POLYWORLD: BENCHMARKING MULTIMODAL REASONING OVER EVOLVING POLYMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have shown capabilities in visual understanding and scientific reasoning. However, existing benchmarks largely focus on static observations, property prediction, or knowledge-based reasoning, and rarely test whether models can reason about how a physical system evolves after an intervention and use the resulting state to decide what to do next. We introduce *PolyWorld*, a multimodal benchmark for physical reasoning over evolving polymers, where structural interventions lead to diverse changes in material properties and behavior. PolyWorld covers 3,181 polymer units and 8,472 polymer structures. We apply controlled interventions and molecular dynamics simulations to generate physical trajectories, and construct 4,178 multimodal questions spanning perception, reasoning, decision-making, and planning. Questions are grounded in polymer structures, physical trajectories, and intervention outcomes, with answers derived from construction records or computed physical quantities. This enables automatic verification without manual answer annotation. We evaluate 13 state-of-the-art MLLMs on PolyWorld. The best-performing model, GPT-5.6 Sol, reaches 44.3% overall accuracy, compared with substantially higher human performance. These results show that reasoning about evolving physical systems remains challenging for current multimodal models.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have demonstrated strong cross-modal reasoning over text, images, and other modalities, while MLLM-based agents further extend these capabilities through planning, tool use, and iterative interaction (OpenAI, 2025; Google, 2026a; Zhang et al., 2024; Wu et al., 2023; Li, 2025; Xi et al., 2025; Xie et al., 2025). These advances are increasingly applied to scientific workflows such as hypothesis generation, experiment planning, data analysis, and computational modeling (Bran et al., 2023; Yamada et al., 2025; Ghareeb et al., 2026). In materials science, multimodal models and scientific agents have begun to support characterization, property prediction, simulation, and design (Lai et al., 2025; Chaudhari et al., 2026; Tang et al., 2026). Polymers, in particular, have motivated recent efforts ranging from multimodal property prediction to tool-augmented design through structure generation and modification (Zhang & Yang, 2025; Huang et al., 2025; Nigam et al., 2026).

Existing benchmarks have begun to evaluate multimodal reasoning in materials science and polymer workflows (Alampara et al., 2025; Niyongabo Rubungo et al., 2025; Zhang et al., 2026; Lai et al., 2025; Liu et al., 2026). Yet, they focus primarily on static observations, property prediction, or characterization, with limited emphasis on how materials change after an intervention. In practical applications, changes in temperature, stress, or environment can alter a material's structure and properties, affecting its performance under different use conditions. For example, a polymer product that performs well at room temperature may exhibit substantially different properties at elevated temperatures. Models therefore need to reason beyond the current state: given an intervention, they must predict how the material evolves and how its properties change over time, providing a basis for materials discovery and design. For scientific agents, such predictions can further inform subsequent decisions and interventions (Liang et al., 2025; Zhu et al., 2026). This dynamic reasoning could make scientific discovery more efficient by helping agents select and prioritize informative experiments under limited budgets.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

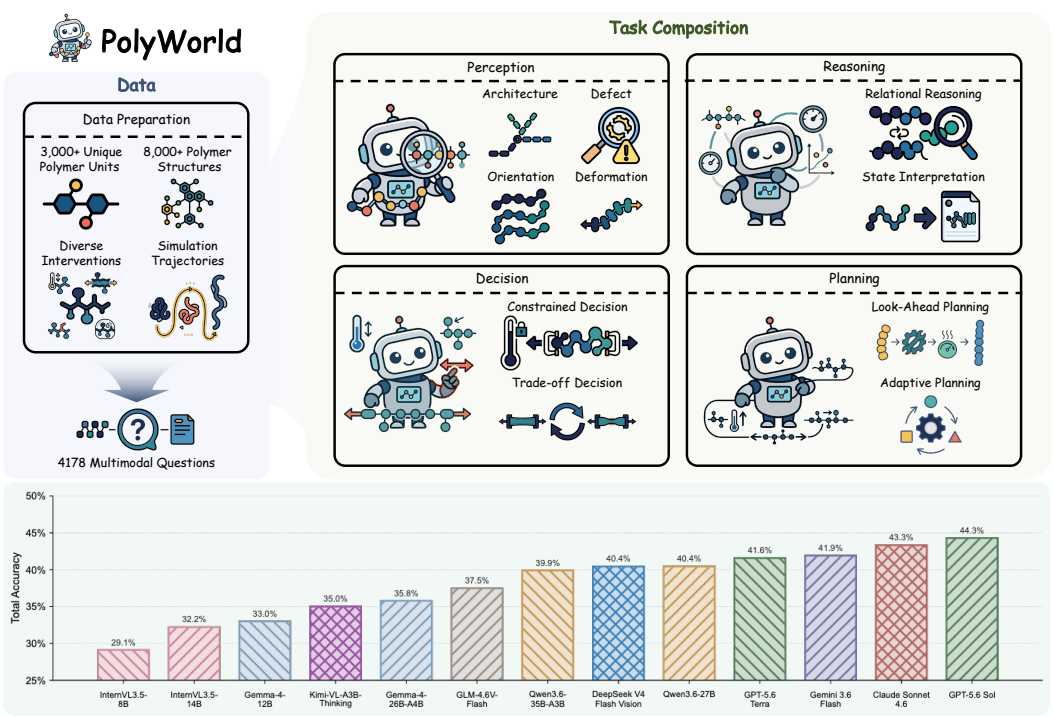


Figure 1: Overview of PolyWorld. Dataset construction (top left), the four evaluation dimensions—Perception, Reasoning, Decision, and Planning (top right), and overall accuracy of state-of-the-art open- and closed-source multimodal models on PolyWorld (bottom).

We introduce *PolyWorld*, a multimodal benchmark for physical reasoning over evolving polymers. Starting from 3,181 unique repeat units, we construct 8,472 3D polymer structures spanning linear, star, branched, and crosslinked architectures. We apply controlled physical interventions and generate changes over time to test whether models can recognize material states, reason about their evolution, and identify temporal relationships. For action-conditioned decision and planning, we further use molecular dynamics (MD) to generate 2,650 physical trajectories under controlled actions, testing whether models can anticipate their outcomes. These trajectories provide explicit evidence of how alternative actions affect subsequent physical evolution, enabling evaluation of decision and planning beyond static state understanding. Overall, PolyWorld contains 4,178 multimodal questions covering polymer structures, physical changes, and action-conditioned outcomes. The answers come directly from construction records, providing programmatically verifiable ground truth without manual answer annotation. We further conduct expert review to assess the quality of the benchmark, focusing on question clarity, and answer unambiguity.

