 only) ensure consistent complexity while maintaining visual clarity for video model processing. Path length distribution analysis ensures balanced difficulty across generated instances.

Raven Progressive Matrices - Rule-Based Pattern Synthesis

Our RPM generator implements systematic pattern creation following cognitive psychology principles:

Rule Categories:

- Shape Progression: Geometric transformations (triangle→square→circle)
- Number Progression: Quantity changes (1→2→3 objects)
- Rotation Patterns: Angular transformations (0°→90°→180°)
- Color Sequences: Hue progressions (red→blue→green)
- Combination Rules: Multiple simultaneous pattern types

Matrix Construction: 3×3 grids with systematic rule application ensuring logical consistency. Bottom-right cell removal creates completion tasks with unique solutions determinable through pattern extrapolation.

Visual Rendering: Custom PIL-based graphics with 150×150 pixel tiles, consistent styling, and clear pattern visibility. Total matrix size 450×450 pixels optimized for model input requirements.

3D Mental Rotation - Voxel Structure Generation

Spatial reasoning tasks employ sophisticated 3D structure generation with camera manipulation:

Voxel Snake Algorithm: Recursive 3D path generation creating connected cube structures. Algorithm parameters: N=8-9 cubes, segment lengths Lmin=2 to Lmax=5, branching probability p_branch=0.2, maximum degree constraints preventing overcrowding.

Spatial Validation: Generated structures must span all three coordinate axes (x,y,z) ensuring genuine 3D complexity. Anti-symmetry checks prevent rotationally equivalent structures reducing task difficulty.

Camera System: Professional 3D rendering with controlled viewpoints:

- Elevation angles: 20-40° (consistent tilt for 3D visibility)
- Azimuth rotations: Exactly 180° horizontal transitions
- Perspective projection with equal aspect ratios
- Consistent lighting and material properties

Rendering Pipeline: Matplotlib 3D with Poly3DCollection faces, consistent coloring (#7070b0), proper edge definition, and 400×400 pixel output resolution.

Sudoku Logical Reasoning - Constraint Satisfaction Implementation

3×3 Sudoku generation employs complete enumeration of valid Latin squares:

Solution Space: Pre-computed catalog of all 12 distinct 3×3 Latin square solutions ensuring mathematical completeness. Random selection provides solution diversity while guaranteeing validity.

Puzzle Construction: Systematic number removal (exactly 1 digit) creating unique completion challenges. Difficulty consistency maintained through uniform removal patterns across all generated instances.

Visual Presentation: Clean grid rendering with matplotlib patches, consistent typography (24pt bold), and clear empty cell indication through light gray backgrounds.

