SpatialLogic-Bench: A Diagnostic Benchmark for Task-Oriented Spatiotemporal Reasoning

Xiaoda Yang[1*], Shenzhou Gao[2*], Can Wang[2*], Jiahe Zhang[2], Menglan Tang[2], Jingyang Xue[2], Sheng Liu[3], Peijian Zhang[2†], Yao Mu[4†], Xiangyu Yue[5]

[1]Zhejiang University
[2]Qingdao University
[3]Karlsruher Institut für Technologie
[4]Shanghai Jiao Tong University
[5]The Chinese University of Hong Kong
xiaodayang@zju.edu.cn, zpj@qdu.edu.cn

Abstract

Vision-Language Models (VLMs) have made significant progress in static perception, but their ability to understand dynamic task-oriented reasoning remains unclear. Existing benchmarks mainly focus on static spatial relationships and lack systematic assessment of dynamic reasoning capabilities. To this end, we propose SpatialLogic-Bench, a novel benchmark designed to evaluate VLMs' understanding of spatiotemporal logic and their ability to assess task progress. The benchmark assesses two critical capabilities: first, fine-grained visual discrimination to accurately perceive subtle physical changes between state frames; second, the logical capacity to connect these changes to task goals and judge whether they indicate progress. To mitigate temporal dependency biases, we introduce a dual-task paradigm, presenting image pairs in both chronological and reversed orders while keeping task descriptions consistent. We construct a multi-scale evaluation system by varying time intervals between frames: smaller intervals test the model's fine-grained perception, while larger intervals demand more sophisticated logical inference. Empirical evaluation reveals that most VLMs experience significant performance degradation on tasks presented in inverse chronological order, indicating an over-reliance on temporal cues rather than robust reasoning abilities. SpatialLogic-Bench clearly exposes critical limitations in current models and provides valuable guidance for improving dynamic spatial perception capabilities.

1 Introduction

Vision-Language Models (VLMs) have made rapid progress in anchoring language and vision in the physical world, with advancements spanning from controllable video generation (Yang et al. 2025c) and sophisticated dialogue systems (Cheng et al. 2025b) to novel prompting techniques (Yan et al. 2025), becoming a crucial cornerstone of spatial intelligence (Gan et al. 2022). However, the key to advancing this field lies in establishing diverse, and challenging evaluation benchmarks to provide guidance.

The evaluation of the spatial perception ability of VLMs has revolved around various tasks, including judging the relative positions between objects, estimating sizes or distances, reasoning about the consistency of scene from different perspectives, and locating specific objects in images (Mao et al. 2016). Although these tasks vary in form, they essentially all directly examine spatial information itself. The core of these traditional evaluations lies in testing whether the model can perceive the objective facts of spatial relationships. This form of testing constitutes the entirety of the conventional evaluation of a model's dynamic spatial cognitive ability (Fu et al. 2024a). However, traditional evaluations have significant limitations. On the one hand, the task design is relatively simple, and relevant indicators can be easily improved by fine-tuning (Pfeiffer et al. 2020). Moreover, the data relies on manual annotation, which is often limited in scale (Krishna et al. 2017). More importantly, such evaluations, which only examine the recognition of spatial information in isolation without associating spatial states with task goals, fail to capture the intrinsic relationship between tasks and the broader physical state space.

To address these limitations, we propose a new evaluation method, SpatialLogic-Bench, which embeds the assessment of spatial perception ability into dynamic task scenarios. Figure 1 provides a conceptual overview of this benchmark. Specifically, this evaluation method requires the model to determine which of the two states, extracted from different stages of a task video, is closer to completion. This design goes beyond the level of "knowing spatial facts" and forces the model to evaluate the relevance between the spatial states and the endpoint of the task against the task goal, thus comprehending "what the spatial state means for the task". This transformation deeply links spatial perception with the practical needs of task reasoning, which is crucial for fields like VLM-empowered embodied intelligence (Shridhar, Manuelli, and Fox 2023) and other complex applications (Yang et al. 2024b; Fu et al. 2024b). In real-world scenarios, the core value of spatial perception lies in providing a basis for judging "how to act" (Savva et al. 2019). To enable this robust evaluation, our benchmark em-

[*]These authors contributed equally.
[†]Corresponding authors.

