FIREWORLDBENCH: BENCHMARKING COMPLEX PHYSICAL WORLD INTELLIGENCE THROUGH COUPLED-FIELD FIRE DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the physical world requires more than object recognition, scene description, and short-term visual prediction, as real-world physical systems involve multiple continuous fields, latent causal mechanisms, partial observations, and intervention-sensitive dynamics. We propose FireWorldBench, a benchmark for evaluating complex physical world intelligence in multimodal large language models and agents through coupled-field fire dynamics. Fire provides a canonical stress-test environment, where multiple interacting physical fields jointly shape observable states and temporal dynamics. FireWorldBench is organized along two complementary axes, a physical capability axis and a fire scenario task axis, jointly covering physical-state understanding, temporal dynamics, causal mechanisms, and intervention reasoning. The benchmark comprises 520 fire-world entries, including 494 controlled simulation worlds and 26 real-world-aligned event groups, spanning 47 scene archetypes across 7 environment families. These entries combine structured textual observations, multiple 2D physical-field visualizations, and 3D event-level scene modeling, yielding 9,074 text-image interleaved question-answer pairs across choice-based and open-ended report-generation formats. FireWorldBench evaluates whether models can infer latent physical states, explain underlying mechanisms, forecast coupled-field evolution, and assess intervention consequences from multimodal partial observations, providing a challenging testbed for complex physical world intelligence.

1 INTRODUCTION

Recent progress in multimodal large language models (MLLMs) and agents has enabled strong capabilities in visual understanding, video interpretation, and tool use. However, whether these models genuinely understand the physical world remains an open question. Existing benchmarks often evaluate physical reasoning through object motion, spatial relations, intuitive mechanics (Riochet et al., 2018; Bakhtin et al., 2019), or short video prediction (Yi et al., 2020; Bear et al., 2021). While valuable, these settings often simplify the physical world into discrete objects, isolated events, or visually recognizable dynamics. In contrast, many real-world physical systems are governed by continuous fields, multi-physics coupling, and latent state variables. Understanding such systems requires models to connect partial observations to physical variables, infer hidden states, and reason about causal mechanisms beyond perception and language plausibility.

We refer to this ability as complex **physical world intelligence**. Unlike object-centric physical reasoning, complex physical world intelligence concerns environments where observable phenomena arise from *the interaction of multiple physical fields and mechanisms.* A physical understanding model need to forecast system evolution, ground heterogeneous observations in latent physical states, interpret coupling across multiple fields, infer the mechanisms behind observed changes, and assess hypothetical interventions under partial observability. This capability is central to MLLMs and agents intended to support scientific analysis, safety-critical monitoring, operational decision-making, embodied manipulation, and real-world problem solving.

In this work, we use **coupled-field fire dynamics** as a representative testbed for complex physical world intelligence. Fire provides a compact yet expressive physical system in which temperature, soot

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.

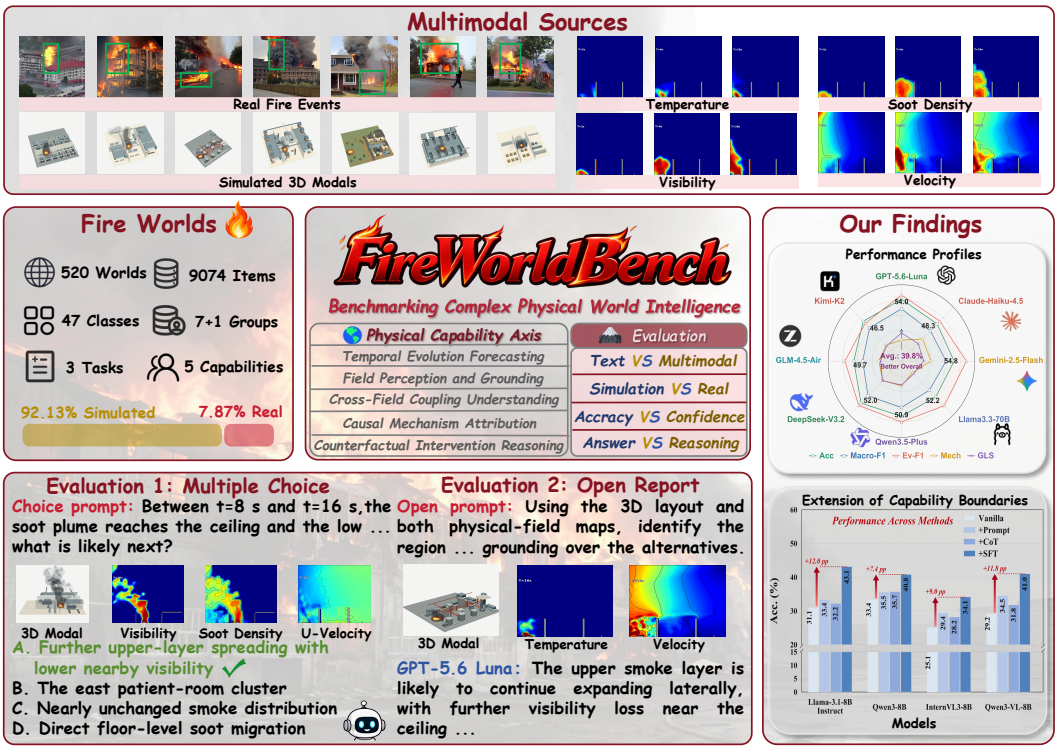


Figure 1: **FireWorldBench** is a reproducible multimodal benchmark with 520 coupled-field fire worlds, 47 scene archetypes across 7 environment families, and 9,074 question-answer instances. It evaluates physical-capability and fire-scenario task axes, revealing current limits in coupled-field understanding, physical explanation, and intervention reasoning.

density, velocity, visibility, carbon-monoxide concentration, and ventilation state interact continuously (McGrattan et al., 2013; Overholt, 2014). These physical fields form a coupled state whose spatial and temporal agreement reflects latent fire conditions, while their joint response to changes in openings, sources, exhaust, or suppression supports intervention reasoning. For example, a local ignition source can generate a thermal plume, evolve into ceiling jet flow, drive smoke-layer formation, alter visibility and toxicity, and respond nonlinearly to ventilation or exhaust interventions. These properties make fire dynamics a natural stress test for whether MLLMs and agents can understand an evolving, coupled, and actionable physical world.

We introduce FireWorldBench, a benchmark for evaluating MLLMs and agents on this form of physical understanding. It is organized along two complementary axes. The physical capability axis progresses through *Temporal Evolution Forecasting, Physical Field Perception and Grounding, Cross-Field Coupling Understanding, Causal Mechanism Attribution, and Counterfactual Intervention Reasoning*, while the fire scenario task axis operationalizes these capabilities through Localized Onset, Coupled Propagation and Critical Transition. This dual-axis design allows FireWorldBench to remain grounded in a real physical domain while targeting general capabilities needed for complex world understanding. FireWorldBench contains 520 fire-world entries, including 494 controlled simulation worlds and 26 real-world-aligned event groups, spanning 47 scene archetypes across 7 environment families. Each entry combines structured textual observations, multiple 2D physical-field visualizations, and 3D event-level scene modeling, yielding 9,074 text-image interleaved QA instances across single-choice, multiple-choice, and open-ended report-generation formats. Beyond evaluation, we compare physical-field prompting, chain-of-thought reasoning, and supervised fine-tuning (SFT) for improving complex physical-world understanding. The SFT-adapted model, **FireWorldGPT**, consistently outperforms its base model across both benchmark axes under the unified evaluation protocol, showing that coupled-field supervision yields stronger gains than input-level prompting or inference-time reasoning alone.

Our contributions are summarized as follows:

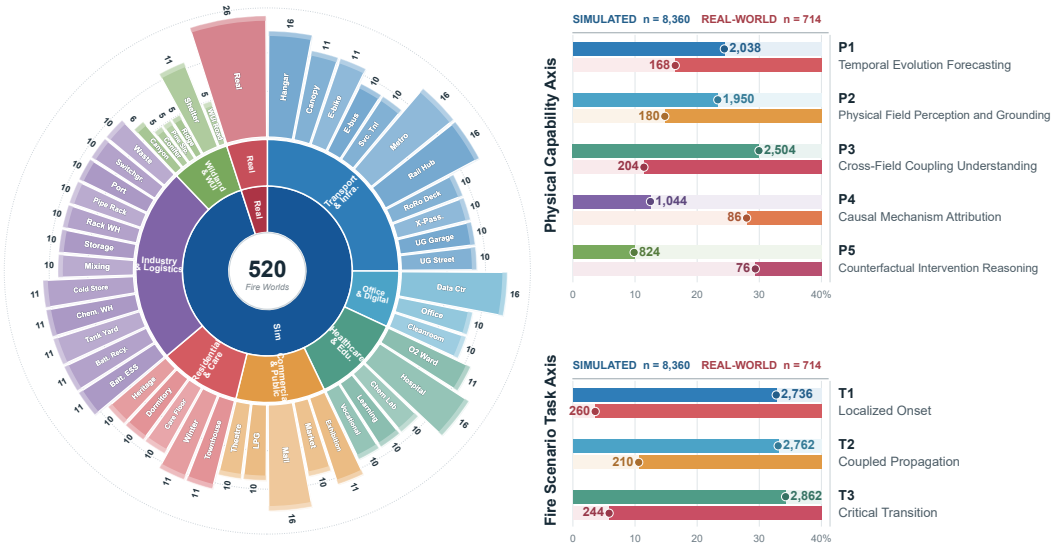


Figure 2: **Overview of Dataset.** The left sunburst summarizes the hierarchical composition of 520 fire incidents, while the right panels characterize the coverage of 8,360 simulated and 714 real-world-aligned QA pairs along the five physical capability axis (P1–P5) and fire scenario task axis (T1–T3).

- **Problem formulation.** We formulate complex physical world intelligence as a benchmark target beyond object-centric physical reasoning, emphasizing coupled continuous fields, latent mechanisms, and counterfactual interventions.
- **Dual-axis benchmark design.** We introduce a dual-axis evaluation design that connects progressive physical capabilities with concrete fire-dynamics tasks, supported by a comprehensive evaluation protocol across observation modalities and answer formats.
- **Comprehensive empirical analysis.** Extensive experiments on simulated and real-world-aligned data reveal current models' capability boundaries, the gap between physical answer and reasoning / explanation, and the limits of parameter scaling.
- **Strategies for improving physical-world understanding.** We systematically compare physical-field prompting, CoT, and supervised fine-tuning, and show that domain adaptation with coupled multi-physics field visualizations enables FireWorldGPT to substantially strengthen the physical-world understanding of its base model.

2 Related Work

Physical World Understanding Benchmarks and Multimodal Reasoning. Understanding the physical world is a core component of visual intelligence. IntPhys tests physical plausibility in videos (Riochet et al., 2018), PHYRE studies action-conditioned 2D puzzles (Bakhtin et al., 2019), CLEVRER evaluates descriptive, explanatory, predictive, and counterfactual reasoning over collision-based scenes (Yi et al., 2020), and Physion compares machine prediction against human judgments in realistic simulations (Bear et al., 2021). Fire dynamics extends this setting to coupled thermal, fluid, chemical, and ventilation-driven processes (McGrattan et al., 2013; Overholt, 2014). Recent multimodal benchmarks cover physical properties and dynamics (Chow et al., 2025), hierarchical spatial cognition (Xu et al., 2025), 3D and embodied spatial generalization (Peng et al., 2026), quantitative physical inference (Li et al., 2026c), unified perception and reasoning across physical scenarios (Han et al., 2026), and causal object-attribute-event dependencies (Tang et al., 2026).

World Models and Simulation-Grounded Physical Reasoning. A parallel line evaluates whether video generators and world models obey physical principles through PhyGenBench (Meng et al., 2024), VideoPhy (Bansal et al., 2024), Physics-IQ (Motamed et al., 2026), WorldBench (Upadhyay et al., 2026), and WorldModelBench (Li et al., 2026a). VisPhyWorld questions whether conventional tasks require falsifiable physical hypotheses (Liang et al., 2026), while Scene Dynamic Field uses simulator-derived dynamic fields for intuitive physics reasoning (Li et al., 2026b). PDEBench

(Takamoto et al., 2022) and FluidsBench (Somasekharan et al., 2025) benchmark PDE-governed systems and fluid surrogates, and Open-CK provides simulation-generated combustion fields (Fei et al., 2025). FireWorldBench instead converts coupled-field simulations into multimodal physical-world understanding tasks.

AI for Fire Dynamics and Smart Firefighting. Fire safety research increasingly applies machine learning to detection, localization, forecasting, and emergency decision support using visual data (Töreyin et al., 2006; Muhammad et al., 2018), sensor streams and fire measurements (Jia et al., 2026), and CFD-generated simulations (Fang & Tam, 2025). Tunnel fire datasets and systems such as PolyUFire (Zhang et al., 2021), extra-wide tunnel fire localization (Zhang et al., 2024), AI-driven digital firefighting (Zhang et al., 2022), and multimodal indoor fire datasets (Jia et al., 2026) demonstrate the practical value of data-driven fire modeling. However, most existing fire AI studies focus on engineering objectives such as classification, localization, forecasting, or control in specific scenarios, rather than evaluating general physical world intelligence. FireWorldBench instead reframes fire dynamics as a stress test for whether multimodal agents can understand coupled physical states, explain their evolution, and assess how interventions alter future outcomes.

3 FIREWORLDBENCH

FireWorldBench comprises diverse coupled-field fire worlds from controlled simulations and real-world-aligned events across varied environments and intervention settings. Figure 2 summarizes scene coverage and instance distributions across both evaluation axes. This design supports systematic evaluation of complex physical-world understanding.

3.1 CAPABILITIES, TASKS, AND INTERFACES

FireWorldBench organizes every evaluation along two complementary axes. The *physical capability axis* progresses from Temporal Evolution Forecasting (P1), Physical Field Perception and Grounding (P2), Cross-Field Coupling Understanding (P3) to Causal Mechanism Attribution (P4) and Counterfactual Intervention Reasoning (P5). From a fire-evolution perspective, the *fire scenario task axis* spans Localized Onset (T1), Coupled Propagation (T2), Critical Transition (T3). Together, the two axes evaluate models from complementary perspectives of physical-world understanding and fire-dynamics understanding, with tasks designed to span distinct levels of reasoning difficulty. The benchmark contains 520 fire-world entries—494 controlled simulation worlds and 26 real-world-aligned event groups—spanning 47 scene archetypes across 7 environment families. Each world combines structured textual observations, multiple 2D physical-field visualizations, and 3D event-level scene assets, yielding 9,074 QA instances across single-choice, multiple-choice, and open-report interfaces. Matched choice-format and open-report questions share the same physical target under matched evaluation conditions, enabling direct comparison between hypothesis selection and evidence-grounded explanation; detailed distributions and representative paired cases are provided in Appendix G.

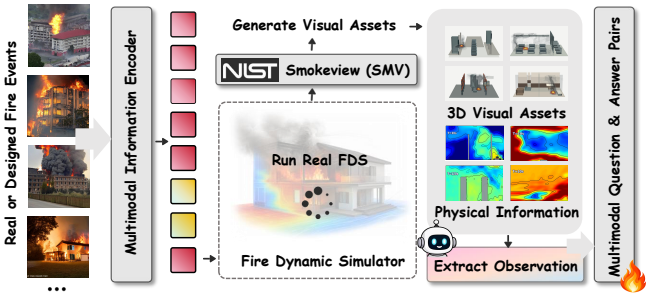


Figure 3: **Construction pipeline of FireWorldBench.** Aligned FDS-derived 3D and physical-field observations are converted into multimodal question–answer pairs.

3.2 DATA COLLECTION AND CURATION

Figure 3 summarizes a world-to-instance pipeline with two complementary branches. The controlled branch starts from structured specifications of scene geometry, ignition source, combustible materials, ventilation, and intervention variables, enabling controlled variation of physical causes and outcomes. The real-world-aligned branch uses selected sequences from MmodalFire, which

contains 65 indoor-fire videos synchronized with six types of physical sensing data (Jia et al., 2026). A multimodal information encoder converts these streams into event-level constraints for aligned FDS reconstruction. Recorded observations anchor event chronology and visible evidence, while the reconstruction supplies dense latent states; observed and simulation-completed information remain explicitly distinguished.

Each specification is simulated with FDS to resolve the spatiotemporal evolution of temperature, soot density, velocity, visibility, and related physical fields. Smokeview and the observation extractor produce co-registered 3D assets, structured physical records, and 2D field visualizations under shared geometry, spatial regions, and timestamps. Retained physical states define the targets, while task-admissible evidence is rendered through matched choice-format and open-report interfaces under the same event, modality, and temporal cutoff. Quality control removes failed or ambiguous simulations, verifies cross-modal alignment and answer consistency, keeps related simulations and paired interventions within the same split, and links every instance to its source world and observation configuration. Full construction, provenance, and quality-control details are provided in Appendix A.

3.3 Evaluation Protocol

Under a unified protocol fixing dataset splits, matched task instances, observation policies, response schemas, and deterministic scoring across both benchmark subsets and interfaces, we evaluate 24 models in three groups: 1) **Frontier closed-source models**: GPT-5.6-Luna, Gemini-2.5-Flash, Claude-Haiku-4.5, and Qwen3.5-Plus-02-15 (OpenAI, 2026; Comanici et al., 2025; Anthropic, 2025; Qwen Team, 2026); 2) **Open-weight text-only LLMs**: Llama-3.1-8B-Instruct, Qwen3-8B, Gemma-3-12B-IT, Gemma-3-27B-IT, Qwen3-32B, Llama-3.3-70B-Instruct, Qwen3-Next-80B-A3B-Instruct, GLM-4.5-Air (106B-A12B), Qwen3-235B-A22B, DeepSeek-V3.2, and Kimi-K2-Instruct (1T-A32B) (Grattafiori et al., 2024; Yang et al., 2025; Gemma Team et al., 2025; Qwen Team, 2025; Zeng et al., 2025; Liu et al., 2025; Team et al., 2025); and 3) **Open-weight multimodal LLMs**: InternVL3-8B, InternVL3-38B, InternVL3-78B, Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct, and Qwen3-VL-235B-A22B-Instruct (Zhu et al., 2025; Bai et al., 2025).