B Failure Modes

B.1 Solving Chess

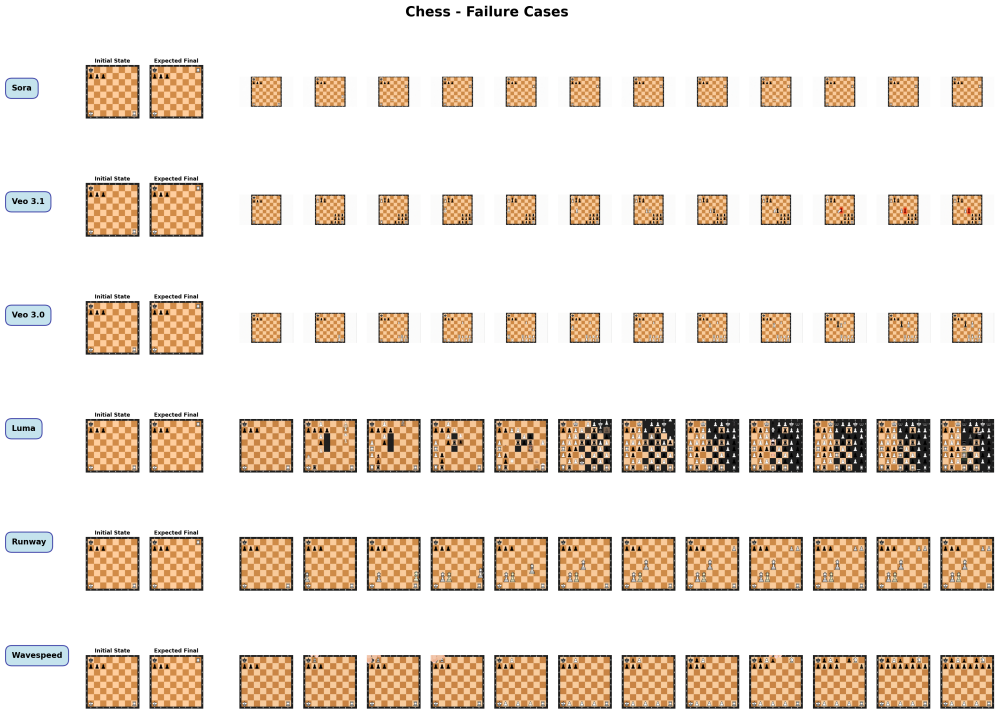

Figure 12: Failure Modes on Chess

B.2 Solving Maze

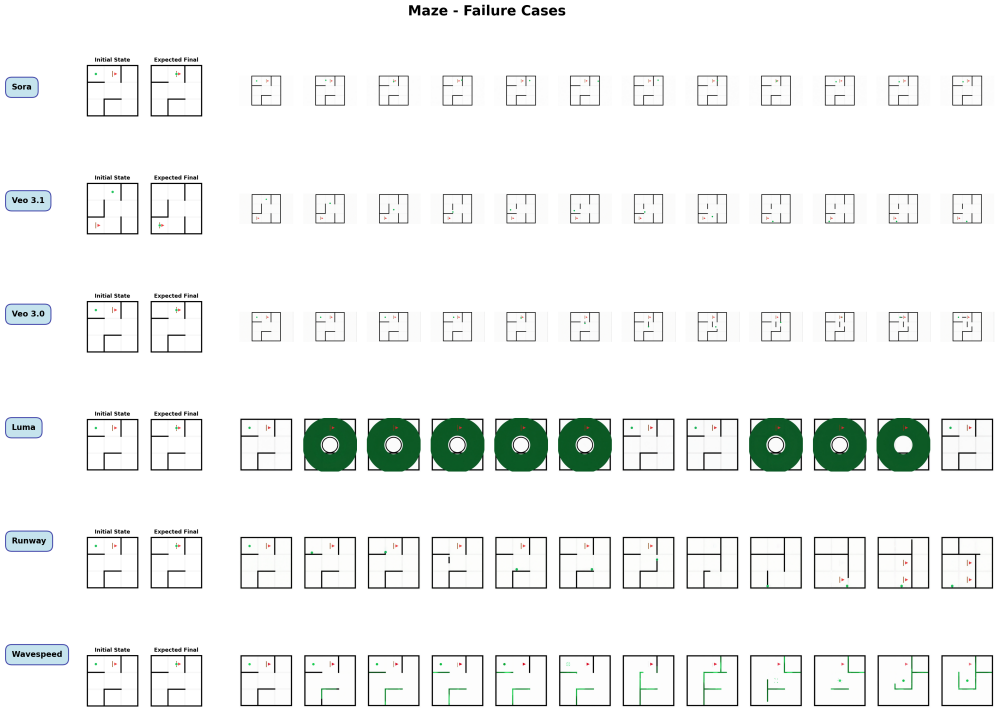

Figure 13: Failure Modes on Maze

B.3 Solving Sudoku