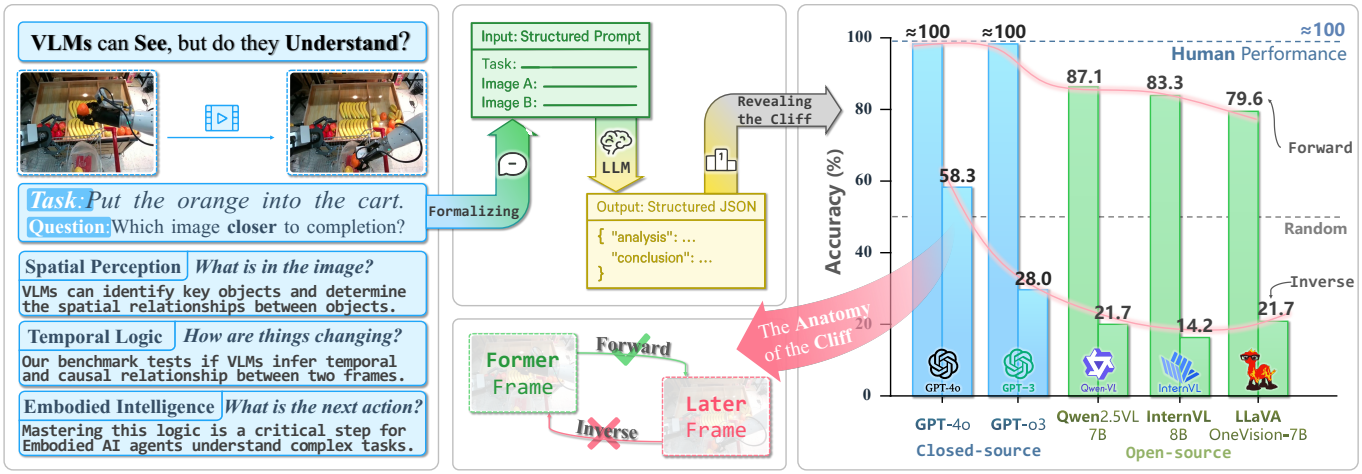

Figure 1: An overview of the benchmark's core concepts, illustrating the progression from data foundation to the dual-task paradigm for evaluating the spatiotemporal reasoning of VLMs.

ploys a multi-granularity sliding window strategy and a pioneering Dual-Task Paradigm that challenges models with both forward and inverse chronological sequences.

The main contributions can be summarized as follows.

- **Novel Evaluation Task.** We propose a novel evaluation task, namely task-oriented spatiotemporal logical reasoning. This task elevates the assessment of spatial intelligence from traditional tasks like static relation recognition and action classification to the far more challenging level of causal logical judgment of task progression.
- **Large-Scale Benchmark.** We design and construct a large-scale evaluation benchmark, SpatialLogic-Bench, based on real-world data. This new benchmark employs a multi-granularity sliding window sampling strategy, enabling a hierarchical and in-depth assessment of the model's reasoning ability at different timescales.
- **Dual-Task Paradigm.** We pioneer a novel Dual-Task Paradigm, where a carefully designed symmetric setting of forward and inverse tasks achieves the effective decoupling and quantification of the model's true logical reasoning ability from its dependence on temporal heuristics, which provides a new comprehensive method for the rigorous and evaluation of the robustness of VLMs.

2 Related Work

2.1 Benchmarks of Spatial Perception Ability

Spatial perception benchmarks have evolved from simple relational understanding (e.g., CLEVR (Johnson et al. 2017)) to complex reasoning in realistic 2D (e.g., GQA (Hudson and Manning 2019)) and 3D scenes (e.g., ScanNet (Dai et al. 2017), Matterport3D (Chang et al. 2017)). A crucial shift towards dynamic interaction occurred with environments like Habitat (Savva et al. 2019) and AI2-THOR (Kolve et al. 2017). However, despite their increasing complexity and recent advances in multi-image reasoning (Yang et al. 2025a) and robustness evaluation (Cheng et al. 2025a), these benchmarks share a fundamental limitation: they primarily test the recognition of spatial facts (e.g., "where is object A?"). They fall short of evaluating task-oriented reasoning—whether a model understands how a spatial change contributes to or hinders a specific goal. SpatialLogic-Bench is designed to fill this gap by shifting the focus from recognizing spatial states to understanding dynamic, task-oriented causal logic.

2.2 Benchmarks for Logical Reasoning Ability

In parallel, benchmarks for logical reasoning have evolved from multi-hop textual navigation (HotpotQA (Yang et al. 2018)) and discrete reasoning (DROP (Dua et al. 2019), ReClor (Yu et al. 2020)) to structured challenges like mathematical problem-solving (GSM8K (Cobbe et al. 2021)) and abstract pattern generalization (ARC (Chollet 2019)). While these benchmarks rigorously probe complex inference, they operate in purely symbolic domains, detached from perceptual input. They fail to address the challenge of grounding logical steps in noisy visual data—a central problem in multimodal learning (Yang et al. 2024a, 2025b; Fu et al. 2025). SpatialLogic-Bench addresses this gap by requiring models to derive causal judgments from visual inputs, bridging the divide between disembodied logic and spatial perception.