We use six deterministic evaluation metrics under matched observation and response settings. *Completion Accuracy* (Acc) measures recovered Gold content, *Macro-F1* (F1) assesses balanced answer performance, and the *Brier score* measures confidence–correctness consistency. Open reports are additionally evaluated by *Evidence-F1* (Evi-F1) for supporting-evidence recovery, *Mechanism Alignment* (Mech) for causal-mechanism agreement, and *Gold-Linked Support* (GLS) for correct prediction slots explicitly supported by evidence or mechanism anchors. All metrics except Brier are higher-is-better, and none relies on an LLM judge. Complete definitions, scoring rules, and invalid-response handling are provided in Appendix B.

3.4 Why Fire Dynamics Is a Testbed for Physical-World Intelligence?

Fire dynamics brings multiple dimensions of physical-world intelligence into a single evolving environment. Heat transfer, buoyancy-driven flow, combustion, smoke and toxic-gas transport, visibility, and ventilation interact continuously, and their spatial distributions jointly determine how hazards emerge and spread. Understanding such a system requires connecting observations to the underlying physical state, identifying the mechanisms that govern its evolution, forecasting future conditions, and assessing the consequences of possible actions. Fire scenarios therefore provide a compact setting for complex physical world intelligence.

Three characteristics make this setting especially demanding. First, observations are partial and heterogeneous: sensors, field maps, scene assets, and textual summaries each reveal different aspects of the same event. A model must locate relevant evidence and integrate it across modalities, regions, and time. Second, the system state is jointly constrained by *coupled multi-physics fields*. Interpreting any field depends on its spatial and temporal relations with the others. Third, physical systems are intervention-sensitive: changes to boundary conditions, source terms, control inputs, or environmental constraints can redirect transport pathways and produce nonlinear downstream effects throughout the system. These properties directly instantiate the requirements of temporal forecasting, evidence grounding, cross-field understanding, causal attribution, and counterfactual intervention reasoning for physical world intelligence.

Table 1: **Overall evaluation along the physical capability axis on the controlled simulation subset.** Results aggregate choice-format and open-report scores within the same event groups and observation track. We report accuracy (Acc), macro-F1 (F1), and Brier score for P1–P5 and their average.

Model	P1			P2			P3			P4			P5			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records																		
Open-weight models: fewer than 20B total parameters																		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	31.93	42.55	0.44	31.35	39.79	0.49	28.38	39.19	0.44	37.16	40.09	0.42	29.38	38.81	0.42	31.64	40.09	0.44
Qwen3-8B (Yang et al., 2025)	46.55	56.68	0.31	35.34	42.53	0.48	33.88	45.05	0.38	42.98	45.58	0.37	31.58	39.89	0.44	38.07	45.95	0.40
Gemma-3-12B-IT (Gemma Team et al., 2025)	34.67	44.06	0.45	38.57	48.96	0.42	33.30	44.88	0.42	34.03	36.60	0.47	27.89	36.01	0.51	33.70	42.10	0.46
Open-weight models: 20–50B total parameters																		
Gemma-3-27B-IT (Gemma Team et al., 2025)	59.04	67.47	0.23	53.40	60.59	0.31	46.56	57.64	0.31	37.08	37.95	0.47	46.85	54.32	0.33	48.59	55.59	0.33
Qwen3-32B (Yang et al., 2025)	58.30	67.64	0.21	45.76	53.59	0.34	40.16	51.84	0.33	42.67	44.80	0.37	39.61	48.33	0.35	45.30	53.24	0.32
Open-weight models: 50–200B total parameters																		
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	47.08	56.89	0.22	44.00	51.33	0.32	37.11	49.34	0.26	38.02	40.23	0.33	27.80	36.88	0.34	38.80	46.94	0.29
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	48.41	58.75	0.33	42.61	51.66	0.41	39.53	51.06	0.39	32.91	34.67	0.53	33.95	42.23	0.48	39.48	47.67	0.43
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	49.53	59.97	0.21	39.76	47.24	0.36	39.84	51.64	0.28	37.49	39.95	0.36	29.56	38.02	0.35	39.23	47.36	0.31
Open-weight models: more than 200B total parameters																		
Qwen3-235B-A22B (Yang et al., 2025)	55.38	65.12	0.20	42.76	49.66	0.39	43.73	55.41	0.29	40.56	42.06	0.37	36.47	44.12	0.35	43.78	51.27	0.32
DeepSeek-V3.2 (Liu et al., 2025)	60.70	68.97	0.18	49.23	58.04	0.28	45.12	56.37	0.24	46.91	48.36	0.31	44.45	52.86	**0.21**	49.28	56.92	0.24
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	61.28	70.30	0.16	51.80	59.80	0.29	44.53	55.98	0.26	40.98	43.70	0.38	38.00	45.91	0.28	47.32	55.14	0.27
Closed-source models																		
GPT-5.6-Luna (OpenAI, 2026)	**66.59**	**74.62**	**0.13**	**54.44**	59.98	0.33	**55.49**	**66.57**	0.20	**52.14**	**53.70**	**0.30**	**48.07**	**54.94**	0.23	**55.34**	**61.96**	0.24
Gemini-2.5-Flash (Comanici et al., 2025)	59.92	68.51	0.23	50.81	59.02	0.34	42.52	54.16	0.34	37.32	40.62	0.47	40.41	48.51	0.34	46.20	54.16	0.35
Claude-Haiku-4.5 (Anthropic, 2025)	56.56	66.30	0.14	53.95	**62.28**	**0.22**	39.92	51.64	**0.20**	42.03	45.77	0.31	31.45	39.12	0.24	44.78	53.02	**0.22**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	60.77	70.07	0.22	47.20	53.97	0.38	42.76	53.98	0.35	40.61	42.89	0.43	44.92	53.66	0.31	47.25	54.91	0.34
I-track: multimodal physics-rendered visual evidence																		
Open-weight multimodal models: 7–8B																		
InternVL3-8B (Zhu et al., 2025)	26.86	28.87	0.46	25.08	28.82	0.41	34.08	36.55	0.43	26.59	28.96	0.42	26.27	31.93	0.45	27.77	31.03	0.43
Qwen3-VL-8B-Instruct (Bai et al., 2025)	31.92	40.28	0.51	21.47	27.38	0.66	32.45	39.98	0.53	24.80	28.70	0.63	36.42	41.94	0.47	29.41	35.66	0.56
Open-weight multimodal models: 20–50B																		
InternVL3-38B (Zhu et al., 2025)	41.94	40.20	0.33	26.66	30.31	0.58	37.95	41.52	0.43	35.55	34.52	0.50	31.94	37.33	0.47	34.81	36.77	0.46
Qwen3-VL-32B-Instruct (Bai et al., 2025)	58.28	66.45	0.18	40.23	**46.74**	0.42	41.50	49.61	0.36	36.91	39.08	0.42	35.82	40.73	0.38	42.55	48.52	0.35
Open-weight multimodal models: more than 50B total parameters																		
InternVL3-78B (Zhu et al., 2025)	45.96	43.66	0.30	37.17	40.59	0.49	35.84	39.34	0.46	33.42	34.56	0.52	41.22	46.72	0.45	38.72	40.98	0.44
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	38.02	44.83	0.34	40.42	46.46	0.42	36.66	44.05	0.41	34.38	37.81	0.42	20.62	23.96	0.36	34.02	39.42	0.39
Closed-source multimodal models																		
GPT-5.6-Luna (OpenAI, 2026)	**60.74**	**68.18**	0.19	**41.11**	46.69	**0.39**	44.09	51.18	0.34	34.40	38.13	0.41	34.61	39.48	0.32	42.99	48.73	0.33
Gemini-2.5-Flash (Comanici et al., 2025)	52.84	59.95	0.28	40.84	46.52	0.46	**45.86**	**53.70**	0.34	**41.12**	**45.48**	0.42	**43.46**	**48.59**	0.36	**44.83**	**50.85**	0.37
Claude-Haiku-4.5 (Anthropic, 2025)	56.04	63.92	**0.17**	34.52	40.72	0.40	40.49	48.33	**0.30**	34.03	38.33	**0.38**	32.47	35.98	**0.21**	39.51	45.46	**0.29**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	44.92	52.06	0.37	40.77	46.73	0.44	40.72	48.23	0.42	32.75	35.84	0.49	37.83	42.77	0.40	39.40	45.12	0.42

4 FIREWORLDGPT

FireWorldBench reveals failures in physical-evidence structuring, causal reasoning, and task-specific reasoning. We study three controlled enhancements: physics-aware prompting for input-level structuring, chain-of-thought reasoning for physical inference, and supervised fine-tuning for parameter-level adaptation. These methods isolate the roles of representation, reasoning, and learning; their implementation is detailed in Appendix E.

Physics-Aware Prompting: Coupled-Field Atlas. Physical-world observations are often fragmented across variables, locations, and time, leaving their underlying correspondence implicit. The *Coupled-Field Atlas* transforms these heterogeneous observations into a coordinate-consistent representation by registering physical variables to a shared spatial partition and temporal index while preserving extrema, gradients, flow directions, cross-field co-location, and temporal changes. This prompt-based organization strengthens the model's unified representation and understanding of complex physical worlds.

Chain-of-Thought Reasoning: Constructing Traceable Physical Inference Chains. Complex physical judgments require evidence identification, state estimation, cross-field association, mechanism attribution, and consequence assessment; models that answer directly often bypass these steps. *Constructing Traceable Physical Inference Chains* organizes physical reasoning as an executable inference program constrained by observational evidence. Without introducing additional observations, it guides a frozen model through an evidence-grounded sequence in which intermediate conclusions are linked to observed physical variables, spatial regions, and temporal indices before producing a structured output. This transforms implicit physical intuition into an auditable and physically constrained inference chain for reasoning from partial observations.

Supervised Fine-Tuning: Internalizing Observation-to-Representation Mapping. Even with explicitly organized evidence or guided reasoning, a foundation model may lack parameterized representations that reliably map multi-physics observations to latent states and coupling mechanisms. Domain-specific fine-tuning therefore *internalizes the mapping from physical observations to physical states directly into model parameters*. FireWorldGPT undergoes 1-epoch autoregressive LoRA fine-

Table 2: **Overall evaluation along the physical capability axis on the real-world-aligned subset.** Results aggregate choice-format and open-report scores within the same event groups and observation track. We report accuracy (Acc), macro-F1 (F1), and Brier score for P1–P5 and their average.

Model	P1			P2			P3			P4			P5			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records																		
Open-weight models: fewer than 20B total parameters																		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	35.16	43.98	0.46	30.85	39.77	0.48	32.95	41.36	0.43	29.25	37.84	0.46	34.49	44.73	0.36	32.54	41.53	0.44
Qwen3-8B (Yang et al., 2025)	37.95	45.41	0.49	35.06	43.30	0.49	48.86	57.95	0.31	39.06	48.25	0.42	34.49	45.00	0.41	39.09	47.98	0.42
Gemma-3-12B-IT (Gemma Team et al., 2025)	43.73	52.13	0.35	44.34	55.99	0.34	39.43	47.23	0.38	40.54	49.12	0.41	38.20	47.50	0.40	41.25	50.40	0.38
Open-weight models: 20–50B total parameters																		
Gemma-3-27B-IT (Gemma Team et al., 2025)	56.94	65.09	0.27	43.15	54.25	0.35	44.13	49.62	0.43	43.64	53.41	0.37	36.30	46.36	0.40	44.83	53.75	0.37
Qwen3-32B (Yang et al., 2025)	62.71	69.47	0.19	45.58	54.70	0.34	52.12	60.53	0.27	39.14	45.63	0.43	38.66	49.26	0.36	47.64	55.92	0.32
Open-weight models: 50–200B total parameters																		
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	66.12	72.94	0.15	48.53	58.81	0.23	47.35	54.02	0.26	49.02	58.66	**0.23**	45.00	55.09	0.22	51.20	59.91	0.22
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	52.98	60.05	0.31	56.27	65.96	0.26	46.44	54.85	0.36	46.14	54.91	0.37	43.05	52.97	0.38	48.98	57.75	0.34
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	66.66	72.41	0.17	53.31	61.62	0.25	51.33	57.50	0.28	45.61	53.28	0.31	35.84	47.31	0.28	50.55	58.43	0.26
Open-weight models: more than 200B total parameters																		
Qwen3-235B-A22B (Yang et al., 2025)	53.80	60.64	0.24	47.33	57.06	0.32	46.41	54.39	0.33	41.86	50.41	0.36	32.96	43.89	0.36	44.47	53.28	0.32
DeepSeek-V3.2 (Liu et al., 2025)	63.36	68.94	0.17	59.50	69.38	0.17	47.23	53.03	0.26	40.52	46.41	0.38	35.60	44.73	0.31	49.24	56.50	0.26
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	61.17	66.02	0.15	60.23	69.06	0.18	46.21	52.58	0.33	41.46	47.22	0.38	38.61	47.59	0.29	49.54	56.49	0.27
Closed-source models																		
GPT-5.6-Luna (OpenAI, 2026)	**75.29**	**78.97**	**0.13**	61.60	69.16	0.22	**53.64**	**62.05**	**0.25**	**60.34**	**63.34**	0.29	49.95	58.80	**0.21**	**60.16**	**66.46**	**0.22**
Gemini-2.5-Flash (Comanici et al., 2025)	69.79	75.14	0.15	57.55	67.14	0.23	48.48	55.08	0.34	48.62	56.25	0.35	47.31	56.30	0.29	54.35	61.98	0.27
Claude-Haiku-4.5 (Anthropic, 2025)	63.97	70.40	0.16	**65.12**	**73.31**	**0.15**	50.53	58.03	0.26	55.27	61.21	0.27	40.37	48.34	0.29	55.05	62.26	0.22
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	64.53	70.50	0.22	57.45	66.52	0.25	50.76	58.26	0.33	53.44	60.29	0.31	**50.55**	**60.83**	0.28	55.35	63.28	0.28
I-track: multimodal physics-rendered visual evidence																		
Open-weight multimodal models: 7–8B																		
InternVL3-8B (Zhu et al., 2025)	19.20	19.93	0.60	28.80	29.70	0.32	26.58	28.41	0.40	29.64	32.66	0.32	30.41	36.25	0.42	26.93	29.39	0.41
Qwen3-VL-8B-Instruct (Bai et al., 2025)	42.36	45.89	0.44	41.76	48.93	0.45	29.00	35.59	0.58	30.50	36.14	0.57	39.79	45.25	0.47	36.68	42.36	0.50
Open-weight multimodal models: 20–50B																		
InternVL3-38B (Zhu et al., 2025)	31.03	29.77	0.45	38.70	39.48	0.43	35.95	37.06	0.46	46.25	45.95	0.40	31.16	36.50	0.48	36.62	37.75	0.44
Qwen3-VL-32B-Instruct (Bai et al., 2025)	44.73	49.08	0.33	53.59	59.92	0.31	39.00	46.09	0.40	50.97	58.03	0.31	**48.62**	**53.16**	0.31	47.38	53.26	0.33
Open-weight multimodal models: more than 50B total parameters																		
InternVL3-78B (Zhu et al., 2025)	27.56	25.01	0.54	37.85	39.70	0.43	34.11	34.62	0.47	46.22	45.98	0.40	46.62	52.16	0.42	38.47	39.50	0.45
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	57.70	59.80	0.28	41.50	47.89	0.43	30.77	38.84	0.48	44.39	49.28	0.42	32.96	38.34	0.41	41.47	46.83	0.40
Closed-source multimodal models																		
GPT-5.6-Luna (OpenAI, 2026)	**59.37**	**61.94**	0.20	**60.50**	**65.92**	**0.20**	**41.30**	47.33	0.31	**53.48**	**58.18**	**0.30**	42.91	46.50	0.27	**51.51**	**55.97**	**0.26**
Gemini-2.5-Flash (Comanici et al., 2025)	58.96	60.70	0.30	49.81	55.00	0.39	37.69	46.88	0.40	47.23	53.86	0.39	44.54	50.66	0.35	47.65	53.42	0.37
Claude-Haiku-4.5 (Anthropic, 2025)	58.35	61.68	**0.19**	51.64	57.26	0.28	41.25	**48.80**	**0.28**	43.58	49.34	0.33	34.91	40.00	**0.24**	45.95	51.41	0.26
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	54.39	56.34	0.35	52.41	58.85	0.34	39.86	47.09	0.43	40.70	47.08	0.42	30.00	36.16	0.41	43.47	49.11	0.39

Table 3: **FireWorldGPT results on controlled simulations.** Rows compare Vanilla, physics-aware prompting, chain-of-thought reasoning, and supervised fine-tuning across four foundation models. Acc. and F1 are higher-is-better, whereas lower Brier is better.

