MORSE: A SUITE OF PROGRAMMATICALLY CONTROLLABLE MULTIMODAL REASONING ENVIRONMENTS WITH STEERABLE DIFFICULTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid progress in vision language models, current multimodal reasoning *development pipelines* are limited by static training imagery, narrow task diversity, and benchmark saturation. We present **MORSE** ("Multimodal Reasoning Suite"), a programmatically controlled collection of *video* reasoning environments with *steerable difficulty* and *verifiable reasoning traces and answers*. The suite comprises: (i) **MORSE-∞**, a simulator that produces unlimited, difficulty steerable instances with reasoning traces; (ii) **MORSE-500**, a curated benchmark of 500 challenging videos covering six complementary reasoning categories and designed to retain headroom as models improve; and (iii) **MORSE-Agent**, which automates generation and curation to reduce human effort over time. Instances are produced via deterministic Python scripts (Manim, Matplotlib, MoviePy), generative video models, and curated real footage, exposing explicit controls over visual complexity, distractors, and temporal span. On **MORSE-500**, the strongest state of the art system achieves 23.6%, with a large gap to human performance at 55.4%, highlighting persistent deficits in abstract and planning categories. We release code, data, seeds, and the evaluation harness to support transparent, reproducible, and forward looking multimodal reasoning research.

1 INTRODUCTION

Progress in multimodal reasoning has been accelerated by ever-larger vision–language models (Bai et al., 2023; Hurst et al., 2024) and broader pretraining corpora (Li et al., 2025). Yet further advances are increasingly limited by the rigidity of available resources. Training corpora are static and rarely capture the temporal and interactive structure of the real world. Evaluation benchmarks quickly saturate, offering little room to probe new capabilities or expose failure modes (Lu et al., 2024). A sustainable path forward requires environments that can generate unlimited diverse problems, expose controllable difficulty, and provide verifiable reasoning traces that allow precise diagnostic analysis.

Existing resources fall short along several axes. First, most datasets focus on static images and short cues (Yue et al., 2023), underrepresenting temporal reasoning and multi-step dependencies. Second, practitioners lack access to scalable video-based resources with explicit difficulty control, making it hard to analyze model performance under varying levels of challenge. Third, current benchmarks generally do not provide ground-truth reasoning chains, leaving failure analysis and interpretability limited. Finally, evaluation sets quickly saturate, with little capacity to scale alongside improving models (Mirzadeh et al., 2024).

We introduce **MORSE** ("Multimodal Reasoning Suite"), a programmatically controllable collection of video reasoning environments with *steerable difficulty* and *verifiable reasoning traces and answers*. This suite contains three components:

- **MORSE-∞**, a data simulator that produces unlimited, difficulty steerable instances with reasoning traces and validated answers. This enables systematic analysis of model performance as task difficulty increases, and provides a rich source of verifiable reasoning chains that can be used for supervision or diagnostic evaluation.

- **MORSE-500**, a curated benchmark of 500 videos sampled from MORSE-∞, spanning six complementary reasoning categories. It preserves headroom through controlled variation and seed based regeneration.
- **MORSE-Agent**, an agentic framework for automatically authoring new video generators. The agent proposes ideas, drafts Python code, receives structured feedback from a vision–language model critic, and iteratively refines its implementation until a functional generator emerges.

Empirical findings and contributions. Evaluation of leading closed and open-source systems on MORSE-500 reveals substantial capability gaps, with the strongest models achieving only 23.6% accuracy compared to 55.4% human performance. Deficits are particularly pronounced in abstract reasoning (23.4% vs 37.5%) and planning tasks (5.0% vs 56.0%), indicating fundamental limitations in compositional reasoning and temporal integration.

This work establishes four foundational contributions to multimodal reasoning research:

- **Unified programmatic ecosystem**: The first integrated framework providing systematic control over complexity dimensions across training generation, rigorous evaluation, and autonomous curation, with deterministic scripts enabling exact reproducibility and comprehensive trace supervision.
- **Scalable data generation**: MORSE-∞ delivers unlimited, difficulty-steerable instances with executable reasoning supervision, exposing explicit controls for visual complexity, distractor density, and temporal dynamics across six reasoning categories to support targeted curriculum design.
- **Challenging video reasoning benchmark**: MORSE-500 provides balanced assessment across temporal reasoning domains while preserving evaluation headroom through systematic difficulty scaling, with all generation scripts and seeds released for exact regeneration and extension.
- **Autonomous ecosystem evolution**: MORSE-Agent automates the complete development lifecycle from generation through difficulty calibration to adaptive curation, implementing a self-improving data ecosystem where enhanced model capabilities drive more sophisticated content generation.

2 MORSE ∞: Infinite Video Generation with Difficulty Control

MORSE-∞ is a generator of unlimited video reasoning problems with controlled difficulty and explicit step-by-step solutions. Unlike static datasets that quickly become outdated, MORSE- can continually produce new instances parameterized by complexity, temporal span, and distractor density. For each generated video, the underlying codebase also produces the ground-truth answer and a reasoning trace: a structured explanation of how the problem is solved, derived from the code itself.