3 SpatialLogic-Bench

3.1 Overview of SpatialLogic-Bench

Our benchmark strategically moves beyond static spatial evaluation by requiring models to reason about changes between two frames from a task video. To systematically control the difficulty and focus of this evaluation, we introduce **window size** as a core mechanism, defined as the temporal interval between the selected frames (I_{start}, I_{end}). This allows us to create a graded assessment: smaller window sizes present subtle physical differences to test fine-grained per-

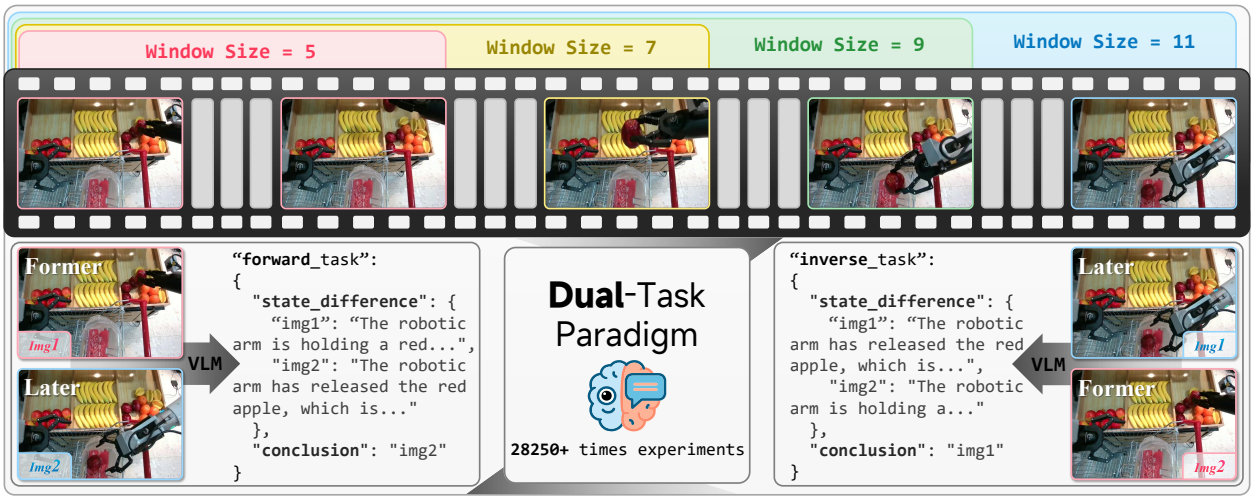

Figure 2: Representative examples from SpatialLogic-Bench across different window sizes (5 to 11). Each example shows an image pair and the required analysis for determining which frame is closer to task completion.

Task	Quantity
Packing in e-commerce	2,624
Iron clothes	2,563
Brush water bottle	2,052
Cook vegetables (oven)	1,920
Flatten shorts	1,804
Packing in supermarket	1,580
Pickup from supermarket	1,512
Wash dishes (dishwasher)	1,187
Hang clothes with hanger	1,082
Brew tea	958
Pack industrial items	944
Open fridge to get food	904
Insert book into bookshelf	816
Sort food	796
Sort clothes	740
Other Tasks (16 total)	*6,768*
Total	**28,250**

Table 1: Distribution of the top 15 tasks in SpatialLogic-Bench, totally 28,250 instances. A full breakdown is in the Supplementary Material (Section G).

ception, while larger ones omit more intermediate steps, demanding macroscopic understanding and logical inference. This multi-scale evaluation is implemented across **28,250 image pairs** from a diverse set of real-world tasks (Table 1), with representative examples clearly shown in Figure 2.

3.2 Benchmark Construction Process

As clearly illustrated in Figure 3, the systematic construction of SpatialLogic-Bench is carefully guided by core principles designed to ensure a rigorous and fair evaluation. These principles permeate the entire process from data source selection to the final construction of the task paradigm.

Data Source (Data Collection). To ensure ecological validity, our benchmark is exclusively built upon the comprehensive AgiBot-World dataset (Bu et al. 2025), which features real-world robotic manipulations. Unlike benchmarks that rely on static images or synthetic simulations, AgiBot-World provides large-scale, diverse, and goal-directed video sequences of physical tasks. This grounding in real-world, continuous state evolution offers a solid foundation for evaluating the understanding of dynamic tasks. Crucially, it serves as the prerequisite for assessing the higher-order spatiotemporal reasoning that our benchmark targets.

Data Curation and Multi-Granularity Sampling. The benchmark employs a data processing pipeline incorporating temporal downsampling and a sliding window mechanism. First, a 10-fold temporal downsampling is uniformly applied to all video streams. This downsampling serves a dual purpose: it filters out high-frequency visual noise (e.g., camera jitters) while ensuring that consecutive frames exhibit significant state changes, thereby forcing models beyond simple pattern matching. Subsequently, a sliding window mechanism traverses the downsampled sequence, extracting only the start and end frames from windows of various preset sizes to form our core image pairs. This "process omission" design is the key to the benchmark's challenge, as it requires the model to perform logical interpolation and causal inference on the unobserved action sequences under the constraints of the given task objectives. Systematically adjusting the window size enables a multi-level evaluation, ranging from fine-grained state identification (small windows) to coarse-grained logical planning (large windows).