We evaluate a broad range of state-of-the-art multimodal large language models on PolyWorld, spanning recent proprietary and open-source systems. Despite their strong performance on general multimodal tasks, current models struggle with the evolving polymer processes captured by our benchmark, with the best-performing model achieving only 44.28% accuracy across 4,178 questions. The challenge grows as models move from recognizing the current polymer state to predicting its evolution, making decisions, and planning future interventions. These results reveal a substantial gap between visual understanding and reliable reasoning about evolving polymer systems, highlighting the need for stronger multimodal scientific agents for polymer science. More broadly, MLLMs need stronger capabilities to reason about evolving physical systems and their consequences.

Our contributions are as follows:

- We introduce *PolyWorld*, a multimodal benchmark for reasoning over evolving polymer systems, constructed from diverse polymer structures, multimodal observations, and physical trajectories generated under controlled interventions.

Table 1: Comparison of PolyWorld with existing scientific and materials-science benchmarks. **Counterfactual Evolution** denotes reasoning about alternative physical evolutions induced by different controlled interventions from the same initial state.

Benchmark	Multimodal	Temporal	Intervention	Counterfactual Evolution
MaCBench (Alampara et al., 2024)	✓	×	×	×
MSQA (Cheung et al., 2025)	×	×	×	×
Matcha (Lai et al., 2025)	✓	×	×	×
MatSciBench (Zhang et al., 2026)	✓	×	×	×
PolyReal (Liu et al., 2026)	✓	×	×	×
MatVQA (Wu et al., 2026)	✓	×	×	×
AtomWorld (Lv et al., 2025)	✓	×	✓	×
PolyWorld (Ours)	✓	✓	✓	✓

- We develop a unified set of tasks spanning perception, reasoning, decision-making, and planning, covering state changes, future evolution, intervention effects, and multi-step action selection across diverse polymer systems and physical processes.
- We conduct a comprehensive evaluation of recent proprietary and open-source multimodal models, revealing substantial gaps in their ability to reason about evolving polymer processes and make reliable decisions based on multimodal observations.

2 Related Work

Materials Science Agents. Multimodal and language-based models are increasingly applied to materials science for property prediction, representation, and discovery (Boyar et al., 2025; Belouadi et al., 2026; Tang et al., 2026). Recent agents integrate language-based reasoning with domain-specific models and tools for materials research. LLaMP combines hierarchical LLM agents with materials databases and computational tools for retrieval, structure editing, and simulation (Chiang et al., 2024), while MatAgent combines LLM-guided reasoning with generative and predictive models for inorganic materials design (Bazgir et al., 2025). MARS further connects multiple LLM agents with domain-specific tools and robotic experiments for closed-loop materials discovery (Shi et al., 2026). These efforts extend language models toward increasingly interactive materials research, but leave limited emphasis on reasoning about how materials evolve under controlled interventions.

Polymer Modeling and AI. Recent work has applied language and multimodal models to polymer property prediction, molecular representation, and structure generation (Qiu et al., 2024; Qiu & Sun, 2024; Huang et al., 2025; Zhang & Yang, 2025). PolyNC uses a polymer-specific prompting framework with a T5-based model for multitask property prediction and classification (Qiu et al., 2024), while PolyTAO explores property-conditioned polymer generation for on-demand reverse design (Qiu & Sun, 2024). Polymer-Agent integrates LLM-based reasoning with polymer property prediction and generative design tools (Nigam et al., 2026). These approaches demonstrate the growing role of AI in understanding and designing polymers, and motivate benchmarks that evaluate whether such models can reason about how polymers evolve under controlled interventions.

Materials Science Benchmarks. Existing benchmarks evaluate materials-science AI across multimodal reasoning, scientific workflows, and structure editing (Alampara et al., 2024; Lv et al., 2025; Lai et al., 2025; Cheung et al., 2025; Zhang et al., 2026; Liu et al., 2026; Wu et al., 2026). MaCBench evaluates multimodal reasoning across chemistry and materials research workflows, from data extraction to experiment execution and analysis (Alampara et al., 2024). MatVQA focuses on visual scientific reasoning over materials images, including microscopy and diffraction patterns (Wu et al., 2026), while MatSciBench covers broad materials-science reasoning across major subfields (Zhang et al., 2026). PolyReal evaluates multimodal reasoning in real-world polymer science workflows (Liu et al., 2026), and AtomWorld evaluates structure editing and spatial reasoning over crystalline materials (Lv et al., 2025). However, existing benchmarks largely focus on established materials knowledge, experimental observations, and documented results, rather than outcomes induced by controlled interventions. PolyWorld complements these settings by introducing physical trajectories generated under controlled interventions, enabling evaluation of how

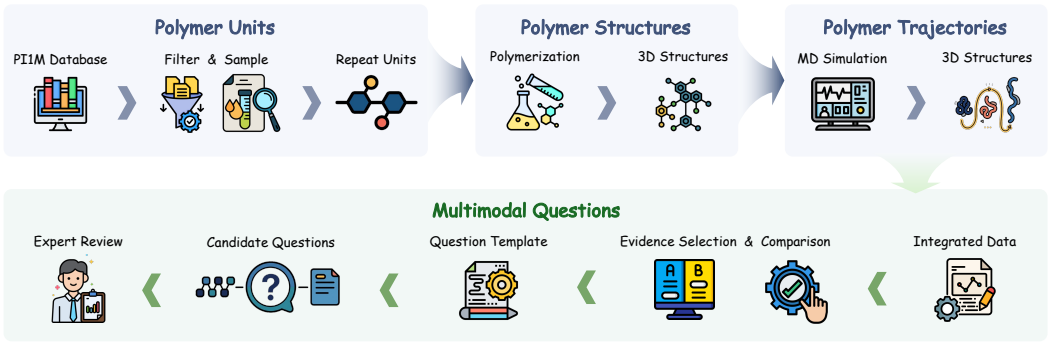


Figure 2: Overview of PolyWorld construction. The top shows the construction of the polymer world, from polymer units and repeat-unit sampling to polymer structures, 3D configurations, and physical trajectories generated by molecular dynamics. The bottom shows the construction of multimodal questions from expert review and task-specific templates, followed by evidence selection, comparison, and A/B testing for verification and integration.

models anticipate and reason about the resulting physical evolution. A detailed comparison with existing benchmarks is provided in Table 1.