2.1 Programmatic Generation Architecture

The generation pipeline synthesizes instances through deterministic Python scripts that define complete world models, render video sequences, and emit verified reasoning chains. Figure 2 illustrates the four-stage process: (1) parameter sampling from controlled difficulty distributions, (2) deterministic world state generation using fixed random seeds, (3) video rendering through Manim/Matplotlib with temporal coherence constraints, and (4) executable reasoning trace generation with comprehensive unit test validation. Each generator encodes domain-specific physics, maintains object persistence across frames, and ensures causal consistency between visual events and reasoning requirements. The modular architecture enables independent control over visual complexity (object count, occlusion patterns), temporal structure (sequence length, event timing), and reasoning depth (inference steps, compositional requirements) while maintaining semantic coherence.

2.2 Reasoning Traces from Code

Since every task is generated from executable code, the complete solution is known by construction. We implement templates that automatically translate code operations into step-by-step natural language explanations. For example, a pathfinding generator produces both the optimal path and a detailed reasoning trace. Figure 2 illustrates this process with a frozen lake navigation example (visualized in Figure 1 top row). The system generates the environment parameters, executes the

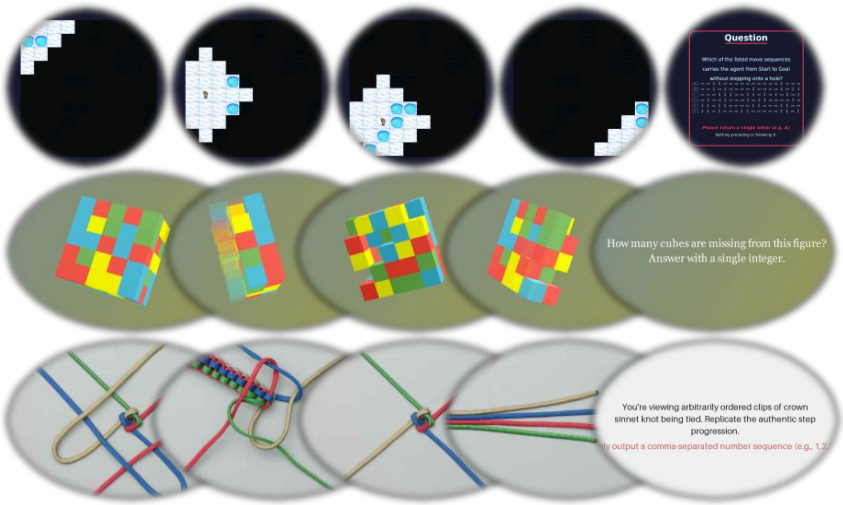


Figure 1: Examples from MORSE. Each row shows sampled frames (visualized as circles to reduce overlapping) from a different video task, ranging from maze navigation to rope tying sequence understanding.

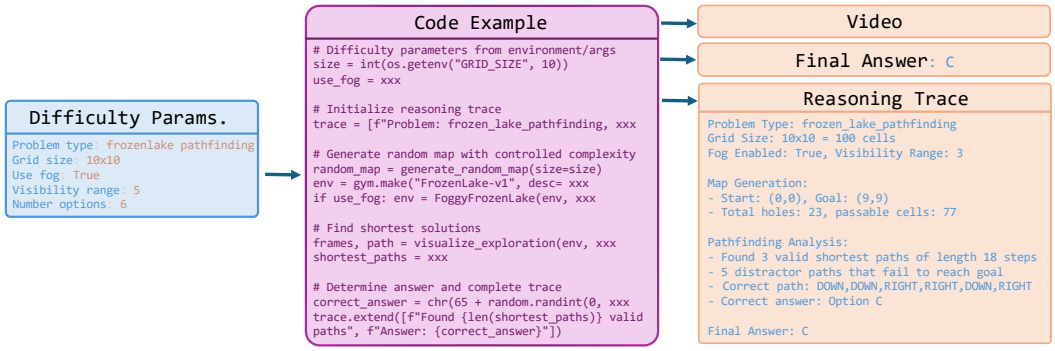


Figure 2: MORSE-∞ generation pipeline showing video synthesis with difficulty control, and reasoning trace generation with validated answer.

pathfinding algorithm, and simultaneously constructs a reasoning trace that explains each decision step.

The reasoning traces include intermediate states, transformation operations, and validation checks that verify consistency at each step. This approach eliminates human annotation errors and enables trace-level supervision during training, providing richer learning signals than traditional input-output pairs (Zelikman et al., 2022; Wang et al., 2022).

2.3 Systematic Difficulty Control

Each reasoning category incorporates systematic difficulty variation through programmatically controlled parameters. Complexity is measured along five orthogonal dimensions: (1) entity complexity (2-15+ objects), (2) reasoning depth (1-5+ inference steps), (3) distractor density (minimal to 8+ irrelevant elements), (4) temporal complexity (static to dynamic concurrent processes), and (5) visual complexity (occlusion patterns, perspective changes, rendering variations).

Figure 4 demonstrates this systematic control across two task categories. As difficulty increases from left to right, performance consistently declines across all evaluated models, validating our parameterization. This controlled degradation enables us to sample appropriately challenging instances for MORSE-500.