Dual-Task Paradigm and Question Formulation. To fundamentally diagnose and eliminate the prevalent "Chronological Bias" (Geirhos et al. 2020)—the tendency

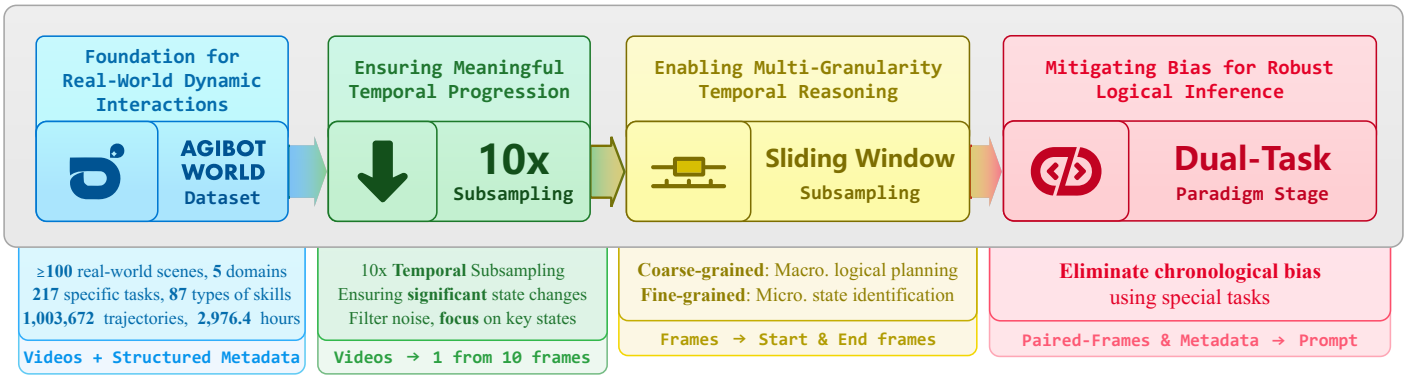

Figure 3: The construction pipeline of SpatialLogic Bench. The process starts from the AgiBot-World dataset, followed by temporal subsampling and a sliding window mechanism to generate multi-granularity samples. Finally, the dual-task paradigm is applied to mitigate chronological bias.

of models to assume the second frame in a sequence is closer to completion—we introduce an innovative "Dual-Task Paradigm." The core mechanism is as follows: **Forward Task:** The model receives an input tuple (I_{start}, I_{end}, x), where I_{start} and I_{end} are the start and end frames, and x is the task description. The ground truth is I_{end}. **Inverse Task:** The image order is reversed to (I_{end}, I_{start}, x). The ground truth remains I_{end}, compelling the model to reason based on task logic rather than sequence order.

The key contribution of this design is to render the input's chronological order an unreliable heuristic. To succeed, a model can no longer rely on the shallow heuristic of "selecting the second image;" instead, it must genuinely reason about the task by integrating its understanding of the goal with the visual evidence from both states. This paradigm thus forces a shift from superficial sequence-based pattern recognition to genuine, task-driven logical reasoning.

4 Experiment

4.1 Evaluation Setup

Evaluated Models. We evaluate a wide range of VLMs, covering closed-source, open-source 2D, and 3D perception models. The evaluated models include major proprietary systems such as GPT-4o (OpenAI 2024) and Gemini-2.5-Pro (Team et al. 2024); leading open-source 2D models like DeepSeek-VL2 (Lu et al. 2024) and LLaVA-OneVision-7B (Li et al. 2024); and 3D-aware models including 3D-LLaVA (Deng et al. 2025) and Video-3D LLM (Zheng, Huang, and Wang 2025). A complete list of evaluated models is provided in the Supplementary Material (Section D).

Evaluation Windows. Our rigorous evaluation employs eight distinct window size categories to systematically vary the temporal interval between input frames. These categories include discrete intervals from 5 to 11, and a final category for all intervals of 12 frames or greater. For each specific category, all models are evaluated on both the forward and inverse tasks as clearly defined in our Dual-Task Paradigm.

Implementation Details. For each specific image pair, models are explicitly prompted to determine which frame is closer to task completion, with the output constrained to a single numerical value to focus on the core judgment. Performance is then measured by accuracy (%), calculated by comparing the model's output to the ground-truth answer.

4.2 Main Results

Our results reveal a trend: model performance is positively correlated with window size (Table 2). Tasks with larger window size, which present obvious state differences, primarily assess logical reasoning about overall task progress, and models generally perform well in this setting. Conversely, tasks with small window size demand meticulous discrimination of subtle changes to test fine-grained spatial perception. In this challenging scenario, even top-performing models struggle, highlighting a limitation across all evaluated models in their ability to capture and reason about subtle state changes in continuous task flows.

Our benchmark reveals a significant performance gap between proprietary and open-source models, with leading closed-source models like Gemini-2.5-Pro and GPT-4o performing exceptionally well. Notably, Gemini-2.5-Pro achieves a perfect 100% accuracy on tasks with the largest window size ($\geq$12), a stark contrast to the generally weaker performance of their open-source counterparts. Even top-tier open-source models like DeepSeek-VL2 and Video3D LLM lag significantly behind, highlighting the considerable challenge for the open-source community to bridge this gap.

