

CAUSALSPATIAL: A BENCHMARK FOR OBJECT-CENTRIC CAUSAL SPATIAL REASONING

Anonymous authors

Paper under double-blind review

ABSTRACT

Humans can look at a static scene and instantly predict what happens next — *will moving this object cause a collision?* We call this ability Causal Spatial Reasoning. However, current multimodal large language models (MLLMs) cannot do this, as they remain largely restricted to static spatial perception, struggling to answer “what-if” questions in a 3D scene. We introduce **CausalSpatial**, a diagnostic benchmark evaluating whether models can anticipate consequences of object motions across four tasks: Collision, Compatibility, Occlusion, and Trajectory. Results expose a severe gap: humans score 84% while GPT-5 achieves only 54%. Why do MLLMs fail? Our analysis uncovers a fundamental deficiency: models over-rely on textual chain-of-thought reasoning that drifts from visual evidence, producing fluent but spatially ungrounded hallucinations. To address this, we propose the **CAUSAL OBJECT WORLD MODEL (COW)**, a framework that externalizes the simulation process by generating videos of hypothetical dynamics. With explicit visual cues of causality, COW enables models to ground their reasoning in physical reality rather than linguistic priors. Code and dataset will be publicly available.

1 INTRODUCTION

Humans can naturally perceive their surroundings and form a mental 3D model of the world. This mental model enables us to simulate possible physical interactions between objects. For example, as shown in Figure 1, in a static scene where a car faces a metal vase, we can use spatial cues such as orientation to predict a potential collision if the car moves forward. We refer to the task of grounding causal inference in spatial reasoning as **Causal Spatial Reasoning**.

This task requires three progressive spatial reasoning capabilities. As illustrated in Figure 1, the process begins with 2D image understanding and 3D static spatial reasoning. In these paradigms, models are tasked with recognizing objects and estimating spatial cues, such as camera perspective, orientation, and distance. While these capabilities form the foundation for further reasoning, they are inherently *static*—answering only “what/where is something” but failing to address “what will happen if...”, which is critical for understanding the physical world. To reason about causality, models must advance to 4D causal spatial reasoning, a comprehensive and dynamic paradigm that integrates spatial understanding with temporal prediction. Rather than analyzing spatial cues in isolation, models must synthesize this information to form coherent mental simulations. This requires first extracting 3D spatial relationships from static images, then simulating hypothetical or counterfactual interactions with temporal consistency (*e.g.*, “The car will collide with the vase if it moves forward.”).

While humans perform causal spatial reasoning effortlessly, whether MLLMs possess this capability remains largely underexplored. Current evaluation frameworks (as listed in Table 1) are limited to 2D and 3D static reasoning or overlook object interactions, failing to assess whether models can reason about how objects dynamically affect each other in the given spatial contexts.

To fill this gap, we propose CausalSpatial, a synthetic benchmark designed to diagnose causal spatial reasoning capabilities from an object-centric perspective. Each evaluation item consists of an image showing multiple objects in a scene, paired with a hypothetical object-level motion context (*e.g.*, If the car moves forward, a collision may occur.). This setup requires the model to move beyond static recognition of the given static image, and simulate the future object interaction outcomes according to the motion context. To provide a comprehensive diagnosis, we define four distinct forms of causal anticipation under different spatial configurations: **Collision**, **Occlusion**, **Compatibility**,

and *Trajectory*. As the first diagnostic benchmark of its kind, CausalSpatial provides the necessary testbed for advancing models toward more robust and causally grounded spatial understanding.

Extensive experiments on CausalSpatial expose a critical disparity between state-of-the-art MLLMs and human capabilities. Our analysis attributes this failure to a fundamental deficiency: with recent improvements in reasoning capability, MLLMs typically over-rely on a textual Chain-of-Thought (CoT) to infer dynamics and reason about causality. However, such textual CoT struggles to maintain consistent alignment with the original visual input, resulting in inaccurate spatial information during reasoning. Consequently, subsequent textual causal reasoning often includes spatial illusion, making models to generate seemingly plausible yet spatially ungrounded predictions.

To address the failure of text-based causal reasoning in maintaining spatial grounding, we propose leveraging world models to provide explicit causality hints with consistent 4D trajectory control. We propose **CAUSAL OBJECT WORLD MODEL (COW)**, an object-centric world model that externalizes the simulation process by rendering hypothetical object dynamics into video. Unlike textual reasoning, which drifts from reality, COW generates realistic future frames based on extracted 4D trajectories. This explicit visual evidence effectively compensates for the models’ inability to perform internal physical simulation, enabling grounded causal inference and demonstrating significant promise in complex spatial reasoning tasks.

In summary, our contribution is three-fold:

- We introduce CausalSpatial, the first object-centric causal spatial reasoning benchmark. It provides a comprehensive and diagnostic testbed for causal spatial reasoning.
- We conduct extensive experiments on CausalSpatial across a diverse set of MLLMs, revealing a significant performance gap between machines and humans. Our analysis indicates that models hallucinate in textual CoT, struggling to conduct consistent causal reasoning anchored to actual 3D configurations.
- We propose **CAUSAL OBJECT WORLD MODEL**, an object-centric video generation framework that maintains 3D consistency by fusing spatial cues and simulates cause-and-effect dynamics via rendering videos, paving the way for leveraging WMs in advanced causal spatial reasoning.

2 CAUSALSPATIAL BENCHMARK

In this section, we formally present CausalSpatial, a diagnostic benchmark designed to assess the capability of MLLMs in grounding causal inference within 3D spatial environments. We begin by formulating the question of causal spatial reasoning (Section 2.1), followed by a detailed taxonomy of the four proposed subtasks (Section 2.2). We then introduce the data collection pipeline (Section 2.3) and conclude with a comprehensive evaluation of current state-of-the-art models (Section 3).

