

MMGR: Multi-Modal Generative Reasoning

Benchmark and Evaluation

Zefan Cai^{1,*} Haoyi Qiu^{2,*} Tianyi Ma^{3,*} Haozhe Zhao^{4,*} Gengze Zhou⁵ Tingting Liao⁶ Xinyan Velocity Yu⁷ Kung-Hsiang Huang⁸
Ke Wan⁹ Shawn Lin¹⁰ Parisa Kordjamshidi³ Minjia Zhang⁴ Wen Xiao⁹ Jiuxiang Gu¹¹ Nanyun Peng² Junjie Hu¹

* Equal contribution

¹University of Wisconsin-Madison ²University of California, Los Angeles ³Michigan State University ⁴University of Illinois Urbana-Champaign
⁵University of Adelaide ⁶MBZUAI ⁷University of Southern California ⁸Salesforce AI Research ⁹Microsoft ¹⁰Kuma Stories Ltd ¹¹Adobe Research

EMNLP 2026 · MAIN CONFERENCE

01 Why MMGR? Visual plausibility does not imply valid reasoning

A generation must solve the original task while preserving its rules, state and causal structure.

Abstract reasoning · Preserve rules

Fixed givens and global constraints expose solutions that quietly change the task.

Embodied navigation · Preserve state

The scene, agent and goal must stay consistent across viewpoints and trajectories.

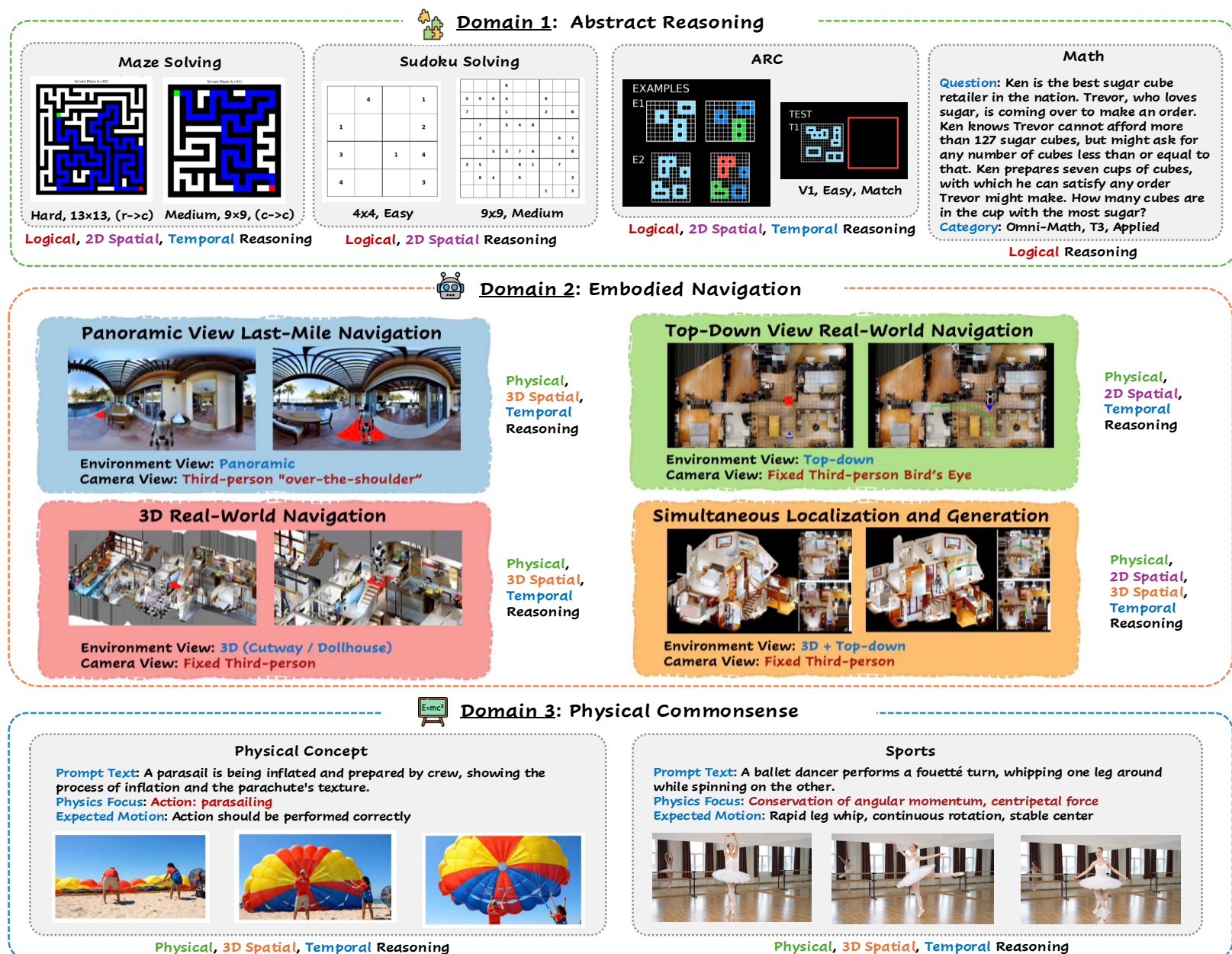
Physical commonsense · Preserve causality