Furthermore, we also identify significant bottlenecks in perceptual capabilities. Among open-source models, 2D-based architectures consistently outperform their 3D counterparts, suggesting that current 3D models struggle to effectively integrate spatial information for high-level semantic reasoning tasks. This challenge in fine-grained perception is not isolated to 3D models; it represents a widespread issue, as evidenced by the degraded performance of all models on

Model	WinSize							
	@5	@6	@7	@8	@9	@10	@11	@≥12
Human Perception	97.3	97.3	98.2	99.1	99.1	**100.0**	**100.0**	**100.0**
Closed-Source Models								
GPT-o4mini	70.0	70.3	72.8	67.7	60.0	67.5	67.7	70.3
GPT-o4mini-high	70.5	72.7	71.8	72.3	75.5	77.3	78.1	78.6
GPT-4o	71.1	71.7	76.7	**77.8**	**79.2**	79.2	79.2	79.2
GPT-o3	67.2	61.4	59.6	61.6	59.6	64.0	62.0	66.6
Gemini-2.5-Pro	**71.3**	**75.0**	**72.8**	75.0	77.2	80.9	**89.7**	**100.0**
Doubao-1.5-pro	64.0	76.0	56.0	46.0	56.9	**86.0**	90.0	96.0
Seed1.6	66.2	66.8	70.6	73.3	83.2	82.9	86.8	98.3
Open-Source Models (2D Perception)								
NVILA-8B	46.5	52.0	46.0	56.1	50.0	50.0	47.5	55.0
Paligemma	43.8	45.1	44.2	45.7	42.3	43.6	42.5	46.1
Qwen2.5-VL-3B	51.3	54.2	57.1	58.8	54.5	59.6	59.6	66.7
Qwen2.5-VL-7B	53.1	55.4	55.0	57.3	55.2	56.1	56.0	56.5
InternVL-8B	47.7	50.6	49.4	47.9	47.1	48.7	48.9	51.0
DeepSeek-VL2	64.4	66.9	65.2	67.8	67.0	69.1	77.4	81.0
LLaVA-OneVision-7B	54.8	55.4	53.5	49.8	52.1	50.6	52.7	50.4
Open-Source Models (3D Perception)								
3D-LLM	27.0	28.0	30.5	30.3	33.9	32.5	36.8	37.7
LL3DA	36.0	33.9	32.9	34.5	36.9	33.3	39.6	36.0
Chat-Scene	30.8	32.5	40.7	49.0	41.6	45.3	43.3	51.0
3D-LLaVA	43.2	46.1	46.4	47.9	49.6	48.8	48.0	47.9
Video-3D LLM	50.0	50.2	50.7	51.1	49.4	50.4	51.1	49.3

Table 2: Main results on the SpatialLogic-Bench. The results highlight the performance gap between closed-source and open-source models, as well as the general trend of higher accuracy with larger window sizes. The columns labeled with @ represent different window sizes (WinSize). Best results per column are in **bold**, and best results within each group are underlined.

small-window tasks. This universal difficulty in integrating subtle visual cues for accurate judgment highlights a fundamental inadequacy affecting nearly all current VLMs.

In summary, our findings highlight several critical areas for future VLM development: improving fine-grained visual discrimination, bridging the performance gap between open-source and proprietary models, and enhancing the synergy between 3D perception and high-level logical reasoning.

4.3 Analysis of Chronological Bias

As illustrated in Figure 4, the performance gap between forward and inverse tasks reveals significant differences in temporal reasoning robustness across models. A few models, notably Gemini-2.5-Pro and 3D-LLaVA, demonstrate remarkable balance. Gemini-2.5-Pro achieves a forward accuracy of 82.7% and an inverse accuracy of 77.7%, while 3D-LLaVA shows a forward accuracy of 45.8% and an inverse accuracy of 49.0%. Although their absolute accuracies differ, this balance implies a more robust temporal understanding that is independent of the reasoning direction.

However, the vast majority of models, such as GPT-4o, GPT-o4mini (OpenAI 2024), and Qwen2.5-VL (Bai et al. 2023), exhibit a clear "sequence bias," where their accuracy on forward tasks is considerably higher than on inverse tasks. For instance, GPT-4o's forward accuracy is 99.8%, while its inverse accuracy plummets to only 53.7%. This steep decline suggests a heavy reliance on temporal heuristics learned from conventional "front-to-back" training data. This issue is particularly acute in models like Paligemma (Google 2024) and InternVL-8B (Chen et al. 2024), whose inverse-task accuracies fall to 4.0% and 15.0% respectively—far below random chance—implying fundamental flaws in their logical reasoning abilities.

Notably, a few models exhibit an "inverse advantage", where their inverse accuracy far exceeds forward accuracy. This rare and intriguing phenomenon is further analyzed in detail in the Supplementary Material (Section A.5).

In summary, balanced performance across forward and inverse sequences is a crucial indicator of a model's advanced temporal reasoning capabilities. Currently, few models exhibit this robustness, as most remain heavily biased towards conventional "forward" logic. This finding highlights a clear direction for future optimization: training models on more diverse and non-chronological sequences is crucial for enhancing their generalization and robustness (Liu et al. 2021).

