CRYSTALBENCH DYNAMICS: A MULTIMODAL BENCHMARK FOR DYNAMIC REASONING IN CRYSTAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding crystal behavior requires connecting structural perturbations to symmetry changes, relaxation responses, and finite-temperature evolution. Existing materials-science benchmarks evaluate knowledge, property prediction, multimodal interpretation, and structure manipulation, but rarely evaluate these connections within a unified framework. We introduce **CrystalBench Dynamics**, a simulation-grounded benchmark comprising 4,016 questions across 18 tasks in text-only and multimodal settings, derived from 1,500 real crystal structures. Our pipeline applies five types of structural perturbation: lattice strain, vacancy creation, atomic substitution, interstitial insertion, and atomic displacement. The perturbed structures then undergo geometry relaxation and finite-temperature molecular dynamics. The tasks span four capability families: Perception, Symmetry Reasoning, Symmetry-Response Reasoning, and Dynamics Prediction, with reference answers computed directly from structures, optimization records, and simulation trajectories. Across the evaluated language and vision-language models, the strongest model achieves only 51.96% task-macro accuracy on the Full Track, more than 20 percentage points below human experts. Enabling thinking improves Gemma-4-31B-IT from 31.26% to 43.04%, demonstrating the benefits of inference-time reasoning while leaving a substantial performance gap. These findings highlight the difficulty current LLMs and MLLMs face in connecting structural perturbations to symmetry changes and subsequent physical evolution, motivating explicit evaluation of these capabilities.

1 INTRODUCTION

Understanding crystalline materials requires connecting structural perturbations to the physical responses they induce. Strain, defects, and atomic displacements can alter local coordination and crystallographic symmetry, while subsequent relaxation and thermal motion may produce further structural changes. Reasoning about these processes requires models to interpret crystal structures, distinguish physical states, and anticipate their evolution. As large language models (LLMs) and multimodal large language models (MLLMs) become scientific assistants, evaluating these capabilities is essential to assessing their usefulness in materials modeling.

Existing benchmarks address several foundations of this reasoning. MaScQA evaluates materials knowledge and scientific question answering (Zaki et al., 2024), while LLM4Mat-Bench assesses property prediction from crystal representations (Rubungo et al., 2024). MaCBench and MatCha examine the interpretation of scientific and materials-characterization images (Alampara et al., 2025; Lai et al., 2025), and AtomWorld evaluates explicit crystal-structure manipulation and spatial reasoning (Lv et al., 2025). Building on these capabilities, an important question remains insufficiently evaluated: can models connect a controlled perturbation to the resulting symmetry changes, structural relaxation, and finite-temperature evolution? Addressing this question requires evaluating both crystallographic reasoning and the ability to relate structural changes to subsequent physical responses.

We introduce **CrystalBench Dynamics**, a benchmark designed around this connection. Starting from 1,500 real crystal structures, our generation pipeline applies five types of perturbation—lattice

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.

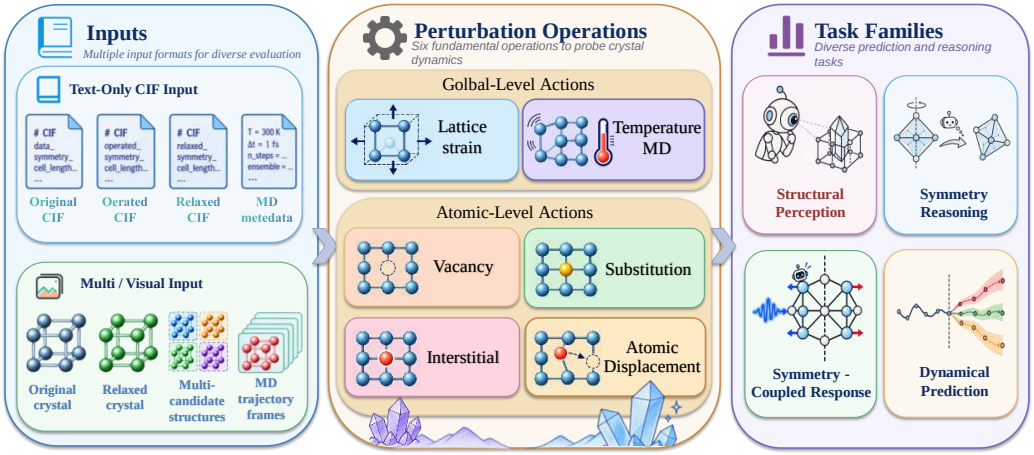


Figure 1: Overview of CrystalBench Dynamics. Starting from textual and visual crystal representations, the generation pipeline applies controlled structural perturbations, relaxation, and finite-temperature simulation to construct 18 tasks across four capability families: Perception, Symmetry Reasoning, Symmetry–Response Reasoning, and Dynamics Prediction.

strain, vacancies, substitutions, interstitial defects, and atomic displacements—followed by geometry relaxation and finite-temperature molecular dynamics. Reference answers are computed directly from the resulting structures, optimization records, and simulation trajectories. This pipeline links each intervention to its simulated consequences, providing a common basis for evaluating structural interpretation and prediction.

As illustrated in Figure 1, CrystalBench Dynamics comprises 4,016 questions across 18 tasks in text-only and multimodal settings. The tasks are organized into four capability families: *Perception*, *Symmetry Reasoning*, *Symmetry-Response Reasoning*, and *Dynamics Prediction*. These families distinguish the reasoning required to interpret structures, analyze symmetry changes, relate symmetry to relaxation responses, and predict dynamical outcomes; they do not impose a strict hierarchy of difficulty. Textual CIF representations and rendered crystal structures provide complementary views of the underlying systems, enabling evaluation across different forms of structural input.

Our evaluation reveals substantial limitations in current models' ability to reason about crystal evolution. The strongest model achieves only 51.96% task-macro accuracy on the Multimodel Track. Additional inference-time reasoning can improve performance: enabling thinking raises Gemma-4-31B-IT from 31.26% to 43.04%. Nevertheless, even the best-performing models remain well below human experts, with pronounced difficulties connecting symmetry changes to relaxation responses and dynamical outcomes. These results identify reasoning about the physical consequences of structural perturbations as a key challenge for current LLMs and MLLMs.

Our main contributions are:

- We introduce **CrystalBench Dynamics**, a benchmark comprising 4,016 questions across 18 tasks in text-only and multimodal settings, organized into four capability tasks.
- We develop a simulation-grounded generation pipeline that connects controlled crystal perturbations with structural relaxation and finite-temperature molecular dynamics, producing programmatically verifiable evaluation targets.
- We systematically evaluate LLMs and MLLMs, demonstrating the benefits of inference-time reasoning while revealing substantial remaining gaps in symmetry-aware response reasoning and dynamics prediction.

Table 1: Comparison with representative materials-science benchmarks. CrystalBench Dynamics combines symmetry reasoning with controlled perturbations, relaxation responses, and dynamical evolution.