2.1 BENCHMARK DESIGN

The design of CausalSpatial aims to provide a diagnostic assessment of MLLMs’ causal spatial reasoning capabilities. We simplify the analysis by focusing on scenarios where one object moves

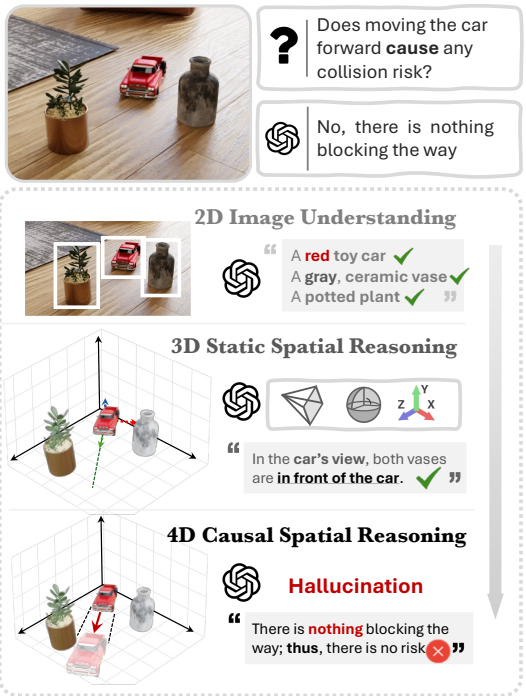


Figure 1: **Causal spatial reasoning task.** The task requires models to progress from 2D image understanding to 3D spatial perception and ultimately to causal spatial reasoning, predicting how actions affect future spatial configurations. Current MLLMs fail at this complex task due to hallucinations.

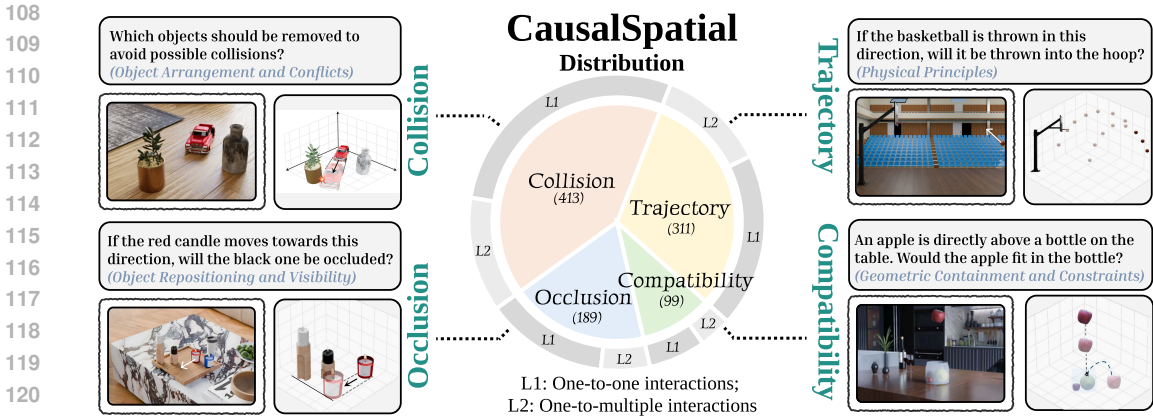


Figure 2: **CausalSpatial tasks.** CausalSpatial encompasses four causal reasoning tasks that require models to anticipate physical outcomes in 3D scenes: Collision, Occlusion, Trajectory, and Compatibility. The number of evaluation entries for each task is listed in the middle. Each task is designed with two difficulty levels, denoted as "L1/L2". All scenes are rendered in Blender to provide realistic 3D environments for physics-grounded evaluation.

hypothetically, and examine how this motion might subsequently affect other objects in the scene. We formalize each scenario as a state-transition problem: given an initial scene configuration, the model must predict the outcome when a specific object undergoes motion. Each query explicitly identifies the moving object to ensure the model performs object-grounded spatial reasoning rather than superficial pattern matching.

Causality formulation. Each query in CausalSpatial consists of an image I_0 capturing a 3D scene at initial state S_0 and a textual motion context T describing a hypothetical causal intervention. The model is tasked with predicting the resulting future spatial state S_{future} after this motion is applied. Unlike traditional VQA, which queries existing static attributes in an image, this task demands reasoning about cause-and-effect relationships: understanding how the described motion (cause) transforms the initial spatial configuration (precondition) into a new spatial arrangement (effect), purely from initial visual and textual inputs and internalized physical world knowledge.

Object-level design. A key feature of CausalSpatial is the object-level nature. In CausalSpatial, each motion context focuses on a specific target object and its potential interaction with the surrounding environment. This setup decouples complex scene dynamics into manageable interaction units, requiring the model to: (1) perceive the precise 3D spatial attributes of objects in the image (*e.g.*, position, orientation, geometry); (2) perform a mental simulation of the hypothetical motion m ; and (3) reason about the causal consequences (*e.g.*, collision, occlusion) resulting from the interaction between the target and other objects.

Improved realism. Unlike previous benchmarks [Wang et al. (2025c); Johnson et al. (2017a); Li et al. (2023)], which focus on simple objects and synthetic scenes, our benchmark is built on realistic scene settings with diverse object assets and lighting conditions. Notably, CausalSpatial includes a real-world subset from manual curation to evaluate model performance on authentic images captured in uncontrolled environments, providing a more challenging testbed specifically designed for real-world causal spatial reasoning.

Benchmark	Evaluation			Design		
	2D	3D	TD	CP	OM	RL
CLEVR-Count [Johnson et al. 2017b]						
CLEVR [Johnson et al. 2017a]	✓					
VQA [Goyal et al. 2017]	✓					✓
GQA [Hudson & Manning 2019a]	✓					✓
CV-Bench [Tong et al. 2024a]		✓				✓
SAT [Ray et al. 2024]	✓	✓	✓			✓
STI-Bench [Li et al. 2025]	✓	✓	✓			✓
Spatial457 [Wang et al. 2025c]	✓	✓		✓		
CausalVQA [Foss et al. 2025]	✓		✓	✓		✓
VSTI-Bench [Fan et al. 2025]	✓	✓	✓			✓
WM-Bench [Gao et al. 2025]	✓	✓	✓			
VSI-Super [Yang et al. 2025]	✓	✓	✓			✓
CausalSpatial	✓	✓	✓	✓	✓	✓

Table 1: **Comparison of benchmarks.** 2D Perception (2D), 3D Static Relation (3D), Temporal Dynamics (TD), Causal Prediction (CP), Object Manipulation (OM), and Realistic (RL).