3 PolyWorld Benchmark

We introduce PolyWorld, a multimodal benchmark for physical reasoning over evolving polymers. The benchmark covers diverse polymer structures and controlled physical trajectories, with tasks spanning material perception, physical reasoning, decision-making, and planning. It contains 4,178 questions grounded in the underlying structures, interventions, and simulation results.

3.1 Data Construction

An overview of the question generation pipeline is shown in Fig. 2. We start from PI1M, a large-scale polymer database containing approximately one million p-SMILES representative of real-world polymers (Ma & Luo, 2020). We treat the polymer representations as candidate repeat units (RUs) and perform basic structure cleaning, removing duplicate entries and structures that cannot be parsed by RDKit (RDKit). To ensure chemical diversity, we represent the remaining structures using Morgan fingerprints (Rogers & Hahn, 2010) and apply MaxMin sampling (Ashton et al., 2002) to select structures that are diverse in fingerprint space. The resulting RUs are then polymerized into four architectures, including linear, branched, star, and network structures, across different polymerization scales to capture diverse polymer configurations while keeping the computational cost of subsequent simulations manageable.

To obtain diverse and physically grounded trajectories, we therefore use molecular dynamics (MD) simulations. To generate physically grounded MD trajectories, we reuse the force-field parameterization and molecular dynamics components of ADEPT-PolyGraphMT (Alosious et al., 2026). Given each initial polymer structure, we run LAMMPS-based simulations under controlled conditions and interventions (Thompson et al., 2022), varying temperature, lateral pressure, and mechanical stretching to produce multiple trajectories from the same initial state. With over 600 CPU hours of computation, we generate more than 2,000 trajectories. This provides a diverse and controlled source of physical evidence for constructing questions about system states, their evolution, and the effects of interventions. We instantiate multimodal questions using task-specific templates and derive gold answers from structural and physical evidence using predefined rules. Human experts then perform a final quality check to ensure the generated questions are well-formed and scientifically meaningful. Fig. 1 provides an overview of how we construct the physical world and the multimodal benchmark. More details are provided in the Appendix A.1.

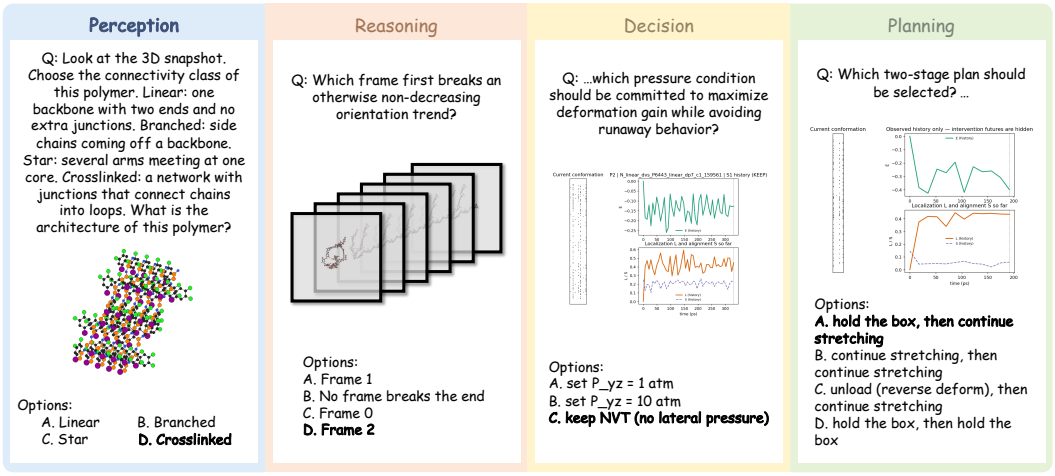


Figure 3: Representative examples from PolyWorld across different tasks. Solving these questions requires understanding the temporal evolution of scientific processes and reasoning about the effects of interventions on evolving physical systems.

3.2 Task Design

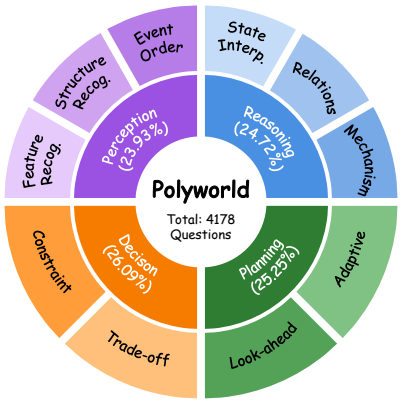
PolyWorld evaluates multimodal physical reasoning at four levels: Perception, Reasoning, Decision, and Planning. These levels progress from identifying the current material state, to interpreting its evolution, to selecting interventions, and finally to planning over future states. Different task types are instantiated from different combinations of structural observations, trajectory frames, physical measurements, and intervention outcomes. Fig. 4 summarizes the benchmark composition and key statistics, including task distribution, polymer structures and trajectories, intervention settings, modalities, and question formats.

Perception. Perception questions test whether models can extract physically meaningful information from polymer structures and trajectory observations. This capability provides the physical state needed for subsequent reasoning and action selection. We consider state, structure, and temporal information. *Feature Recognition* covers chain orientation and deformation localization, while *Structure Recognition* covers architecture, defects, and topology. *Event Ordering* further tests whether models can identify the temporal order of physical events from trajectory observations.

Reasoning. Reasoning questions test whether models can connect observations and physical measurements to understand how a material evolves. This capability goes beyond recognizing individual states and requires models to integrate evidence across frames and modalities. We consider three forms of reasoning. *State Interpretation* focuses on identifying the physical state or transition represented by an observation. *Relational Reasoning* requires comparing quantities, trends, or events across observations. *Mechanistic Reasoning* further requires connecting structural and physical evidence to explain an observed outcome.

Decision. Decision questions test whether models can use physical understanding to select appropriate interventions. Given the current state of a polymer, it may need to choose how to change the temperature or pressure while respecting physical constraints or balancing competing outcomes. We consider two task types: *Constrained Decision*, which requires selecting an intervention that satisfies a specified physical constraint, and *Trade-off Decision*, which requires balancing competing physical outcomes.