Method	Llama-3.1-8B-Instruct			Qwen3-8B			InternVL3-8B			Qwen3-VL-8B		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
Vanilla	31.1	32.9	0.77	33.4	35.7	0.80	25.1	26.4	0.59	29.2	29.5	0.79
+Prompt	33.4 ↑2.3	34.5 ↑1.6	0.80 ↑0.03	35.5 ↑2.1	37.0 ↑1.3	0.82 ↑0.02	29.4 ↑4.3	29.4 ↑3.0	0.56 ↓0.03	34.5 ↑5.3	33.7 ↑4.2	**0.78** ↓0.01
+CoT	32.2 ↑1.1	33.6 ↑0.7	**0.70** ↓0.07	35.7 ↑2.3	37.1 ↑1.4	**0.79** ↓0.01	28.2 ↑3.1	30.0 ↑3.6	**0.45** ↓0.14	31.8 ↑2.6	33.9 ↑4.4	0.82 ↑0.03
+SFT	**43.1** ↑12.0	**44.6** ↑11.7	0.89 ↑0.12	**40.8** ↑7.4	**40.8** ↑5.1	0.84 ↑0.04	**34.1** ↑9.0	**37.3** ↑10.9	0.87 ↑0.28	**41.0** ↑11.8	**39.3** ↑9.8	0.82 ↑0.03

tuning on valid observations and standardized structured answers from non-evaluation data, organized by capability progression. At inference time, it uses the standard foundation-model prompt without an explicit reasoning program or external tool. Isolated from Prompt and CoT, this parameter-level intervention tests whether domain adaptation enables more stable representations and reasoning over complex physical worlds.

5 Experiments

5.1 Main Results

Controlled Simulation Results. Table 1 summarizes performance on the controlled simulation subset across the physical capability axis. The S-track evaluates reasoning from structured sensor and physical records, while the I-track uses multimodal physics-rendered visual evidence. Accuracy and Macro-F1 measure predictive quality, while lower Brier scores indicate better calibration. GPT-5.6-Luna leads the S-track (55.34 Acc / 61.96 F1), Gemini-2.5-Flash leads the I-track (44.83 Acc / 50.85 F1), and Claude-Haiku-4.5 achieves the lowest average Brier score on both tracks, revealing clear capability gaps in complex physical world understanding.

[0]Our project is available at: `https://anonymous.4open.science/r/FireWorldBench/`.

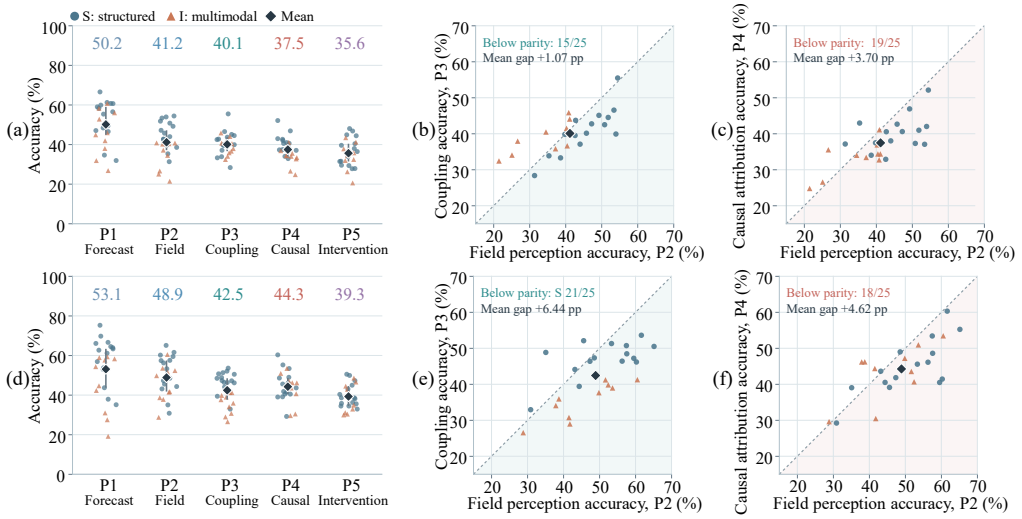


Figure 4: **Physical capability profiles and perception–mechanism gaps.** Performance across physical capabilities, cross-field coupling, and mechanism attribution on the controlled-simulation (top) and real-world-aligned (bottom) subsets.

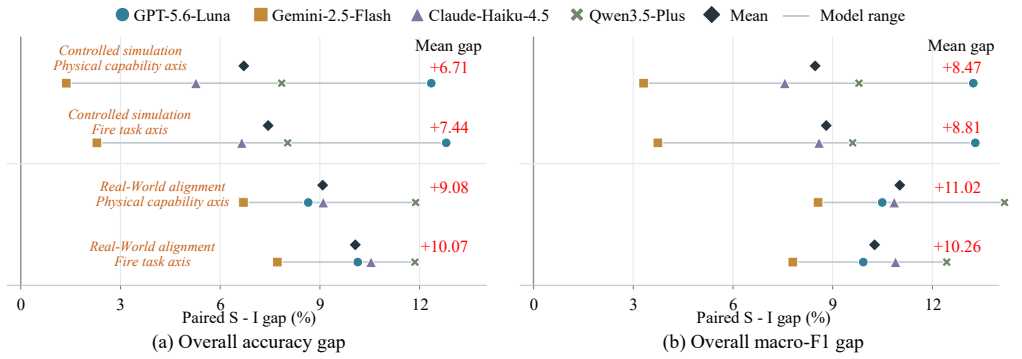


Figure 5: **Paired performance gaps between structured and multimodal observations.** S-track minus I-track differences in (a) Accuracy and (b) Macro-F1 across capability and fire-task axes.

Real-World Alignment Results. Table 2 summarizes performance on the real-world-aligned subset. GPT-5.6-Luna leads both the S-track (60.16 Acc / 66.46 F1 / 0.22 Brier) and I-track (51.51 Acc / 55.97 F1 / 0.26 Brier). Across the four frontier closed models, structured records outperform physics-rendered visual evidence by 9.09 Acc / 11.02 F1 on average, highlighting persistent difficulties in reasoning from real-world-aligned field visualizations.

FireWorldGPT Results. Table 3 compares Vanilla models with physics-aware prompting, chain-of-thought reasoning, and supervised fine-tuning. Prompting and CoT provide moderate gains, while SFT delivers the most consistent improvements in accuracy and Macro-F1, though calibration remains challenging. These results show that FireWorldGPT can internalize mappings from multi-physics observations to physically grounded representations through structured supervision.

5.2 In-Depth Analysis

Physical Understanding Gaps. Figure 4 shows that mechanism attribution trails field perception by 3.70 and 4.62 points on the controlled and real-world-aligned subsets, while cross-field coupling gaps are 1.07 and 6.44 points, indicating that field recognition does not reliably translate into interaction or mechanism understanding. Figure 5 shows that structured observations consistently outperform multimodal evidence, with accuracy gaps increasing from 6.71–7.44 points on controlled simulations to 9.08–10.07 on real-world-aligned data, with similar trends in macro-F1.

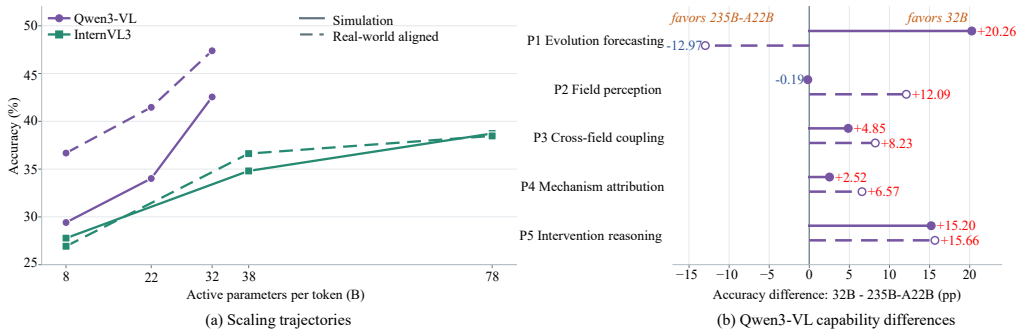


Figure 6: **Active-parameter scaling yields limited returns in physical-world understanding.** (a) I-track accuracy across Qwen3-VL and InternVL3 checkpoints. (b) Capability-wise advantage of the 32B model over the 22B-active checkpoint.

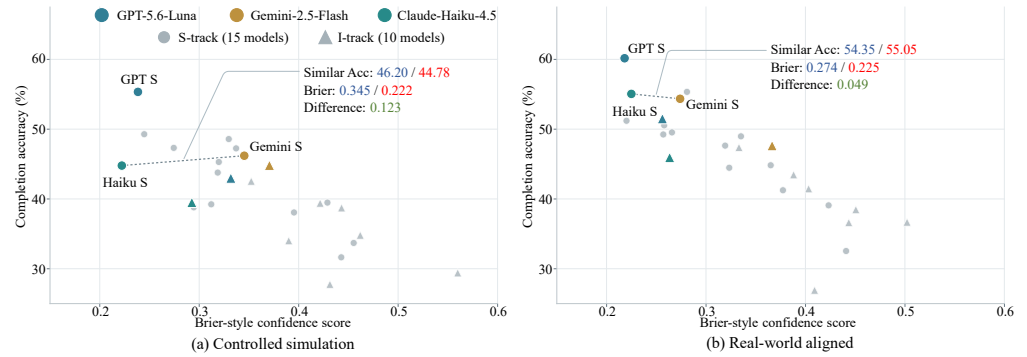


Figure 7: **Physical correctness and confidence quality.** Completion accuracy versus Brier-style confidence scores on the physical capability axis for (a) controlled and (b) real-world-aligned data.

Scaling and Confidence Limitations. Figure 6 shows diminishing returns from scaling: the 32B model broadly outperforms the 22B-active checkpoint, while scaling from 38B to 78B yields only marginal gains. Together with three frontier closed-source models, this suggests that without targeted training on coupled multi-physics data, model scaling alone provides limited returns for physical-world understanding. Figure 7 shows that Gemini-2.5-Flash and Claude-Haiku-4.5 have comparable accuracies on both subsets but substantially different Brier-style scores, showing that correctness and confidence reliability should both be considered.

Expert-Validated Questions and Answers. Appendix F shows that question clarity and observation sufficiency are both 93.0%, Gold-answer and mechanism acceptance reach 96.67% and 99.0%, and expert exact agreement reaches 97.0% on answers and 97.76% on output slots. These results support the reliability of the audited questions and Gold references.

6 Conclusion

We introduced FireWorldBench, a benchmark for complex physical-world intelligence using coupled-field fire dynamics as a canonical stress test. By combining controlled simulations and real-world-aligned events, structured records and multi-physics visualizations, and matched choice and open-report interfaces, FireWorldBench evaluates models along physical-capability and fire-task axes. Our results expose persistent boundaries: current models struggle to transform observable fields into coupled states and causal mechanisms, remain sensitive to observation modality, and show limitations in explanation, confidence calibration, and scaling. Controlled enhancement studies show that supervised adaptation is most effective, enabling FireWorldGPT to map multi-physics observations to physically grounded states and mechanisms. FireWorldBench thus provides a reproducible foundation for advancing multimodal models toward reliable understanding of complex physical worlds.

REFERENCES

Anthropic. Claude Haiku 4.5 system card. System card, October 2025. URL https://www.anthropic.com/system-cards.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4191ef5f6c1576762869ac49281130c9-Abstract.html.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. URL https://arxiv.org/abs/2406.03520.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. *CoRR*, abs/2106.08261, 2021. URL https://arxiv.org/abs/2106.08261.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q6a9W6kzv5.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

Hongqiang Fang and Wai Cheong Tam. Fire data generator (FD-Gen) v1.0.0. Technical Report NIST Technical Note 2332, National Institute of Standards and Technology, 2025. URL https://doi.org/10.6028/NIST.TN.2332.

Zaige Fei, Fan Xu, Junyuan Mao, Yuxuan Liang, Qingsong Wen, Kun Wang, Hao Wu, and Yang Wang. Open-ck: A large multi-physics fields coupling benchmarks in combustion kinetics. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=A23C57icJt.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Minghao Han, Dingkang Yang, Yue Jiang, Yizhou Liu, and Lihua Zhang. Exploring physical intelligence emergence via omni-modal architecture and physical data engine. *arXiv preprint arXiv:2602.07064*, 2026. URL https://arxiv.org/abs/2602.07064.

Yang Jia, Yihan Guo, Yetang Chen, Xinmeng Zhang, Gang Wang, and Qixing Zhang. MmodalFire: A continuous multimodal dataset comprising video and physical sensing data for detecting indoor fires. *Scientific Data*, 13:489, 2026. doi: 10.1038/s41597-026-06810-6. URL https://doi.org/10.1038/s41597-026-06810-6.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Nanxi Li, Xiang Wang, Yuanjie Chen, Haode Zhang, Hong Li, and Yong-Lu Li. Beyond static vision: Scene dynamic field unlocks intuitive physics understanding in multimodal large language models. *arXiv preprint arXiv:2604.03302*, 2026b. doi: 10.48550/arXiv.2604.03302. URL https://arxiv.org/abs/2604.03302.

Puyin Li, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-Fei, and Ehsan Adeli. QUANTIPHY: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33174–33184, 2026c. URL https://openaccess.thecvf.com/content/CVPR2026/html/Puyin_QUANTIPHY_A_Quantitative_Benchmark_Evaluating_Physical_Reasoning_Abilities_of_Vision-Language_CVPR_2026_paper.html.

Jiarong Liang, Max Ku, Ka-Hei Hui, Ping Nie, and Wenhu Chen. Visphyworld: Probing physical reasoning via code-driven video reconstruction. *arXiv preprint arXiv:2602.13294*, 2026. URL https://arxiv.org/abs/2602.13294.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Kevin B. McGrattan, Randall McDermott, Craig Weinschenk, Kristopher Overholt, Simo Hostikka, and Jason Floyd. Fire dynamics simulator technical reference guide, volume 1: Mathematical model. Technical Report

NIST Special Publication 1018, Sixth Edition, National Institute of Standards and Technology, 2013. URL https://doi.org/10.6028/NIST.SP.1018e6.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. URL https://arxiv.org/abs/2410.05363.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958. IEEE, 2026. URL https://arxiv.org/abs/2501.09038.

Khan Muhammad, Jamil Ahmad, and Sung Wook Baik. Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing*, 288:30–42, 2018. doi: 10.1016/j.neucom.2017.04.083. URL https://doi.org/10.1016/j.neucom.2017.04.083.

OpenAI. GPT-5.6 System Card. System card, OpenAI, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6. Accessed: 2026-09-18.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Kristopher J. Overholt. Verification and validation of commonly used empirical correlations for fire scenarios. Technical Report NIST Special Publication 1169, National Institute of Standards and Technology, 2014. URL https://doi.org/10.6028/NIST.SP.1169.

Haosong Peng, Hao Li, Jiaqi Chen, Yuhao Pan, Runmao Yao, Yalun Dai, Fushuo Huo, Fangzhou Hong, Zhaoxi Chen, Haozhao Wang, et al. Spatialbench: Is your spatial foundation model an all-round player? *arXiv preprint arXiv:2605.27367*, 2026.

Qwen Team. Qwen3-Next-80B-A3B-Instruct. Official model card, 2025. URL https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018. doi: 10.48550/arXiv.1803.07616. URL https://arxiv.org/abs/1803.07616.

Nithin Somasekharan, Ling Yue, Yadi Cao, Weichao Li, Patrick Emami, Pochinapeddi Sai Bhargav, Anurag Acharya, Xingyu Xie, and Shaowu Pan. Cfdllmbench: A benchmark suite for evaluating large language models in computational fluid dynamics. *arXiv preprint arXiv:2509.20374*, 2025. URL CFDLLMBench: ABenchmarkSuiteforEvaluatingLargeLanguageModelsinComputationalFluidDynamics.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *Advances in neural information processing systems*, 35:1596–1611, 2022.

Tianyi Tang, Zhuoyi Lin, Zeyu Feng, Tianyi Ma, Yew-Soon Ong, Ivor Tsang, and Haiyan Yin. Causal scaffolding for physical reasoning: A benchmark for causally-informed physical world understanding in VLMs. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Datasets and Benchmarks Track*, 2026. URL https://arxiv.org/abs/2606.05966.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. URL https://arxiv.org/abs/2507.20534.

B. Uğur Töreyin, Yiğithan Dedeoğlu, Uğur Güdükbay, and A. Enis Çetin. Computer vision based method for real-time fire and flame detection. *Pattern Recognition Letters*, 27(1):49–58, 2006. doi: 10.1016/j.patrec.2005.06.015. URL https://doi.org/10.1016/j.patrec.2005.06.015.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Yunhao Ba, Alex Wong, Celso M. de Melo, and Achuta Kadambi. Worldbench: Benchmarking physical understanding of world models by isolating physics concepts. *arXiv preprint arXiv:2601.21282*, 2026. URL https://arxiv.org/abs/2601.21282.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025. URL https://arxiv.org/abs/2511.21471v4.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Aohan Zeng, Xin Lv, Qinkai Zheng, et al. GLM-4.5: Agentic, reasoning, and coding foundation models. *arXiv preprint arXiv:2508.06471*, 2025. URL https://arxiv.org/abs/2508.06471.

Tianhang Zhang, Zilong Wang, Yanfu Zeng, Xiqiang Wu, Xinyan Huang, and Fu Xiao. Building artificial-intelligence digital fire (AID-Fire) system: A real-scale demonstration. *Journal of Building Engineering*, 62: 105363, 2022. doi: 10.1016/j.jobe.2022.105363. URL https://doi.org/10.1016/j.jobe.2022.105363.

Xiaoning Zhang, Xiqiang Wu, Younggi Park, Tianhang Zhang, Xinyan Huang, Fu Xiao, and Asif Usmani. Perspectives of big experimental database and artificial intelligence in tunnel fire research. *Tunnelling and Underground Space Technology*, 108:103691, 2021. doi: 10.1016/j.tust.2020.103691. URL https://doi.org/10.1016/j.tust.2020.103691.