2.2 TAXONOMY

To systematically evaluate the causal spatial reasoning capability, as shown in Figure 2, we design four types of hypothetical anticipation tasks. Each task reflects a distinct ability.

Collision anticipation evaluates the capability to reason about *object arrangement and conflicts*. In this task, models must predict: (1) whether a collision would occur if a target object moves forward, and (2) whether the collision outcome would change if another specific object were removed from the scene. Successfully addressing this task demands precise recognition of the spatial attributes of the target object (*e.g.*, orientation, size), and the spatial relationship among different objects (*e.g.*, distance, relative position).

Occlusion anticipation evaluates the capability to reason about *how object repositioning alters the visible scene structure*. In this task, objects are arranged in a 3D scene with complex depth relationships. Models must predict how moving a target object affects the visibility of other objects. Success requires understanding the camera perspective, analyzing spatial occlusion relationships between objects, and predicting how spatial adjustments affect visibility.

Trajectory anticipation evaluates the ability to predict object motion under *physical principles and causal constraints*. In this task, target objects move according to basic physical laws such as gravity, bouncing, and momentum transfer. Models must predict how objects interact with the environment or other objects (*e.g.*, whether a moving billiard ball will collide with others and where it will travel).

Compatibility anticipation evaluates the capability to reason about *geometric containment and spatial constraints*. In this task, models are given a hypothetical placement (*e.g.*, “If I want to put an apple into the bowl...”), and are required to assess containment feasibility. Specifically, models must determine whether issues such as overflow, spatial overlap, geometric penetration, or physical instability would occur. Success requires fine-grained understanding of object geometry, including size, shape, and volume relationships.

Two levels of difficulty. For each category, we define two difficulty levels: **(L1) One-to-one reasoning**, which focuses on reasoning about interactions between a single pair of objects, and **(L2) One-to-multiple reasoning**, which requires comprehending how a target object simultaneously relates to multiple others within the scene. This hierarchical formulation enables systematic evaluation of both basic pairwise reasoning and the model’s ability to generalize towards more complex, compositional multi-object interactions.

2.3 DATA COLLECTION AND ANNOTATION

To comprehensively evaluate the causal spatial reasoning capabilities of MLLMs, we construct CausalSpatial using publicly available realistic 3D scenes and assets. Specifically, we curate a diverse collection of 3D object assets spanning over 50 object categories. To ensure visual realism and diversity, we manually refine the texture of each object asset. Each scene is accompanied by thorough annotations, including RGB images, depth maps, and oriented bounding boxes, which enable accurate question generation.

For each scene, we design multiple action prompts that specify hypothetical or counterfactual object motions. These motion events are physically simulated and rendered using Blender’s physics engine, ensuring all outcomes adhere to real-world dynamics and physical plausibility. Each evaluation instance undergoes rigorous human verification to ensure causal correctness and quality. Full construction details are provided in the Appendix D.

3 BENCHMARK EVALUATION

In this section, we detail the experiment setup in Section 3.1 and then report quantitative results and key observations on CausalSpatial in Section 3.2 and Section 3.3. We discuss failure cases and possible improvement directions in Section 3.4. Our analysis highlights the substantial gap between current MLLMs and human intelligence.

Model	Level 1					Level 2					Avg.	Rank
	Col.	Com.	Occ.	Traj.	Avg.	Col.	Com.	Occ.	Traj.	Avg.		
Spatial Reasoning MLLMs												
Spatial-VLM Chen et al. (2024)	20.98	47.62	48.15	46.88	36.24 (3.83)	15.75	41.67	24.07	37.96	28.25 (7.17)	33.50	15
Spatial-MLLM Wu et al. (2025a)	41.96	55.56	48.15	54.17	47.93 (1.91)	24.41	44.44	12.96	24.82	24.56 (0.72)	40.00	12
Open-source MLLMs												
LLaVAOneVision1.5-8B-I Li et al. (2024)	43.01	42.86	38.52	29.17	38.17 (15.91)	15.38	41.67	20.37	38.69	27.84 (23.30)	34.62	14
Qwen2.5-VL-7B-I Team (2025c)	41.75	42.86	42.22	44.27	42.66 (2.21)	30.40	41.67	24.07	38.69	33.79 (5.73)	39.61	13
Qwen3-VL-2B-I Team (2025c)	47.43	47.62	49.63	43.23	46.69 (0.88)	19.35	38.89	22.22	37.96	28.98 (1.79)	40.60	11
Qwen3-VL-4B-I Team (2025c)	47.55	50.79	48.15	44.79	47.19 (6.92)	24.41	44.44	16.67	62.77	40.11 (10.39)	44.76	8
Qwen3-VL-8B-I Team (2025c)	41.40	66.67	51.11	47.92	47.55 (2.95)	22.76	36.11	29.63	50.36	35.85 (7.89)	43.53	9
Qwen3-VL-8B-T Team (2025c)	42.74	63.49	48.15	55.73	49.43 (1.03)	23.75	44.44	22.22	34.31	29.71 (1.79)	42.65	10
Qwen3-VL-30B-A3B-T Team (2025c)	47.02	55.56	48.15	47.40	48.15 (1.47)	30.16	52.78	31.48	52.55	41.32 (1.08)	45.80	6
Qwen3-VL-30B-A3B-T Team (2025c)	52.99	58.73	51.11	54.17	53.49 (0.00)	35.09	63.89	46.30	48.18	44.79 (0.36)	50.50	3
Closed-source MLLMs												
GPT5 OpenAI (2025a)	49.65	61.90	59.26	54.17	54.00 (1.47)	40.94	50.00	62.96	64.96	54.52 (1.08)	54.17	1
GPT5-min OpenAI (2025a)	37.76	42.86	48.15	44.79	42.31 (6.19)	35.43	47.22	50.00	67.88	51.41 (3.94)	45.44	7
Gemini2.5 Pro Team (2025b)	43.62	63.49	60.00	53.12	51.44 (0.00)	30.65	52.78	48.15	64.96	48.85 (0.36)	50.55	2
Gemini2.5 Flash Team (2025a)	46.21	50.79	48.89	56.25	50.02 (0.15)	42.31	58.33	25.93	55.47	46.53 (1.79)	48.82	4
Sonnet 3.7 Anthropic (2025)	55.00	49.21	45.56	46.35	50.12 (5.04)	29.41	48.61	34.26	60.58	44.16 (17.29)	48.07	5
Human	86.84	80.00	92.50	82.50	85.44	78.57	97.22	97.50	62.50	83.54	84.49	-
Random	23.61	31.75	32.35	38.54	30.34	8.54	33.33	20.37	32.14	24.50	28.6	-