Planning. Planning questions extend decision-making from a single intervention to a sequence of interventions. Models must anticipate future states and choose actions accordingly, especially when experiments are costly and resources are limited. We consider *Look-ahead Planning*, which requires selecting a sequence of interventions based on their future consequences, and *Adaptive Planning*, which requires updating the next intervention after observing an intermediate state.



Benchmark Statistics	
Questions	4,178
Polymer Units	3,181
Polymer Structures	8,472
Physical Trajectories	2,650
Polymer Architectures	4
Intervention Settings	15
Primary Tasks	4
Modalities	Image / Video / Text
Question Format	Multiple-choice / Numerical

Figure 4: **PolyWorld benchmark composition and statistics.** (a) Distribution of the 4,178 questions across four primary tasks. (b) Key statistics of the benchmark, including its polymer units, structures, physical trajectories, modalities, and question formats.

4 Experiments

We evaluate state-of-the-art MLLMs on PolyWorld to assess their ability to reason about evolving physical systems. The experiments characterize current model capabilities across different task types, modalities, and levels of reasoning.

4.1 Experimental Setup

4.1.1 Baselines.

We evaluate a broad range of recent MLLMs, covering both proprietary and open-source models. The selection spans different model sizes, architectures, and reasoning configurations, from frontier models to open-source systems with lightweight and larger-scale variants. This provides a broad evaluation of MLLMs and allows us to examine how performance varies across model families.

Proprietary models. We evaluate models from several major model families, including GPT-5.6 Terra and GPT-5.6 Sol (OpenAI, 2026), Gemini 3.6 Flash (Google, 2026b), and Claude Sonnet 4.6. (Anthropic, 2026) These models represent recent frontier MLLMs from different providers and cover different model scales and reasoning configurations. Together, they provide a strong reference for assessing the capabilities of current proprietary systems on evolving physical reasoning.

Open-source models. We evaluate InternVL3.5-8B and InternVL3.5-14B (Wang et al., 2025), Gemma 4 12B and Gemma 4 26B A4B (Team et al., 2026), Kimi-VL-A3B-Thinking (Team et al., 2025), GLM-4.6V-Flash (Hong et al., 2025), Qwen3.6-27B and Qwen3.6-35B-A3B (Yang et al., 2025), and DeepSeek-V4-Flash-Vision (Xu et al., 2026). These models cover a range of model scales, providing open-source baselines for comparison with proprietary systems.

Human Expert Baseline. To provide a human reference point, we evaluate PolyWorld with domain experts familiar with polymer science and molecular simulation. Each expert independently answers a subset of the benchmark questions under the same task format as the MLLMs. On the same subset, human experts achieve substantially higher accuracy than the best-performing model. Details are provided in Appendix A.4.

4.1.2 Implementation details.

All models are evaluated in a zero-shot setting without fine-tuning or few-shot demonstrations. We use a temperature of 0 and a maximum output length of 32K tokens for all models. For models with a dedicated reasoning or thinking mode, we enable it with the default medium reasoning budget. For API-based models, we use the standard inference settings exposed by their official APIs and set the temperature and output length whenever these parameters are available. All open-source

Table 2: Comparison of accuracy (%) of frontier MLLMs on PolyWorld. We report results across four task categories: Perception, Reasoning, Decision, and Planning, together with human baseline. SI, RR, and MR denote State Interpretation, Relational Reasoning, and Mechanistic Reasoning; D1/D2 denote Constrained and Trade-off Decision; P1/P2 denote Look-ahead and Adaptive Planning. Best and second-best results within each group are shown in **bold** and underline, respectively.

Model	Overall	Perception			Reasoning			Decision		Planning	
		Feature	Structure	Event	SI	RR	MR	D1	D2	P1	P2
Human Expert	71.25	–	–	–	–	–	–	–	–	–	–
Closed-source models											
GPT-5.6 Sol	**44.28**	55.03	**56.06**	**64.78**	52.63	**44.12**	**84.21**	44.67	24.60	50.98	23.85
Claude Sonnet 4.6	43.32	**56.51**	45.53	61.64	47.81	40.11	82.46	40.45	**27.80**	**59.10**	22.82
Gemini 3.6 Flash	41.91	44.97	55.27	58.49	51.32	37.83	71.93	**50.12**	27.66	43.46	**27.18**
GPT-5.6 Terra	41.57	**56.51**	45.73	60.38	**53.51**	38.90	71.93	46.90	25.91	46.02	23.85
Open-source models											
Qwen3.6-27B	**40.45**	51.18	**53.88**	**61.64**	52.63	36.76	45.61	32.75	29.99	38.80	33.59
DeepSeek V4 Flash Vision	40.43	52.37	42.54	54.09	**56.58**	**38.24**	**56.14**	**44.17**	27.07	46.47	23.59
Qwen3.6-35B-A3B	39.90	53.55	46.92	55.35	47.81	36.23	35.09	35.24	29.69	43.61	32.31
GLM-4.6V-Flash	37.48	52.96	43.14	40.25	48.25	36.50	29.82	29.28	24.75	**48.57**	24.36
Gemma 4 26B A4B	35.78	24.56	30.62	54.09	38.16	33.96	35.09	40.20	**31.15**	44.51	**35.64**
Kimi-VL-A3B-Thinking	34.99	36.09	33.20	49.06	50.00	32.49	36.84	34.00	27.51	38.35	34.87
Gemma 4 12B	32.98	**54.44**	30.02	38.99	42.11	27.81	15.79	35.98	28.97	31.43	29.49
InternVL3.5-14B	32.22	40.53	30.62	57.86	39.04	26.74	40.35	28.54	25.04	39.10	26.67
InternVL3.5-8B	29.10	41.42	25.05	55.35	31.58	25.13	35.09	35.73	22.27	29.77	22.31

models except DeepSeek-V4-Flash-Vision are deployed locally using vLLM (Kwon et al., 2023), while DeepSeek-V4-Flash-Vision and proprietary models are accessed through their official APIs. We use the same prompts and task inputs across models to ensure a consistent evaluation protocol.