4.4 VLM Hallucination

Beyond chronological bias, we also observe instances of "hallucination" (Ji et al. 2023), where models generate action descriptions that directly contradict the visual evidence.

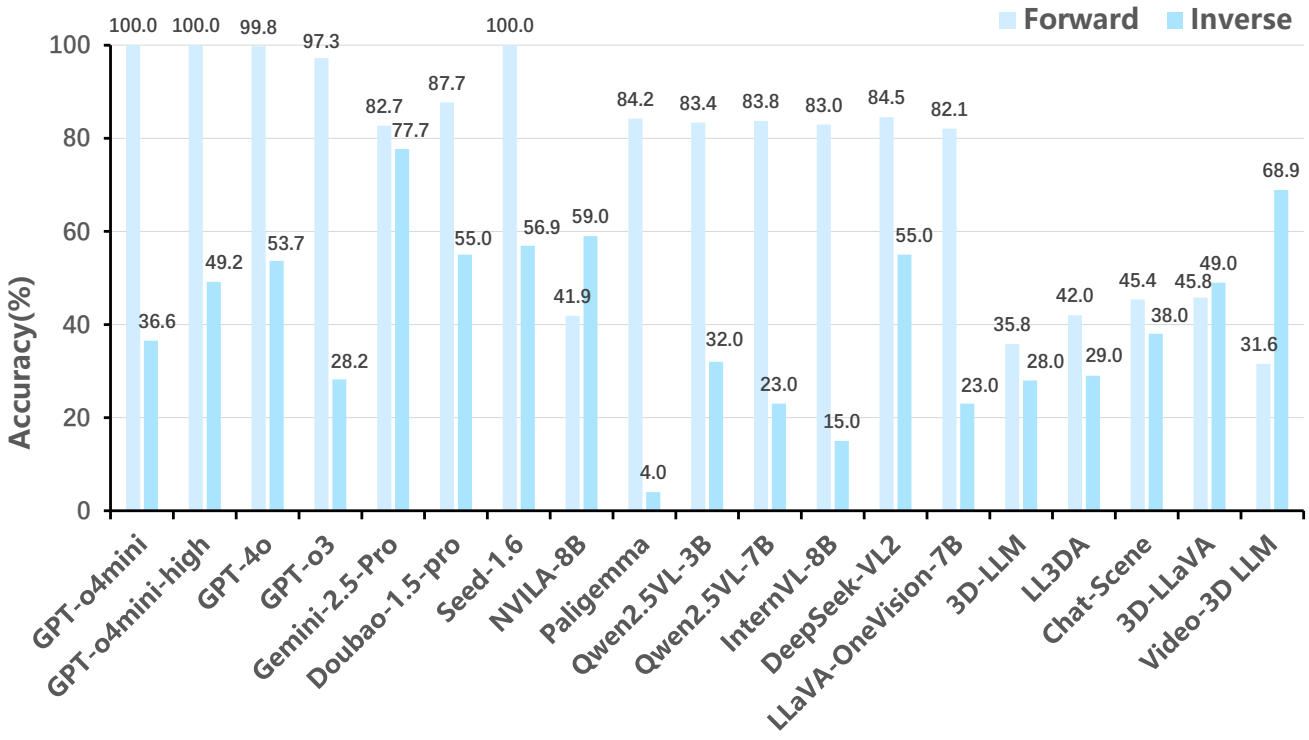

Figure 4: Comparison of model performance on forward and inverse tasks. The chart displays the accuracy of various models, revealing a significant performance gap between forward and inverse reasoning for most models.

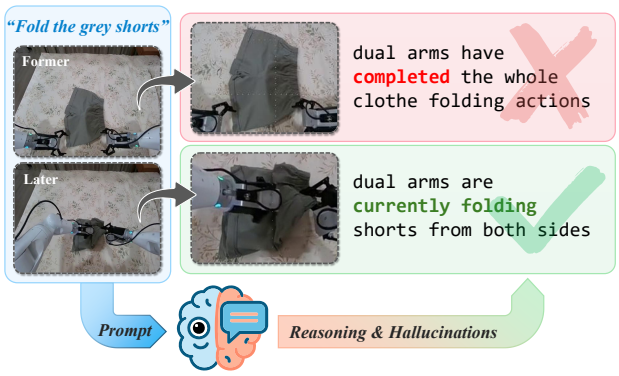

Figure 5: Failure cases illustrating hallucination, highlighting limitations in fine-grained visual understanding.

As illustrated in Figure 5, this failure is particularly prevalent in our challenging inverse tasks, where the conflict between a model's internal priors and the visual input can trigger such flawed reasoning; we discuss this mechanism in detail in the Supplementary Material (Section A.4). Ultimately, these cases demonstrate that even advanced VLMs suffer from fundamental limitations in fine-grained visual understanding that result in inaccurate perceptions rather than visual facts, a critical failure of visual grounding.