Table 2: **The performance of MLLMs on CausalSpatial.** *Col.* denotes Collision anticipation. *Com.* denotes Compatibility anticipation. *Occ.* denotes Occlusion anticipation. *Traj.* denotes Trajectory anticipation. *Avg.* denotes Average scores. **Best** & **Second best**. The gray figures denote the Not Sure Rate (NSR) defined in §3.1. A clear performance gap between humans and MLLMs.

3.1 EVALUATION SETUP

Evaluated MLLMs. Our evaluation spans a diverse set of state-of-the-art MLLMs, including spatial-experts (Spatial-VLM [Chen et al. \(2024\)](#) and Spatial-MLLM [Wu et al. \(2025a\)](#)), as well as strongest general MLLMs, including Qwen3-VL series [Team \(2025c\)](#); [Bai et al. \(2025\)](#)), GPT-5 series [OpenAI \(2025a\)](#), Gemini 2.5 series [Team \(2025a\)](#), and Claude [Anthropic \(2025\)](#). Additionally, we recruited three independent human evaluators to complete the CausalSpatial benchmark. We also report random guessing scores to contextualize the difficulty of CausalSpatial.

Metrics. Standard binary evaluation paradigms often incentivize hallucinations by rewarding forced guessing over acknowledging uncertainty [OpenAI \(2025c\)](#). In spatial reasoning, there exists an inherent ambiguity in complex spatial dynamics, triggering the hallucination of MLLMs. To mitigate this bias, we adopt two complementary metrics: (1) **True Positive Rate (TPR)**, which measures the proportion of correctly answered questions among all samples, reflecting the model’s grounded reasoning accuracy; and (2) **Not Sure Rate (NSR)**, which quantifies the proportion of responses where the model chooses “Not Sure” option. This dual-metric setup allows us to decouple confident errors from epistemic uncertainty, providing a clearer view of the model’s behavioral calibration.

Evaluation details. We prompt the models to reason over the question and put their answer in the {“Reasoning”:..., “Answer”:...} JSON format. The maximum output length is 8192 tokens to ensure completing reasoning traces. Each question is repeated five times to reduce randomness. More details are presented in Appendix [E](#).

3.2 RESULT ANALYSIS

Main results. The results demonstrate a critical disparity between MLLMs and humans. As detailed in Table [2](#), humans maintain high accuracy across Level 1 (85.44%) and Level 2 (83.54%). However, MLLMs struggle to bridge this gap. Even the state-of-the-art GPT5 only achieves 54.17%, indicating causal spatial reasoning remains a significant challenge for current models.

In the open-source landscape, the Qwen3-VL series demonstrates superior efficiency. Qwen3-VL-30B-A3B-T shows the best performance (53.49% on Level 1 and 50.50% on Level 2), narrowing the gap with closed-source MLLMs. The compact Qwen3-VL-2B-I notably outperforms significantly larger models from other families, such as Qwen2.5-VL-7B (39.61%) and LLaVA-OneVision-8B (34.62%). Furthermore, specialized spatial reasoning MLLMs generally underperform compared to general-purpose MLLMs, highlighting the unique difficulty of our causal spatial tasks.

Different model sizes. As detailed in Figure [4](#), we observe limited returns from model scaling within the Qwen3-VL series. The instruction-tuned variants exhibit performance plateaus, with the 4B,

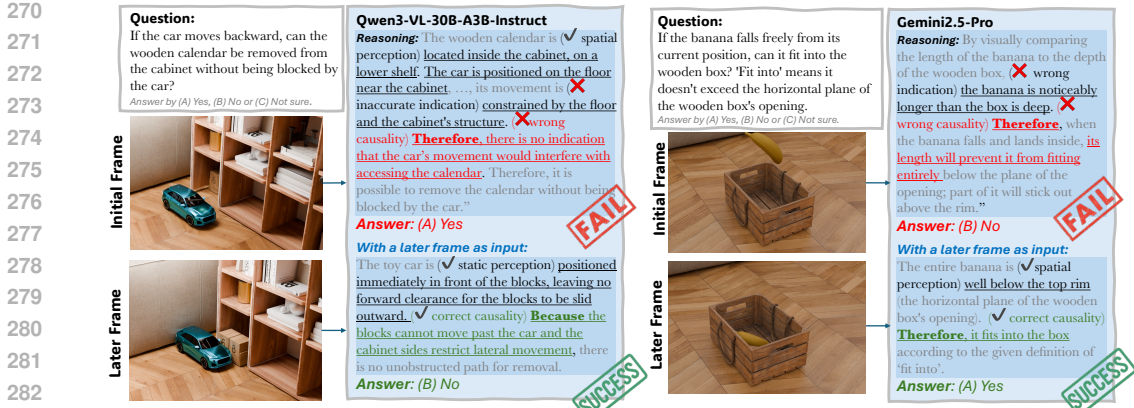


Figure 3: **Why do SoTA MLLMs fail in CausalSpatial?** These examples illustrate how MLLMs often produce lengthy and seemingly coherent explanations while failing to ground their reasoning in the visual evidence. In the car-cabinet scenario (Left), the model ignores the visible placement of the toy car directly blocking the wooden calendar and instead follows a generic linguistic pattern about object removal. In the banana-box example (Right), the model asserts that the banana exceeds the box depth—an inference inconsistent with the image. Across such failure modes, the reasoning chains are verbose but inaccurate, revealing that current MLLMs fail to simulate the motion process by solely relying on the textual reasoning.