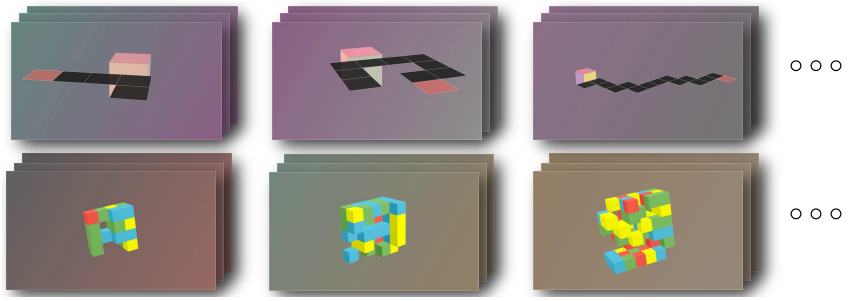
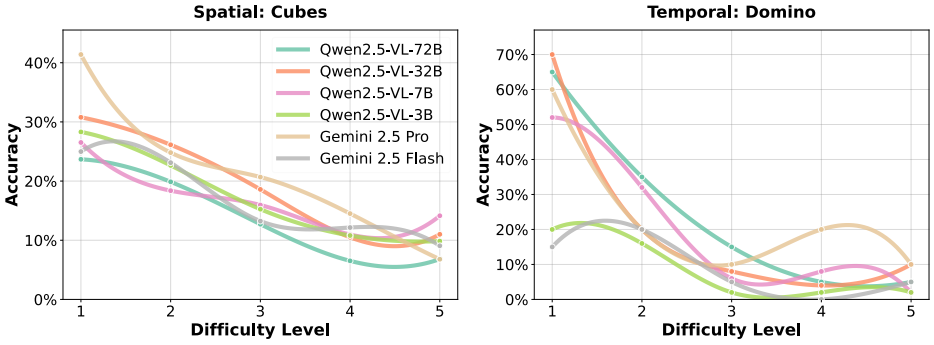


Figure 3: Programmatic difficulty control in MORSE. Each row demonstrates a series of videos with increasing complexity from left to right. For example the number of cube rotations and the size of the cubes controls the reasoning depth and visual complexity.



(a) Spatial Cube Problems

(b) Temporal Domino Problems

Figure 4: Model performance decreases systematically with increasing task difficulty across reasoning categories.

The programmatic foundation enables systematic expansion: new instances can be generated with specified difficulty profiles, and category-specific stress tests can be developed as model capabilities advance.

2.4 Multi-Domain Reasoning Coverage

MORSE-∞ generates data across six complementary reasoning categories, each with specialized generators that maintain domain-specific constraints while respecting unified difficulty parameterization. **Mathematical reasoning** involves geometric transformations, algebraic relationships, and quantitative analysis with controlled complexity scaling. **Abstract reasoning** creates pattern recognition tasks based on ARC-AGI principles with systematic control over pattern complexity and transformation depth. **Spatial reasoning** synthesizes 3D transformation tasks, perspective changes, and spatial relationship problems with parameterized geometric complexity. **Temporal reasoning** produces sequence understanding tasks with controlled temporal dependencies and event complexity. **Physical reasoning** generates intuitive physics scenarios with systematic control over physical complexity and causal chain length. **Planning reasoning** creates multi-step goal-directed tasks with parameterized planning horizon and environmental complexity. Each domain generator maintains semantic coherence while ensuring systematic curriculum progression across all reasoning types, with domain-specific validation to ensure generated instances remain meaningful and solvable.

3 MORSE-500: Evaluating Multimodal Reasoning Limits

Based on MORSE ∞, we sample 500 challenging problems to test the limits of current AI capable of multi-modal reasoning, and release this test set as a benchmark.

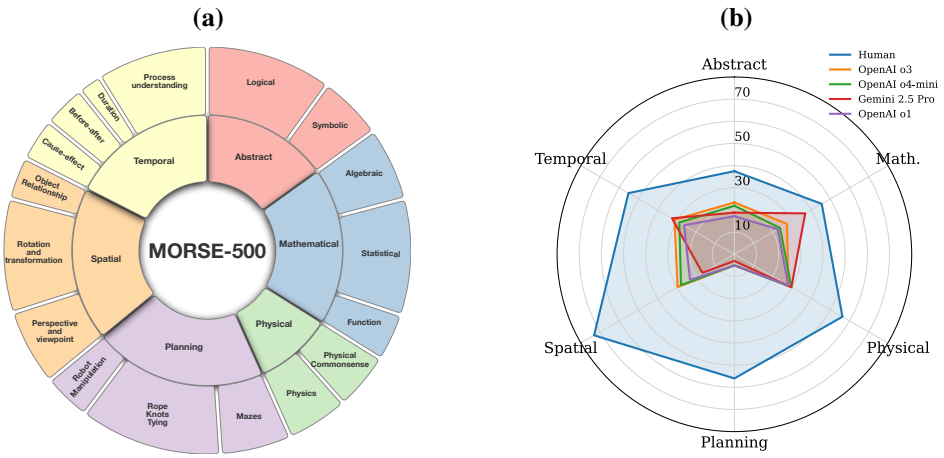


Figure 5: **(a)** Task distribution of MORSE. **(b)** Performance of the best-performing models on MORSE.