4.2 Main Results

Human experts remain substantially ahead. As shown in Table 2, human experts achieve an overall accuracy of 71.25%, compared with 44.28% for the best-performing model, GPT-5.6 Sol, leaving a gap of 26.97%. Claude Sonnet 4.6 and Gemini 3.6 Flash achieve 43.32% and 41.91%, respectively. These results show that frontier MLLMs still fall substantially behind human experts in the dynamic scientific understanding, temporal reasoning, and decision-making evaluated by PolyWorld, indicating that the benchmark remains challenging for current systems. These findings also highlight the need for general-purpose MLLMs to reason and act under evolving physical conditions.

Closed-source models lead overall. The top four models are closed-source systems: GPT-5.6 Sol (44.28%), Claude Sonnet 4.6 (43.32%), Gemini 3.6 Flash (41.91%), and GPT-5.6 Terra (41.57%). The strongest open-source models, Qwen3.6-27B and DeepSeek V4 Flash Vision, achieve 40.45% and 40.43%, respectively. Closed-source models perform better on most evaluated subtasks, while on some remaining subtasks, open-source models achieve higher accuracy. The differences are particularly noticeable on intervention-based tasks, where models may favor different interventions under the same underlying physical conditions.

Performance varies substantially across task categories. Perception is relatively stronger, with the best accuracy reaching 64.78% on Event Ordering, 56.51% on Feature Recognition, and 56.06% on Structure Recognition. Within Reasoning, Mechanistic Reasoning reaches 84.21%, while State Interpretation and Relational Reasoning reach 56.58% and 44.12%, respectively. Decision shows a similar gap between Constrained Decision (50.12%) and Trade-off Decision (31.15%). For Planning, Look-ahead Planning reaches 59.10%, compared with 35.64% for Adaptive Planning. Overall, models perform substantially worse on relational reasoning, trade-off decision-making, and adaptive planning under evolving conditions.

Larger variants generally achieve higher accuracy. Within the same model family, larger variants tend to perform better. InternVL3.5-14B achieves 32.22%, compared with 29.10% for InternVL3.5-8B, while Gemma 4 26B A4B reaches 35.78%, compared with 32.98% for Gemma 4 12B. GPT-5.6 Sol also outperforms GPT-5.6 Terra, with accuracies of 44.28% and 41.57%, respectively. The Qwen3.6 results show that this relationship is not strictly monotonic across different model configurations: Qwen3.6-27B achieves 40.45%, slightly above the 39.90% of Qwen3.6-35B-A3B.

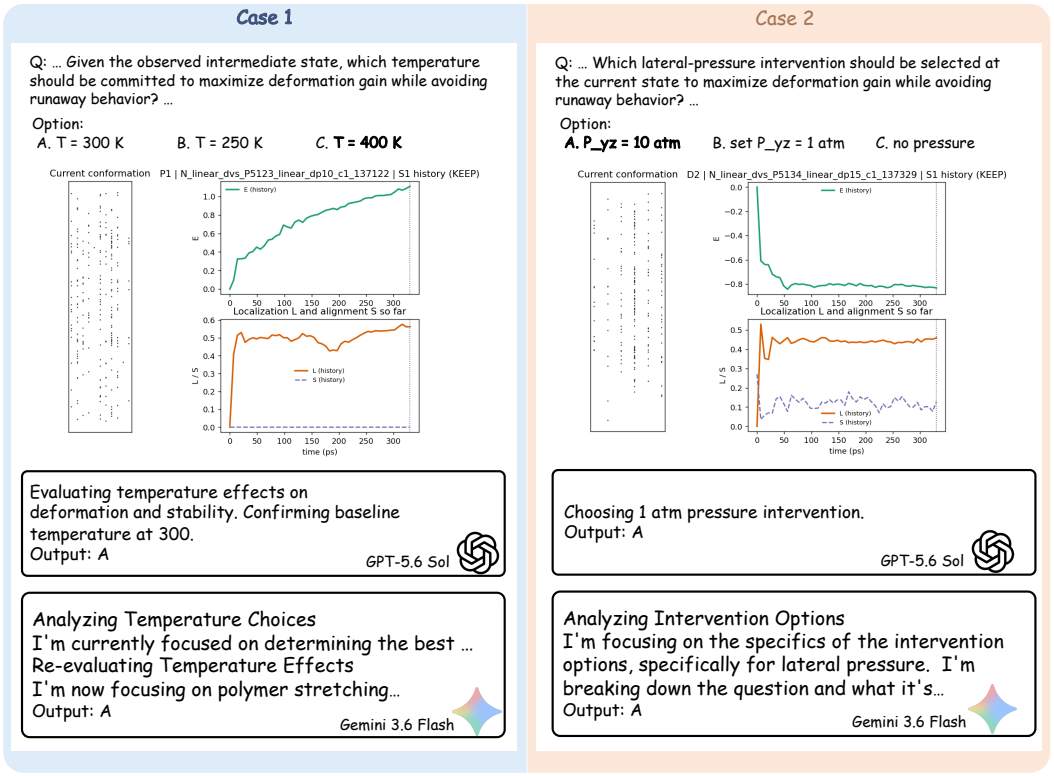


Figure 5: Representative failure cases from PolyWorld. Each case presents the question context, responses, and final answers from GPT-5.6 Sol and Gemini 3.6 Flash.

PolyWorld reveals persistent challenges in reasoning over evolving scientific processes. Current MLLMs still struggle to understand how scientific systems evolve over time, especially in relational reasoning, trade-off decision-making, and adaptive planning. While closed-source models lead overall, open-source models remain competitive on several subtasks. These results point to important gaps in temporal understanding, causal and mechanistic reasoning, long-horizon planning, and decision-making under evolving conditions.

4.3 Analysis

We further inspect incorrect predictions across Perception, Reasoning, Decision, and Planning to characterize recurring failure patterns. In Perception, errors often involve fine-grained structural or temporal distinctions that require identifying subtle changes in the observed system. For example, models may miss a local chain break in a dense structure, assign an incorrect topological role to an atom, or confuse the first occurrence of an event with a later stage where the same phenomenon becomes more pronounced. These patterns may reflect the difficulty of grounding fine-grained structural and temporal evidence in visually complex observations, consistent with recent findings that fine-grained perception and visual grounding remain challenging for multimodal models (Yan et al., 2026). In Reasoning, errors more often arise when models need to distinguish closely related physical factors or spatial patterns. For example, models may conflate strong molecular orientation with localized deformation, even when the two properties are explicitly distinguished in the question. These cases indicate that models can struggle to maintain fine-grained distinctions when multiple physical attributes are present simultaneously.