Generated motion must follow valid dynamics, beyond a convincing visual appearance.

02 Benchmark design & evaluation

3 domains · 10 tasks · 1,853 samples

Three complementary reasoning domains · Fig. 2



Task coverage and reasoning abilities · Table 1

Benchmark Task Domain	Phy.	Log.	3D Spa.	2D Spa.	Temp.	# Samples
D1: Abstract Reasoning						
Maze (Section C)	✓			✓	✓	240
Sudoku (Section D)	✓			✓		300
ARC (Section E)	✓			✓	✓	456
Math (Section F)	✓					327
D2: Embodied Navigation						
Last-Mile Nav. (Section H)	✓		✓		✓	120
Top-down View Nav. (Section I)	✓			✓	✓	120
3D R.-W. Nav. (Section J)	✓		✓		✓	120
SLAG (Section K)	✓		✓	✓	✓	120
D3: Physical Commonsense						
Physical Concepts (Section L)	✓		✓		✓	25
Sports (Section L)	✓		✓		✓	25
Total Evaluation Samples						1,853

Task success requires every check to pass.

1 · Integrity

Givens, scene, agent, goal

2 · Outcome

Correct target or answer

3 · Evidence

Video steps / final image

Maze/Sudoku: pixel + OCR · Other tasks: VLM + sampled human checks

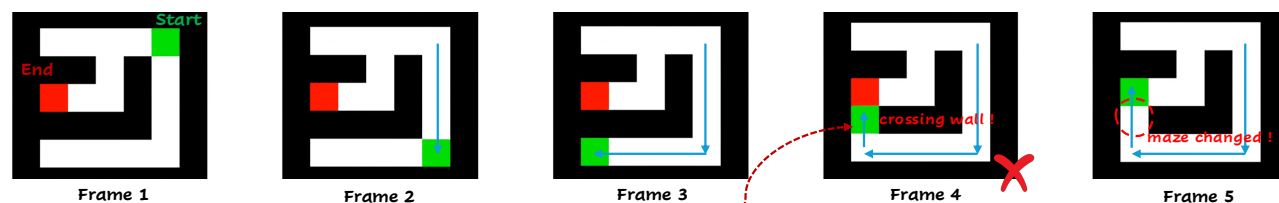
03 Results: symbolic failures, navigation gaps and physical plausibility

Table 4 · Zero-shot task success (%) · Video: process + final state; image: final state · Bold = best · Navigation rankings are evaluator-dependent