3.1 Evaluation Settings

Models and Baselines. We evaluate a diverse set of vision-language models spanning both proprietary and open-source architectures. Closed-source models include Gemini-2.5-Pro (Google, 2025), Gemini-2.5-Flash, Gemini-2.0-Flash variants, and Gemini-1.5-Pro from Google DeepMind, alongside OpenAI's o3 (OpenAI, 2025), GPT-4o (Hurst et al., 2024), o1 (Jaech et al., 2024), and o4-mini (OpenAI, 2025). Open-source models encompass the Qwen2.5-VL family (Bai et al., 2023) (3B, 7B, 32B, 72B) with and without quantization (AWQ), Qwen2.5-Omni-7B (Xu et al., 2025), LLaVA-NeXT-Video-7B (liu, 2024), MiniCPM-o-2_6 (Yao et al., 2024), InternVL3-8B (Zhu et al., 2025), and Gemma-3 (Gemma Team et al., 2025).

Evaluation Protocol and Metrics. Models with native video support received full video clips, while image-only models used frames sampled at 2fps with maximum 32-frame context. All models received the minimal prompt: "`Answer the question in this video.`" with no few-shot examples to isolate intrinsic reasoning abilities. For models without video API support (OpenAI models), we provided downsampled image frames (512px maximum side length). We report accuracy as the primary metric, extracting answers using LLM-based parsing and string matching following MathVista (Lu et al., 2024).

3.2 Quantitative Results

Table 1 presents accuracy across reasoning categories for all evaluated models. Overall performance remains substantially below human-level capabilities, with even the strongest systems averaging below 25% accuracy—a significant gap compared to human performance at 55.4%.

Proprietary Model Performance. Among proprietary models, OpenAI's o3 achieves the highest overall score of 23.6%, demonstrating relatively balanced performance across categories with particular strength in temporal reasoning (31.2%). Gemini-2.5-Pro follows closely at 21.8%, exhibiting notable proficiency in mathematical reasoning (36.9%) and temporal understanding (32.5%), but struggling significantly with abstract reasoning (18.8%) and planning tasks (3.0%). Interestingly, while Gemini-2.5-Flash achieves similar mathematical performance (35.7%), it shows markedly weaker performance in abstract reasoning (9.4%) and planning (1.0%), suggesting that model scale and optimization strategies significantly impact reasoning capabilities across different cognitive domains.

The performance patterns reveal interesting trade-offs: models optimized for mathematical reasoning tend to excel in structured, rule-based tasks but struggle with open-ended abstract reasoning. Conversely, models with stronger general reasoning capabilities (like o3) show more balanced performance but may sacrifice peak performance in specific domains.

Table 1: Accuracy (%) on MORSE across all six reasoning categories and overall average. Models are organized by closed-source and open-source categories.

Model	All	Abstract	Math.	Physical	Planning	Spatial	Temporal
Human	55.4	37.5	45.5	56.3	56.0	73.1	55.2
Closed-source Models							
o3	23.6	23.4	27.4	28.1	5.0	29.6	31.2
o4-mini	22.2	21.9	23.8	29.7	5.0	27.8	28.7
Gemini-2.5-Pro	21.8	18.8	36.9	29.7	3.0	16.7	32.5
o1	19.8	17.2	22.6	28.1	5.0	23.1	26.2
Gemini-2.5-Flash	19.2	9.4	35.7	28.1	1.0	24.1	18.8
Gemini-1.5-Pro	18.8	12.5	21.4	26.6	1.0	26.9	26.2
GPT-4o	17.4	17.2	20.2	34.4	4.0	12.0	25.0
Gemini-2.0-Flash	16.0	12.5	29.8	28.1	0.0	13.0	18.8
Gemini-2.0-Flash-Lite	14.2	17.2	21.4	21.9	2.0	14.8	12.5
Open-source Models							
Qwen2.5-VL-72B	17.8	6.2	21.4	34.4	1.0	22.2	25.0
Qwen2.5-VL-32B-AWQ	16.8	14.1	23.8	34.4	1.0	15.7	18.8
Qwen2.5-VL-72B-AWQ	16.4	12.5	11.9	29.7	2.0	27.8	16.2
Qwen2.5-VL-32B	15.6	9.4	19.0	29.7	2.0	16.7	21.2
Gemma-3-27b	14.6	20.3	20.2	25.0	1.0	13.0	15.0
MiniCPM-o-2.6	11.6	4.7	10.7	23.4	1.0	16.7	15.0
Qwen2.5-Omni-7B	11.4	6.2	9.5	21.9	2.0	15.7	15.0
Qwen2.5-VL-7B	11.2	7.8	11.9	25.0	2.0	12.0	12.5
InternVL3-8B	7.8	6.2	6.0	14.1	1.0	11.1	10.0
Qwen2.5-VL-3B	7.6	9.4	3.6	18.8	1.0	9.3	7.5
LLaVA-NeXT-Video-7B	5.0	1.6	11.9	6.2	0.0	5.6	5.0
Chat-UniVi-7B	1.0	1.6	1.2	0.0	2.0	0.0	1.2

Open-Source Model Analysis. The open-source landscape demonstrates a clear scaling relationship between model size and reasoning performance. Among the Qwen2.5-VL family, the 72B model achieves 17.8% overall accuracy, substantially outperforming smaller variants (32B: 16.8%, 7B: 11.2%, 3B: 7.6%). However, quantization effects are mixed: while the 72B-AWQ model shows slightly lower overall performance (16.4%) compared to its full-precision counterpart, the 32B-AWQ variant actually outperforms the standard 32B model (16.8% vs. 15.6%), suggesting that quantization impacts vary with model scale.