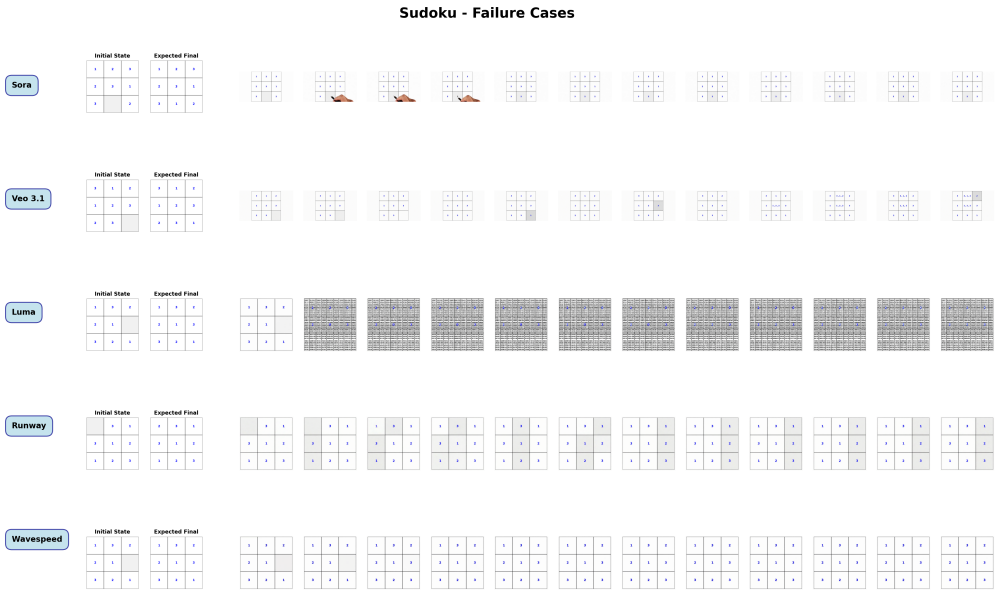

Figure 14: Failure Modes on Sudoku

B.4 Solving Mental Rotation

Figure 15: Failure Modes on Mental Rotation

B.5 Solving Raven's Matrices

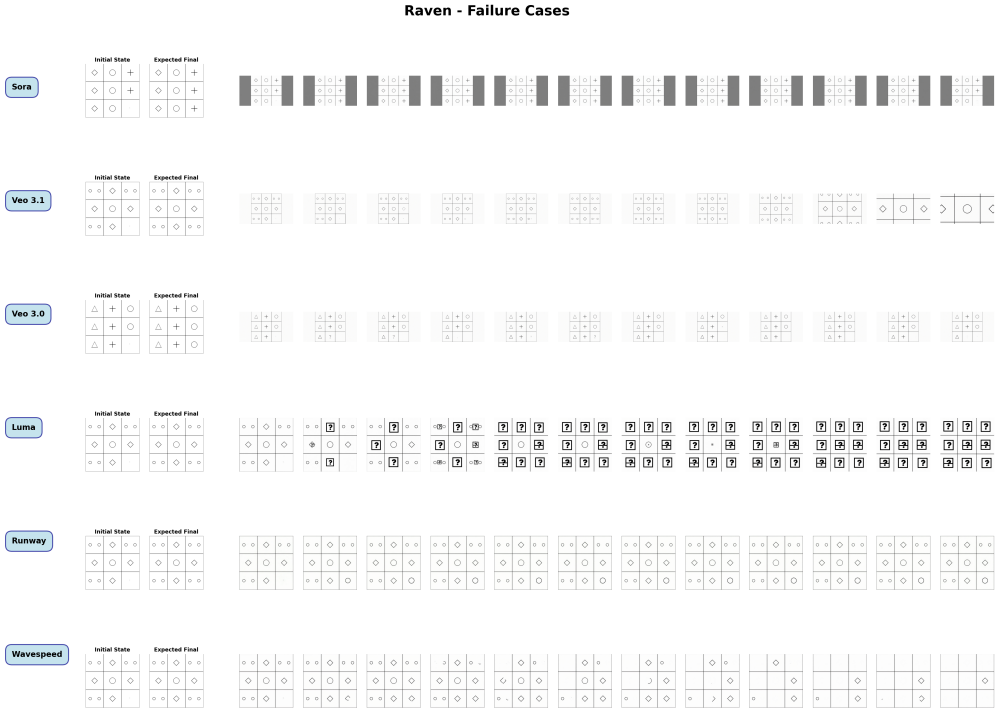

Figure 16: Failure Modes on Raven's Matrices