Task	Video Models			Image Models					LLM Baselines	
	Veo-3	Sora-2	Wan-2.2	Nano-banana	Nano-banana Pro	GPT-4o-image	GPT-image-1.5	Qwen-image	Gemini-3-Flash	Gemini-3-Pro
D1: Abstract Reasoning										
Maze	3.67	0.00	0.00	3.67	9.17	0.00	0.97	0.14	55.67	59.33
Sudoku	0.00	0.00	0.00	8.60	7.00	1.33	6.78	0.00	65.53	63.75
ARC	4.80	11.67	0.15	8.15	28.07	0.00	12.72	2.15	—	—
Math	10.61	10.56	0.00	11.47	72.69	26.24	25.52	11.22	71.38	74.14
D2: Embodied Navigation										
Last-Mile Nav.	60.00	0.00	14.17	74.17	75.83	0.00	55.84	16.67	—	—
Top-down View Nav.	19.49	3.39	5.09	11.11	33.05	3.39	26.27	5.08	—	—
3D R.-W. Nav.	22.50	0.00	24.17	79.17	85.00	13.33	77.92	38.33	—	—
SLAG	11.02	12.50	0.85	28.79	37.29	16.67	31.36	6.78	—	—
D3: Physical Commonsense										
Physical Concepts	41.67	76.00	26.67	—	—	—	—	—	—	—
Sports	60.00	64.00	21.33	—	—	—	—	—	—	—

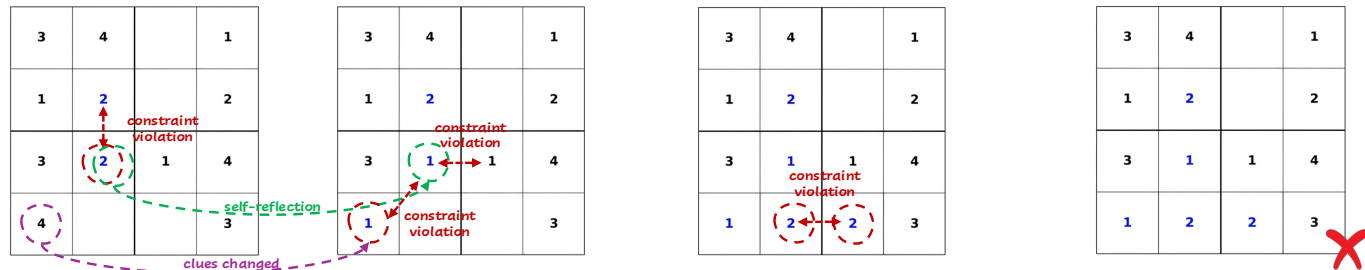
04 Case studies: where the generated solution breaks

D1 · Maze



Walls are crossed and the maze changes, even though the target is reached.

D1 · Sudoku



Repeated revisions still violate constraints and change a given clue.

D2 · 3D navigation



The agent reaches the target after the staircase geometry changes.

D2 · SLAG



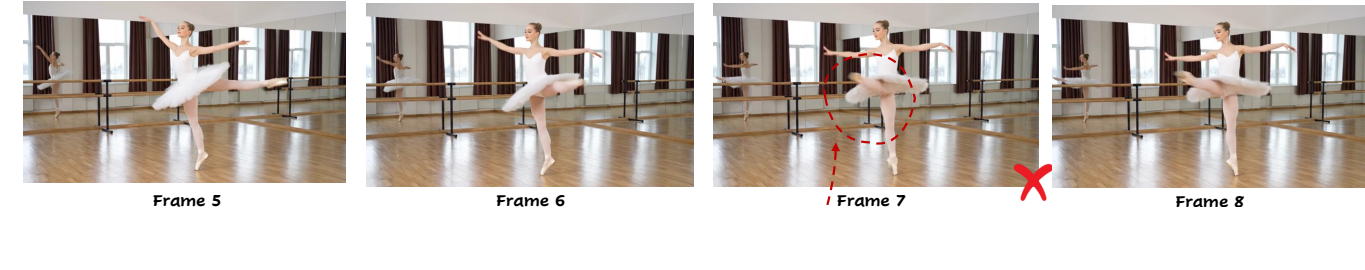
The 3D rollout reaches the goal while the 2D trajectory drifts away.