Specialized models show domain-specific strengths: MiniCPM-o-2.6, despite lower overall performance (11.6%), demonstrates competitive physical reasoning capabilities (23.4%), indicating that targeted optimization can yield focused improvements. However, highly specialized models like LLaVA-NeXT-Video-7B, despite being designed for video understanding, achieve only 5.0% overall accuracy, highlighting the gap between video comprehension and video-based reasoning.

Category-Specific Performance Patterns. Performance varies dramatically across reasoning categories, revealing systematic weaknesses in current models. Mathematical reasoning shows the highest performance, with several systems exceeding 20% accuracy due to the structured nature of mathematical problems and their prevalence in training data. Physical reasoning demonstrates moderate competency (20-35% for top performers), suggesting intuitive physics concepts are partially captured in training paradigms, though the gap from human performance (56.3%) remains substantial. Spatial and temporal reasoning show moderate but inconsistent performance across models, with some displaying surprising deficits (e.g., Gemini-2.5-Pro's 16.7% spatial accuracy despite strong mathematical performance). Abstract reasoning proves most challenging, with even the best performers struggling to exceed 25% accuracy, suggesting fundamental limitations in pattern recognition, analogical thinking, and rule induction—core components of general intelligence. Most concerning, planning tasks show near-random performance across all models (0-5% accuracy), indicating critical gaps in multi-step reasoning and goal-directed behavior with significant implications for real-world deployment in autonomous systems.

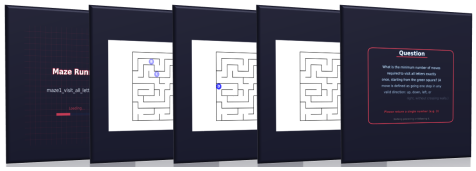
Implications and Model Limitations. The uniformly low performance across all reasoning categories, particularly in abstract reasoning and planning, suggests that current multimodal models suffer from fundamental architectural limitations rather than mere training inefficiencies. The inability to perform multi-step reasoning, maintain temporal coherence, and engage in abstract pattern matching indicates that these models may be primarily engaging in sophisticated pattern matching rather than genuine reasoning.

Furthermore, the substantial human-model performance gaps (30+ percentage points in most categories) underscore that achieving human-level multimodal reasoning remains a significant challenge. The particularly poor performance on planning tasks raises questions about the suitability of current models for autonomous decision-making applications.

3.3 Diagnostic Analysis of Model Limitations

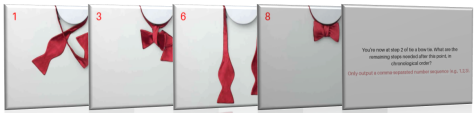
MORSE-500 enables systematic diagnosis of current model limitations across reasoning domains. We analyze performance on particularly challenging task categories to identify specific failure modes.

Figure 6 illustrates four demanding task types: Mazes require optimal pathfinding under partial visibility, testing temporal reasoning and spatial memory. Rope Knots involve sequence reconstruction from randomized tying steps with visual transformations. Physical Commonsense tasks distinguish realistic from AI-generated videos, probing physical intuition. ARC-AGI-2 adaptations test abstract pattern completion and counting under complex visual transformations.



(a) Planning Reasoning - Maze with Different Endpoints

(b) Physical Commonsense - Tennis Ball Rolling



(c) Planning Reasoning - Knots of Tying a Bow Tie

(d) Abstract Reasoning - ARC-AGI-2

Figure 6: Example videos for challenging tasks: Mazes, Rope Knots, Physical Commonsense, and ARC-AGI-2.

Table 2 reveals systematic limitations across all model categories. While humans achieve substantial performance (47.1-63.6% across tasks), even frontier models struggle dramatically: o3 reaches only 3.8-14.0%, and Gemini 2.5 Pro scores near-zero (0.0-4.2%). This performance gap highlights critical weaknesses in temporal reasoning, physical intuition, and abstract pattern recognition—fundamental capabilities for robust multimodal reasoning.

Model	Mazes	Rope Knots	Physical Commonsense	ARC-AGI-2
Human	58.3	53.8	63.6	47.1
o3	10.0	3.8	13.6	14.0
Gemini 2.5 Pro	0.0	1.3	4.2	0.0

Table 2: Performance of top frontier models across 4 challenging tasks.

These diagnostic results demonstrate MORSE-500's value in identifying specific architectural and training limitations, providing clear targets for future model development across reasoning categories.

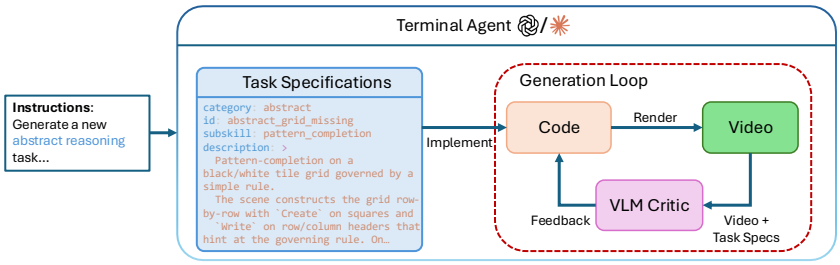


Figure 7: Workflow of MORSE-Agent. The agent proposes an idea, drafts code, receives feedback from a VLM critic, and refines its code iteratively until a stable generator is produced.