Zhen Zhang, Liang Wang, Songlin Liu, and Yunfei Yin. Intelligent fire location detection approach for extrawide immersed tunnels. *Expert Systems with Applications*, 239:122251, 2024. doi: 10.1016/j.eswa.2023.122251. URL https://doi.org/10.1016/j.eswa.2023.122251.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

Appendix Contents

A Benchmark Construction Details . . . 14
A.1 Construction Scope and Provenance . . . 14
A.2 Controlled Fire-World Specification . . . 14
A.3 Real-World-Aligned Event Reconstruction . . . 14
A.4 Simulation and Multimodal Observation Construction . . . 14
A.5 Gold-First Logical-Item Construction . . . 15
A.6 Quality Control and Split Isolation . . . 15
A.7 Traceability and Reproducibility . . . 15
B Evaluation Protocol and Metrics . . . 16
B.1 Deterministic Scoring Protocol . . . 16
B.2 Completion Accuracy . . . 16
B.3 Interface-Specific Macro-F1 . . . 17
B.4 Evidence F1 . . . 19
B.5 Mechanism Alignment . . . 19
B.6 Brier Score . . . 20
B.7 Gold-Linked Support . . . 21
B.8 Aggregation and Invalid Responses . . . 21
C FireWorldBench Additional Results . . . 22
C.1 Fire Scenario Task Axis Results . . . 23
C.2 Controlled Simulation Results . . . 24
C.3 Real-World-Aligned Results . . . 28
C.4 Open-Report Explanation Metrics . . . 32
C.5 Environment-family analysis . . . 36
D FireWorldGPT Additional Results . . . 37
D.1 Multiple-Choice Results . . . 38
D.2 Open-Report Results . . . 38
E FireWorldGPT Implementation Details . . . 38
E.1 Controlled Enhancement Protocol . . . 38
E.2 Physics-Aware Prompting . . . 39
E.3 Chain-of-Thought Reasoning . . . 39
E.4 Supervised Fine-Tuning . . . 39
E.5 Paired Evaluation and Reproducibility . . . 40
F Expert-Validated Questions and Answers. . . . 40
G Qualitative Benchmark Cases and Multimodal Fire-World Visualizations . . . 40

A Benchmark Construction Details

A.1 Construction Scope and Provenance

FireWorldBench is constructed through two provenance-preserving branches that share a common representation and question-generation protocol, as summarized in Figure 3. The controlled branch contains 494 designed simulation worlds, while the real-world-aligned branch contains 26 event groups reconstructed from observed experimental fires. Together, they cover 47 scene archetypes across the seven environment families summarized in Figure 2 and yield 9,074 question–answer instances: 8,360 from controlled simulations and 714 from real-world-aligned events. The two branches differ in how their initial conditions are obtained, but converge on the same construction invariant: every released instance is traceable to a versioned fire-world specification, a retained physical state, an explicitly delimited observation, and a deterministic Gold target. This invariant allows the benchmark to compare controlled and observed settings without conflating recorded evidence with simulation-completed latent state.

Each instance is annotated along the two complementary axes introduced in Section 3. The physical capability axis comprises Temporal Evolution Forecasting (P1), Physical Field Perception and Grounding (P2), Cross-Field Coupling Understanding (P3), Causal Mechanism Attribution (P4), and Counterfactual Intervention Reasoning (P5). The fire scenario task axis comprises Localized Onset (T1), Coupled Propagation (T2), and Critical Transition (T3). These labels are assigned to the same logical target and provide two cross-cutting views of the resulting dataset rather than two independently generated collections.

A.2 Controlled Fire-World Specification

Each controlled world begins from a structured, versioned specification that defines scene geometry, materials, ignition source and location, heat-release-rate behavior, openings, ventilation, obstacles, sensors, cameras, event timing, and, where applicable, a designated intervention variable. The 47 scene archetypes instantiate the seven environment families through physically meaningful variation in confinement, connectivity, combustible loading, source placement, ambient or mechanical flow, and propagation conditions. Within an archetype, alternative topologies, source configurations, and ventilation regimes produce distinct physical systems rather than cosmetic reskins of a fixed scene.

Objects that provide semantic evidence are registered to their physical counterparts in the simulation. Furniture, vehicles, equipment, storage racks, vegetation, and structural elements therefore enter the solver as flow obstacles, thermal boundaries, radiative occluders, combustible components, or ventilation and control boundaries as appropriate; assets that are purely decorative are excluded from benchmark evidence. This binding prevents the visual scene from implying physical structure that is absent from the underlying world state.

Counterfactual worlds are generated from a shared base specification. A paired intervention preserves the geometry, source world, observation configuration, and all nondesignated conditions while changing only the declared control variable, such as an opening, exhaust setting, or suppression condition. Differences between the resulting trajectories can therefore be attributed to the intervention rather than to an uncontrolled scene change.

A.3 Real-World-Aligned Event Reconstruction

The real-world-aligned branch begins from selected sequences in MmodalFire, a collection of 65 indoor-fire videos synchronized with six types of physical sensing data (Jia et al., 2026). A multimodal information encoder synchronizes the recorded streams and converts the observed chronology, scene configuration, and sensor evidence into event-level constraints for an aligned Fire Dynamics Simulator (FDS) reconstruction. The recorded frames and measurements remain the source of observed evidence; the aligned reconstruction supplies the dense spatial fields and intervention-sensitive state needed to instantiate the same diagnostic, forecasting, and intervention tasks used for controlled worlds.

We preserve this provenance distinction throughout construction. Every retained asset is marked as either observed or simulation-completed, and reconstructed fields are not presented as direct measurements. The observed sequence anchors event timing and appearance, whereas the FDS realization provides a co-registered latent representation of temperature, soot density, velocity, visibility, streamwise flow, and related fields. This separation is essential to real-world alignment: it permits physically structured evaluation around recorded events without treating simulator-derived quantities as independently observed ground truth.

A.4 Simulation and Multimodal Observation Construction

Designed and aligned specifications are executed with FDS (McGrattan et al., 2013; Overholt, 2014) to resolve the spatiotemporal evolution of the coupled fire state. The construction targets a nominal mesh spacing of $0.5\,\mathrm{m}$ or finer, with local refinement around ignition regions and narrow openings when required by geometric

resolution or numerical stability. Because the scene families differ substantially in scale and topology, the exact mesh, solver configuration, event duration, and output cadence are stored with each world rather than imposed as a single global setting. Failed, unstable, or incomplete executions are rejected before any question is generated.

Accepted solver outputs are converted into three co-registered representations: structured textual observations, two-dimensional physical-field visualizations, and three-dimensional event-level scene assets. NIST Smokeview produces the time-indexed three-dimensional views, while an observation extractor derives sensor traces, region-level summaries, and two-dimensional fields from the same retained solver state. Shared world coordinates, region identifiers, and timestamps align all representations. Rendering manifests fix the camera, color mapping, temporal sampling, obstacle mask, and source assets; an accompanying sidecar records the configuration required to reconstruct each visual observation.

The benchmark exposes these representations through two observation tracks. The S-track provides only task-authorized sensor traces, field summaries, spatial regions, and timestamps. The I-track provides the corresponding physics-rendered visual evidence and neutral scene context. The two tracks are generated from the same underlying state wherever both are available, but their evidence contracts remain separate: S-track serialization cannot state the target label, and I-track images cannot expose privileged identifiers, filenames, captions, or hidden geometry. Task-specific temporal cutoffs further determine which timestamps are observable. Forecasting instances reveal no post-cutoff state, state-reconstruction instances withhold the queried variables, and intervention instances expose only the factual or counterfactual branch permitted by the task contract.

A.5 Gold-First Logical-Item Construction

Question construction proceeds from the retained physical state to the target and only then to the observable prompt. For each eligible world and cutoff, deterministic task rules first derive a complete, normalized prediction from the solver record. This private construction record includes the required answer fields, the physical evidence supporting them, and the corresponding mechanism or intervention anchors. The builder then selects only task-admissible observations and assigns the resulting logical item to one physical-capability label and one fire-task label. This target-first order prevents answers from being inferred from rendering artifacts, scene names, or manually written descriptions.

Each track-conditioned logical item is instantiated once in a choice format and once as an open report, with the world, observation, temporal cutoff, target, and split held fixed. Choice questions may require either one option or a set of options. Their alternatives encode competing but physically plausible hypotheses at the same semantic granularity as the Gold answer; option positions, wording, and lengths are audited for template and frequency shortcuts. Open reports expose no candidates and require a structured `prediction` together with a concise `conclusion`, supporting `evidence`, a causal `mechanism`, and `confidence`. Both interfaces are therefore grounded in the same physical target while testing recognition and explicit physical commitment under different answer constraints.

A.6 Quality Control and Split Isolation

Quality control is applied before an item enters the released benchmark. At the world level, we reject incomplete runs, unstable numerical outputs, missing source fields, and construction records that cannot be traced to their specifications. At the item level, we remove cases with insufficient admissible evidence, tied or nonunique physical targets, nonidentifiable intervention effects, or visual observations that do not support the requested inference. Automated consistency checks verify field and timestamp alignment, target derivation, answer uniqueness, option validity, open-report schema completeness, and agreement between the paired answer interfaces. When one member of an answer-format pair fails validation, the complete pair is removed so that format comparisons retain a matched population.

Splits are frozen at the event-group level rather than at the rendered-instance level. All temporal windows, observation variants, answer formats, and counterfactual siblings derived from the same source world or event group remain in the same partition. This prevents a model from encountering the same base dynamics or an alternative intervention branch across training and test sets. Difficulty and template checks use only training-side event groups; revisions may alter question sampling or presentation but never the retained FDS state or physical target used by the sealed evaluation set.

A.7 Traceability and Reproducibility

Every world and instance carries persistent identifiers that link the released question to its source specification and construction record. For controlled worlds, the retained bundle includes the FDS input, execution log, device traces, event records, planar fields, three-dimensional outputs, rendering metadata, and quality-control status. For real-world-aligned events, the manifest additionally records the observed source sequence, synchronization metadata, event constraints, and the boundary between observed and simulation-completed assets. Instance-level

records bind the world identifier, split, axis labels, task contract, observation track, temporal cutoff, answer format, Gold fields, and evidence provenance.

The construction release freezes builder versions, asset inventories, configuration files, and hashes of core artifacts. The rebuild procedure begins from the versioned specifications, regenerates solver-derived observations and question records, and compares the resulting manifests and core data files with the frozen release. Because Gold targets are derived from retained physical states and evaluation is performed offline, the benchmark remains auditable at the level of worlds, observations, logical items, and final scores. Detailed metric definitions and invalid-response handling are provided separately in Appendix B.

B Evaluation Protocol and Metrics

B.1 Deterministic Scoring Protocol

All FireWorldBench metrics are computed by a frozen deterministic evaluator, without an LLM judge. Let $\mathcal{D}$ denote an evaluation set and $i \in \mathcal{D}$ an item. Each item has a response type $q_i \in \{\text{choice}, \text{open}\}$, a gold target y_i, a model prediction $\hat{y}_i$, and an optional confidence $c_i \in [0, 1]$. A choice-format response specifies a set of selected options. An open report contains a structured `prediction` object together with `conclusion`, `evidence`, `mechanism`, and `confidence`. When the Gold record provides `answer_fields`, only those required paths are scored; otherwise, all recursively flattened Gold paths under `prediction` are used. Paths outside this task-specific scoring schema are ignored.

Importantly, the structured fields used by the evaluator are not arbitrary formatting tokens introduced for convenient string matching. They are determined during Gold-first construction from the retained physical state and the task contract, before model responses are observed. Required fields encode the physical commitments that are necessary to solve the corresponding task, such as the relevant spatial region, physical variable, temporal trend, causal driver, risk state, prediction target, intervention outcome, or forecast horizon. Thus, matching a required field means recovering a task-defined physical state or relation under a frozen output ontology. Schema compliance alone receives no correctness credit: once a response is parseable, its score is determined by whether these physically meaningful target values are recovered.

The evaluator first normalizes the response and then applies the same parsing and scoring rules to every model. Unparseable responses, missing required content, illegal choice identifiers, and responses whose status is not `ok` receive deterministic failure scores rather than being repaired, re-prompted, or judged semantically. We report six metrics: Completion Accuracy (Acc), Macro-F1 (F1), Brier score, Evidence-F1 (Evi-F1), Mechanism Alignment (Mech), and Gold-Linked Support (GLS). Acc, F1, and Brier apply to both response interfaces; Evi-F1, Mech, and GLS apply only to open reports. Higher is better for all metrics except Brier.

Acc and Macro-F1 measure whether the required physical target is recovered; Evi-F1 examines whether the model identifies the physical regions and variables that support that target; Mech examines whether the generated explanation recovers the annotated mechanism content; GLS explicitly couples target correctness with Gold-linked physical support; and Brier measures whether confidence is consistent with complete physical correctness. No single metric is interpreted as sufficient evidence of physical understanding in isolation. Their joint use is designed to distinguish a plausible final answer from an answer that is physically grounded, mechanistically supported, and reliably expressed.

Throughout this appendix, $\mathbb{1}[\cdot]$ denotes the indicator function. Any metric ratio with a zero denominator is defined as zero unless stated otherwise.

B.2 Completion Accuracy

Completion Accuracy measures the fraction of required answer content recovered by the prediction and permits partial credit. For a choice-format item, let $\mathcal{C}_i$ and $\widehat{\mathcal{C}}_i$ denote the Gold and predicted option sets. The item-level completion ratio is their Jaccard overlap:

$$r_i^{\text{choice}} = \frac{|\widehat{\mathcal{C}}_i \cap \mathcal{C}_i|}{|\widehat{\mathcal{C}}_i \cup \mathcal{C}_i|}. \tag{1}$$

Choice identifiers are normalized to uppercase letters. Duplicate or nonalphabetic identifiers invalidate the response, and an empty union yields zero. This formulation penalizes both omitted and extraneous selections without rewarding the large number of unselected false options.

For the choice interface, the selected option set represents a commitment among physically competing hypotheses constructed at the same semantic granularity. Completion therefore measures how much of the required physical

conclusion is recovered while penalizing both missing and unsupported physical claims. It is not computed from superficial overlap with the question text.

For an open report, let $\mathcal{S}_i$ denote the set of required flattened output slots, and let y_{is} and $\hat{y}_{is}$ be the Gold and predicted values for slot s. A slot is correct only when it is present and its value exactly matches the Gold value under the evaluator's frozen normalization:

$$r_i^{\text{open}} = \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} \mathbb{1}[s \in \widehat{\mathcal{S}}_i \wedge \hat{y}_{is} = y_{is}]. \tag{2}$$

Here, $\widehat{\mathcal{S}}_i$ is the set of predicted flattened paths. A missing required slot is incorrect. Nested outputs, such as horizon-specific temperature or visibility trends, are scored at their flattened paths. Composite slot values are canonicalized and then treated as atomic values for exact comparison.

The exact slot comparison should therefore be interpreted as an exact comparison of *physical commitments under a controlled ontology*, rather than as a generic natural-language exact-match metric. For example, correctly identifying a region while assigning the wrong physical trend, causal driver, intervention consequence, or temporal horizon does not constitute the same physical state and is intentionally scored as incorrect. Conversely, linguistic material outside the required prediction paths does not improve Completion Accuracy. This design makes the structured report a deterministic test of whether the model has recovered the physical quantities and relations required by the task, while leaving explanation quality to the complementary metrics defined below.

The interface-appropriate ratio is denoted by r_i:

$$r_i = \mathbb{1}[q_i = \text{choice}] r_i^{\text{choice}} + \mathbb{1}[q_i = \text{open}] r_i^{\text{open}}. \tag{3}$$

Completion Accuracy is the item average:

$$\text{Acc}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} r_i. \tag{4}$$

Acc is therefore not whole-item exact match. Exact completion is used only as the binary event underlying the Brier score.

Accordingly, Acc answers the first and most direct evaluation question in FireWorldBench: *does the model recover the physical state, outcome, or intervention consequence that the available observations support?* It deliberately permits partial credit because many benchmark targets contain multiple jointly required physical commitments, while the subsequent metrics determine whether those commitments are balanced across target types and grounded in the appropriate evidence and mechanisms.

B.3 Interface-Specific Macro-F1

Macro-F1 complements item-averaged Acc, but its construction follows the semantics of each response interface. Choice-format F1 balances complete answer-set outcomes, whereas open-report F1 balances required structured output slots. The F1 columns in all tables use the corresponding definition below.

The role of Macro-F1 is distinct from Acc. Physical-world tasks contain heterogeneous states and outcomes whose empirical frequencies need not be balanced. A model that repeatedly predicts a common fire state, dominant trend, or frequent answer structure may obtain nontrivial item-averaged completion without demonstrating uniformly reliable physical reasoning. Macro-F1 therefore asks whether performance remains balanced across distinct physical answer outcomes or structured physical fields, reducing the extent to which frequent or easier outcomes dominate the reported score.

Choice-format Macro-F1. For each choice item, the sorted Gold option set is serialized as one categorical label u_i, and the predicted set is serialized as $\hat{u}_i$. Thus, two option sets that differ by even one option are distinct labels. Let $\mathcal{L}_{\mathcal{D}}$ be the union of Gold and predicted exact-set labels in the choice-format evaluation cell. For each label $\ell \in \mathcal{L}_{\mathcal{D}}$, the confusion counts are

$$\text{TP}_\ell = \sum_{i \in \mathcal{D}} \mathbb{1}[u_i = \ell \wedge \hat{u}_i = \ell]. \tag{5}$$

$$\text{FP}_\ell = \sum_{i \in \mathcal{D}} \mathbb{1}[u_i \neq \ell \wedge \hat{u}_i = \ell]. \tag{6}$$

$$\mathrm{FN}_\ell = \sum_{i \in \mathcal{D}} \mathbb{1}[u_i = \ell \wedge \hat{u}_i \neq \ell]. \tag{7}$$

The F1 score for an exact-set label is

$$F1_\ell = \frac{2\mathrm{TP}_\ell}{2\mathrm{TP}_\ell + \mathrm{FP}_\ell + \mathrm{FN}_\ell}. \tag{8}$$

Choice-format Macro-F1 gives equal weight to all exact-set labels:

$$\mathrm{MacroF1}_{\mathrm{choice}}(\mathcal{D}) = \frac{1}{|\mathcal{L}_\mathcal{D}|} \sum_{\ell \in \mathcal{L}_\mathcal{D}} F1_\ell. \tag{9}$$

A failed or missing choice response is represented by a dedicated missing prediction label, ensuring that it contributes a false negative for the Gold exact-set class rather than disappearing from the computation.