Benchmark	Capability	Multi.	Crystal	Symm.	Perturb.	Relax.	Dyn.
MaScQA (Zaki et al., 2024)	Materials QA	✗	✗	✗	✗	✗	✗
LLM4Mat-Bench (Rubungo et al., 2024)	Property Prediction	✗	✓	✗	✗	✗	✗
MaCBench (Alampara et al., 2025)	Multimodal Reasoning	✓	✗	✗	✗	✗	✗
OPENXRD (Vosoughi et al., 2026)	Symmetry QA	✗	✓	✓	✗	✗	✗
AtomWorld (Lv et al., 2025)	Structure Editing	✗	✓	✗	✓	✗	✗
EvoMD-LLM (Tang et al., 2026)	Trajectory Prediction	✗	✗	✗	✗	✗	✓
Ours	**Symmetry-Aware Crystal Dynamics**	✓	✓	✓	✓	✓	✓

2 Related Work

Materials-science benchmarks. Materials-science benchmarks increasingly assess capabilities beyond factual knowledge, including property prediction, multimodal interpretation, and scientific tool use. MaScQA evaluates domain knowledge and numerical calculation (Zaki et al., 2024), while LLM4Mat-Bench assesses property prediction from compositions, textual descriptions, and CIFs (Rubungo et al., 2024). MatSciBench broadens the evaluation of materials reasoning, and MatTools examines the use of computational tools (Zhang et al., 2025; Liu et al., 2025). Multimodal benchmarks such as MaCBench, MatCha, and MatVQA cover scientific image interpretation, materials characterization, and structure–property reasoning (Alampara et al., 2025; Lai et al., 2025; Wu et al., 2025). MaterialFigBench and OmniMatBench further expand figure-based evaluation across materials-science domains (Yoshitake et al., 2026; Liu et al., 2026).

Crystallographic reasoning. Crystallography-focused benchmarks emphasize explicit structural representations. AtomWorld evaluates spatial reasoning through CIF-based structure editing (Lv et al., 2025), OPENXRD assesses crystallographic knowledge and contextual reasoning (Vosoughi et al., 2026), and CrystalXRD-Bench evaluates peak indexing using CIFs and XRD patterns (Xu et al., 2026). These benchmarks assess the interpretation and manipulation of crystal structures, motivating further evaluation of whether models can connect structural interventions to subsequent physical responses.

Scientific agents and dynamical prediction. Materials-specific models and agents support crystallographic data and computational workflows. LLaMat incorporates crystallographic representations (Ahlawat et al., 2026), while LLaMP supports database access and the orchestration of simulations, including molecular dynamics (Chiang et al., 2025). EvoMD-LLM addresses dynamical prediction by modeling molecular-species evolution from symbolic reactive-MD trajectories (Tang et al., 2026). These efforts advance workflow automation and trajectory-based prediction, but leave open the joint evaluation of crystallographic reasoning and post-perturbation structural evolution.

As summarized in Table 1, CrystalBench Dynamics links controlled perturbations, symmetry changes, structural relaxation, and finite-temperature evolution within a common evaluation framework. Its tasks assess whether models can relate successive structural states and anticipate the consequences of an intervention, unifying structural interpretation and physical-response reasoning.

3 CrystalBench Dynamics

3.1 Benchmark Overview and Task Organization

CrystalBench Dynamics evaluates whether models can reason about crystal symmetry under controlled perturbations and subsequent physical evolution. It connects changes in structural organization to the responses that emerge during relaxation and finite-temperature dynamics. The benchmark combines five types of structural perturbation, lattice strain, vacancy creation, atomic substitution, interstitial insertion, and atomic displacement, with geometry relaxation and finite-temperature molecular dynamics. This framework supports the evaluation of symmetry-related reasoning across local changes in atomic environments, global lattice distortions, and thermal evolution.

Table 2: Capability and task organization of CrystalBench Dynamics. **[T]** and **[V]** denote text-only and visual/multimodal tasks, respectively.

	Perception	**Symmetry Reasoning**	**Symmetry–Response Reasoning**	**Dynamics Prediction**
Scientific capabilities	• Visual evidence • Defect/geometry recognition	• Space-group identification • Symmetry-change reasoning • Directional reasoning	• Relaxation outcome • Symmetry–stability coupling	• Optimization outcome • Thermal evolution
Tasks	1. **[V]** Defect Location 2. **[V]** Defect Intensity 3. **[V]** Operation Type 4. **[V]** Relaxed Volume	1. **[T]** Space Group 2. **[T]** Symmetry Change 3. **[T]** Strain Action 4. **[V]** Relaxation Direction	1. **[T]** Relaxation Basin 2. **[T]** Lowest Energy 3. **[T]** Maximum RMSD 4. **[T]** Fragility Direction 5. **[V]** Distortion Intensity	1. **[T]** Optimizer Steps 2. **[T]** Instability 3. **[T]** MD RMSD 4. **[V]** Thermal Volume 5. **[V]** Thermal Stability

Task organization. Table 2 organizes the 18 tasks into four complementary families: perceiving structural changes, interpreting symmetry and geometric transformations, relating perturbed structures to relaxation responses, and predicting optimization and thermal outcomes. Together, these families connect structural recognition with reasoning about crystal evolution. Some tasks directly assess crystallographic symmetry, while others examine the geometric evidence and physical responses that support a broader understanding of structural change.

Perception assesses whether models can extract structural information from rendered crystal structures. *Defect Location*, *Defect Intensity*, and *Operation Type* evaluate the ability to identify the location, extent, and type of a structural modification. *Relaxed Volume* tests the comparison of per-atom volumes across relaxed structures. These tasks establish whether models can recognize local defects and geometric differences that inform subsequent symmetry and response reasoning.

Symmetry reasoning. evaluates crystallographic symmetry and its relationship to geometric transformations. *Space Group* tests symmetry identification in individual structures, while *Symmetry Change* assesses how symmetry differs across structural transformations. *Strain Action* and *Relaxation Direction* examine lattice deformations and the direction of structural change. Together, these tasks connect the characterization of individual configurations with the interpretation of transformations between them.

Symmetry–Response Reasoning evaluates whether models can relate perturbed structures to their relaxation outcomes. *Relaxation Basin* tests inference of the basin reached during relaxation. *Lowest Energy* and *Max RMSD* assess comparisons of energetic outcomes and the magnitude of relaxation-induced atomic rearrangement, respectively. *Fragility Direction* and *Distortion Intensity* examine the direction and severity of structural responses. This family probes the connection between structural organization and physical response beyond identifying symmetry changes alone.

Dynamics prediction assesses numerical optimization outcomes and finite-temperature structural evolution. *Optimizer Steps* tests the prediction of optimization effort, while *Instability* evaluates the anticipation of structural instability. *MD RMSD*, *Thermal Volume*, and *Thermal Stability* assess predictions of atomic displacements, volume changes, and structural stability under thermal perturbations. These tasks complement explicit symmetry analysis by testing whether models can anticipate structural persistence and change beyond the initial configuration.