In Decision and Planning, we observe recurring patterns in intervention selection across different physical states and intervention settings. GPT-5.6 Sol and Gemini 3.6 Flash exhibit similar incorrect choices in several representative cases, as shown in Fig. 5. Beyond these cases, related errors also recur across groups of questions that share the same underlying physical concept. For example, models repeatedly confuse *chain break* with the absence of a defect, while another recurring pattern

Table 3: Ablation results on thinking mode and text-only input. The left table compares NoThink and Think accuracy, while the right table compares full and text-only input. Δ denotes the absolute change in percentage points (pp).

Thinking-mode ablation

Model	NoThink	Think	Δ
GLM-4.6V-Flash	33.30%	37.65%	+4.35
Qwen3.6-27B	36.55%	40.45%	+3.90
Gemma 4 26B A4B	34.35%	35.78%	+1.43

Text-only ablation

Model	Full	Text-only	Δ
Qwen3.6-27B	40.45%	28.60%	-11.85
Sonnet 4.6	43.32%	20.06%	-23.26
GPT-5.6 Terra	41.57%	16.01%	-25.56

is confusing *uniform stretching* with a *localized band*. These observations suggest that model failures are not limited to individual question instances, while the variation across models indicates that different models may struggle with different physical distinctions.

4.4 Ablation

Thinking mode. We conduct an ablation study to examine the effect of explicit thinking on PolyWorld performance. As shown in Table 3, enabling the Think setting improves all three evaluated models across different model families, but the magnitude of the gain varies across models. GLM-4.6V-Flash improves from 33.30% to 37.65%, corresponding to a 4.35 percentage-point gain, while Qwen3.6-27B improves from 36.55% to 40.45%, a 3.90-point gain. Gemma 4 26B A4B shows a smaller but consistent improvement from 34.35% to 35.78%, with a 1.43-point gain. These results show that enabling the thinking mode provides a consistent benefit across different model families, although the magnitude of the improvement varies.

Text-only input. We further conduct a text-only ablation to assess the contribution of visual inputs to PolyWorld. Removing the visual inputs reduces accuracy for all three models, but the magnitude of the degradation differs substantially. Qwen3.6-27B drops from 40.45% to 28.60% (11.85 percentage points), whereas Claude Sonnet 4.6 and GPT-5.6 Terra drop by 23.26 and 25.56%, respectively, from 43.32% to 20.06% and from 41.57% to 16.01%. The larger degradation for Claude Sonnet 4.6 and GPT-5.6 Terra is partly associated with their substantially higher rates of empty responses under text-only input, while Qwen3.6-27B produces no empty responses in this setting.

5 Discussion

5.1 Limitations

PolyWorld has several limitations. First, its physical trajectories are generated using molecular dynamics simulations, which provide controlled and reproducible physical processes but may not capture all aspects of real-world experiments. Second, the intervention space is defined by the simulation protocols and conditions considered in the benchmark. While PolyWorld covers multiple structural and environmental interventions, it does not exhaust the possible ways of manipulating polymer systems or the range of longer-horizon processes that may arise in practice. Third, PolyWorld focuses on polymer systems, and the extent to which the observed challenges generalize to other scientific domains remains to be studied.

5.2 Conclusion

We introduced PolyWorld, a multimodal benchmark for physical reasoning over evolving polymers. Built from 3,181 polymer units, 8,472 polymer structures, and 2,650 physical trajectories, PolyWorld evaluates 4,178 questions spanning perception, physical reasoning, decision-making, and planning. Evaluation of state-of-the-art MLLMs reveals a substantial gap from human experts: while models perform reasonably well on perception-oriented tasks, they remain limited in interpreting evolving physical states, understanding mechanisms, and making effective decisions and plans. These results highlight the challenge of reasoning about how physical systems evolve under interventions and motivate models that can connect physical observations, dynamics, and actions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with software implementation and paper preparation. AI-assisted coding tools, including Cursor and Codex, were used during the development and refinement of experimental code. Generative AI tools were also used to create and refine scientific icons and visual elements, and to improve the readability and presentation of the manuscript. All AI-assisted code, experimental results, figures, and manuscript content were reviewed and verified by the authors. We take full responsibility for the final content of this work, including all claims, results, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

PolyWorld is constructed using molecular dynamics simulations and publicly available scientific knowledge. The human evaluation only involves researchers answering benchmark questions, and no sensitive personal information is collected.

REPRODUCIBILITY STATEMENT

We describe the benchmark construction, trajectory generation, task definitions, question formats, and evaluation protocol in the paper and supplementary material. Details of the simulation setup are provided in Appendix A.1, question-format composition in Appendix A.3, model variants in Appendix A.2, and the human evaluation protocol in Appendix A.4. These details are intended to make the benchmark construction and reported evaluation procedures reproducible.

REFERENCES

Nawaf Alampara, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, Mara Schilling-Wilhelmi, NM Anoop Krishnan, and Kevin Maik Jablonka. Macbench: a multimodal chemistry and materials science benchmark. In *Proceedings of the 38th conference on neural information processing systems (neurips 2024)*, volume 3, 2024.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature computational science*, 5(10):952–961, 2025.

Sobin Alosious, Yuhan Liu, Jiaxin Xu, Gang Liu, Renzheng Zhang, Meng Jiang, and Tengfei Luo. Adept-polygraphmt: automated molecular simulation and multi-task multi-fidelity machine learning for polymer property generation and prediction. *Digital Discovery*, 2026.

Anthropic. Introducing claude sonnet 4.6, 2026. URL `https://www.anthropic.com/news/claude-sonnet-4-6`.

Mark Ashton, John Barnard, Florence Casset, Michael Charlton, Geoffrey Downs, Dominique Gorse, John Holliday, Roger Lahana, and Peter Willett. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quantitative Structure-Activity Relationships*, 21(6):598–604, 2002.

Adib Bazgir, Yuwen Zhang, et al. Matagent: a human-in-the-loop multi-agent llm framework for accelerating the material science discovery cycle. *AI for Accelerated Materials Design-ICLR 2025*, pp. 1–35, 2025.