4 MORSE AGENT: AUTONOMOUS PROGRAM GENERATION

While MORSE-∞ provides an infinite stream of video reasoning tasks once generators are authored, building those generators is itself non-trivial. Each generator must define sampling rules, render coherent video scenes, and produce corresponding solution traces. To scale beyond hand-written scripts, we introduce MORSE-Agent, an agentic framework for automatically writing new video generators.

4.1 AGENTIC CODE GENERATION LOOP

MORSE-Agent is a terminal agent that operates in an iterative workflow inspired by program synthesis with feedback (Austin et al., 2021; Ellis et al., 2023):

1. **Idea Proposal** – The agent draws on the six-category taxonomy as scaffolding for proposing new generator ideas. The taxonomy is provided as input guidance, and example code serve as demonstrations.
2. **Initial Implementation** – Using example code snippets and templates, the agent drafts a Python program that implements the proposed generator.
3. **VLM Critic Feedback** – The draft is executed, and the resulting video along with the code are passed to a vision–language model critic. The critic evaluates correctness, coherence, and alignment with the intended reasoning category, and provides structured feedback. Here we use Gemini 2.5 Pro as the critic.
4. **Refinement** – The agent revises its code based on the critic's feedback, aiming to fix bugs, improve clarity, or better match the task description.
5. **Iteration** – Steps 2–4 repeat for several rounds until the generator consistently produces valid videos with correct answers and solution traces.

4.2 GENERATED TASK EXAMPLES

Figure 8 demonstrates MORSE-Agent's iterative refinement process on an abstract reasoning task, showing how the system identifies and corrects critical issues like premature answer revelation. MORSE-Agent reduces the manual burden of generator authoring and enables rapid expansion of the MORSE suite. By leveraging multimodal feedback, it creates generators that are both syntactically correct and semantically meaningful. However, its effectiveness depends on the quality of the VLM critic, and it may require several iterations to converge to a valid generator. MORSE-Agent should be viewed as an *enabler* of MORSE-∞, bootstrapping new generators efficiently.

5 RELATED WORK

Controllable Complexity Evaluation. Recent work reveals systematic limitations in reasoning models through controlled complexity manipulation. GSM-Symbolic (Mirzadeh et al., 2024) shows that LLMs exhibit significant variance when mathematical problem templates are systematically

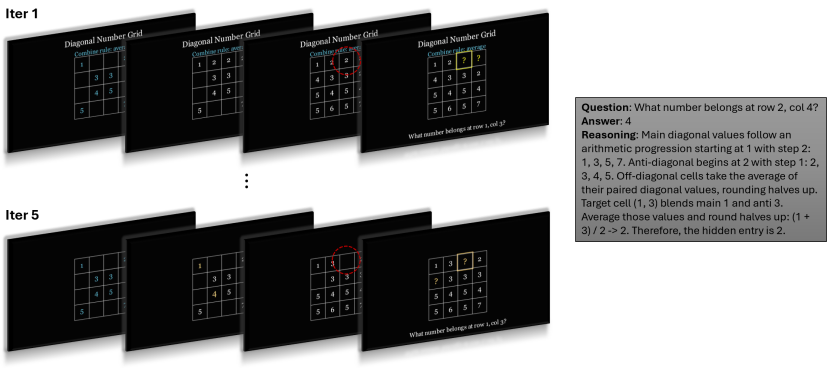


Figure 8: Example generation process for an abstract reasoning task. Code refinement is able to fix a critical issue with the answer being revealed in the middle of the video.

varied, exposing brittleness in seemingly robust capabilities. Shojaee et al.(Shojaee et al., 2025) use controllable puzzle environments to demonstrate "accuracy collapse" in Large Reasoning Models beyond specific complexity thresholds. OMEGA(Sun et al., 2025) evaluates three generalization axes: exploratory (complex instances within domains), compositional (combining distinct skills), and transformative (novel strategies). Game-RL (Tong et al., 2025) synthesizes verifiable game tasks with controllable difficulty across 30 games, demonstrating out-of-domain generalization from game-specific training. These approaches establish the foundation for our programmatic difficulty control in multimodal video reasoning.

Multimodal Reasoning Benchmarks. Existing benchmarks fall into two categories with complementary limitations. Static image benchmarks like MathVista (Lu et al., 2024), MMMU (Yue et al., 2023), and ARC-AGI (Chollet et al., 2025) provide strong reasoning challenges but ignore temporal dynamics and quickly saturate. Video understanding benchmarks like Video-MME (Fu et al., 2024), MVBench (Li et al., 2024), Mementos (Wang et al., 2024b), and (Song et al., 2025) incorporate temporal information but focus primarily on perception and retrieval rather than systematic reasoning evaluation. MORSE-500 bridges this gap by combining the systematic reasoning challenges of static benchmarks with the temporal complexity of video understanding, while providing programmatic difficulty control to prevent saturation. Our approach enables precise modulation of complexity dimensions and generates unlimited instances for continuous evaluation as models improve.