Modality-aware evaluation. The benchmark contains 18 tasks: 10 text-only tasks and 8 visual/multimodal tasks. Tasks are grouped by their role in evaluating crystal symmetry dynamics, with input modality treated as a separate evaluation condition. Visual tasks therefore contribute not only to perception but also to symmetry reasoning, relaxation-response assessment, and thermal prediction. This organization supports analysis of whether model limitations arise in extracting structural evidence, interpreting symmetry transformations, or relating structural organization to physical evolution.

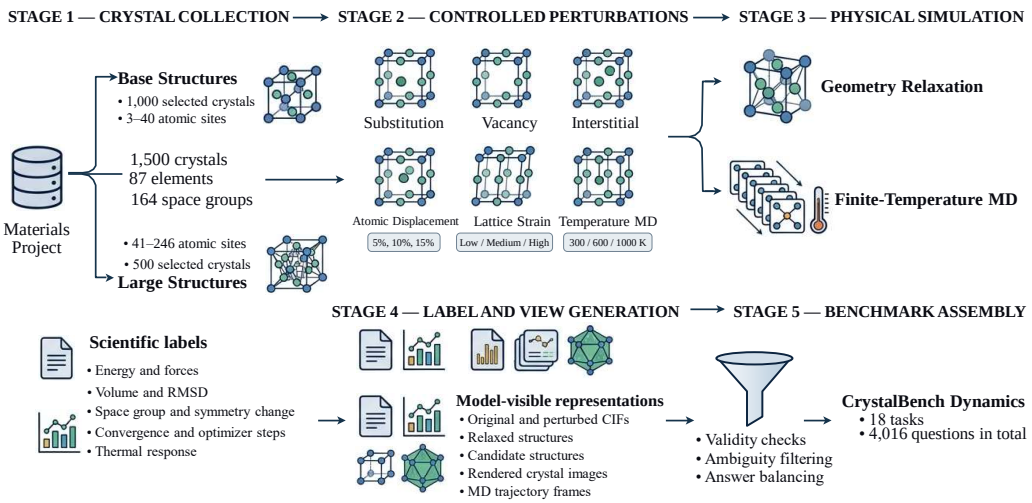


Figure 2: Data generation pipeline of CrystalBench Dynamics. Structures sampled from Materials Project are subjected to controlled perturbations, followed by MACE-based geometry relaxation or finite-temperature molecular dynamics. Simulation-derived labels are converted into text-only and multimodal questions, followed by validity filtering, ambiguity removal, and answer balancing.

3.2 Benchmark Generation Pipeline

CrystalBench Dynamics is constructed through a simulation-grounded pipeline that transforms crystal structures into controlled perturbations, physical trajectories, scientific labels, and task-specific questions. Figure 2 summarizes the complete procedure, while Tables 3 and 4 report the structural coverage and simulation configuration.

Crystal collection. We construct CrystalBench Dynamics from Materials Project and we construct custom subsets from the Materials Project. The base subset contains 1,000 structures with 3–40 atomic sites, sampled from 117,578 candidate materials. To increase structural complexity, we additionally introduce 500 crystals with 41–246 sites, selected from a separate pool of 34,833 candidates. The resulting collection contains 1,500 structures spanning 87 elements, 1,489 chemical systems, and 164 space groups. This combination preserves broad coverage of relatively compact crystals while extending the benchmark to structures with substantially larger unit cells.

Table 3: Structure selection and coverage in CrystalBench Dynamics.

Statistic	Value
Candidate structures	152,411
Selected structures	1,500
Base structures (3–40 sites)	1,000
Large structures (41–246 sites)	500
Unique elements	87
Chemical systems	1,489
Space groups	164

Controlled perturbations. We apply five types of structural perturbation: atomic substitution, vacancy creation, interstitial insertion, atomic displacement, and lattice strain. Finite-temperature molecular dynamics further probes thermal responses. For the four site-based operations, low-, medium-, and high-intensity perturbations target 5%, 10%, and 15% of the atomic sites, respectively. Lattice strain spans isotropic, uniaxial, biaxial, shear, and general symmetric deformation modes, with normal and shear strain amplitudes assigned corresponding low, medium, and high intensity levels.

Geometry relaxation and molecular dynamics. We relax perturbed structures using the medium variant of MACE-MP-0 in float64 precision, coupled with the ASE BFGS optimizer and `FrechetCellFilter`. Optimization stops when the maximum force falls below $0.03\,\mathrm{eV\,\AA^{-1}}$ or the 1,000-step limit is reached. We generate finite-temperature trajectories at 300, 600, and 1,000 K using fixed-cell NVT dynamics with a Berendsen thermostat. Each simulation runs for 5,000 steps

Table 4: Simulation protocol used for benchmark generation.

Setting	Configuration
Potential	MACE-MP-0, medium model, float64, dispersion disabled
Structural operations	Substitution, vacancy, interstitial, displacement, and strain
Site perturbations	5%, 10%, and 15% target-site fractions
Strain amplitudes	1–5% normal strain and 0.5–3% shear strain
Geometry optimizer	ASE BFGS with `FrechetCellFilter`
Convergence	Maximum force below $0.03\,\mathrm{eV\,\AA^{-1}}$; at most 1,000 steps
MD ensemble	Fixed-cell NVT with a Berendsen thermostat
Temperatures	300, 600, and 1,000 K
MD duration	5,000 steps at 1 fs per step, corresponding to 5 ps
Thermostat coupling	100 fs
Frame sampling	Every 100 steps; 52 stored frames per completed trajectory

with a timestep of 1 fs, and structures are saved every 100 steps. Full simulation settings are provided in Table 4.

Generation scale. The pipeline performs 22,500 geometry relaxations, of which 22,290 converge under the prescribed criterion. It additionally produces 4,480 completed MD trajectories containing 232,960 stored frames and 22.4 million integration steps. The MD jobs require 23.6 and 43.2 hours for the base and large subsets, respectively, when distributed over eight NVIDIA H100 80 GB GPUs, corresponding to 534.6 allocated GPU-hours. The retained CPU job timestamps indicate approximately 247.1 hours of wall-clock time for geometry relaxation; CPU core-hours were not recorded.

Question construction. Geometry relaxation provides energies, forces, volumes, optimizer steps, convergence states, relaxed structures, and displacement-based labels. Molecular dynamics additionally provides temperature traces, trajectory frames, RMSDs, and thermal-response statistics. These outputs are converted into CIF-based text inputs and rendered visual/multimodal inputs. Task-specific questions undergo validity checks, ambiguity filtering, and answer balancing before inclusion.

The final benchmark contains 18 tasks and 4,016 questions. The base-structure subset contributes 2,861 questions, and the large-structure extension contributes 1,155 questions. Across both subsets, 2,660 questions use text-only inputs and 1,356 use visual/multimodal inputs.