Jonas Belouadi, Tamy Boubekeur, and Adrien Kaiser. Multimat: Multimodal program synthesis for procedural materials using large multimodal models. In *International Conference on Learning Representations*, volume 2026, pp. 17507–17528, 2026.

Onur Boyar, Indra Priyadarsini, Seiji Takeda, and Lisa Hamada. Llm-fusion: A novel multimodal fusion model for accelerated material discovery. *arXiv preprint arXiv:2503.01022*, 2025.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Akshat Chaudhari, Janghoon Ock, and Amir Barati Farimani. Modular large language model agents for multi-task computational materials science. *Communications Materials*, 7(1):131, 2026.

Jerry Junyang Cheung, Shiyao Shen, Yuchen Zhuang, Yinghao Li, Rampi Ramprasad, and Chao Zhang. Msqa: Benchmarking llms on graduate-level materials science reasoning and knowledge. *arXiv preprint arXiv:2505.23982*, 2025.

Yuan Chiang, Elvis Hsieh, Chia-Hong Chou, and Janosh Riebesell. Llamp: Large language model made powerful for high-fidelity materials knowledge retrieval and distillation. *arXiv preprint arXiv:2401.17244*, 2024.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Dmytro Shved, Gavin J Gyimesi, Jon M Laurent, Samantha M Wright, Muhammed T Razzak, et al. A multi-agent system for automating scientific discovery. *Nature*, pp. 1–3, 2026.

Google. Gemini 3.5 flash-lite, 2026a. URL `https://deepmind.google/models/gemini/flash-lite/`.

Google. Gemini 3.6 flash, 2026b. URL `https://deepmind.google/models/model-cards/gemini-3-6-flash/`.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Qi Huang, Yedi Li, Lei Zhu, Qibin Zhao, and Wenjie Yu. Unified multimodal multidomain polymer representation for property prediction. *npj Computational Materials*, 11(1):153, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Zhengzhao Lai, Youbin Zheng, Zhenyang Cai, Haonan Lyu, Jinpu Yang, Hongqing Liang, Yan Hu, and Benyou Wang. Can multimodal llms see materials clearly? a multimodal benchmark on materials characterization, 2025. URL `https://arxiv.org/abs/2509.09307`.

Xinzhe Li. A review of prominent paradigms for llm-based agents: Tool use, planning (including rag), and feedback learning. In *Proceedings of the 31st international conference on computational linguistics*, pp. 9760–9779, 2025.

Haotong Liang, Chuangye Wang, Heshan Yu, Dylan Kirsch, Rohit Pant, Austin McDannald, A Gilad Kusne, Ji-Cheng Zhao, and Ichiro Takeuchi. Real-time experiment-theory closed-loop interaction for autonomous materials science. *Science advances*, 11(27):eadu7426, 2025.

Wanhao Liu, Weida Wang, Jiaqing Xie, Suorong Yang, Jue Wang, Benteng Chen, Guangtao Mei, Zonglin Yang, Shufei Zhang, Yuchun Mo, et al. Polyreal: A benchmark for real-world polymer science workflows. *arXiv preprint arXiv:2604.02934*, 2026.

Taoyuze Lv, Alexander Chen, Fengyu Xie, Chu Wu, Jeffrey Meng, Dongzhan Zhou, Yingheng Wang, Bram Hoex, Zhicheng Zhong, and Tong Xie. Atomworld: A benchmark for evaluating spatial reasoning in large language models on crystalline materials. *arXiv preprint arXiv:2510.04704*, 2025.

Ruimin Ma and Tengfei Luo. Pi1m: a benchmark database for polymer informatics. *Journal of Chemical Information and Modeling*, 60(10):4684–4690, 2020.

Vani Nigam, Achuth Chandrasekhar, and Amir Barati Farimani. Polymer-agent: Large language model agent for polymer design. *Journal of Chemical Information and Modeling*, 66(9):5145–5154, 2026.

Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. Llm4mat-bench: benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6(2):020501, 2025.

OpenAI. Introducing gpt 5, 2025. URL https://openai.com/index/introducing-gpt-5/.

OpenAI. Introducing gpt-6 sol and luna, 2026. URL https://openai.com/index/introducing-gpt-6-sol-and-luna/.

Haoke Qiu and Zhao-Yan Sun. On-demand reverse design of polymers with polytao. *Npj Computational Materials*, 10(1):273, 2024.

Haoke Qiu, Lunyang Liu, Xuepeng Qiu, Xuemin Dai, Xiangling Ji, and Zhao-Yan Sun. Polync: a natural and chemical language model for the prediction of unified polymer properties. *Chemical Science*, 15(2):534–544, 2024.

RDKit. Rdkit: Open-source cheminformatics. https://www.rdkit.org.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Tongyu Shi, Yutang Li, Zhanlong Wang, Wenhe Xu, Guolai Jiang, Dawei Dai, Jie Zhou, Hao Huang, Rui He, Seeram Ramakrishna, et al. Knowledge-driven autonomous materials research via collaborative multi-agent and robotic system. *Matter*, 9(2), 2026.

Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steven Farrell, Benjamin Erichson, Michael W Mahoney, Andy Nonaka, and Zhi Jackie Yao. A multimodal large language model for materials science. *Nature Machine Intelligence*, pp. 1–14, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Aidan P Thompson, H Metin Aktulga, Richard Berger, Dan S Bolintineanu, W Michael Brown, Paul S Crozier, Pieter J In't Veld, Axel Kohlmeyer, Stan G Moore, Trung Dac Nguyen, et al. Lammps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer physics communications*, 271:108171, 2022.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE international conference on big data (bigdata)*, pp. 2247–2256. IEEE, 2023.

Sifan Wu, Huan Zhang, Yizhan Li, Farshid Effaty, Hongyuan Mei, Amirreza Ataei, and Bang Liu. Seeing beyond words: Matvqa for challenging visual-scientific reasoning in materials science. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23599–23614, 2026.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China information sciences*, 68(2):121101, 2025.