Model	TPR	NSR
Qwen3-VL-2B-Instruct	57.76	3.45
Qwen3-VL-4B-Instruct	48.28	20.69
Qwen3-VL-8B-Instruct	56.90	5.17
Qwen3-VL-8B-Thinking	62.07	7.76
Qwen3-VL-30B-A3B-Instruct	63.79	2.59
Qwen3-VL-30B-A3B-Thinking	68.10	0.00
Human	93.10	-

Table 3: **Model performance on real-world subset.** TPR (True Positive Rate), NSR (Not Sure Rate). The results show a significant gap between MLLMs and humans.

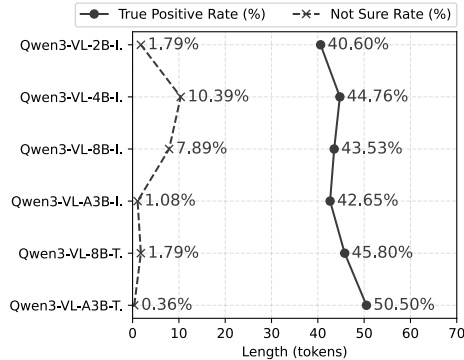


Figure 4: **Statistics for model results.** Larger models exhibit significantly lower NSR.

8B, and 30B-A3B models achieving similar scores of 44.76%, 43.53%, and 45.80%, respectively. This saturation reveals that model size scaling, though effective for general tasks, does not directly translate to improvements in causal spatial reasoning.

At the same time, we observe a decoupling between model confidence and accuracy performance. Larger models become increasingly decisive, often unwarranted. As shown in Figure 4, scaling model parameters dramatically reduces uncertainty. The Not Sure Rate (NSR) falls from 18.77% on Qwen3-4B-I to nearly zero (0.10%) on Qwen3-30B-A3B-T. This trend is further exacerbated by the thinking mode, where the NSR drops even more aggressively (e.g., from 10.57% (I) to 2.37% (T) on Qwen3-8B, from 2.84% (I) to 0.10% (T) on Qwen3-30B-A3B). However, this decisiveness is deceptive. The sharp decrease in uncertainty is not accompanied by a corresponding leap in accuracy.

3.3 EXTENSION TO REAL-WORLD SUBSET

To validate generalization beyond synthetic data, we conduct evaluation on a manually curated real-world subset. Results in Section 3.2 reveal that the challenges identified in controlled settings transfer to authentic, noisy environments. While humans maintain strong performance at 93.10% TPR, confirming that CausalSpatial tasks remain well-defined and solvable in real scenarios, the leading model Qwen3-30B-A3B-Thinking falls short by 25%. This substantial gap highlights the persistent difficulty MLLMs encounter when performing causal reasoning in real-world scenes.

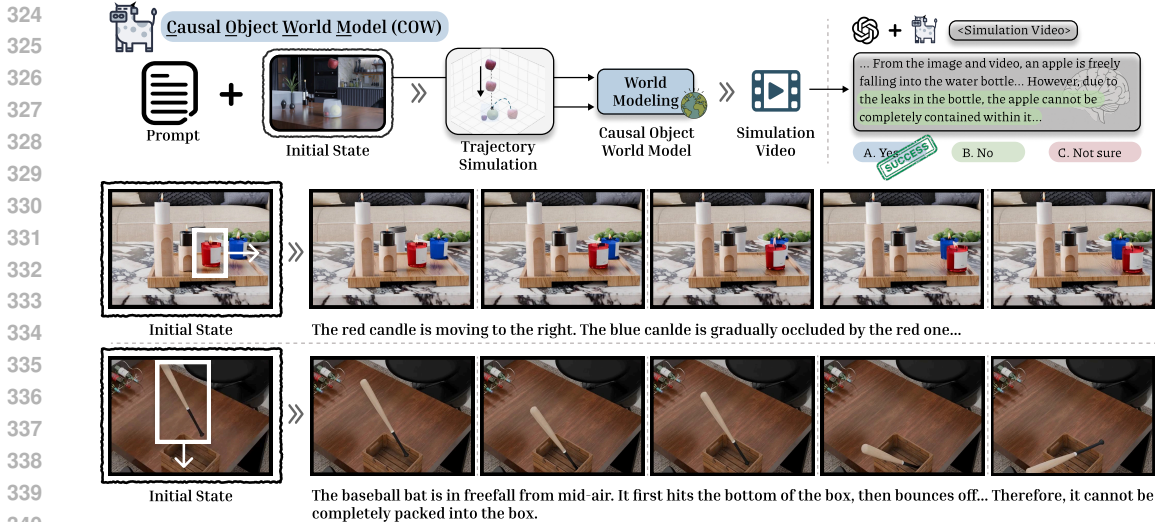


Figure 5: (Above) **COW pipeline overview**: COW is a 4D trajectory-controlled video generation method that enhances spatial reasoning by explicitly rendering object dynamics through object-centric video generation. (Below) **Case demonstration**: COW produces physically and visually plausible simulation videos that offer clear visual cues for improved causal spatial inference.

We observe a correlation between model capability and confidence. Qwen3-VL-4B-I exhibits the highest uncertainty with a Not Sure Rate of 20.69%. However, as capabilities scale, models become increasingly decisive. The SOTA model, Qwen3-VL-30B-A3B-T, drops to 0, indicating that advanced models can form definite conclusions even in complex real-world settings, although their reasoning accuracy still requires improvement to match human levels.

3.4 FAILURE ANALYSIS & IMPROVE DIRECTIONS

Despite the challenging nature of causal spatial reasoning revealed by our benchmark, understanding *why* current MLLMs fail and *how* to address these failures provides valuable insights for future progress. We conduct a detailed failure analysis and discuss promising directions for improvement.