4 Experiments

4.1 Experimental Setup

4.1.1 Models

We evaluate four open-weight multimodal models: Qwen3.8-27B (Qwen Team, 2026), InternVL3.5-38B (Wang et al., 2025), Gemma-4-12B-IT (Gemma Team, 2026), and Gemma-4-31B-IT (Gemma Team, 2026). We additionally evaluate GPT-5.5 (OpenAI, 2026), Claude Sonnet 5 (Anthropic, 2026), and Gemini 3.1 Pro Preview (Google Gemini Team, 2026) as proprietary multimodal models, together with DeepSeek V4 Pro (DeepSeek-AI, 2026) on the text-only track. All models with complete text and visual coverage are evaluated on both the 10-task Text Track and the 18-task Multimodel Track. DeepSeek V4 Pro is reported only on the Text Track.

The main comparison uses direct inference for each model, except GPT-5.5, which is evaluated with low reasoning effort. For Gemma-4-31B-IT, we additionally report its native thinking mode to measure the effect of inference-time reasoning under the same benchmark. The expert evaluation was conducted by researchers with expertise in computational chemistry, standard simulation methods, and the interpretation of characterization data (Appendix C).

4.1.2 EVALUATION PROTOCOL

Inputs. No model is fine-tuned on CrystalBench Dynamics. Each model receives the benchmark instruction, candidate answers, and either geometry-only CIF content or the associated crystal images. Provenance information, simulation logs, energies, optimizer statistics, derived target values, and gold labels are removed from the model-visible input. For multimodal tasks, images are supplied in the order specified by the benchmark record.

Inference and scoring. We collect one response for every model–question pair using the model's official interface and chat template. Because the evaluated systems expose different reasoning controls and serving requirements, context limits, output budgets, and decoding parameters are configured per model and documented in the Appendix. Each question requires a single option label. A prediction is correct only when the extracted final label matches the gold answer; missing, malformed, unavailable, or truncated answers are counted as incorrect. Raw responses and generation metadata are retained for auditing.

4.1.3 METRICS

Task-macro accuracy. Accuracy is computed separately for each task after pooling correct counts and question counts across the base and large-structure subsets. A capability-family score is the unweighted mean of its constituent task accuracies. The Text-Track and Multimodel-Track scores are subsequently computed as unweighted means over their 10 and 18 tasks, respectively. This task-macro aggregation prevents tasks with more questions from dominating the benchmark score.

Reported statistics. We report capability-family accuracy, overall task-macro accuracy, $\Delta_{\mathrm{Rand.}}$, and N_{win}. Here, $\Delta_{\mathrm{Rand.}}$ denotes the absolute improvement over the track-specific random baseline, while N_{win} counts the capability families in which a model achieves the highest reported score. The task-macro random baselines are 25.83% for the Text Track and 26.39% for the Multimodel Track. Question-level correct counts and per-task accuracies for all evaluated models are provided in Appendix; these supplementary statistics are not used to rank models.

4.2 MAIN RESULTS

Tables 5 and 6 report task-macro accuracy under the Text and Multimodel Tracks, respectively. The Text Track provides a controlled comparison over the same 10 text-only tasks, whereas the Multimodel Track evaluates end-to-end performance over all 18 tasks. The two averages are computed over different task sets and are therefore interpreted separately.

GPT-5.5 achieves the strongest overall performance. On the Text Track, GPT-5.5 obtains 57.32% average accuracy, exceeding the random baseline by 31.49 points and leading all three eligible capability families. Gemma-4-31B-IT with thinking and Gemini 3.1 Pro Preview follow at 48.42% and 46.94%, respectively. On the Multimodel Track, GPT-5.5 reaches 51.96% and leads Symmetry Reasoning, Symmetry–Response Reasoning, and Dynamics Prediction. Gemini achieves the strongest Perception result at 49.60% and ranks second overall at 45.42%. Together, GPT-5.5 and Gemini account for all four capability-level wins.

A clear gap remains between direct open-weight and leading proprietary models. The direct open-weight models obtain Multimodel-Track averages between 29.88% and 32.21%, substantially below GPT-5.5 and Gemini. Their largest limitation is Symmetry Reasoning, where most scores remain close to the 27.08% random baseline. Thinking-enabled Gemma narrows this gap and becomes the strongest open-weight configuration at 43.04%, exceeding Claude Sonnet 5 while remaining 8.92 points behind GPT-5.5.

Perception does not imply physical reasoning. Claude Sonnet 5 achieves 44.71% on Perception but only 27.65% on Dynamics Prediction, slightly below the corresponding random baseline. A similar disparity appears across several direct open-weight models, whose visual recognition performance exceeds their symmetry and dynamics scores. These results indicate that identifying visible structural differences is considerably easier than relating them to relaxation, stability, and finite-temperature evolution.

Table 5: Results on the **Text Track**, consisting of 10 text-only tasks and 2,660 questions. $\Delta_{\mathrm{Rand.}}$ denotes the absolute improvement in average accuracy over the random baseline. N_{win} counts the capability families in which a model achieves the best result.

Model	Symmetry Reasoning	Symmetry–Response Reasoning	Dynamics Prediction	Avg.	$\Delta_{\mathrm{Rand.}}$	N_{win}
Random	25.00	25.00	27.78	25.83	–	–
Expert	–	–	–	63.33	–	–
Open-Weight Models						
Qwen3.8-27B	28.12	36.79	31.82	32.70	+6.87	0
InternVL3.5-38B	30.57	32.32	33.16	32.05	+6.21	0
Gemma-4-12B-IT (direct)	21.73	34.19	23.25	27.17	+1.34	0
Gemma-4-31B-IT (direct)†	26.86	27.76	26.22	27.03	+1.20	0
Gemma-4-31B-IT (thinking)†	48.51	49.07	47.46	48.42	+22.58	0
Proprietary/Closed-Source Models						
GPT-5.5 (low)	**62.91**	**52.25**	**58.50**	**57.32**	**+31.49**	**3**
Claude Sonnet 5	40.38	43.36	23.42	36.49	+10.65	0
Gemini 3.1 Pro Preview	56.97	47.59	36.04	46.94	+21.11	0
DeepSeek V4 Pro	28.92	30.20	32.30	30.45	+4.61	0

Table 6: Results on the **Multimodel Track**, consisting of all 18 tasks and 4,016 questions. Only models with valid results for both textual and visual inputs are included. $\Delta_{\mathrm{Rand.}}$ denotes the absolute improvement in average accuracy over the random baseline. N_{win} counts the capability families in which a model achieves the best result.