D3 · Coffee grinding



Beans become fine grounds without showing the required crushing interaction.

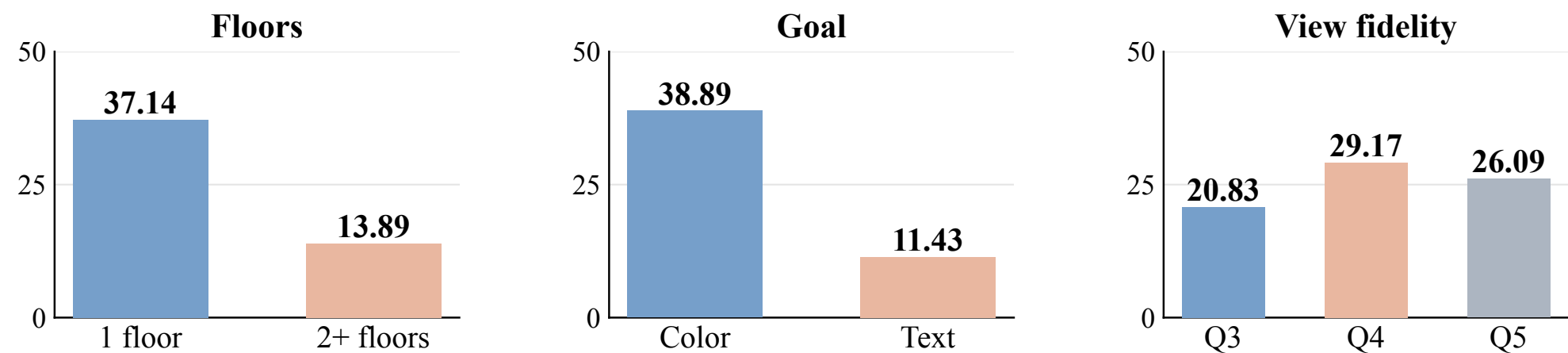
D3 · Ballet rotation



The generated turn distorts the body and misses the required leg-whip dynamics.

05 Ablation diagnostics & evaluation reliability

Setting sensitivity · Veo-3 Top-down overall success (%)



Appendix diagnostic subsets; floor, goal and fidelity groups use their reported settings.

Human-evaluator agreement (%) · By task and model

Task · Evaluator	Veo-3	Sora-2	Wan-2.2	Nano-banana Pro
Maze · Pixel	95.56	100.00	100.00	—
Sudoku · OCR	100.00	100.00	100.00	—
ARC · VLM	94.90	—	—	—
Last-Mile Nav. · VLM	53.50	—	—	30.00
Top-down Nav. · VLM	79.15	—	—	51.69
3D R.-W. Nav. · VLM	75.33	—	—	21.67
SLAG · VLM	88.45	—	—	62.39
Physics · VLM	65.31	—	—	—

Reported reliability subsets; raw agreement is not task success and is sensitive to class imbalance. Navigation rankings require human validation.

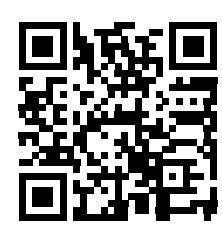
Evaluate generated media as solutions under constraints.

Video Sudoku: **0%** across all three models.

Physical Concepts: Sora-2 **76%**.

Inspect the preserved state, valid steps and final outcome.

EMNLP 2026 · Main Conference



PROJECT
Project website



CODE
GitHub repository



DATA
Dataset



VIDEO CASES
Watch the cases