Open-report Slot-Macro-F1. For open reports, Macro-F1 is computed from exact (slot, value) matches and then macro-averaged over structured output slots. Let $\mathcal{S}_\mathcal{D}$ be the union of required flattened slots in an open-report evaluation cell, and let $\mathcal{I}_s$ contain the items for which slot s is required. For each slot, an exactly matched value contributes one true positive:

$$\mathrm{TP}_s = \sum_{i \in \mathcal{I}_s} \mathbb{1}[s \in \widehat{\mathcal{S}}_i \wedge \hat{y}_{is} = y_{is}]. \tag{10}$$

A present but incorrect value contributes one false positive:

$$\mathrm{FP}_s = \sum_{i \in \mathcal{I}_s} \mathbb{1}[s \in \widehat{\mathcal{S}}_i \wedge \hat{y}_{is} \neq y_{is}]. \tag{11}$$

The same incorrect value also misses the Gold value and therefore contributes one false negative; an omitted required slot contributes one false negative but no false positive:

$$\mathrm{FN}_s = \sum_{i \in \mathcal{I}_s} \mathbb{1}[s \notin \widehat{\mathcal{S}}_i \vee (s \in \widehat{\mathcal{S}}_i \wedge \hat{y}_{is} \neq y_{is})]. \tag{12}$$

The slot-level exact-match F1 is

$$F1_s = \frac{2\mathrm{TP}_s}{2\mathrm{TP}_s + \mathrm{FP}_s + \mathrm{FN}_s}. \tag{13}$$

The reported open-report F1 is the unweighted mean over scored slots:

$$\text{Slot-MacroF1}(\mathcal{D}) = \frac{1}{|\mathcal{S}_\mathcal{D}|} \sum_{s \in \mathcal{S}_\mathcal{D}} \frac{2\mathrm{TP}_s}{2\mathrm{TP}_s + \mathrm{FP}_s + \mathrm{FN}_s}. \tag{14}$$

This definition assigns equal weight to heterogeneous slots such as `location`, `trend`, `driver`, and `risk`, irrespective of how frequently each slot occurs or how many categorical values it can take. In particular, it does not macro-average the raw value labels within a slot. Extra paths outside the task-specific scored slot set do not affect this metric.

This equal weighting is particularly important for evaluating physical-world understanding. The scored slots correspond to qualitatively different aspects of a physical state: identifying *where* an event occurs is not interchangeable with identifying *how* a field evolves, *what* drives the change, or *which* region becomes hazardous. Slot-Macro-F1 prevents a model from compensating for systematic failure on one such physical component by performing well on a more frequent component. Together, Acc and Macro-F1 therefore measure both the amount of physical target content recovered and the breadth with which that recovery extends across heterogeneous physical commitments.

B.4 EVIDENCE F1

Evidence F1 is a deterministic proxy for whether an open report identifies the Gold-linked spatial regions and physical variables supporting its prediction.

The motivation is that a physically correct conclusion should be grounded in the observations that make that conclusion identifiable. In FireWorldBench, Gold evidence anchors are derived from the same retained physical state used to construct the target. They encode which spatial region and which physical variable provide task-relevant support, rather than generic keywords expected to appear in a fluent fire description. Consequently, Evidence F1 asks whether the model attends to the physically relevant part of the world state rather than merely producing a plausible final answer.

Let t_i be the lowercased concatenation of the generated `conclusion`, `evidence`, and `mechanism` fields:

$$t_i = \text{lower}\left(\text{concat}(\texttt{conclusion}_i, \texttt{evidence}_i, \texttt{mechanism}_i)\right). \tag{15}$$

Each Gold `evidence_claim` yields an anchor $a = (r_a, v_a)$, where r_a is its lowercased region identifier and v_a is the first physical variable matched from the frozen vocabulary

$$\mathcal{V} = \{\texttt{temperature}, \texttt{visibility}, \texttt{soot}, \texttt{flow}, \texttt{velocity}, \texttt{risk}, \texttt{fire}, \texttt{smoke}, \texttt{sensor}, \texttt{ventilation}\}. \tag{16}$$

Let $\mathcal{A}_i$ be the Gold anchor collection and $K(a)$ the set of nonempty components of anchor a. Let $\text{hit}(x, t_i)$ denote the evaluator's deterministic match for component x. The anchor-level evidence recall is

$$R_i^{\text{Ev}} = \frac{1}{\max(1, |\mathcal{A}_i|)} \sum_{a \in \mathcal{A}_i} \frac{1}{\max(1, |K(a)|)} \sum_{x \in K(a)} \mathbb{1}[\text{hit}(x, t_i)]. \tag{17}$$

Let $\mathcal{M}_i$ contain all case-insensitive region identifiers matching `R\d+` and all vocabulary terms in $\mathcal{V}$ extracted from t_i. Let $\mathcal{G}_i$ collect the corresponding nonempty Gold anchor components:

$$\mathcal{G}_i = \bigcup_{a \in \mathcal{A}_i} K(a). \tag{18}$$

Evidence precision penalizes unsupported region or variable mentions:

$$P_i^{\text{Ev}} = \frac{|\mathcal{M}_i \cap \mathcal{G}_i|}{\max(1, |\mathcal{M}_i|)}. \tag{19}$$

The item-level Evidence F1 is

$$\text{Ev}_i = \frac{2 P_i^{\text{Ev}} R_i^{\text{Ev}}}{P_i^{\text{Ev}} + R_i^{\text{Ev}}}. \tag{20}$$

Evidence F1 rewards recovery of annotated anchors while penalizing unsupported mentions. It is a rule-based region–variable proxy, not semantic tuple matching or model-based judging; mentioning a correct region and a correct variable does not by itself establish that the generated text relates them correctly.

This limitation is deliberate and defines the role of Evi-F1 within the complete protocol. Evi-F1 is not intended to establish causal reasoning by itself; it tests the preceding requirement that the model has located the physical evidence on which such reasoning should be based. A report that mentions the wrong region or the wrong physical field cannot be considered well grounded even if its final answer happens to be correct. Whether the recovered evidence is accompanied by the appropriate mechanism and actually supports the final physical commitment is evaluated separately by Mech and GLS. Thus, Evi-F1 forms the evidence-grounding stage of the overall evaluation chain.

B.5 MECHANISM ALIGNMENT

Mechanism Alignment measures recall of the normalized lexical content in the Gold causal mechanism.

Whereas Evi-F1 asks *what physical evidence is relevant*, Mechanism Alignment asks whether the explanation recovers the physical process used by the Gold construction to connect that evidence to the target state or outcome.

The Gold mechanism statements are attached to the task-specific physical target during benchmark construction and describe physically meaningful processes such as transport, buoyancy-driven motion, ventilation effects, field propagation, accumulation, or intervention-induced changes. The resulting score therefore provides a deterministic indication of whether the generated explanation contains the mechanism concepts required by the reference physical interpretation.

The tokenizer lowercases the text and extracts tokens matching `[a-z][a-z0-9_]+`. Let g_i^{mech} be the concatenation of all Gold `mechanism_claims.statement` strings and m_i the generated `mechanism`. The Gold token set is

$$\mathcal{G}_i^{\mathrm{mech}} = \mathrm{toks}(g_i^{\mathrm{mech}}) \setminus \mathcal{W}. \tag{21}$$

The generated token set is

$$\mathcal{M}_i^{\mathrm{mech}} = \mathrm{toks}(m_i). \tag{22}$$

The frozen Gold-side stop-word set is

$$\mathcal{W} = \{\texttt{the}, \texttt{and}, \texttt{with}, \texttt{that}, \texttt{this}, \texttt{from}, \texttt{into}, \texttt{first}\}. \tag{23}$$

The item score is Gold-token recall:

$$\mathrm{Mech}_i = \frac{|\mathcal{G}_i^{\mathrm{mech}} \cap \mathcal{M}_i^{\mathrm{mech}}|}{|\mathcal{G}_i^{\mathrm{mech}}|}. \tag{24}$$

Because Mech is recall-only, additional irrelevant mechanism tokens are not directly penalized. It should therefore be interpreted jointly with Evidence F1 and GLS rather than as an independent measure of semantic or causal faithfulness.

Accordingly, we do not treat lexical mechanism recovery alone as sufficient evidence of causal understanding. Its value comes from its position between evidence grounding and Gold-linked support: Evi-F1 verifies that the report recovers the relevant observable variables and regions, Mech tests whether the annotated physical mechanism is represented in the explanation, and GLS subsequently requires a correct physical prediction to be explicitly supported by task-aware evidence or mechanism anchors. This decomposition makes it possible to distinguish three qualitatively different failures: selecting the wrong evidence, describing the wrong mechanism, and reaching a prediction that is not supported by the physical explanation.

B.6 Brier Score

Brier measures agreement between reported confidence and exact whole-item completion. Its binary target is not the partial-credit completion ratio but the event that every required answer component is correct:

$$z_i = \mathbb{1}[r_i = 1]. \tag{25}$$

The response schema requires a numeric confidence value $c_i \in [0, 1]$ for every scored item, and all responses included in the reported Brier results satisfy this validity requirement. The Brier score is therefore computed over the complete evaluation set:

$$\mathrm{Brier}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} (c_i - z_i)^2. \tag{26}$$

An open report that correctly recovers eight of nine required slots contributes $8/9$ to Acc but has $z_i = 0$; assigning high confidence to that incomplete report is therefore strongly penalized. Lower Brier is better. Because the target itself depends on exact completion, this metric measures confidence–completion consistency rather than correctness-independent calibration.

Confidence quality constitutes a complementary aspect of reliable physical-world reasoning. Two models may recover a similar fraction of physical targets while differing substantially in whether they recognize when their complete physical interpretation is correct. This distinction is particularly relevant for partially observed and intervention-sensitive systems, where an incorrect but highly confident physical judgment is qualitatively different from an uncertain one. Brier is therefore not used as another measure of physical-answer accuracy; instead, it measures whether a model's stated confidence is consistent with the strict event that all required physical commitments for an item have been recovered.

B.7 GOLD-LINKED SUPPORT

GLS measures whether correct open-report predictions are explicitly supported by task-aware evidence or mechanism anchors.

GLS provides the final link between answer correctness and physical justification. Acc can establish that a model produced the required target, Evi-F1 can establish that relevant evidence appears in the report, and Mech can establish that reference mechanism content is recovered; none of these conditions alone guarantees that the *correct prediction is supported by the corresponding physical evidence*. GLS therefore evaluates their conjunction at the level of required prediction slots.

For each required slot $s \in \mathcal{S}_i$, define a correctness indicator using the frozen equivalence function:

$$\mathrm{corr}_{is} = \mathbb{1}[s \in \widehat{\mathcal{S}}_i]\mathbb{1}[\mathrm{equiv}(\hat{y}_{is}, y_{is})]. \tag{27}$$

Let g_i be the concatenated Gold anchor text and $\phi_s(y_{is}, t_i, g_i)$ the frozen task-aware support predicate. The supported-correctness indicator is

$$\mathrm{supp}_{is} = \mathrm{corr}_{is}\, \phi_s(y_{is}, t_i, g_i). \tag{28}$$

The item-level GLS score is

$$\mathrm{GLS}_i = \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} \mathrm{supp}_{is}. \tag{29}$$

For equiv, numeric values are compared under floating-point equality, whereas nonnumeric values are compared after the evaluator removes nonalphanumeric characters. The predicate ϕ_s uses word-boundary matching and frozen aliases, such as mapping `thermal`, `heat`, and `hot` to temperature-related support and mapping `up` to increasing-trend expressions. It applies the following slot-dependent rules:

1. A region-valued slot, or a path containing `region`, is supported only if its value occurs in both the Gold anchors and the generated explanation.
2. A path ending in `_trend` requires the explanation to mention both the corresponding variable and a compatible trend term. A nested `10s`, `30s`, or `60s` path additionally requires the matching horizon.
3. Paths containing `signal`, `driver`, `variable`, or `target` require the predicted value to occur in both the Gold anchors and the explanation.
4. A numeric Gold value requires the corresponding number to appear in the explanation, or the explanation to contain `all regions` or `across regions`.
5. All remaining slots require the normalized Gold value to occur in both the Gold anchors and the generated explanation.

These slot-dependent predicates are designed to reflect the semantics of the physical commitment being evaluated. A spatial prediction must be supported in the corresponding region; a temporal prediction must identify both the physical variable and a compatible change direction, together with the relevant horizon when applicable; and a driver or target prediction must be explicitly connected to the corresponding physical quantity. The evaluator therefore does not treat all textual overlap as equivalent support. Instead, the required support condition changes according to the physical role of the predicted field.

The denominator includes all required Gold slots, not only correctly predicted slots. GLS is therefore a joint correctness-and-support rate, rather than support conditioned on correctness.

This conjunction is central to the interpretation of GLS. A fluent explanation containing correct physical terminology cannot obtain GLS credit for an incorrect prediction, and a correct prediction cannot obtain GLS credit when its required physical support is absent. GLS thus operationalizes the benchmark's strongest report-level requirement: the model must not only reach the correct physical commitment, but also expose evidence or mechanism information that is consistent with why that commitment is correct under the Gold physical state.

B.8 AGGREGATION AND INVALID RESPONSES

Items are partitioned by benchmark split, axis label, observation track, and response interface before scoring. Acc, interface-specific F1, and Brier are reported for both choice-format and open-report cells. The three explanation-specific metrics are arithmetic means over open-report items only:

$$\mathrm{Ev}(\mathcal{D}_{\mathrm{open}}) = \frac{1}{|\mathcal{D}_{\mathrm{open}}|} \sum_{i \in \mathcal{D}_{\mathrm{open}}} \mathrm{Ev}_i. \tag{30}$$

$$\mathrm{Mech}(\mathcal{D}_{\mathrm{open}}) = \frac{1}{|\mathcal{D}_{\mathrm{open}}|} \sum_{i \in \mathcal{D}_{\mathrm{open}}} \mathrm{Mech}_i. \tag{31}$$

$$\mathrm{GLS}(\mathcal{D}_{\mathrm{open}}) = \frac{1}{|\mathcal{D}_{\mathrm{open}}|} \sum_{i \in \mathcal{D}_{\mathrm{open}}} \mathrm{GLS}_i. \tag{32}$$

The evaluator's `overall` entry recomputes each metric over all eligible items and is therefore instance-weighted. By contrast, the Average columns in the paper first compute cell-level scores and then assign equal weight to the five physical capability dimensions P1–P5 or the three fire scenario task dimensions T1–T3. For any reported axis containing K cells, this aggregation is

$$\mathrm{AvgAxis}(M) = \frac{1}{K} \sum_{k=1}^{K} M(\mathcal{D}_k). \tag{33}$$

Here, M denotes any reported metric and $\mathcal{D}_k$ the evaluation cell for the kth axis category under a fixed split, track, and response interface. This cell-level averaging prevents categories with more instances from dominating the axis summary.

Equal weighting across axis categories is also important for the intended physical interpretation of the benchmark. It prevents a capability or fire-task category containing more generated instances from dominating the overall assessment and ensures that forecasting, grounding, coupling, mechanism attribution, intervention reasoning, and their corresponding operational tasks remain visible as distinct sources of model strength or failure.

Failed or missing predictions receive $r_i = 0$. For choice-format F1, a failed response is assigned the missing prediction label. For open-report Slot-Macro-F1, every required Gold slot in a failed response contributes one false negative and no true positive. Failed open reports receive zero for Evidence F1, Mechanism Alignment, and GLS. If multiple responses share a `qa_id`, a response with `status=ok` is retained preferentially. All six metrics are reported separately and are not collapsed into a single composite score.

We deliberately retain the six measurements separately because they diagnose different stages of physical-world understanding rather than different estimators of the same quantity. High Acc or F1 with low Evi-F1 indicates that a model can reach physical answers without reliably recovering the annotated evidence; high evidence recovery with low Mech indicates difficulty connecting observations through the intended physical process; high answer and evidence scores with low GLS indicate that the final commitment is not consistently linked to its required support; and Brier distinguishes answer quality from confidence reliability. The complete protocol therefore evaluates physical-world understanding through a set of mutually constraining observations rather than through final-answer correctness alone.

In this sense, the structured fields and deterministic matching rules serve as an auditable interface between the latent physical state of the benchmark and the model's externally observable commitments. Correct field recovery is a necessary condition for claiming that the model has produced the corresponding task-defined physical judgment, while Evi-F1, Mech, and GLS test progressively stronger requirements on how that judgment is grounded and justified. This layered design allows FireWorldBench to separate recognition, physical-state recovery, evidence grounding, mechanism recovery, evidence-supported reasoning, and confidence reliability while avoiding reliance on an opaque model-based evaluator.

C FireWorldBench Additional Results

The main paper summarizes FireWorldBench performance by averaging the multiple-choice and open-report interfaces within each evaluation cell. In this section, we provide the corresponding interface-specific results to make the benchmark behavior fully transparent. We retain the same model grouping, observation tracks, evaluation metrics, and axis-wise aggregation protocol as in the main paper, but report multiple-choice and open-report performance separately.

Throughout this appendix, we follow the canonical taxonomy defined in the main text. The physical capability axis progresses from Temporal Evolution Forecasting (P1), Physical Field Perception and Grounding (P2), Cross-Field Coupling Understanding (P3), and Causal Mechanism Attribution (P4), to Counterfactual Intervention Reasoning (P5). The fire scenario task axis consists of Localized Onset (T1), Coupled Propagation (T2), and Critical Transition (T3). All fire-axis results below are ordered according to this canonical T1–T3 definition.