Model	Perception	Symmetry Reasoning	Symmetry–Response Reasoning	Dynamics Prediction	Avg.	$\Delta_{\mathrm{Rand.}}$	N_{win}
Random	25.00	27.08	25.00	28.33	26.39	–	–
Expert	–	–	–	–	73.33	–	–
Open-Weight Models							
Qwen3.8-27B	32.10	27.69	37.15	30.98	32.21	+5.82	0
InternVL3.5-38B	33.43	30.57	30.20	31.66	31.40	+5.01	0
Gemma-4-12B-IT (direct)	34.07	23.77	34.34	26.97	29.88	+3.49	0
Gemma-4-31B-IT (direct)†	34.88	28.13	31.01	31.11	31.26	+4.87	0
Gemma-4-31B-IT (thinking)†	36.95	42.98	46.72	44.28	43.04	+16.65	0
Proprietary/API Models							
GPT-5.5 (low)	46.11	**61.07**	**49.03**	**52.28**	**51.96**	**+25.57**	**3**
Claude Sonnet 5	44.71	39.66	41.80	27.65	38.04	+11.65	0
Gemini 3.1 Pro Preview	**49.60**	52.10	44.46	37.70	45.42	+19.03	1

Substantial headroom remains. The best Multimodel-Track average is 51.96%, and no model dominates every capability family. Even the strongest capability-level results range from 49.03% on Symmetry–Response Reasoning to 61.07% on Symmetry Reasoning. Even the top-performing GPT-5.5 (51.96%) lags behind human experts (73.33%) by more than 20 percentage points, demonstrating that our benchmark remains highly challenging for current state-of-the-art models. CrystalBench Dynamics therefore remains far from saturated, particularly for tasks that require models to connect symmetry changes with physical response and subsequent structural evolution.

4.3 Analysis

Thinking substantially improves reasoning. Enabling thinking raises Gemma-4-31B-IT from 27.03% to 48.42% on the Text Track and from 31.26% to 43.04% on the Multimodel Track. The Multimodel-Track improvement is concentrated in Symmetry Reasoning, Symmetry–Response Reasoning, and Dynamics Prediction, which gain 14.85, 15.71, and 13.17 percentage points, respectively. Perception gains only 2.07 points. This pattern suggests that additional inference-time reasoning is particularly useful for relating structural changes to symmetry, relaxation, and dynamical outcomes.

Table 7: Post-hoc outcomes; simulation statistics were hidden from models.

Candidate	Perturbation	Steps	ΔE (eV)	Outcome
A	Vacancy	82	-0.22	Converged
B	Displacement	32	-0.27	Converged
C	Substitution	33	-0.19	Converged
D	**Displacement**	**1,000**	$\mathbf{-16.14}$	**Unstable**

Comparison across model scales within the Gemma family. Under direct inference, Gemma-4-31B-IT scores 31.26% on the Multimodel Track, compared with 29.88% for Gemma-4-12B-IT. This 1.38-point overall advantage masks uneven changes across families: the larger model performs better on Symmetry Reasoning and Dynamics Prediction, but worse on Symmetry–Response Reasoning. In this comparison, increased model scale brings a modest aggregate gain without consistently improving the reasoning capabilities assessed by the benchmark.

Proprietary models generally outperform open-weight models. GPT-5.5 and Gemini 3.1 Pro Preview achieve the two highest Multimodel-Track scores, at 51.96% and 45.42%, while all direct open-weight configurations remain below 33%. The gap narrows when inference-time reasoning is enabled: Gemma-4-31B-IT with thinking reaches 43.04%, exceeding Claude Sonnet 5 at 38.04%. The results therefore show a broad advantage for the leading proprietary systems, while also demonstrating that inference settings substantially affect comparisons between model groups.

Case study: predicting structural instability. Instability prediction is the hardest Dynamics Prediction task, averaging 28.93% across eight Full-Track configurations, only 3.93 points above random. Models receive only initial CIFs and operation labels. In the example in Table 7, candidates B and D share the same operation and displacement magnitude, yet only D becomes unstable. Only three of seven primary Full-Track settings select D. The remaining models choose different converged candidates, suggesting no single shared distractor bias. This case illustrates the central challenge of CrystalBench Dynamics: recognizing a perturbation does not imply an ability to predict its subsequent structural response.

5 Limitations and Future Work

CrystalBench evaluates crystal-structure reasoning under controlled perturbations, but it remains limited by the coverage of its source structures, perturbation operators, simulation conditions, and machine-learned potential. In addition, its multiple-choice format may not fully reflect open-ended scientific reasoning, and rendered views inevitably omit part of the three-dimensional structural information. Future work will expand the chemical and structural space, incorporate longer and more diverse simulations and higher-fidelity calculations, and extend the benchmark toward open-ended prediction, structure generation, and tool-augmented reasoning.

6 Conclusion

We introduce **CrystalBench Dynamics**, a benchmark comprising 4,016 questions across 18 tasks for evaluating crystal-structure understanding under controlled perturbations, structural relaxation, and finite-temperature dynamics. Built from 1,500 real crystal structures, the benchmark applies lattice strain, vacancy creation, substitution, interstitial insertion, atomic displacement, and temperature-driven molecular dynamics. Its simulation-grounded labels enable evaluation beyond static recognition, testing whether models can connect perturbations to symmetry changes, relaxation outcomes, structural stability, and dynamical evolution. The benchmark organizes these tasks into perception, symmetry reasoning, symmetry– response reasoning, and dynamics prediction, with separate text and multimodal tracks. Our results show that proprietary models generally lead open-weight models, while thinking substantially improves scientific reasoning in Gemma-4-31B-IT. Nevertheless, the best Multimodel-Track accuracy remains 51.96%, leaving substantial room for models that can reliably connect crystal symmetry, structural response, and dynamical evolution.

AI USE STATEMENT

We used OpenAI Codex to assist with the implementation and debugging of the benchmark-generation, experimental and evaluation code. The authors designed the benchmark taxonomy, crystal perturbation and simulation pipeline, task-generation procedures, and ground-truth logic. Reference answers were computed programmatically from crystal structures, optimization records, and molecular-dynamics outputs, rather than generated through LLM inference. We used ChatGPT to assist with language polishing, manuscript editing. The authors reviewed all AI-assisted code and content and take responsibility for the final text, results, code, and artifacts presented in this work.

REFERENCES

Dhruv Ahlawat, Vaibhav Mishra, Somaditya Singh, Mohd Zaki, Vaibhav Bihani, Hargun Singh Grover, Biswajit Mishra, Santiago Miret, Mausam, and N. M. Anoop Krishnan. A family of large language models for materials research with insights into model adaptability in continued pretraining. *Nature Machine Intelligence*, 8(3):435–448, 2026.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5:952–961, 2025. doi: 10.1038/s43588-025-00836-3.

Anthropic. Introducing Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, June 2026. Accessed: 2026-09-26.

Yuan Chiang, Elvis Hsieh, Chia-Hong Chou, and Janosh Riebesell. LLaMP: Large language model made powerful for high-fidelity materials knowledge retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25189–25221, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1280.

DeepSeek-AI. DeepSeek-V4-Pro ga release. https://api-docs.deepseek.com/news/news260813/, August 2026. Accessed: 2026-09-26.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Google Gemini Team. Gemini 3.1 Pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, February 2026. Accessed: 2026-09-26.

Zhengzhao Lai, Youbin Zheng, Zhenyang Cai, Haonan Lyu, Jingpu Yang, Hong-Qing Liang, Yan Hu, and Benyou Wang. Can multimodal LLMs see materials clearly? a multimodal benchmark on materials characterization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4384–4404. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.235.