6 Conclusion

We introduce MORSE (Multimodal Reasoning Suite), a programmatically controlled collection of video reasoning environments with three integrated components: MORSE-∞ for unlimited difficulty-steerable instance generation, MORSE-500 as a challenging benchmark with headroom for model improvement, and MORSE-Agent for automated generator creation. Our evaluation reveals substantial gaps in current AI capabilities. On MORSE-500, state-of-the-art systems achieve only 23.6% accuracy versus 55.4% human performance, with severe deficits in abstract reasoning (23.4% vs 37.5%) and planning tasks (5.0% vs 56.0%). These systematic failures indicate fundamental limitations in compositional reasoning and temporal integration. MORSE's programmatic foundation enables precise control over visual complexity, distractors, and temporal dynamics with verifiable ground truth. Unlike static benchmarks that saturate quickly, the suite's scalable difficulty generation ensures continued relevance as models advance, with modular design supporting extension to additional reasoning domains.

Limitations. Current generators emphasize synthetic environments; incorporating naturalistic scenarios while preserving verifiability remains challenging. MORSE-Agent depends on critic quality and may require iterative refinement. **Broader Impact.** MORSE enables targeted analysis of reasoning failures and supports development of more robust multimodal AI systems for real-world applications.

LLM Usage: Language models were used to improve text clarity and readability.

REFERENCES

Llava-next: Open large multimodal models. 2024. URL https://github.com/LLaVA-VL/LLaVA-NeXT.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022. URL https://arxiv.org/abs/2204.14198.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics, 2023. URL https://arxiv.org/abs/2304.08488.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2023. URL https://arxiv.org/abs/2308.12966.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. arXiv preprint arXiv:1606.01540, 2016. URL https://arxiv.org/abs/1606.01540.

John Bissell Carroll. Human cognitive abilities: A survey of factor-analytic studies. Number 1. Cambridge university press, 1993. URL https://www.cambridge.org/core/books/human-cognitive-abilities/F83D5EADF14A453F6350FF3DD39631C8.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. Advances in Neural Information Processing Systems, 37:53168–53197, 2024. URL https://arxiv.org/abs/2411.04998.

François Chollet. On the measure of intelligence. arXiv preprint arXiv:1911.01547, 2019. URL https://arxiv.org/abs/1911.01547.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. arXiv preprint arXiv:2505.11831, 2025. URL https://arxiv.org/abs/2505.11831.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. arXiv preprint arXiv:2501.16411, 2025. URL https://arxiv.org/abs/2501.16411.

Stanislas Dehaene. The number sense: How the mind creates mathematics. OUP USA, 2011.

Kevin Ellis, Lionel Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lore Anaya Pozo, Luke Hewitt, Armando Solar-Lezama, and Joshua B Tenenbaum. Dreamcoder: growing generalizable, interpretable knowledge with wake–sleep bayesian program learning. Philosophical Transactions of the Royal Society A, 381(2251):20220050, 2023.

Jonathan St B. T. Evans and Keith E. Stanovich. Dual-process theories of higher cognition: Advancing the debate. Perspectives on psychological science, 8(3):223–241, 2013. doi: 10.1177/1745691612460685. URL https://doi.org/10.1177/1745691612460685.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075, 2024. URL https://arxiv.org/abs/2405.21075.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025. URL https://arxiv.org/abs/2503.19786.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2): 155–170, 1983. doi: 10.1207/s15516709cog0702_3. URL https://doi.org/10.1207/s15516709cog0702_3.

Anastasis Germanidis. Introducing gen-3 alpha: A new frontier for video generation. https://runwayml.com/research/introducing-gen-3-alpha, 2024. Runway Research blog, Accessed 22 May 2025.

Google. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, 2025. Accessed: 2025-05-16.

Google DeepMind. State-of-the-art video and image generation with veo 2 and imagen 3. https://blog.google/technology/google-labs/video-image-generation-update-december-2024/, 2024. Accessed 22 May 2025.

Grog. Animated knots. https://www.animatedknots.com, 2025. Accessed: 2025-05-16.

Hailuo AI. Transform idea to visual with ai. https://hailuoai.video/, 2025. Accessed 22 May 2025.

Siddhant Haldar, Vaibhav Mathur, Denis Yarats, and Lerrel Pinto. Watch and match: Supercharging imitation with regularized optimal transport, 2023. URL https://arxiv.org/abs/2206.15469.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025), 2025. URL https://arxiv.org/abs/2501.05444.

Keith J. Holyoak and Robert G. Morrison, editors. The Oxford handbook of thinking and reasoning. Oxford University Press, 2013. ISBN 9780199730013. URL https://academic.oup.com/edited-volume/34559.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024. URL https://arxiv.org/abs/2410.21276.

Michael Igorevich Ivanitskiy, Rusheb Shah, Alex F. Spies, Tilman Räuker, Dan Valentine, Can Rager, Lucia Quirke, Chris Mathwin, Guillaume Corlouer, Cecilia Diniz Behn, and Samy Wu Fung. A configurable library for generating and manipulating maze datasets, 2023. URL https://arxiv.org/abs/2309.10498.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024. URL https://arxiv.org/abs/2412.16720.

Kling AI. Kling ai: Next-generation ai creative studio. https://klingai.com/, 2025. Accessed 22 May 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023), volume 202 of Proceedings of Machine Learning Research, pages 19109–19133. PMLR, jul 2023. URL https://arxiv.org/abs/2301.12597.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 22195–22206, 2024. doi: 10.1109/CVPR52733.2024.02049. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. arXiv preprint arXiv:2501.14818, 2025.