4.5 Conditional Reasoning

To further diagnose the root cause of model failures, we designed an ablation study to decouple logical reasoning from spatial perception. In this experiment, we provided models not only with the image pair and task description but also with an explicit textual analysis of the key state changes (i.e., the ground-truth chain of thought). This intervention is designed to effectively isolate the models' core logical reasoning capabilities from any perceptual ambiguities.

The results, presented in Table 3, reveal a stark divergence in failure modes. For leading closed-source models like GPT-4o, performance on the challenging inverse task improves dramatically. This suggests their primary limitation is a conquerable perceptual bottleneck; once visual ambiguity is removed, their logical engines perform robustly.

In stark contrast, many open-source models, such as Paligemma and InternVL-8B, fail catastrophically even with clear textual guidance. This demonstrates that their failure is not merely perceptual but stems from a fundamental deficit in logical processing. This reveals there is no single root cause for failure: some models struggle to see, while others struggle to reason. A detailed methodology and full results are available in the Supplementary Material (Section C).

5 Future Work

The limitations identified in our study motivate our primary future work: developing a novel training paradigm centered on a new Chain-of-Thought (CoT) (Wei et al. 2022)

Model	Analysis as Prompt			Analysis Excluded		
	Forward	Inverse	Avg.	Forward	Inverse	Avg.
Closed-Source Models						
GPT-o4mini	99.5	98.6	99.04	100.0	40.5	70.3
GPT-o4mini-high	98.2	96.2	97.1	100.0	57.2	78.6
GPT-4o	97.6	95.2	96.4	**100.0**	58.3	79.2
GPT-o3	-	-	-	**100.0**	33.3	66.6
Gemini-2.5-Pro	**100.0**	**98.0**	**99.0**	**100.0**	**100.0**	**100.0**
Doubao-1.5-pro	96.8	93.3	95.1	92.0	100.0	96.0
Seed1.6	98.6	96.7	97.7	**100.0**	96.5	98.3
Open-Source Models (2D Perception)						
NVILA-8B	47.6	85.4	66.5	46.4	63.6	55.0
Paligemma	89.8	12.2	51.0	83.8	8.3	46.1
Qwen2.5VL-7B	85.7	27.0	56.3	85.4	20.1	52.8
InternVL-8B	92.3	21.5	56.9	86.2	15.8	51.0
DeepSeek-VL2	84.7	20.6	52.7	93.1	69.0	81.1
LLaVA-OneVision-7B	84.7	32.5	58.62	80.0	20.8	50.4
Open-Source Models (3D Perception)						
3D-LLM	36.2	36.0	36.1	39.7	35.8	37.8
LL3DA	46.0	26.3	36.1	45.8	26.2	36.0
Chat-Scene	-	-	-	58.1	43.9	51.0
3D-LLaVA	47.4	65.1	56.2	47.3	48.5	47.9
Video-3D LLM	40.2	78.5	59.4	31.1	67.4	49.3

Table 3: Ablation study results comparing performance with and without the ground-truth analysis provided in the prompt. "Analysis as Prompt" refers to the new condition, while "Analysis Excluded" corresponds to the main results. All results are on the WinSize$\geq$12 setting. Best results within each group are underlined, and best overall results per column are in **bold**.

dataset. To construct this dataset efficiently, we will leverage our extensive video resources, expanding them quadratically by combining different sliding window sizes. We will then compare captioning methods to generate high-quality, structured reasoning steps for each training instance. This approach provides both a scalable and low-cost pathway to creating a large-scale CoT dataset that is specifically designed to enhance robust spatiotemporal reasoning.

This CoT-based training simulates the model's interaction with the world, functioning similarly to an offline reinforcement learning strategy (Levine et al. 2020) with dense rewards. By training on a pre-collected dataset of state comparisons, the model will learn to evaluate state quality and infer logical action sequences without needing real-time interaction. The ultimate goal is to systematically integrate the model's spatial perception and logical reasoning capabilities. The target model trained under this paradigm will be guided to first generate a potential action stream or thought process. It will then learn to critically evaluate the logical coherence of these steps before committing to a final decision. We anticipate this paradigm will not only enhance robustness and decision-making accuracy but also equip the model to generate structured reasoning to overcome hallucinations. This represents a crucial step towards developing agents that not only act correctly but can also explain why their actions are logical, a hallmark of true intelligence.

6 Conclusion

We have presented SpatialLogic-Bench, a large-scale benchmark designed to evaluate the crucial, yet largely undertested, ability of Vision-Language Models to reason about task-oriented spatiotemporal logic, moving beyond the simple recognition of static spatial facts. Leveraging a multi-scale evaluation system and an innovative dual-task paradigm on 28,250 real-world image pairs, we revealed a striking "performance cliff"—a catastrophic drop in accuracy on inverse-chronological tasks that contradicts the near-perfect scores often seen on forward tasks. This core finding provides compelling evidence that even state-of-the-art models heavily rely on superficial temporal heuristics rather than performing genuine causal reasoning about task progress. Furthermore, our analysis also exposed a universal struggle with fine-grained visual discrimination and, more profoundly, a fundamental deficit in logical processing itself—a weakness that better vision alone cannot solve. By systematically isolating and quantifying these distinct failure modes, SpatialLogic-Bench provides the community with a vital diagnostic tool that exposes the brittleness of current models. It charts a clear and actionable path forward: developing VLMs that can overcome these identified shortcomings to achieve the robust, grounded reasoning truly required for embodied intelligence in the physical world.