Siyu Liu, Jiamin Xu, Beilin Ye, Bo Hu, David J. Srolovitz, and Tongqi Wen. Mattools: Benchmarking large language models for materials science tools. *arXiv preprint arXiv:2505.10852*, 2025.

Wanhao Liu, Jiaqing Xie, Qian Tan, Weida Wang, Jue Wang, Ran Sun, Zhuo Yang, Wanli Ouyang, Lei Bai, Tianfan Fu, Lu Chen, Xin Chen, and Yuqiang Li. Omnimatbench: A human-calibrated multimodal reasoning benchmark across 19 materials science subfields. *arXiv preprint arXiv:2605.29833*, 2026.

Taoyuze Lv, Alexander Chen, Fengyu Xie, Chu Wu, Jeffrey Meng, Dongzhan Zhou, 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.

OpenAI. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, April 2026. Accessed: 2026-09-26.

Qwen Team. Qwen3.8-27B model card. https://huggingface.co/Qwen/Qwen3.8-27B, August 2026. Accessed: 2026-09-26.

Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. Llm4mat-bench: Benchmarking large language models for materials property prediction. *arXiv preprint arXiv:2411.00177*, 2024.

Zhichen Tang, Zhengzheng Dang, Yulin Chen, Jixin Wu, Haiwen Li, and Yanming Wang. Evomd-llm: Learning the language of species evolution in reactive molecular dynamics. *arXiv preprint arXiv:2605.29394*, 2026.

Ali Vosoughi, Ayoub Shahnazari, Yufeng Xi, Zeliang Zhang, Griffin Hess, Chenliang Xu, and Niaz Abdolrahim. OPENXRD: A comprehensive benchmark framework for llm/mllm xrd question answering. *Digital Discovery*, 5:1991–2015, 2026. doi: 10.1039/D5DD00519A.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Sifan Wu, Huan Zhang, Yizhan Li, Farshid Effaty, Amirreza Ataei, and Bang Liu. Seeing beyond words: Matvqa for challenging visual-scientific reasoning in materials science. *arXiv preprint arXiv:2505.18319*, 2025.

Chengliang Xu, Xiaogang Li, Peiyao Xiao, Beng Wang, Hu Wei, and Bing Zhao. Crystalxrd-bench: Benchmarking vision-language models for xrd peak indexing across diverse crystalline materials. *arXiv preprint arXiv:2605.29446*, 2026.

Michiko Yoshitake, Yuta Suzuki, Ryo Igarashi, Yoshitaka Ushiku, and Keisuke Nagato. Materialfigbench: Benchmark dataset with figures for evaluating college-level materials science problem-solving abilities of multimodal large language models. *arXiv preprint arXiv:2603.11414*, 2026.

Mohd Zaki, Jayadeva, Mausam, and N. M. Anoop Krishnan. Mascqa: Investigating materials science knowledge of large language models. *Digital Discovery*, 3:313–327, 2024. doi: 10.1039/D3DD00188A.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and Wei Wang. Matscibench: Benchmarking the reasoning ability of large language models in materials science. *arXiv preprint arXiv:2510.12171*, 2025.

A Task Specifications and Ground Truth

All labels in CrystalBench Dynamics are computed from structures, simulation records, relaxation outputs, and molecular-dynamics trajectories. The final and large-final subsets are merged in the counts below.

Programmatic ground truth. The benchmark contains 18 tasks and 4,016 questions after merging the final and large-final subsets. Ground-truth labels are not written by language models. Text tasks use CIF structures, action metadata, relaxation logs, energies, RMSD values, space-group analysis, or molecular-dynamics outputs. Visual tasks are rendered from the same structures or trajectories, but their answers are assigned from the corresponding programmatic quantities before rendering.

Implementation-sensitive tasks. For *Symmetry Change*, the label is the space-group transition between the initial and final CIFs, computed with pymatgen's `SpacegroupAnalyzer` backed by spglib. The same tolerance settings are used across all examples. The sampled cases cover unchanged structures, P1-to-non-P1 transitions, non-P1-to-P1 transitions, and transitions between different non-P1 groups.

Table 8: CrystalBench Dynamics tasks. Counts combine the final and large-final subsets. T: text; V: visual/multimodal.

Task	Mod.	#Opt.	#Q	Description
Perception				
Defect Location	V	4	157	Locate the defect.
Defect Intensity	V	4	204	Classify defect concentration.
Operation Type	V	4	165	Identify the applied perturbation.
Relaxed Volume	V	4	181	Compare relaxed volume per atom.
Symmetry Reasoning				
Space Group	T	4	330	Identify the space group.
Symmetry Change	T	4	218	Compare initial and final symmetry.
Strain Action	T	4	335	Identify the lattice strain.
Relaxation Direction	V	3	144	Find the largest lattice-axis change.
Symmetry–Response Reasoning				
Relaxation Basin	T	4	302	Match a relaxed structure to its parent.
Lowest Energy	T	4	394	Find the lowest-energy relaxed structure.
Max RMSD	T	4	135	Find the largest relaxation RMSD.
Fragility Direction	T	4	274	Find the direction of greatest fragility.
Distortion Intensity	V	4	166	Classify relaxation-induced distortion.
Dynamics Prediction				
Optimizer Steps	T	4	119	Find the longest optimization.
Instability	T	4	337	Predict structural instability.
MD RMSD	T	3	216	Compare final MD displacements.
Thermal Volume	V	3	180	Compare trajectory envelope volumes.
Thermal Stability	V	4	159	Order hidden-temperature trajectories.
Total			**4,016**	

For *Relaxation Basin*, the correct option is the true action-modified initial structure that produced the target relaxed CIF in the simulation record. Candidate sets are screened so that the true parent is also the nearest candidate to the target relaxed structure under the internal geometry distance, reducing ambiguity among visually similar perturbations.

For *Max RMSD*, the answer is the option with the largest relaxation RMSD between the action-modified initial CIF and the relaxed CIF. RMSD is computed using periodic minimum-image displacements in fractional coordinates and then converted to Angstroms with the relaxed lattice.

For *MD RMSD*, all candidates are completed fixed-cell NVT trajectories starting from the same initial structure. The answer is the temperature run whose final CIF has the largest observed RMSD from the initial CIF, computed with atom-order-preserving periodic minimum-image displacement.

For *Thermal Stability*, the visual panels show trajectory frames for three hidden temperature conditions. The label is assigned from trajectory metadata and corresponds to the candidate ordering from lowest to highest temperature, with stability reflected in trajectory-derived displacement statistics rather than manual visual judgment.

B Dataset Construction and Simulation Details

Structure selection. We sampled crystal structures from the non-deprecated Materials Project collection, excluding GNoMe entries. The base subset contains 1,000 structures with 3–40 atoms, selected with random seed 2026; the large subset contains 500 structures with 41–246 atoms, selected with seed 20260909. Sampling preserves the candidate pool's atom-count distribution through per-size quotas and encourages diversity in elements, chemical systems, compositions, and space groups within each quota. The respective candidate pools contained 117,578 and 34,833 structures.