Why do SoTA MLLMs fail in CausalSpatial? Figure 3 illustrates representative failure cases from Qwen3-VL-30B-A3B-Instruct [Team (2025c)] and Gemini2.5-Pro [Team (2025b)]. Initially, MLLMs generally succeed at identifying static object positions and spatial relationships. However, a critical disconnect occurs during the subsequent reasoning phase. These visual cues are either diluted or misinterpreted as the textual chain progresses. In the car-cabinet example, the model produces a coherent explanation but does not consider the car’s visible position in front of the wooden calendar, resulting in an incorrect answer. In the banana-box example, the model concludes that the banana is longer than the depth of the box, which is not supported by the image.

Consistent with prior studies [Chen et al. (2025); OpenAI (2025c)], we observe that MLLMs prefer to reason based on linguistic priors rather than grounding their logic in the image. Consequently, when conducting causal spatial reasoning, models tend to drift toward generic textual CoT, which includes remarkable hallucinations, neglecting the specific geometric constraints present in the visual input.

This analysis reveals that: *MLLMs fail to ground their reasoning in the geometry evidence, drifting toward linguistically plausible but spatially inaccurate explanations with inaccurate prediction.*

How to anchor reasoning in 3D constraints? World Models (WMs) offer a promising solution, as they are fundamentally designed to capture causal outcomes and temporally consistent dynamics given hypothetical actions. However, existing world models [Lu et al. (2025); Wan et al. (2025)] are limited to scene exploration, primarily generating static environments with moving viewpoints, and thus overlook object-level spatial consistency critical for causal reasoning. Accordingly, we argue that a world model maintaining strict object-level spatial consistency can serve as a reliable visual grounding mechanism, enabling MLLMs to perform accurate causal spatial reasoning.

4 COW: CAUSAL OBJECT WORLD MODEL

Based on the above discussions, we propose **CAUSAL OBJECT WORLD MODEL**, a framework designed for fine-grained control over 4D object trajectories. The pipeline design is described in Section 4.1. We evaluate the effectiveness on CausalSpatial and discuss the future direction in Section 4.2. Further details are available in Appendix G.

4.1 COW PIPELINE

The queries from CausalSpatial consist of an image representing the initial state and a motion context implying an intervention. COW processes this information, externalizes causality modeling by trajectory simulation and trajectory-conditioned video generation, providing rich visual cues for MLLM reasoning. The pipeline includes three key steps.

4D trajectory simulation. We extend the spatial state into a 4D representation using a 6-dimensional vector $Obj_i = \{x, y, z, v_x, v_y, v_z\}$. The initial position $\{(x, y, z)\}$ (predict the bounding box) and $\vec{v} = \{v_x, v_y, v_z\}$ (predict the object orientation or direction indicator in images) are estimated via prompting MLLMs. We then construct the synthetic trajectory τ by iteratively updating the object’s position, modulated by external physical factors, such as gravity.

Controllable video generation. We employ ATI Wang et al. (2025a) as the world model to render physically consistent dynamics. The predicted trajectories τ_i and motion descriptions m_i serve as control signals to condition a diffusion-based generator, synthesizing coherent frames: $F_{later} = W_{\theta}(F_{init}|\tau_i^t, d_i)$. This design uses rich priors of large-scale video models, effectively bridging physical modeling with data-driven imagination for realistic simulation.

Assist MLLMs in reasoning. We integrate COW with an MLLM to enable temporally grounded reasoning. The MLLM jointly interprets the simulated video and the text query to generate informed answers, effectively moving beyond static perception to support motion process prediction.

4.2 EVALUATION WITH COW

Experiment setup. We test the representative model, Qwen3-VL-30B-A3B-Instruct, to evaluate the effectiveness of COW. Our evaluation also includes a subset of Spatial 457 Wang et al. (2025c) that is associated with causal spatial reasoning. We use three initial frames from generated videos as additional visual cues for subsequent reasoning. All images are resized to 512 pixels. Experiments are repeated three times to ensure robustness.

Main results. Figure 6 demonstrates a consistent performance boost on CausalSpatial. Introducing merely three frames from the generated video yields a notable accuracy improvement, 2.40% on Collision and 2.20% on Trajectory. Crucially, COW only provides a limited number of simulated frames to MLLMs, which can already improve the model’s performance in spatial reasoning without requiring extra training or datasets.

However, there is a slight decrease in the Occlusion task. A possible reason lies in the continuous nature of the Occlusion task (*e.g.*, the target object may be occluded during the initial stages, yet eventually revealed in later frames.) Since the first few video frames only simulate the immediate future, they provide limited visual cues for subsequent reasoning required by MLLMs.

For Spatial457 Wang et al. (2025c), which is an external benchmark for evaluation, introducing simulated frames from COW leads to an accuracy

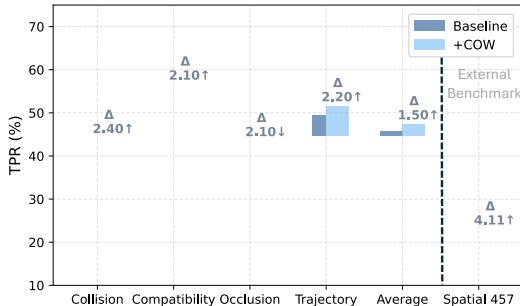


Figure 6: **Quantitative results of COW.** The visual simulation leads to performance gains in CausalSpatial, highlighting the potential of explicit vision simulation. The improvement on an external benchmark (Spatial457 Wang et al. (2025c)) further validates the effectiveness and generalization ability.

improvement of 4.11%. This improvement confirms that COW is effective and generalizable under various data distributions.

Overall, the results demonstrate that COW provides a robust visual grounding mechanism that helps models move beyond linguistic priors and ground their reasoning in actual 3D constraints of a scene.

4.3 DISCUSSION

As illustrated in Figure 3, the additional simulated images play a pivotal role in anchoring the initial state implied by the question image, thus facilitating reasoning about the subsequent dynamics. In the car-cabinet example, the model correctly interprets the spatial arrangement, identifying the visible blocking between the car and the wooden cabinet. Similarly, in the banana-box example, COW successfully simulates the future state where the banana lies in the box, enabling the model to accurately assess their size relationship and avoid the previous erroneous conclusion about depth incompatibility. These cases demonstrate a clear pattern: when equipped with simulated future states, the model effectively suppresses hallucinations by grounding its reasoning in accurate visual dynamics rather than unreliable linguistic priors. This visual grounding mechanism proves crucial for maintaining reasoning accuracy.