Table 4: **Overall evaluation along the fire scenario task axis on the controlled simulation subset.** Models are evaluated by averaging matched choice-format and open-report scores under the same simulated fire-world entry, observation track, task dimension, and metric. Each task reports accuracy (Acc), macro-F1 (F1), and Brier score, with Average summarizing performance across T1–T3.

Model	T1			T2			T3			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records												
Open-weight models: fewer than 20B total parameters												
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	32.55	40.39	0.46	29.81	39.07	0.45	32.57	41.68	0.43	31.64	40.38	0.45
Qwen3-8B (Yang et al., 2025)	37.53	43.84	0.44	35.61	45.19	0.39	44.50	52.78	0.33	39.21	47.27	0.39
Gemma-3-12B-IT (Gemma Team et al., 2025)	37.82	46.69	0.43	29.27	38.84	0.47	34.58	42.19	0.46	33.89	42.57	0.46
Open-weight models: 20–50B total parameters												
Gemma-3-27B-IT (Gemma Team et al., 2025)	46.67	52.04	0.39	39.57	49.21	0.38	57.87	64.41	0.24	48.04	55.22	0.34
Qwen3-32B (Yang et al., 2025)	43.62	50.38	0.36	38.83	48.69	0.36	55.51	63.15	0.23	45.99	54.07	0.32
Open-weight models: 50–200B total parameters												
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	41.75	48.70	0.32	34.07	43.85	0.31	44.15	52.24	0.24	39.99	48.26	0.29
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	41.19	48.28	0.43	31.81	41.05	0.50	46.86	54.91	0.35	39.95	48.08	0.43
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	38.74	45.40	0.37	35.97	45.91	0.32	46.47	54.87	0.23	40.39	48.73	0.31
Open-weight models: more than 200B total parameters												
Qwen3-235B-A22B (Yang et al., 2025)	41.45	47.47	0.39	39.72	49.26	0.34	52.60	60.24	0.22	44.59	52.32	0.32
DeepSeek-V3.2 (Liu et al., 2025)	46.32	52.69	0.31	45.43	54.62	0.26	58.47	65.22	0.18	50.07	57.51	0.25
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	47.46	53.20	0.35	41.50	51.17	0.30	57.55	64.63	0.18	48.83	56.33	0.27
Closed-source models												
GPT-5.6-Luna (OpenAI, 2026)	**49.62**	54.05	0.35	**56.51**	**64.91**	0.23	**63.90**	**70.05**	**0.14**	**56.68**	**63.00**	0.24
Gemini-2.5-Flash (Comanici et al., 2025)	46.64	52.80	0.40	37.95	47.53	0.40	56.96	63.87	0.24	47.18	54.73	0.35
Claude-Haiku-4.5 (Anthropic, 2025)	47.53	**54.08**	**0.30**	40.93	50.77	**0.22**	51.96	59.99	0.15	46.80	54.94	**0.22**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	44.96	49.98	0.40	39.36	49.08	0.39	58.84	66.26	0.23	47.72	55.11	0.34
I-track: multimodal physics-rendered visual evidence												
Open-weight multimodal models: 7–8B												
InternVL3-8B (Zhu et al., 2025)	26.91	37.90	**0.36**	29.93	42.42	0.47	26.66	36.83	0.46	27.83	39.05	0.43
Qwen3-VL-8B-Instruct (Bai et al., 2025)	23.34	28.32	0.66	28.15	35.27	0.56	34.35	40.86	0.49	28.61	34.82	0.57
Open-weight multimodal models: 20–50B												
InternVL3-38B (Zhu et al., 2025)	30.47	39.48	0.56	35.59	48.39	0.46	38.46	50.12	0.38	34.84	45.99	0.46
Qwen3-VL-32B-Instruct (Bai et al., 2025)	40.53	46.13	0.42	36.96	44.45	0.39	50.86	57.52	0.25	42.78	49.37	0.35
Open-weight multimodal models: more than 50B total parameters												
InternVL3-78B (Zhu et al., 2025)	37.70	**49.35**	0.50	33.51	46.06	0.48	44.32	55.89	0.35	38.51	50.43	0.44
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	38.63	43.18	0.46	35.58	42.82	0.38	32.93	37.58	0.35	35.71	41.19	0.40
Closed-source multimodal models												
GPT-5.6-Luna (OpenAI, 2026)	40.77	44.86	0.42	39.21	46.07	0.34	**51.67**	**58.21**	0.24	43.88	49.71	0.33
Gemini-2.5-Flash (Comanici et al., 2025)	**42.42**	46.45	0.45	**42.67**	**50.52**	0.37	49.59	56.00	0.31	**44.89**	**50.99**	0.37
Claude-Haiku-4.5 (Anthropic, 2025)	36.62	41.29	0.41	35.96	43.54	**0.32**	47.86	54.22	**0.18**	40.15	46.35	**0.30**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	41.99	45.81	0.43	34.62	41.89	0.47	42.46	48.83	0.38	39.69	45.51	0.42

For Acc. and F1, higher values indicate better performance, whereas lower Brier scores indicate better calibration. The reported Average columns follow the same protocol as the main paper: scores are first computed independently for each evaluation dimension and are then averaged with equal weight across P1–P5 for the physical capability axis or T1–T3 for the fire scenario task axis. This prevents dimensions containing more evaluation instances from disproportionately dominating the axis-level summary.

C.1 Fire Scenario Task Axis Results

Table 4 provides an integrated evaluation of complex physical world understanding on the controlled simulation subset along the fire scenario task axis. It aggregates choice-format and open-report results under matched simulated fire-world entries, observation tracks, task dimensions, and metrics, thereby reducing interface-specific variance and focusing on task-level physical competence. The results show that task-level physical understanding remains strongly observation-track dependent, even under controlled conditions. GPT-5.6-Luna leads the S-track (56.68 Acc / 63.00 F1), whereas Gemini-2.5-Flash leads the I-track (44.89 Acc / 50.99 F1); Claude-Haiku-4.5 achieves the lowest average Brier score on both tracks.

Takeaway: Even in clean, controlled fire worlds, current multimodal large language models do not exhibit robust complex physical world intelligence: clear limitations persist in both physical-capability reasoning and task-level problem solving.

Table 5 summarizes fire-world task performance on the real-world-aligned subset. Performance remains markedly dependent on the observation track: GPT-5.6-Luna attains the highest average predictive scores on both structured records (62.88 Acc / 62.52 F1) and physics-rendered visual evidence (52.74 Acc / 52.60 F1). For the four frontier

Table 5: **Overall evaluation along the fire scenario task axis on the real-world-aligned subset.** Models are evaluated by averaging matched choice-format and open-report scores under the same real-world-aligned event group, observation track, task dimension, and metric. Each task reports accuracy (Acc), macro-F1 (F1), and Brier score, with Average summarizing performance across T1–T3.

Model	T1			T2			T3			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records												
Open-weight models: fewer than 20B total parameters												
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	33.84	34.05	0.44	25.69	30.04	0.50	34.96	37.76	0.43	31.50	33.95	0.46
Qwen3-8B (Yang et al., 2025)	39.12	40.15	0.45	39.94	41.73	0.39	36.93	41.02	0.47	38.66	40.97	0.44
Gemma-3-12B-IT (Gemma Team et al., 2025)	45.82	46.33	0.34	35.67	39.71	0.43	42.10	43.52	0.37	41.20	43.18	0.38
Open-weight models: 20–50B total parameters												
Gemma-3-27B-IT (Gemma Team et al., 2025)	43.95	44.05	0.36	42.89	43.80	0.40	50.85	52.98	0.31	45.89	46.94	0.36
Qwen3-32B (Yang et al., 2025)	41.95	42.86	0.40	48.58	49.92	0.30	55.61	56.71	0.24	48.71	49.83	0.31
Open-weight models: 50–200B total parameters												
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	49.77	51.74	0.24	46.42	46.78	0.24	59.90	61.69	0.17	52.03	53.40	0.22
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	51.17	53.27	0.32	49.21	51.72	0.32	50.06	51.65	0.33	50.14	52.21	0.32
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	49.17	50.86	0.30	51.31	52.90	0.25	57.56	59.20	0.20	52.68	54.32	0.25
Open-weight models: more than 200B total parameters												
Qwen3-235B-A22B (Yang et al., 2025)	45.42	44.88	0.35	44.52	46.75	0.33	47.66	49.45	0.28	45.86	47.03	0.32
DeepSeek-V3.2 (Liu et al., 2025)	51.30	52.01	0.27	47.40	48.16	0.27	55.17	56.68	0.21	51.29	52.28	0.25
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	51.98	52.99	0.29	47.48	48.59	0.27	54.51	54.46	0.19	51.32	52.01	0.25
Closed-source models												
GPT-5.6-Luna (OpenAI, 2026)	56.24	56.44	0.29	**64.60**	**63.22**	**0.19**	**67.82**	**67.90**	**0.15**	**62.88**	**62.52**	0.21
Gemini-2.5-Flash (Comanici et al., 2025)	52.48	52.96	0.32	52.04	52.61	0.27	63.16	62.79	0.20	55.89	56.12	0.26
Claude-Haiku-4.5 (Anthropic, 2025)	**58.65**	**58.91**	**0.23**	57.95	57.56	0.19	57.01	59.29	0.20	57.87	58.58	**0.21**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	52.10	52.60	0.33	58.44	60.00	0.24	60.41	62.29	0.24	56.98	58.30	0.27
I-track: multimodal physics-rendered visual evidence												
Open-weight multimodal models: 7–8B												
InternVL3-8B (Zhu et al., 2025)	28.14	39.70	0.30	28.66	40.63	0.38	22.88	28.35	0.54	26.56	36.23	0.41
Qwen3-VL-8B-Instruct (Bai et al., 2025)	37.31	39.62	0.52	30.56	31.11	0.55	41.52	43.70	0.45	36.46	38.14	0.50
Open-weight multimodal models: 20–50B												
InternVL3-38B (Zhu et al., 2025)	35.19	45.65	0.51	45.75	**56.25**	0.35	31.08	38.86	0.46	37.34	46.92	0.44
Qwen3-VL-32B-Instruct (Bai et al., 2025)	53.06	52.88	0.32	43.22	43.25	0.36	46.01	46.17	0.32	47.43	47.43	0.33
Open-weight multimodal models: more than 50B total parameters												
InternVL3-78B (Zhu et al., 2025)	37.49	49.96	0.47	41.63	52.44	0.39	33.82	42.26	0.50	37.64	48.22	0.45
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	38.73	38.82	0.48	39.60	40.56	0.40	49.59	50.84	0.32	42.64	43.40	0.40
Closed-source multimodal models												
GPT-5.6-Luna (OpenAI, 2026)	**55.50**	**55.36**	**0.28**	**48.75**	48.96	0.26	53.98	53.48	0.22	**52.74**	**52.60**	**0.25**
Gemini-2.5-Flash (Comanici et al., 2025)	49.43	49.87	0.40	40.86	40.63	0.39	**54.24**	**54.46**	0.32	48.17	48.32	0.37
Claude-Haiku-4.5 (Anthropic, 2025)	47.86	48.26	0.34	43.45	43.89	**0.26**	50.67	50.92	**0.20**	47.33	47.69	0.27
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	49.74	50.16	0.37	39.24	39.38	0.41	46.39	48.07	0.37	45.12	45.87	0.38

closed models assessed under both tracks, structured records yield an average gain of 10.06 Acc and 10.26 F1. The resulting gap shows that converting visualized multi-field observations into dependable fire-task decisions remains a central limitation.

Takeaway: Even on relatively simple real-world-aligned events, current models retain clear boundaries in physical-world understanding: they reason more reliably from structured records than from visualized fields, which remain difficult to map into stable latent-state, mechanism, and intervention judgments.

C.2 Controlled Simulation Results

We first decompose performance on the controlled simulation subset. Tables 6 and 7 report results along the physical capability axis, while Tables 8 and 9 provide the complementary decomposition along the fire scenario task axis. For each axis, we report the structured-observation S-track and the multimodal image-observation I-track separately. This decomposition exposes whether the aggregate trends reported in the main paper persist when models must either select among candidate hypotheses or construct a physical-world report without answer candidates.

Table 6: **Choice-format evaluation along the physical capability axis.** Models are evaluated on the controlled simulation subset under the structured-observation track (S-track) and multimodal image-observation track (I-track). Each dimension reports accuracy (Acc), macro-F1 (F1), and Brier score; Average is the equal-weight mean across P1–P5.

Model	P1			P2			P3			P4			P5			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records																		
Open-weight models: fewer than 20B total parameters																		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	23.33	34.36	0.59	22.21	31.01	0.65	23.52	34.43	0.58	28.83	40.15	0.52	26.59	37.92	0.49	24.90	35.57	0.56
Qwen3-8B (Yang et al., 2025)	38.40	49.22	0.33	30.75	38.93	0.55	27.70	38.56	0.47	39.31	50.16	0.39	33.70	44.03	0.44	33.97	44.18	0.44
Gemma-3-12B-IT (Gemma Team et al., 2025)	22.14	31.59	0.64	32.70	43.56	0.52	32.36	45.62	0.48	31.25	42.09	0.53	28.65	38.52	0.57	29.42	40.28	0.55
Open-weight models: 20–50B total parameters																		
Gemma-3-27B-IT (Gemma Team et al., 2025)	73.23	78.88	0.17	**57.11**	**64.73**	0.31	53.04	64.37	0.29	33.91	41.14	0.55	**61.92**	**68.99**	0.26	55.84	63.62	0.32
Qwen3-32B (Yang et al., 2025)	62.61	70.99	0.18	43.31	52.80	0.35	38.38	50.58	0.35	41.31	51.15	0.36	42.98	53.16	0.33	45.72	55.74	0.31
Open-weight models: 50–200B total parameters																		
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	40.04	50.20	0.24	44.53	54.04	0.30	31.25	43.66	0.31	33.29	43.55	0.34	27.32	37.48	0.36	35.29	45.79	0.31
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	37.80	48.37	0.40	41.01	51.79	0.42	36.92	49.27	0.42	31.48	40.23	0.54	33.41	42.44	0.49	36.12	46.42	0.45
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	43.24	54.48	0.21	36.74	46.40	0.33	37.71	50.39	0.27	34.19	44.61	0.35	32.57	42.72	0.30	36.89	47.72	0.29
Open-weight models: more than 200B total parameters																		
Qwen3-235B-A22B (Yang et al., 2025)	57.51	66.67	0.16	43.30	51.71	0.39	43.38	55.52	0.30	39.49	48.34	0.37	36.50	45.33	0.35	44.04	53.51	0.31
DeepSeek-V3.2 (Liu et al., 2025)	64.44	71.81	0.13	48.09	58.01	0.27	45.03	56.10	0.24	45.64	53.26	0.29	46.65	55.60	0.19	49.97	58.96	0.23
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	63.99	72.41	0.13	50.89	59.63	0.30	43.43	54.88	0.27	42.18	50.43	0.35	41.89	49.84	0.26	48.48	57.44	0.26
Closed-source models																		
GPT-5.6-Luna (OpenAI, 2026)	**75.06**	**80.52**	0.10	54.43	59.50	0.33	**57.74**	**68.16**	0.20	**57.24**	**63.38**	0.24	58.35	64.65	**0.18**	**60.56**	**67.24**	0.21
Gemini-2.5-Flash (Comanici et al., 2025)	59.40	67.96	0.21	50.76	59.57	0.36	41.15	53.33	0.35	40.03	48.68	0.44	42.85	51.90	0.34	46.84	56.29	0.34
Claude-Haiku-4.5 (Anthropic, 2025)	56.64	66.54	**0.09**	53.52	63.53	**0.19**	41.52	53.59	**0.16**	45.63	54.69	**0.23**	36.94	45.52	0.19	46.85	56.77	**0.17**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	62.22	71.28	0.20	50.14	58.46	0.34	40.88	52.25	0.38	39.22	47.98	0.42	50.50	60.45	0.29	48.59	58.08	0.33
I-track: multimodal physics-rendered visual evidence																		
Open-weight multimodal models: 7–8B																		
InternVL3-8B (Zhu et al., 2025)	34.07	42.03	0.37	28.20	38.72	**0.24**	33.31	45.26	0.39	26.53	36.83	0.31	29.20	40.54	0.22	30.26	40.68	0.31
Qwen3-VL-8B-Instruct (Bai et al., 2025)	31.53	38.43	0.50	26.70	35.56	0.61	27.79	38.09	0.56	22.90	30.09	0.64	29.93	40.96	0.52	27.77	36.63	0.56
Open-weight multimodal models: 20–50B																		
InternVL3-38B (Zhu et al., 2025)	44.26	51.54	0.36	33.99	42.91	0.53	40.47	51.59	0.41	30.60	36.58	0.60	36.38	47.15	0.37	37.14	45.95	0.46
Qwen3-VL-32B-Instruct (Bai et al., 2025)	55.42	61.18	0.23	42.99	52.06	0.38	42.29	53.02	0.33	38.61	46.01	0.43	32.47	42.29	0.31	42.36	50.91	0.34
Open-weight multimodal models: more than 50B total parameters																		
InternVL3-78B (Zhu et al., 2025)	47.66	54.21	0.34	38.65	48.26	0.48	40.39	51.46	0.40	34.14	42.12	0.54	32.86	43.86	0.51	38.74	47.98	0.45
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	46.92	52.24	0.35	45.02	53.31	0.38	38.17	48.29	0.40	38.86	45.00	0.47	19.99	26.67	0.36	37.79	45.10	0.39
Closed-source multimodal models																		
GPT-5.6-Luna (OpenAI, 2026)	**59.46**	**63.86**	0.28	45.73	53.33	0.38	45.57	55.01	0.35	38.27	44.44	0.40	**45.88**	**55.64**	0.27	46.98	54.46	0.34
Gemini-2.5-Flash (Comanici et al., 2025)	55.36	60.10	0.29	**48.29**	**56.47**	0.35	**49.05**	**59.56**	0.27	**47.97**	**55.60**	0.34	40.67	50.92	0.31	**48.27**	**56.53**	0.31
Claude-Haiku-4.5 (Anthropic, 2025)	54.26	59.45	**0.22**	39.06	48.13	0.31	40.56	51.03	**0.24**	37.14	45.61	**0.31**	27.44	34.47	**0.13**	39.69	47.74	**0.24**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	48.35	54.28	0.36	44.33	53.25	0.38	39.95	50.13	0.40	34.44	42.51	0.47	29.40	39.29	0.39	39.29	47.89	0.40

Physical capability decomposition. The multiple-choice results already reveal a substantial capability gradient. On the S-track, the model-averaged Acc. decreases from 52.00 on P1 to 38.87 on P4, with only a small recovery to 40.05 on P5. The trend is therefore not perfectly monotonic for every capability, but higher-level coupled-field and mechanism-oriented reasoning is consistently harder than temporal prediction for the model population as a whole. Strong individual models can partially resist this degradation: GPT-5.6-Luna maintains comparatively high performance across the axis, while Gemma-3-27B-IT is particularly competitive on several physical capability levels. Nevertheless, no single model dominates every capability and every metric.