Table 9: Geometry-relaxation and molecular-dynamics settings.

Stage	Setting
Relaxation	ASE BFGS with `FrechetCellFilter`
Force threshold	0.03 eV Å^{-1}
Maximum optimizer steps	1,000
MD ensemble	Fixed-cell NVT, Berendsen thermostat
Target temperatures	300, 600, and 1,000 K
Thermostat coupling time	100 fs
Time step and duration	1 fs; 5,000 steps (5 ps)
Trajectory sampling	Every 100 steps, plus an initial write

Structural perturbations. For a structure with N atoms, substitution, vacancy, interstitial, and displacement levels target fractions $c \in \{0.05, 0.10, 0.15\}$, with $n = \max(1, \mathrm{round}(cN))$ affected sites, subject to operation-specific limits. Substitution sites are sampled across element types and replaced by randomly selected elements from a fixed pool. Vacancy sites are separated by at least $\max(3.0\ \text{Å}, 1.2d_{\min})$, where $d_{\min}$ is the original structure's minimum interatomic distance; samples for which this constraint cannot be met are skipped. Interstitial species are drawn from $\{\mathrm{H}, \mathrm{Li}, \mathrm{Na}, \mathrm{Mg}, \mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{F}\}$. Each inserted site maximizes its minimum periodic distance to existing atoms over an 8^3 fractional-coordinate grid and 512 random candidates. Displaced atoms move along random unit vectors by distances sampled log-uniformly from 0.03 to 0.30 Å.

Lattice strain uses one of five sampled modes: isotropic, uniaxial, biaxial, shear, or general symmetric strain. The low, medium, and high levels have maximum normal-strain amplitudes of 1%, 3%, and 5%, and shear amplitudes of 0.5%, 1.5%, and 3%, respectively. The lattice is transformed by $I + \boldsymbol{\epsilon}$ while fractional atomic coordinates are retained. Perturbations use deterministic seeds derived from 20260811 and the material, operation, and level indices. Structures with fewer than two atoms, a minimum interatomic distance below 0.75 Å, non-positive volume, or a lattice condition number above 100 are rejected before relaxation.

Simulation protocol. Both simulations use MACE-MP-0 (medium model; `mace-torch` v0.3.16) through ASE v3.29.0, with float64 precision and dispersion disabled. Geometry optimization relaxes atomic positions and the cell. Temperature-only MD starts from the sampled Materials Project structure and keeps the cell fixed. Table 9 lists the remaining settings.

Symmetry assignment. Space groups are computed from structure geometry using `pymatgen`'s `SpacegroupAnalyzer` with a positional tolerance of 0.01 Å and an angular tolerance of 5°. Symmetry-change labels compare the space groups assigned to the structures before and after the operation under the same tolerances.

Retention and failure handling. Table 10 separates execution failures from completed but unconverged optimizations. A relaxation is retained as converged only when the optimizer reports convergence within the 1,000-step limit. Failed perturbation or simulation records and unconverged optimizations are excluded from labels requiring a converged relaxed structure. The MD run logs record 4,500 successful runs; 4,480 complete trajectories are present in the current data snapshot. The records available here do not establish why the other 20 logged runs are absent, so they are reported separately rather than classified as MD failures.

Trajectory-frame count. The MD writer records the initial configuration explicitly at step 0. ASE also invokes the attached sampling callback at step 0, so the initial configuration appears twice. It then stores steps 100,200,...,5000. Thus each retained trajectory contains $2+5000/100 = 52$ stored frames representing 51 distinct time points. Across 4,480 retained trajectories, this gives 232,960 stored frames.

Table 10: Structure and simulation counts. "Base" and "Large" refer to the two structure subsets.

Stage	Base	Large	Total
MP candidates	117,578	34,833	152,411
Structures selected	1,000	500	1,500
Relaxations planned	15,000	7,500	22,500
Failed before result	1	19	20
Completed, unconverged	35	155	190
Converged and retained	14,964	7,326	22,290
MD runs planned	3,000	1,500	4,500
Successful in run logs	3,000	1,500	4,500
Trajectories retained	3,000	1,480	4,480
Logged runs absent from snapshot	0	20	20

C Tasks for Expert Evaluation

To assess whether the benchmark tasks are meaningfully solvable by domain experts, we constructed a separate expert-evaluation subset from the final CrystalBench task pool. The subset contains 90 multiple-choice questions, including 60 text/CIF-based tasks and 30 multimodal image-based tasks. The questions were sampled from both the standard and large benchmark splits, with 60 questions from `final_benchmark` and 30 questions from `final_large-benchmark`. The sampled subset was controlled such that the GPT-5.5 accuracy on the same questions was 42.22%, within the target range of 26–45%, ensuring that the set was neither trivially easy nor overwhelmingly difficult for current frontier models.

The text/CIF portion covers ten structure- and simulation-centered task types: space-group recognition, symmetry-transition reasoning, lattice-strain action identification, lowest-energy relaxed-structure selection, optimizer-step prediction, relaxation-basin inverse matching, maximum relaxation-RMSD prediction, instability prediction, structural-fragility direction prediction, and temperature-MD observed-RMSD comparison. Each text task was provided with the necessary CIF files, a Chinese task statement, answer options, and task-specific notes describing the relevant benchmark assumptions, such as RMSD conventions, space-group tolerances, fixed-cell NVT conditions, or hidden simulation quantities.

The multimodal portion covers visual crystal-structure interpretation tasks, including defect-intensity recognition, defect-location recognition, operation recognition from rendered structures, relaxation-direction recognition, relaxation-distortion intensity estimation, relaxed volume-per-atom comparison, temperature-envelope volume-change comparison, and temperature-stability ranking. For these tasks, evaluators were provided with rendered crystal-structure images, element-color legends, view descriptions, atom-size encoding rules, answer options, and Chinese task statements. For multi-image trajectory tasks, the frame layout and candidate-image correspondence were explicitly specified.

The human evaluators were three PhD students specializing in computational physical chemistry, familiar with standard atomistic simulation methods and the interpretation of structural characterization data. Evaluators were instructed to answer using only the provided task materials, including CIF files, rendered images, element-color legends, and written task descriptions. They were explicitly instructed not to use file names, material identifiers, hidden metadata, external databases, or answer-containing organizer files as evidence. Each question was scored by exact match against the benchmark answer. On this expert-evaluation subset, the human evaluators answered 66 out of 90 questions correctly, corresponding to an accuracy of 73.33%.

D Evaluation Prompt and Representative Tasks

To make the expert-evaluation protocol transparent, we report the standardized instruction shown to the evaluators and four representative questions from the evaluation subset. The examples comprise two text/CIF tasks and two multimodal tasks. Each evaluator received the Chinese translation of the

Table 11: Composition of the expert-evaluation subset.