Consequently, the quality and physical plausibility of generated videos are critical to reasoning accuracy. COW maintains precise geometric consistency by grounding video generation in explicit object trajectories, ensuring that simulated dynamics remain spatially coherent throughout the temporal sequence. However, how to guarantee this consistency in additional geometric attributes such as volume, shape deformation, and inter-object collision dynamics remains a challenging open question for future exploration.

5 RELATED WORK

Causal Spatial Reasoning. Existing benchmarks largely focus on static 3D relations (Johnson et al. (2017c); Li et al. (2023); Hudson & Manning (2019b); Tong et al. (2024b); Liu et al. (2023); Ma et al. (2025) or passive motion tracking (Li et al. (2025); Zhou et al. (2025b); Parmar et al. (2025); Lin et al. (2025), often relying on unrealistic scenes (Wang et al. (2025c); Qi et al. (2025); Liu et al. (2025); Wu et al. (2025b)). However, true spatial intelligence demands *causal reasoning*, predicting “what will happen” rather than merely observing the state. To this end, we propose CausalSpatial. Unlike prior descriptive tasks, CausalSpatial emphasizes causal prediction from object-level manipulations, requiring MLLMs to simulate physical dynamics.

Video Generation & World Models. While world models demonstrate strong simulation capabilities in robotics and driving (Ha & Schmidhuber (2018); Wan et al. (2025); Wang et al. (2025a); Cui et al. (2025); OpenAI (2025b); Deepmind (2025), they predominantly prioritize global dynamics (Alonso et al. (2024); Yu et al. (2025b) or ego-motion (Du et al. (2023); Ko et al. (2023); Yang et al. (2023), lacking fine-grained object controllability. Current object-centric approaches (Wu et al. (2023b); Locatello et al. (2020); Didolkar et al. (2025) are typically limited to synthetic settings. We introduce **CAUSAL OBJECT WORLD MODEL (COW)**, an object-centric pipeline enabling trajectory-controlled motion generation to support precise causal reasoning.

6 CONCLUSION

In this paper, we define a higher-level spatial reasoning paradigm, Causal Spatial Reasoning, a significant step to bridge the virtual reasoning and the physical world. To better investigate this question, we present CausalSpatial, the first-of-this-kind benchmark designed to evaluate causal spatial reasoning capabilities. Extensive experiments reveal that current models struggle to form coherent mental simulations for advanced reasoning. To address this, we introduced COW, a world-model-based framework that generates controllable object-level motion simulation to support temporally grounded reasoning. The experimental improvements highlight the potential of explicit causal imaginations, pointing the direction for future exploration.

REFERENCES

- 486
487
488 Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and
489 François Fleuret. Diffusion for world modeling: Visual details matter in atari. *NeurIPS*, 2024.
- 490 Anthropic. Claude sonnet 3.7, 2025. URL <https://www.anthropic.com/news/claude-3-7-sonnet>.
- 491
492
493 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang,
494 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,
495 2025.
- 496 Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia.
497 Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, pp.
498 14455–14465, 2024.
- 499
500 Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor
501 Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention
502 mechanism perspective on focus areas. *ICML*, 2025.
- 503 An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang,
504 and Sifei Liu. Spatialrpt: Grounded spatial reasoning in vision language model. *NeurIPS*, 2024.
505
- 506 Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan
507 Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners.
508 *arXiv preprint arXiv:2510.26583*, 2025.
- 509 Google Deepmind. Veo3.1, 2025. URL <https://deepmind.google/models/veo/>.
- 510
511 Aniket Didolkar, Andrii Zadaianchuk, Rabiul Awal, Maximilian Seitzer, Efstratios Gavves, and
512 Aishwarya Agrawal. Ctrl-o: language-controllable object-centric visual representation learning. In
513 *CVPR*, 2025.
- 514
515 Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and
516 Pieter Abbeel. Learning universal policies via text-guided video generation. *NeurIPS*, 2023.
- 517 Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi
518 Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-
519 aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.
- 520
521 Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao.
522 Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint*
523 *arXiv:2506.09943*, 2025.
- 524 Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun,
525 Gautham Kishore, Bo Ai, et al. Do vision-language models have internal world models? towards
526 an atomic evaluation. *arXiv preprint arXiv:2506.21876*, 2025.
- 527
528 Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang,
529 and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile
530 controllability. *NeurIPS*, 2024. URL <https://arxiv.org/abs/2405.17398>.
- 531
532 Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa
533 matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.
- 534
535 Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld:
536 a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*,
537 2025.
- 538
539 David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.
- Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *ICLR*, 2021.

- 540 Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton,
541 and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint*
542 *arXiv:2309.17080*, 2023.
- 543 Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning
544 and compositional question answering. In *CVPR*, 2019a.
- 545 Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning
546 and compositional question answering. In *CVPR*, 2019b.
- 547 Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and
548 Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual
549 reasoning. In *CVPR*, 2017a.
- 550 Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and
551 Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual
552 reasoning. In *CVPR*, 2017b.
- 553 Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and
554 Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual
555 reasoning. In *CVPR*, 2017c.
- 556 Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from
557 actionless videos through dense correspondences. *arXiv preprint arXiv:2310.08576*, 2023.
- 558 Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A
559 world model for indoor navigation. In *ICCV*, 2021.
- 560 Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open*
561 *Review*, 62(1):1–62, 2022.
- 562 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan
563 Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint*
564 *arXiv:2408.03326*, 2024.
- 565 Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench:
566 Are mlms ready for precise spatial-temporal world understanding? *ICCV*, 2025.
- 567 Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin
568 Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness
569 in visual reasoning. In *CVPR*, 2023.
- 570 Xiaoxing Lian, Aidong Yang, Jun Zhu, Peng Wang, and Yue Zhang. Imagine in space: Exploring the
571 frontier of spatial intelligence and reasoning efficiency in vision language models. *arXiv preprint*
572 *arXiv:2511.13782*, 2025.
- 573 Jieru Lin, Zhiwei Yu, and Börje F Karlsson. Switch: Benchmarking modeling and handling of
574 tangible interfaces in long-horizon embodied scenarios. *arXiv preprint arXiv:2511.17649*, 2025.
- 575 Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu
576 Liu, Aniket Bera, and Zhaoshuo Li. Scenethesis: A language and vision agentic framework for 3d
577 scene generation. *arXiv preprint arXiv:2505.02836*, 2025.
- 578 Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *ACL*, 2023.
- 579 Zhaochen Liu, Kaiwen Gao, Shuyi Liang, Bin Xiao, Limeng Qiao, Lin Ma, and Tingting Jiang.
580 Beyond the visible: Benchmarking occlusion perception in multimodal large language models.
581 *arXiv preprint arXiv:2508.04059*, 2025.
- 582 Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold,
583 Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention.
584 *NeurIPS*, 2020.