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Bu, Q.; Cai, J.; Chen, L.; Cui, X.; Ding, Y.; Feng, S.; Gao, S.; He, X.; Hu, X.; Huang, X.; et al. 2025. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Cheng, X.; Fu, D.; Wen, C.; Yu, S.; Wang, Z.; Ji, S.; Arora, S.; Jin, T.; Watanabe, S.; and Zhao, Z. 2025a. AHa-Bench: Benchmarking Audio Hallucinations in Large Audio-Language Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Cheng, X.; Fu, D.; Yang, X.; Fang, M.; Hu, R.; Lu, J.; Jionghao, B.; Wang, Z.; Ji, S.; Huang, R.; Li, L.; Chen, Y.; Jin, T.; and Zhao, Z. 2025b. OmniChat: Enhancing Spoken Dialogue Systems with Scalable Synthetic Data for Diverse Scenarios. arXiv:2501.01384.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Deng, J.; He, T.; Jiang, L.; Wang, T.; Dayoub, F.; and Reid, I. 2025. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3772–3782.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Fu, C.; Zhang, Y.-F.; Yin, S.; Li, B.; Fang, X.; Zhao, S.; Duan, H.; Sun, X.; Liu, Z.; Wang, L.; et al. 2024a. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Fu, D.; Cheng, X.; Li, L.; Yang, X.; Yang, L.; and Jin, T. 2025. PACHAT: Persona-Aware Speech Assistant for Multi-party Dialogue. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics.

Fu, D.; Cheng, X.; Yang, X.; Hanting, W.; Zhao, Z.; and Jin, T. 2024b. Boosting Speech Recognition Robustness to Modality-Distortion with Contrast-Augmented Prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3838–3847. New York, NY, USA: Association for Computing Machinery.

Gan, Z.; Li, L.; Li, C.; Wang, L.; Liu, Z.; Gao, J.; et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4): 163–352.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Google. 2024. Introducing PaliGemma, Gemma 2, and an Upgraded Responsible AI Toolkit. https://developers.googleblog.com/en/gemma-family-and-toolkit-expansion-io-2024/. Accessed: 2025-07-30.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2901–2910.

Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Deitke, M.; Ehsani, K.; Gordon, D.; Zhu, Y.; et al. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1): 32–73.

Levine, S.; Kumar, A.; Tucker, G.; and Fu, J. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*.

Liu, J.; Shen, Z.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-07-30.

Pfeiffer, J.; Rücklé, A.; Poth, C.; Kamath, A.; Vulić, I.; Ruder, S.; Cho, K.; and Gurevych, I. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 46–54. Association for Computational Linguistics.

Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9339–9347.

Shridhar, M.; Manuelli, L.; and Fox, D. 2023. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, 785–799. PMLR.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Yan, W.; Lin, W.; Guo, Z.; Wang, Y.; Feng, F.; Yang, X.; Wang, Z.; and Jin, T. 2025. Diff-Prompt: Diffusion-driven Prompt Generator with Mask Supervision. In *The Thirteenth International Conference on Learning Representations*.

Yang, S.; Xu, R.; Xie, Y.; Yang, S.; Li, M.; Lin, J.; Zhu, C.; Chen, X.; Duan, H.; Yue, X.; et al. 2025a. MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence. *arXiv preprint arXiv:2505.23764*.

Yang, X.; Cheng, X.; Duan, J.; Qiu, H.; Hong, M.; Fang, M.; Ji, S.; Zuo, J.; Hong, Z.; Zhang, Z.; et al. 2024a. AudioVSR: Enhancing Video Speech Recognition with Audio Data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15352–15361.

Yang, X.; Cheng, X.; Fang, M.; Qiu, H.; Ma, Y.; Lu, J.; Duan, J.; Cai, S.; Wang, Z.; Hu, R.; et al. 2025b. Multimodal Conditional Retrieval with High Controllability. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 3577–3585.

Yang, X.; Cheng, X.; Fu, D.; Fang, M.; Zuo, J.; Ji, S.; Zhao, Z.; and Tao, J. 2024b. Synctalklip: Highly synchronized lip-readable speaker generation with multi-task learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8149–8158.

Yang, X.; Xu, J.; Luan, K.; Zhan, X.; Qiu, H.; Shi, S.; Li, H.; Yang, S.; Zhang, L.; Yu, C.; et al. 2025c. OmniCam: Unified Multimodal Video Generation via Camera Control. *arXiv preprint arXiv:2504.02312*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Zheng, D.; Huang, S.; and Wang, L. 2025. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8995–9006.