Task modality	Number of questions	Human correct
Text/CIF	60	38
Multimodal/image	30	28
Total	90	66

task statement, the corresponding options, and all required anonymized CIF files and/or rendered images. The original English task instruction was retained for cross-checking. File names were normalized and did not expose material identifiers, operation labels, temperatures, or other hidden target variables.

Common evaluator prompt.

> Answer the following multiple-choice question using only the supplied CIF files and/or rendered crystal-structure images. For CIF-based questions, inspect the lattice parameters, cell angles, fractional coordinates, and periodic geometry. For image-based questions, use only the visible geometry, element-color legend, unit-cell outline, relative atom sizes, and the stated frame order. Do not infer the answer from file names, material identifiers, option order, hidden metadata, hidden numerical targets, or external databases. Select exactly one option and return only its letter.

Task-specific information needed for a physically meaningful comparison was provided alongside this common instruction. For example, molecular-dynamics questions stated the ensemble, cell constraint, temperature, and number of steps, while image-sequence questions specified the viewing direction, frame order, shared coordinate scale, and atom-size encoding.

D.0.1 TEXT/CIF EXAMPLES

Example T1: lattice-strain action reasoning. The evaluator was given the original CIF, the strained CIF, and the relaxed CIF and was asked:

> Given the original CIF and the relaxed CIF obtained after applying a lattice strain operation, identify the lattice strain operation that was applied before relaxation. Compare changes in lattice-vector lengths and cell angles to determine the dominant strain direction or plane. Return exactly one option letter.

The options were (A) tensile uniaxial strain along the b axis, (B) biaxial strain in the a–c axes with tensile a and compressive c, (C) isotropic tensile strain along all lattice axes, and (D) shear strain in the a–b plane. The reference answer was **A**; the expert answer was **A**, and GPT-5.5 also answered **A**.

Example T2: optimizer-step prediction. The evaluator was given four action-modified CIF structures derived from the same original material and was asked:

> Predict which modified structure will require the largest number of optimizer steps to converge during geometry relaxation. The optimizer step counts are hidden and used only for grading. Use only the broad operation type and the geometry in each CIF. Return exactly one option letter.

The options were (A) atom substitution, (B) vacancy creation, (C) atomic displacement, and (D) lattice strain. The reference answer was **B**; the expert answered **B**, whereas GPT-5.5 answered **A**.

D.0.2 MULTIMODAL EXAMPLES

Example M1: temperature-trajectory envelope change. Three hidden-temperature candidates were shown for the same base material. Each candidate image contained the final five trajectory frames from left to right, rendered with a common coordinate scale and viewing direction. The evaluator was asked:

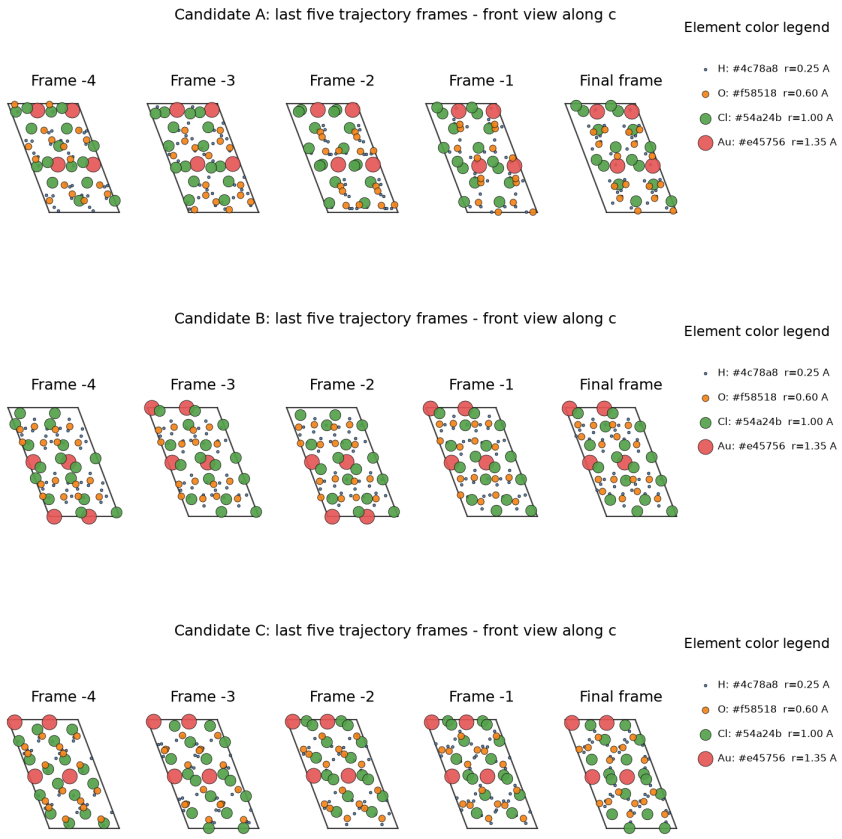


Figure 3: Representative multimodal temperature-trajectory question. Each row corresponds to one candidate condition and shows five consecutive frames. The reference answer was B; both the expert and GPT-5.5 selected B.

> Which candidate shows the largest absolute change in visible atom-cloud envelope volume across its five frames? Use only the visible trajectory frames, lattice-cell outlines, shared coordinate scale, view description, element-color legend, and atom-size encoding. The numerical volume changes and temperature labels are hidden. Return only A, B, or C.

Example M2: realized defect intensity. The evaluator was shown a rendered pair containing the original and final relaxed structures, together with the element-color legend, and was asked:

> What is the approximate realized countable defect intensity in the final relaxed structure? Atom-marker areas are proportional to the squares of the corresponding atomic radii. Use the concentration ranges stated in the options and return exactly one option letter.

The options were (A) medium, approximately 7.5–12.5%; (B) low, approximately 0–7.5%; (C) no countable defect introduced; and (D) high, at least approximately 12.5%. The reference answer was **B**; the expert answered **B**, whereas GPT-5.5 answered **A**.

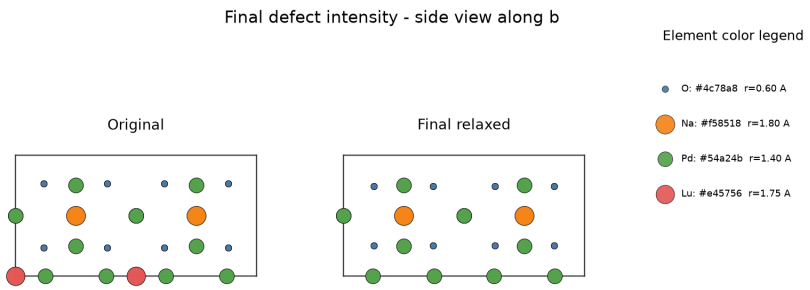


Figure 4: Representative multimodal defect-intensity question. The image juxtaposes the original and final relaxed structures and includes the element-color and relative-size legend.