Junlin Xie, Zhihong Chen, Ruifei Zhang, and Guanbin Li. Large multimodal agents: A survey. *Visual Intelligence*, 3(1):24, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Hao Yan, Xingchen Liu, Hao Wang, Zhenbiao Cao, Handong Zheng, Liang Yin, Xinxing Su, Zihao Chen, Jihao Wu, Minghui Liao, et al. Visuriddles: Fine-grained perception is a primary bottleneck for multimodal large language models in abstract visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 136213–136234, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *Findings of the association for computational linguistics: ACL 2024*, pp. 12401–12430, 2024.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, et al. Matscibench: Benchmarking the reasoning ability of large language models in materials science. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 12868–12879, 2026.

Tianren Zhang and Dai-Bei Yang. Multimodal machine learning with large language embedding model for polymer property prediction. *Chemistry of Materials*, 37(18):7002–7013, 2025.

Jiayi Zhu, Longhan Zhang, Yizhang Zhu, Xiaotian Lin, Yifan Wu, Shimin Di, Bang Liu, Yuyu Luo, and Tongyi Zhang. A survey of agentic materials science and engineering: where are we and where are we going? *Journal of Materials Informatics*, 6(2):N–A, 2026.

A Appendix

A.1 Trajectory Generation

Simulation setup. We generate physical trajectories using a LAMMPS-based molecular dynamics pipeline built on the force-field parameterization and simulation components of ADEPT-PolyGraphMT (Alosious et al., 2026). ADEPT serves as the simulation and force-field parameterization stack rather than a question-generation system. Following its parameterization convention, we primarily use GAFF2 and fall back to the Dreiding force field when required elements are not covered. The resulting polymer systems are simulated in LAMMPS under controlled thermodynamic and mechanical conditions. We initialize each episode from an initial polymer configuration, including linear, branched, star, and crosslinked architectures, and record both atomic trajectories and physical observables throughout the simulation.

Action-conditioned trajectories. For action-conditioned decision and planning tasks, we construct counterfactual branches from a shared physical state. For the nested action setting, each episode first follows a shared prefix consisting of uniaxial stretching to a strain of approximately eight percent. From the resulting checkpoint, we apply one of three actions: *continue*, *hold*, or *unload*, and simulate each branch for approximately 250 picoseconds. We then branch each resulting state again using the same three actions, producing nine complete trajectories for each episode. Each trajectory contains an atomic trajectory file and time-dependent observables, including energy, localization, and structural measurements. The hidden future trajectories allow questions to expose only the state available at the decision point while deriving the gold answer from the subsequent simulated outcomes.

Temperature and pressure trajectories. We additionally construct intervention forks for temperature- and pressure-conditioned questions. Starting from the same checkpoint, corresponding to a baseline temperature of 300 Kelvin, we independently modify the temperature or lateral pressure and continue the simulation after a short re-equilibration period. Temperature interventions use 250, 300, and 400 Kelvin, while pressure interventions use controlled pressures of 1 and 10 atmospheres. The subsequent mechanical action is kept fixed so that the effect of the selected intervention

can be isolated. Gold answers are determined from predefined rules based on the resulting physical outcomes, such as maximizing deformation-related gain while avoiding runaway behavior.

Trajectory-based question construction. These simulations provide controlled alternative evolutions from shared initial states, allowing us to construct questions that require models to reason about intervention effects and their future consequences. The trajectory generator produces both atomic configurations and physical observables, which are used by our question-mining pipeline to instantiate decision and planning questions. Visual inputs are rendered from the simulated configurations and observables, while gold answers are derived from predefined comparison rules over the corresponding trajectory branches. Importantly, the trajectories used in PolyWorld are generated specifically for this benchmark rather than taken from an existing public trajectory dataset.

A.2 Model Versions

To ensure reproducibility, we specify the exact model variants used in our evaluation. We report the model names used by the corresponding model providers or repositories and retain version suffixes when they identify a specific released model. For open-source models, we evaluate the corresponding publicly released checkpoints rather than an unspecified model family.

For Kimi-VL, we use `Kimi-VL-A3B-Thinking`, downloaded from its official Hugging Face repository (Team et al., 2025). For DeepSeek, we use `DeepSeek-V4-Flash-Vision-Exp`, retaining the `-Exp` suffix to identify the specific experimental vision model used in our evaluation (Xu et al., 2026). For the remaining models, we similarly report the specific model variants used in our experiments, including model size and mixture-of-experts configuration when applicable. These details help distinguish models within the same family, which may differ in architecture, parameter configuration, reasoning mode, or release version.

A.3 Question Format Composition

PolyWorld contains 4,178 questions in total, consisting primarily of multiple-choice questions (MCQs) with a small number of numerical fill-in-the-blank questions. Specifically, 2,841 questions (68.00%) use four-option MCQs, 1,237 questions (29.61%) use three-option MCQs, and 100 questions (2.39%) use numerical fill-in-the-blank questions.

The question format varies across the four primary task categories. Perception contains 900 four-option MCQs and 100 fill-in-the-blank questions, for a total of 1,000 questions. Reasoning contains 1,033 four-option MCQs. Decision contains 472 four-option and 618 three-option MCQs, while Planning contains 436 four-option and 619 three-option MCQs. Thus, all Reasoning questions and most Perception questions use four-option MCQs, whereas Decision and Planning include both three- and four-option formats.

The 100 fill-in-the-blank questions all belong to *Structure Recognition* under Perception, specifically the *Topology* task. These questions require models to infer discrete structural quantities, such as the number of chain ends, and provide the answer as an integer rather than selecting from predefined options. This format is used when a numerical structural answer is more appropriate than a fixed set of candidate choices.

A.4 Human Evaluation Details

To provide an independent human reference for model performance, we randomly sample 80 questions from PolyWorld for human evaluation. We use a subset to keep the evaluation feasible for domain researchers while providing a common set of questions for comparison between humans and models. The subset is not intended to provide a statistically representative estimate of human performance on the full benchmark. Two PhD students with materials-science backgrounds and experience in polymer science and molecular simulation independently answer 40 questions each, using the same task format provided to the MLLMs. Across the 80 questions, they answer 57 correctly, corresponding to an accuracy of 71.25%. We also evaluate a selected model on the same 80-question subset for comparison. The model answers 47 questions correctly, achieving an accuracy of 58.75%. This comparison provides a subset-level reference for contextualizing model performance against human performance. Table 4 summarizes the results.

Setting	Correct	Accuracy
Human experts (subset)	57/80	71.25%
Best model on same subset	47/80	58.75%

Table 4: Human evaluation on the 80-question subset.