The I-track presents a different profile. Performance is generally more variable across capability levels and models, indicating that converting physics-rendered visual evidence into the latent physical state required by the task introduces an additional source of difficulty. In particular, models that are competitive when operating on structured sensor records do not necessarily retain the same relative advantage under visual observation. This interface- and modality-dependent behavior supports the use of both tracks rather than treating multimodal rendering as a direct substitute for structured physical measurements.

Table 7: **Open-report evaluation along the physical capability axis.** Models are evaluated on the controlled simulation subset under the structured-observation track (S-track) and multimodal image-observation track (I-track). Each dimension reports accuracy (Acc), macro-F1 (F1), and Brier score; Average is the equal-weight mean across P1–P5.

Model	P1			P2			P3			P4			P5			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records																		
Open-weight models: fewer than 20B total parameters																		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	40.53	50.73	0.30	40.49	48.57	0.33	33.25	43.95	0.30	45.48	40.02	**0.32**	32.17	39.71	0.36	38.38	44.60	0.32
Qwen3-8B (Yang et al., 2025)	54.71	64.14	0.29	39.94	46.13	0.42	40.05	51.53	0.29	46.64	40.99	0.35	29.46	35.76	0.43	42.16	47.71	0.35
Gemma-3-12B-IT (Gemma Team et al., 2025)	47.21	56.54	0.26	44.44	54.36	0.32	34.25	44.15	0.36	36.82	31.11	0.41	27.13	33.49	0.45	37.97	43.93	0.36
Open-weight models: 20–50B total parameters																		
Gemma-3-27B-IT (Gemma Team et al., 2025)	44.84	56.05	0.29	49.69	56.45	0.32	40.09	50.91	0.33	40.24	34.77	0.38	31.78	39.65	0.40	41.33	47.57	0.34
Qwen3-32B (Yang et al., 2025)	53.99	64.29	0.23	48.21	54.38	0.33	41.93	53.10	0.30	44.03	38.45	0.39	36.24	43.49	0.37	44.88	50.74	0.33
Open-weight models: 50–200B total parameters																		
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	54.11	63.57	0.20	43.46	48.62	0.35	42.97	55.03	0.21	42.76	36.92	0.32	28.29	36.28	0.33	42.32	48.08	0.28
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	59.01	**69.13**	0.26	44.20	51.52	0.40	42.13	52.86	0.36	34.33	29.11	0.53	34.50	42.03	0.47	42.83	48.93	0.40
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	55.81	65.46	0.22	42.78	48.08	0.38	41.97	52.88	0.28	40.79	35.28	0.37	26.55	33.31	0.40	41.58	47.00	0.33
Open-weight models: more than 200B total parameters																		
Qwen3-235B-A22B (Yang et al., 2025)	53.25	63.56	0.23	42.22	47.60	0.39	44.08	55.30	0.28	41.64	35.78	0.38	36.43	42.91	0.34	43.52	49.03	0.32
DeepSeek-V3.2 (Liu et al., 2025)	56.96	66.12	0.22	50.37	58.07	0.29	45.20	56.63	0.24	**48.18**	43.45	0.33	**42.25**	**50.12**	**0.24**	48.59	54.88	**0.26**
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	58.56	68.18	0.20	52.72	59.97	0.28	45.64	57.07	0.25	39.79	36.98	0.40	34.11	41.98	0.30	46.16	52.84	0.29
Closed-source models																		
GPT-5.6-Luna (OpenAI, 2026)	58.12	68.72	**0.17**	**54.44**	60.47	0.32	**53.24**	**64.99**	**0.21**	47.03	**44.03**	0.37	37.79	45.23	0.27	**50.12**	**56.69**	0.27
Gemini-2.5-Flash (Comanici et al., 2025)	**60.45**	69.07	0.25	50.86	58.46	0.33	43.88	54.98	0.33	34.61	32.56	0.51	37.98	45.12	0.34	45.56	52.04	0.35
Claude-Haiku-4.5 (Anthropic, 2025)	56.48	66.06	0.18	54.38	**61.04**	**0.25**	38.33	49.68	0.24	38.42	36.85	0.40	25.97	32.73	0.29	42.72	49.27	0.27
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	59.31	68.86	0.24	44.26	49.47	0.42	44.64	55.71	0.32	42.00	37.81	0.43	39.34	46.86	0.32	45.91	51.74	0.35
I-track: multimodal physics-rendered visual evidence																		
Open-weight multimodal models: 7–8B																		
InternVL3-8B (Zhu et al., 2025)	19.65	15.70	0.54	21.96	18.92	0.58	34.84	27.83	0.47	26.64	21.09	0.53	23.33	23.33	0.68	25.28	21.37	0.56
Qwen3-VL-8B-Instruct (Bai et al., 2025)	32.31	42.14	0.52	16.24	19.19	0.72	37.10	41.86	0.49	26.71	27.31	0.62	42.92	42.92	0.43	31.06	34.68	0.56
Open-weight multimodal models: 20–50B																		
InternVL3-38B (Zhu et al., 2025)	39.61	28.87	0.29	19.33	17.71	0.63	35.44	31.45	0.45	**40.50**	32.45	0.40	27.50	27.50	0.57	32.48	27.60	0.47
Qwen3-VL-32B-Instruct (Bai et al., 2025)	61.13	71.73	0.14	**37.48**	**41.42**	0.45	40.72	46.20	0.39	35.20	32.15	0.42	39.17	39.17	0.44	**42.74**	**46.13**	0.37
Open-weight multimodal models: more than 50B total parameters																		
InternVL3-78B (Zhu et al., 2025)	44.26	33.11	0.26	35.70	32.93	0.51	31.30	27.21	0.51	32.71	27.01	0.49	**49.58**	**49.58**	0.39	38.71	33.97	0.43
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	29.11	37.41	0.33	35.83	39.61	0.46	35.14	39.82	0.41	29.91	30.63	**0.38**	21.25	21.25	0.37	30.25	33.74	0.39
Closed-source multimodal models																		
GPT-5.6-Luna (OpenAI, 2026)	**62.02**	**72.50**	**0.11**	36.49	40.04	**0.41**	42.61	47.35	**0.34**	30.53	31.82	0.41	23.33	23.33	0.37	39.00	43.01	**0.33**
Gemini-2.5-Flash (Comanici et al., 2025)	50.33	59.79	0.27	33.40	36.57	0.56	**42.68**	**47.85**	0.40	34.27	**35.37**	0.51	46.25	46.25	0.40	41.39	45.17	0.43
Claude-Haiku-4.5 (Anthropic, 2025)	57.82	68.39	0.11	29.98	33.31	0.49	40.42	45.63	0.36	30.92	31.05	0.45	37.50	37.50	**0.29**	39.33	43.18	0.34
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	41.49	49.84	0.38	37.21	40.20	0.50	41.48	46.33	0.44	31.07	29.16	0.50	46.25	46.25	0.40	39.50	42.36	0.44

Open-report evaluation sharpens this separation. On the S-track, the model-averaged Acc. declines from 54.22 on P1 to 33.33 on P5, showing that the difficulty of the physical capability hierarchy becomes more visible when candidate answers are removed. The strongest models remain substantially better at temporal evolution and field-level interpretation than at constructing reliable counterfactual physical commitments. The I-track is more irregular: its mean performance drops sharply between P1 and P2 and remains substantially below the strongest S-track results over most higher-level capabilities. Thus, open reporting does not simply reproduce the ranking observed under multiple choice; it exposes additional failures in state recovery and physical commitment that can be hidden by candidate-conditioned answering.

Taken together, the two response interfaces suggest that FireWorldBench measures more than answer selection accuracy. Multiple-choice evaluation captures whether a model can recognize a physically plausible hypothesis, whereas open-report evaluation additionally requires the model to recover, organize, and explicitly commit to the underlying physical state. The widening gap at the more demanding capability levels is therefore consistent with the central benchmark observation that recognition of a correct physical hypothesis is easier than independently reconstructing the reasoning needed to support it.

Fire scenario task decomposition. The fire scenario axis provides a complementary view of the same models. Unlike the physical capability axis, the three task categories are not expected to form a strictly monotonic empirical accuracy sequence for every model, because they represent distinct stages of fire evolution: Localized Onset (T1), Coupled Propagation (T2), and Critical Transition (T3). The relevant question is therefore not whether performance decreases at every adjacent T-level, but whether models exhibit stable competence across these distinct stages of fire-world evolution.

Table 8: **Choice-format evaluation along the fire scenario task axis.** Models are evaluated on the controlled simulation subset under the structured-observation track (S-track) and multimodal image-observation track (I-track). Each dimension reports accuracy (Acc), macro-F1 (F1), and Brier score; Average is the equal-weight mean across T1–T3.

Model	T1			T2			T3			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records												
Open-weight models: fewer than 20B total parameters												
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	24.67	34.56	0.59	25.86	36.87	0.56	24.09	35.19	0.57	24.87	35.54	0.57
Qwen3-8B (Yang et al., 2025)	33.76	42.57	0.50	32.74	44.07	0.42	37.31	48.01	0.35	34.60	44.88	0.43
Gemma-3-12B-IT (Gemma Team et al., 2025)	32.84	43.70	0.52	31.19	43.49	0.51	23.65	33.20	0.62	29.23	40.13	0.55
Open-weight models: 20–50B total parameters												
Gemma-3-27B-IT (Gemma Team et al., 2025)	47.45	52.90	0.44	45.69	57.44	0.36	70.60	76.58	0.19	54.58	62.31	0.33
Qwen3-32B (Yang et al., 2025)	41.50	50.29	0.37	40.53	52.64	0.33	58.05	66.84	0.22	46.69	56.59	0.31
Open-weight models: 50–200B total parameters												
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	41.58	51.36	0.31	30.38	41.98	0.32	37.08	47.25	0.27	36.35	46.86	0.30
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	39.53	48.99	0.45	32.35	43.78	0.48	36.78	46.99	0.42	36.22	46.59	0.45
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	35.71	45.07	0.36	36.32	48.72	0.28	40.76	51.75	0.23	37.60	48.51	0.29
Open-weight models: more than 200B total parameters												
Qwen3-235B-A22B (Yang et al., 2025)	40.92	48.81	0.41	42.68	54.23	0.30	52.63	61.71	0.21	45.41	54.92	0.30
DeepSeek-V3.2 (Liu et al., 2025)	45.37	53.52	0.31	46.99	57.54	0.23	60.30	68.04	0.14	50.89	59.70	0.23
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	47.53	54.69	0.35	42.73	54.30	0.27	58.86	67.17	0.16	49.71	58.72	0.26
Closed-source models												
GPT-5.6-Luna (OpenAI, 2026)	**50.78**	56.07	0.34	**62.36**	**71.33**	0.17	**71.18**	**76.83**	0.12	**61.44**	**68.08**	0.21
Gemini-2.5-Flash (Comanici et al., 2025)	48.40	56.41	0.39	38.67	50.20	0.39	55.55	64.23	0.24	47.54	56.95	0.34
Claude-Haiku-4.5 (Anthropic, 2025)	50.08	**58.79**	**0.23**	43.33	55.13	**0.17**	52.06	61.65	**0.12**	48.49	58.52	**0.17**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	45.87	52.95	0.39	40.01	51.79	0.38	59.50	68.76	0.22	48.46	57.83	0.33
I-track: multimodal physics-rendered visual evidence												
Open-weight multimodal models: 7–8B												
InternVL3-8B (Zhu et al., 2025)	25.96	36.04	**0.22**	32.61	44.34	0.40	32.38	41.51	0.32	30.32	40.63	0.31
Qwen3-VL-8B-Instruct (Bai et al., 2025)	26.02	33.64	0.63	25.73	35.68	0.57	30.98	39.31	0.51	27.58	36.21	0.57
Open-weight multimodal models: 20–50B												
InternVL3-38B (Zhu et al., 2025)	32.49	39.51	0.57	37.50	47.86	0.46	41.52	50.02	0.37	37.17	45.80	0.47
Qwen3-VL-32B-Instruct (Bai et al., 2025)	42.68	50.64	0.41	40.16	50.34	0.35	47.45	54.62	0.26	43.43	51.87	0.34
Open-weight multimodal models: more than 50B total parameters												
InternVL3-78B (Zhu et al., 2025)	36.05	44.38	0.52	39.52	50.31	0.42	42.52	50.61	0.40	39.36	48.43	0.45
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	40.93	47.49	0.45	40.89	50.70	0.38	37.56	43.36	0.35	39.79	47.18	0.39
Closed-source multimodal models												
GPT-5.6-Luna (OpenAI, 2026)	44.87	51.54	0.41	41.88	50.68	0.34	**54.75**	**61.00**	0.27	47.17	54.41	0.34
Gemini-2.5-Flash (Comanici et al., 2025)	**48.37**	**55.72**	0.37	**48.50**	**58.63**	0.28	50.26	56.91	0.30	**49.04**	**57.09**	0.31
Claude-Haiku-4.5 (Anthropic, 2025)	39.11	47.56	0.34	38.78	48.98	**0.24**	44.95	50.77	**0.19**	40.95	49.10	**0.26**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	42.56	50.70	0.40	37.20	47.16	0.42	41.77	49.07	0.37	40.51	48.98	0.40

The multiple-choice results show pronounced stage-dependent variation. Performance at Localized Onset does not necessarily transfer to Coupled Propagation or Critical Transition. In the reported controlled-simulation results, the model-level means vary across T1–T3 on both observation tracks, and several models exchange relative positions between the three stages. This reinforces the motivation for retaining the three fire-evolution categories rather than collapsing fire reasoning into a single overall task.

Table 9: **Open-report evaluation along the fire scenario task axis.** Models are evaluated on the controlled simulation subset under the structured-observation track (S-track) and multimodal image-observation track (I-track). Each dimension reports accuracy (Acc), macro-F1 (F1), and Brier score; Average is the equal-weight mean across T1–T3.

Model	T1			T2			T3			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records												
Open-weight models: fewer than 20B total parameters												
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	40.42	46.22	0.33	33.75	41.26	0.34	41.05	48.17	0.30	38.41	45.22	0.32
Qwen3-8B (Yang et al., 2025)	41.30	45.11	0.38	38.47	46.31	0.36	51.69	57.54	0.30	43.82	49.65	0.35
Gemma-3-12B-IT (Gemma Team et al., 2025)	42.79	49.67	0.35	27.35	34.18	0.44	45.50	51.18	0.29	38.55	45.01	0.36
Open-weight models: 20–50B total parameters												
Gemma-3-27B-IT (Gemma Team et al., 2025)	45.89	51.17	0.35	33.45	40.97	0.39	45.14	52.24	0.29	41.49	48.13	0.34
Qwen3-32B (Yang et al., 2025)	45.74	50.47	0.34	37.13	44.73	0.39	52.97	59.46	0.24	45.28	51.55	0.33
Open-weight models: 50–200B total parameters												
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	41.92	46.04	0.33	37.76	45.72	0.30	51.22	57.23	0.21	43.63	49.66	0.28
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	42.84	47.56	0.41	31.27	38.31	0.51	56.93	62.83	0.28	43.68	49.57	0.40
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	41.76	45.73	0.38	35.62	43.09	0.36	52.17	57.99	0.24	43.18	48.94	0.33
Open-weight models: more than 200B total parameters												
Qwen3-235B-A22B (Yang et al., 2025)	41.97	46.13	0.37	36.76	44.28	0.37	52.56	58.76	0.23	43.76	49.72	0.33
DeepSeek-V3.2 (Liu et al., 2025)	47.26	51.85	**0.32**	43.87	51.70	0.30	56.63	62.40	0.21	49.25	55.32	0.27
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	47.38	51.71	0.35	40.27	48.03	0.32	56.23	62.09	0.20	47.96	53.94	0.29
Closed-source models												
GPT-5.6-Luna (OpenAI, 2026)	**48.45**	**52.02**	0.35	**50.66**	**58.49**	0.28	56.62	63.26	**0.17**	**51.91**	**57.92**	**0.27**
Gemini-2.5-Flash (Comanici et al., 2025)	44.88	49.18	0.41	37.23	44.86	0.41	**58.36**	63.50	0.25	46.82	52.51	0.36
Claude-Haiku-4.5 (Anthropic, 2025)	44.97	49.37	0.36	38.52	46.40	**0.27**	51.85	58.32	0.19	45.11	51.36	0.27
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	44.04	47.01	0.41	38.71	46.37	0.41	58.17	**63.75**	0.24	46.97	52.38	0.35
I-track: multimodal physics-rendered visual evidence												
Open-weight multimodal models: 7–8B												
InternVL3-8B (Zhu et al., 2025)	27.85	39.75	0.51	27.25	40.50	0.54	20.93	32.15	0.59	25.34	37.47	0.54
Qwen3-VL-8B-Instruct (Bai et al., 2025)	20.66	23.00	0.69	30.56	34.86	0.55	37.71	42.41	0.47	29.64	33.42	0.57
Open-weight multimodal models: 20–50B												
InternVL3-38B (Zhu et al., 2025)	28.44	39.44	0.54	33.67	**48.91**	0.46	35.40	50.22	0.39	32.50	46.19	0.46
Qwen3-VL-32B-Instruct (Bai et al., 2025)	38.37	41.62	**0.43**	33.76	38.56	0.44	**54.27**	60.42	0.23	**42.13**	46.87	0.37
Open-weight multimodal models: more than 50B total parameters												
InternVL3-78B (Zhu et al., 2025)	39.35	**54.31**	0.47	27.49	41.80	0.54	46.11	**61.16**	0.31	37.65	**52.42**	0.44
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	36.32	38.87	0.47	30.27	34.94	0.37	28.30	31.80	0.35	31.63	35.20	0.40
Closed-source multimodal models												
GPT-5.6-Luna (OpenAI, 2026)	36.66	38.18	0.43	36.54	41.45	**0.34**	48.58	55.42	0.20	40.59	45.02	**0.32**
Gemini-2.5-Flash (Comanici et al., 2025)	36.47	37.18	0.54	**36.83**	42.40	0.46	48.91	55.09	0.31	40.74	44.89	0.43
Claude-Haiku-4.5 (Anthropic, 2025)	34.12	35.01	0.47	33.14	38.09	0.40	50.77	57.66	**0.17**	39.34	43.59	0.35
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	**41.41**	40.91	0.46	32.04	36.61	0.51	43.14	48.59	0.38	38.86	42.04	0.45

The open-report results further amplify this stage dependence. Removing candidate answers changes both absolute performance and the relative strengths of the models across Localized Onset, Coupled Propagation, and Critical Transition. In particular, the reported model-level mean Acc. is lowest at T2 and highest at T3 on both observation tracks, showing that independent reconstruction of coupled propagation is especially challenging in the controlled subset. Closed-source systems remain among the strongest models overall, but their advantages are not uniform across all three fire stages or observation tracks. Likewise, several open-weight models are competitive at individual stages while remaining substantially weaker at others.