- 594 Xiaoxiao Long, Qingrui Zhao, Kaiwen Zhang, Zihao Zhang, Dingrui Wang, Yumeng Liu, Zhengjie
595 Shu, Yi Lu, Shouzheng Wang, Xinzhe Wei, et al. A survey: Learning embodied intelligence from
596 physical simulators and world models. *arXiv preprint arXiv:2507.00917*, 2025.
- 597
598 Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel
599 Khashabi, Rama Chellappa, Alan Yuille, et al. Genex: Generating an explorable world. *ICLR*,
600 2025.
- 601 Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan
602 Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *ICCV*, 2025.
- 603 OpenAI. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>,
604 2025a.
- 605
606 OpenAI. Sora2 is here, 2025b. URL [https://openai.com/zh-Hans-CN/index/
607 sora-2/](https://openai.com/zh-Hans-CN/index/sora-2/).
- 608 OpenAI. Why language models hallucinate, 2025c. URL [https://openai.com/index/
609 why-language-models-hallucinate/](https://openai.com/index/why-language-models-hallucinate/).
- 610
611 J Parker-Holder, P Ball, J Bruce, V Dasagi, K Holsheimer, C Kaplanis, A Moufarek, G Scully,
612 J Shar, J Shi, et al. Genie 2: A large-scale foundation world model. URL: [https://deepmind.
613 google/discover/blog/genie-2-a-large-scale-foundation-world-model](https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model), 2024.
- 614 Paritosh Parmar, Eric Peh, and Basura Fernando. Chainreaction! structured approach with causal
615 chains as intermediate representations for improved and explainable causal video question answer-
616 ing. *arXiv preprint arXiv:2508.21010*, 2025.
- 617
618 Yu Qi, Haibo Zhao, Ziyu Guo, Siyuan Ma, Ziyang Chen, Yaokun Han, Renrui Zhang, Zitiantao Lin,
619 Shiji Xin, Yijian Huang, et al. Bear: Benchmarking and enhancing multimodal language models
620 for atomic embodied capabilities. *arXiv preprint arXiv:2510.08759*, 2025.
- 621 Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha
622 Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training
623 for multimodal language models. *COLM*, 2024.
- 624 Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter
625 Abbeel. Masked world models for visual control. In *CoRL*, pp. 1332–1344. PMLR, 2023.
- 626
627 Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang,
628 Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision
629 language models elicits generalization to spatial reasoning. *EMNLP*, 2024.
- 630 Gemini Google Team. Gemini2.5 flash: Best for fast performance on everyday tasks. [https:
631 //deepmind.google/models/gemini/flash/](https://deepmind.google/models/gemini/flash/), 2025a.
- 632
633 Google Gemini Team. Gemini 2.5 pro best for coding and highly complex tasks, 2025b. URL
634 <https://deepmind.google/models/gemini/pro/>.
- 635 Qwen Team. Qwen3-vl. <https://github.com/QwenLM/Qwen3-VL>, 2025c.
- 636
637 Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha
638 Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully
639 open, vision-centric exploration of multimodal llms. *NeurIPS*, 2024a.
- 640 Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha
641 Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open,
642 vision-centric exploration of multimodal llms. *NIPS*, 2024b.
- 643
644 Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu,
645 Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models.
646 *arXiv preprint arXiv:2503.20314*, 2025.
- 647 Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any
trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025a.

- 648 Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahao Zhang, Yuxiang Guo, Xijun Liu, Rama
649 Chellappa, Cheng Peng, Alan Yuille, et al. Evoworld: Evolving panoramic world generation with
650 explicit 3d memory. *arXiv preprint arXiv:2510.01183*, 2025b.
651
- 652 Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille.
653 Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In
654 *CVPR*, 2025c.
655
- 656 Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mlm: Boosting mllm capabilities
657 in visual-based spatial intelligence. *NeurIPS*, 2025a.
658
- 659 Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. SpatialScore:
660 Towards unified evaluation for multimodal spatial understanding. *arXiv preprint arXiv:2505.17012*,
661 2025b.
662
- 663 Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer:
664 World models for physical robot learning. In *CoRL*, pp. 2226–2240. PMLR, 2023a.
665
- 666 Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised
667 visual dynamics simulation with object-centric models. *ICLR*, 2023b.
668
- 669 Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter
670 Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6,
671 2023.
672
- 673 Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong,
674 Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video.
675 *arXiv preprint arXiv:2511.04670*, 2025.
676
- 677 Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. Wonderworld:
678 Interactive 3d scene generation from a single image. *CVPR*, 2025a.
679
- 680 Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating
681 new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025b.
682
- 683 Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad
684 Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. Combo: compositional world models for
685 embodied multi-agent cooperation. *arXiv preprint arXiv:2404.10775*, 2024.
686
- 687 Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and
688 Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *ArXiv*, abs/2403.09631,
689 2024. URL <https://api.semanticscholar.org/CorpusID:268385444>.
690
- 691 Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan.
692 Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.
693
- 694 Fenzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A
695 comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025a.
696
- 697 Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra,
698 Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal
699 awareness in vision language models. In *ICCV*, 2025b.
700
- 701 Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, D. Y. Yeung, and Chuang Gan. Robodreamer:
Learning compositional world models for robot imagination. *ArXiv*, abs/2404.12377, 2024. URL
<https://api.semanticscholar.org/CorpusID:269214356>.