Pan Lu, Michel Galley, Hao Cheng, Kai-Wei Chang, and Jianfeng Gao. Learn from science textbooks: Retrieval-augmented science question answering. arXiv preprint arXiv:2207.05275, 2022. URL https://arxiv.org/abs/2207.05275.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In International Conference on Learning Representations (ICLR 2024), 2024. URL https://arxiv.org/abs/2310.02255.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 2200–2209, 2021. doi: 10.1109/WACV48630.2021.00224. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Michael McCloskey, Allyson Washburn, and Linda Felch. Intuitive physics: the straight-down belief and its origin. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9 (4):636–649, 1983. doi: 10.1037/0278-7393.9.4.636. URL https://doi.org/10.1037/0278-7393.9.4.636.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv preprint arXiv:2410.05229, 2024.

Ankush Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 947–952. IEEE, 2019.

Akira Miyake, Naomi P. Friedman, Michael J. Emerson, Alexander H. Witzki, Amy Howerter, and Tor D. Wager. The unity and diversity of executive functions and their contributions to complex ¨frontal lobe¨ tasks: A latent variable analysis. Cognitive psychology, 41(1):49–100, 2000. doi: 10.1006/cogp.1999.0734. URL https://doi.org/10.1006/cogp.1999.0734.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? arXiv preprint arXiv:2501.09038, 2025. URL https://arxiv.org/abs/2501.09038.

OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

OpenAI. Sora is here. https://openai.com/index/sora-is-here/, 2024. Accessed 22 May 2025.

OpenAI. Introducing o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025. Accessed: 2025-05-16.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. arXiv preprint arXiv:2504.16074, 2025. URL https://arxiv.org/abs/2504.16074.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. Advances in Neural Information Processing Systems, 36:38154–38180, 2023. URL https://arxiv.org/abs/2303.17580.

Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. Science, 171(3972):701–703, 1971. doi: 10.1126/science.171.3972.701. URL https://doi.org/10.1126/science.171.3972.701.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv preprint arXiv:2506.06941, 2025.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In Conference on Robot Learning (CoRL 2022), 2023. URL https://proceedings.mlr.press/v205/shridhar23a.html. arXiv preprint year 2022.

Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Textvqa: Towards image-text reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8317–8326, 2019. doi: 10.1109/CVPR.2019.00852. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html. Consolidated entry for TextVQA CVPR paper.

Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. arXiv preprint arXiv:2504.14693, 2025.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization. arXiv preprint arXiv:2506.18880, 2025.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Code2logic: Game-code-driven data synthesis for enhancing vlms general reasoning. arXiv preprint arXiv:2505.13886, 2025.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025. URL https://arxiv.org/abs/2503.20314.

Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play, 2023. URL https://arxiv.org/abs/2302.12422.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems, 37:95095–95169, 2024a. URL https://arxiv.org/abs/2402.14804.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. arXiv preprint arXiv:2401.10529, 2024b. URL https://arxiv.org/abs/2401.10529.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems, 37:28828–28857, 2024. URL https://arxiv.org/abs/2407.15754.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: personalized robot assistance with large language models. Autonomous Robots, 47(8):1087–1102, nov 2023. ISSN 1573-7527. doi: 10.1007/s10514-023-10139-z. URL https://doi.org/10.1007/s10514-023-10139-z.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, Shichao Fan, Xinhua Wang, Fei Liao, Zhen Zhao, Guangyu Li, Zhao Jin, Lecheng Wang, Jilei Mao, Ning Liu, Pei Ren, Qiang Zhang, Yaoxu Lyu, Mengzhen Liu, Jingyang He, Yulin Luo, Zeyu Gao, Chenxuan Li, Chenyang Gu, Yankai Fu, Di Wu, Xingyu Wang, Sixiang Chen, Zhenyu Wang, Pengju An, Siyuan Qian, Shanghang Zhang, and Jian Tang. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation, 2025. URL https://arxiv.org/abs/2412.13877.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, et al. Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. arXiv preprint arXiv:2505.19099, 2025. URL https://arxiv.org/abs/2505.19099.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215, 2025. URL https://arxiv.org/abs/2503.20215.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024. URL https://arxiv.org/abs/2408.01800.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502, 2023. URL https://arxiv.org/abs/2311.16502.

Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event segmentation in perception and memory. Trends in Cognitive Sciences, 11(2):80–86, 2007. doi: 10.1016/j.tics.2006.12.001. URL https://doi.org/10.1016/j.tics.2006.12.001.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. Advances in Neural Information Processing Systems, 35:15476–15488, 2022.

Haotian Zhang, Jingyang Li, Zhiwei Li, Zhang Kanchor, Zhengkai Liu, Chen-Yu Lee, Chunyuan Su, Chunyuan Li, Percy Liang, and Ahmed H. Awadallah. Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023. URL https://arxiv.org/abs/2302.00923.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025. URL https://arxiv.org/abs/2504.10479.

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors, 2023. URL https://arxiv.org/abs/2210.11339.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. arXiv preprint arXiv:2411.00836, 2024. URL https://arxiv.org/abs/2411.00836.