These results provide a more detailed interpretation of the averaged tables in the main paper. The aggregate score is useful for summarizing overall benchmark competence, but it hides important structure: a model may identify a localized onset yet fail to reconstruct coupled propagation or a critical transition, or may perform well when evidence is supplied as structured records but fail to reconstruct the same physical state from rendered observations. The two-axis design therefore exposes qualitatively different failure modes that cannot be recovered from a single task-level average.

C.3 Real-World-Aligned Results

We next provide the same decomposition on the real-world-aligned subset. Tables 10 and 11 report the physical capability axis, and Tables 12 and 13 report the fire scenario task axis. The objective of this evaluation is not to reproduce the controlled simulation setting exactly, but to test whether the capability boundaries observed under controlled conditions remain visible when the benchmark is grounded in real-world-aligned fire evidence.

As in the controlled simulation evaluation, choice and open-report results are kept separate. This distinction is particularly important in the real-world-aligned setting, where observation ambiguity and incomplete evidence

Table 10: **Choice-format evaluation along the physical capability axis.** Models are evaluated on the real-world-aligned subset under the structured-observation track (S-track) and multimodal image-observation track (I-track). Each dimension reports accuracy (Acc), macro-F1 (F1), and Brier score; Average is the equal-weight mean across P1–P5.

Model	P1			P2			P3			P4			P5			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records																		
Open-weight models: fewer than 20B total parameters																		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	20.89	30.74	0.66	22.16	30.93	0.65	28.03	40.00	0.51	28.50	40.17	0.52	33.80	46.67	0.41	26.68	37.70	0.55
Qwen3-8B (Yang et al., 2025)	23.10	33.78	0.63	32.40	40.87	0.54	37.12	48.18	0.41	30.62	43.00	0.52	37.50	51.11	0.43	32.15	43.39	0.50
Gemma-3-12B-IT (Gemma Team et al., 2025)	42.08	52.41	0.42	43.92	55.69	0.39	37.95	49.00	0.43	42.75	55.25	0.39	43.06	53.89	0.39	41.95	53.25	0.40
Open-weight models: 20–50B total parameters																		
Gemma-3-27B-IT (Gemma Team et al., 2025)	43.83	55.08	0.38	43.80	54.73	0.37	38.26	44.24	0.54	42.29	53.33	0.39	44.81	55.50	0.35	42.60	52.58	0.40
Qwen3-32B (Yang et al., 2025)	53.69	62.33	0.23	47.29	56.81	0.33	48.18	57.88	0.28	38.29	47.26	0.41	40.28	50.74	0.38	45.55	55.00	0.33
Open-weight models: 50–200B total parameters																		
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	56.38	66.30	0.17	44.34	55.85	0.26	47.73	54.85	0.26	39.71	51.58	0.28	47.41	58.52	0.22	47.11	57.42	0.24
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	38.44	48.25	0.41	53.21	63.75	0.28	36.82	46.06	0.44	44.79	57.58	0.33	45.37	55.93	0.35	43.73	54.31	0.36
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	61.14	69.27	0.17	57.08	65.79	0.21	**51.14**	57.27	0.27	43.71	55.07	0.27	40.19	52.96	0.21	50.65	60.07	0.22
Open-weight models: more than 200B total parameters																		
Qwen3-235B-A22B (Yang et al., 2025)	52.88	62.18	0.23	46.24	55.23	0.36	41.29	50.15	0.37	37.04	46.83	0.43	36.30	48.89	0.37	42.75	52.66	0.35
DeepSeek-V3.2 (Liu et al., 2025)	63.45	71.18	**0.11**	54.68	65.47	0.18	49.02	55.15	**0.21**	37.71	44.83	0.35	41.57	51.67	0.22	49.29	57.66	0.21
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	55.15	60.47	0.16	58.41	67.33	0.18	42.42	48.79	0.38	45.42	52.93	0.32	40.19	49.07	0.25	48.32	55.72	0.26
Closed-source models																		
GPT-5.6-Luna (OpenAI, 2026)	**74.93**	**79.26**	**0.11**	63.66	70.12	0.19	48.18	56.36	0.29	**65.67**	**69.67**	**0.21**	57.31	65.93	**0.13**	**61.95**	**68.27**	**0.19**
Gemini-2.5-Flash (Comanici et al., 2025)	67.40	73.80	0.15	57.61	66.74	0.26	50.00	57.42	0.34	49.75	58.75	0.35	**61.30**	**68.15**	0.22	57.21	64.97	0.26
Claude-Haiku-4.5 (Anthropic, 2025)	57.46	66.09	0.16	**64.80**	**72.03**	**0.15**	48.03	55.61	0.33	58.04	65.17	0.21	51.11	61.11	0.20	55.89	64.00	0.21
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	57.64	65.57	0.26	61.72	70.61	0.22	50.00	**58.79**	0.33	51.87	60.08	0.31	51.11	63.33	0.26	54.47	63.68	0.28
I-track: multimodal physics-rendered visual evidence																		
Open-weight multimodal models: 7–8B																		
InternVL3-8B (Zhu et al., 2025)	24.23	27.32	0.49	29.44	38.81	**0.15**	28.58	40.17	0.29	33.52	45.76	**0.15**	30.83	42.50	0.23	29.32	38.91	0.26
Qwen3-VL-8B-Instruct (Bai et al., 2025)	29.67	34.31	0.51	37.59	46.67	0.49	22.58	30.25	0.65	29.17	36.82	0.57	29.58	40.50	0.51	29.72	37.71	0.55
Open-weight multimodal models: 20–50B																		
InternVL3-38B (Zhu et al., 2025)	40.73	44.72	0.44	48.52	56.67	0.39	34.83	43.68	0.51	44.77	52.12	0.44	37.33	48.00	0.37	41.24	49.04	0.43
Qwen3-VL-32B-Instruct (Bai et al., 2025)	33.74	36.59	0.42	57.93	65.41	0.27	35.92	43.67	0.43	48.14	54.24	0.36	52.25	61.33	0.21	45.60	52.25	0.34
Open-weight multimodal models: more than 50B total parameters																		
InternVL3-78B (Zhu et al., 2025)	24.47	26.67	0.64	42.37	53.56	0.42	36.96	45.50	0.48	40.91	48.79	0.47	28.25	39.33	0.56	34.59	42.77	0.52
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	60.16	63.25	0.28	51.15	57.85	0.35	29.88	39.00	0.48	48.64	53.33	0.39	30.92	41.67	0.37	44.15	51.02	0.37
Closed-source multimodal models																		
GPT-5.6-Luna (OpenAI, 2026)	60.24	62.60	0.21	**65.44**	**70.52**	0.20	**45.92**	**52.83**	0.28	**59.24**	**64.09**	0.24	**55.83**	63.00	0.16	**57.33**	**62.61**	0.22
Gemini-2.5-Flash (Comanici et al., 2025)	**60.77**	62.28	0.32	55.19	60.59	0.37	38.29	48.33	0.41	43.71	50.91	0.46	54.08	**66.33**	0.20	50.41	57.69	0.35
Claude-Haiku-4.5 (Anthropic, 2025)	60.37	**63.74**	**0.19**	61.44	66.96	0.20	38.75	47.17	**0.27**	50.04	57.32	0.23	49.83	60.00	**0.12**	52.09	59.04	**0.20**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	50.77	53.50	0.39	54.81	62.37	0.30	36.79	43.67	0.46	36.70	43.94	0.45	40.00	52.33	0.28	43.81	51.16	0.38

can make open-ended physical-state reconstruction considerably more difficult than selecting among predefined candidate hypotheses.

Physical capability decomposition. The real-world-aligned results preserve the broad capability separation found in controlled simulations, although individual cells are more variable. For the S-track choice setting, the model-averaged Acc. is highest at P1 and remains lower on the more demanding coupled-field and mechanism-oriented capabilities. The I-track shows an even less uniform profile, reflecting the additional difficulty introduced by multimodal observation and the smaller, heterogeneous real-world-aligned evaluation set.

The important pattern is therefore not a perfectly monotonic decrease for every cell, but the absence of a model that is uniformly strong across all five physical capabilities and both observation modalities. Several models that perform competitively on early physical interpretation lose substantial performance on coupling, attribution, or counterfactual reasoning, while others exhibit the opposite trade-off on particular cells.

Table 11: **Open-report evaluation along the physical capability axis.** Models are evaluated on the real-world-aligned subset under the structured-observation track (S-track) and multimodal image-observation track (I-track). Each dimension reports accuracy (Acc), macro-F1 (F1), and Brier score; Average is the equal-weight mean across P1–P5.

Model	P1			P2			P3			P4			P5			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records																		
Open-weight models: fewer than 20B total parameters																		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	49.42	57.22	0.26	39.55	48.61	0.32	37.88	42.73	0.36	30.00	35.50	0.40	35.19	42.78	0.32	38.41	45.37	0.33
Qwen3-8B (Yang et al., 2025)	52.79	57.04	0.35	37.73	45.72	0.43	**60.61**	**67.73**	0.20	47.50	53.50	0.33	31.48	38.89	0.39	46.02	52.58	0.34
Gemma-3-12B-IT (Gemma Team et al., 2025)	45.39	51.86	0.29	44.77	56.30	0.29	40.91	45.45	0.34	38.33	43.00	0.44	33.33	41.11	0.40	40.55	47.54	0.35
Open-weight models: 20–50B total parameters																		
Gemma-3-27B-IT (Gemma Team et al., 2025)	70.04	75.11	0.17	42.50	53.78	0.33	50.00	55.00	0.33	45.00	53.50	0.35	27.78	37.22	0.45	47.06	54.92	0.33
Qwen3-32B (Yang et al., 2025)	71.73	76.60	0.15	43.86	52.60	0.36	56.06	63.18	0.25	40.00	44.00	0.45	37.04	47.78	0.34	49.74	56.83	0.31
Open-weight models: 50–200B total parameters																		
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	**75.86**	**79.59**	**0.13**	52.73	61.77	0.20	46.97	53.18	0.26	**58.33**	**65.75**	**0.19**	42.59	51.67	**0.23**	55.30	62.39	**0.20**
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	67.52	71.84	0.21	59.32	68.17	0.24	56.06	63.64	0.27	47.50	52.25	0.41	40.74	50.00	0.41	54.23	61.18	0.31
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	72.18	75.56	0.17	49.55	57.44	0.29	51.52	57.73	0.29	47.50	51.50	0.36	31.48	41.67	0.35	50.45	56.78	0.29
Open-weight models: more than 200B total parameters																		
Qwen3-235B-A22B (Yang et al., 2025)	54.73	59.11	0.25	48.41	58.89	0.29	51.52	58.64	0.28	46.67	54.00	0.30	29.63	38.89	0.34	46.19	53.91	0.29
DeepSeek-V3.2 (Liu et al., 2025)	63.27	66.70	0.22	64.32	73.28	0.16	45.45	50.91	0.30	43.33	48.00	0.42	29.63	37.78	0.40	49.20	55.33	0.30
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	67.19	71.57	0.14	62.05	70.79	0.17	50.00	56.36	0.27	37.50	41.50	0.43	37.04	46.11	0.34	50.76	57.27	0.27
Closed-source models																		
GPT-5.6-Luna (OpenAI, 2026)	75.65	78.69	0.14	59.55	68.19	0.25	59.09	**67.73**	0.21	55.00	57.00	0.36	42.59	51.67	0.29	**58.38**	**64.66**	0.25
Gemini-2.5-Flash (Comanici et al., 2025)	72.18	76.49	0.16	57.50	67.55	0.20	46.97	52.73	0.34	47.50	53.75	0.35	33.33	44.44	0.36	51.50	58.99	0.28
Claude-Haiku-4.5 (Anthropic, 2025)	70.49	74.71	0.15	**65.45**	**74.60**	**0.14**	53.03	60.45	**0.19**	52.50	57.25	0.33	29.63	35.56	0.38	54.22	60.51	0.24
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	71.43	75.42	0.18	53.18	62.43	0.28	51.52	57.73	0.34	55.00	60.50	0.32	**50.00**	**58.33**	0.31	56.23	62.88	0.29
I-track: multimodal physics-rendered visual evidence																		
Open-weight multimodal models: 7–8B																		
InternVL3-8B (Zhu et al., 2025)	14.18	12.53	0.71	28.15	20.59	0.48	24.58	16.64	0.51	25.76	19.55	0.49	30.00	30.00	0.60	24.53	19.86	0.56
Qwen3-VL-8B-Instruct (Bai et al., 2025)	55.04	57.46	0.37	45.93	51.19	0.41	35.42	40.92	0.50	31.82	35.45	0.57	50.00	50.00	0.44	43.64	47.00	0.46
Open-weight multimodal models: 20–50B																		
InternVL3-38B (Zhu et al., 2025)	21.33	14.83	0.47	28.89	22.30	0.46	37.08	30.44	0.41	47.73	39.77	0.36	25.00	25.00	0.59	32.01	26.47	0.46
Qwen3-VL-32B-Instruct (Bai et al., 2025)	55.72	**61.57**	0.24	49.26	54.44	0.35	42.08	48.50	0.38	**53.79**	**61.82**	**0.26**	45.00	45.00	0.42	**49.17**	**54.27**	0.33
Open-weight multimodal models: more than 50B total parameters																		
InternVL3-78B (Zhu et al., 2025)	30.66	23.35	0.44	33.33	25.85	0.44	31.25	23.75	0.45	51.52	43.18	0.33	**65.00**	**65.00**	**0.27**	42.35	36.23	0.38
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	55.24	56.36	0.27	31.85	37.93	0.51	31.67	38.67	0.47	40.15	45.23	0.45	35.00	35.00	0.45	38.78	42.64	0.43
Closed-source multimodal models																		
GPT-5.6-Luna (OpenAI, 2026)	**58.50**	61.27	0.19	**55.56**	**61.33**	**0.20**	36.67	41.83	0.34	47.73	52.27	0.36	30.00	30.00	0.38	45.69	49.34	**0.29**
Gemini-2.5-Flash (Comanici et al., 2025)	57.15	59.13	0.28	44.44	49.41	0.40	37.08	45.42	0.39	50.76	56.82	0.33	35.00	35.00	0.51	44.89	49.16	0.38
Claude-Haiku-4.5 (Anthropic, 2025)	56.33	59.62	**0.19**	41.85	47.56	0.36	**43.75**	50.42	**0.29**	37.12	41.36	0.43	20.00	20.00	0.36	39.81	43.79	0.32
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	58.00	59.19	0.30	50.00	55.33	0.37	42.92	**50.50**	0.39	44.70	50.23	0.39	20.00	20.00	0.54	43.12	47.05	0.40

The open-report setting produces a substantially clearer difficulty gradient. For the S-track, the model-averaged Acc. decreases from 65.32 on P1 to 52.03 on P2, 50.51 on P3, 46.11 on P4, and 35.43 on P5. This progression is especially informative because models cannot rely on candidate answers: they must directly recover the relevant physical state from the available evidence and express the required commitment. The large decline toward counterfactual intervention reasoning indicates that successful perception or forecasting does not automatically translate into reliable reasoning about how the system would respond under a hypothetical change.

The I-track again exhibits stronger variance, but the same high-level conclusion holds. Visual access to a physical field does not guarantee that the model can translate that field into a consistent latent representation suitable for coupled reasoning, mechanism attribution, or intervention. The real-world-aligned results therefore support rather than replace the controlled simulation analysis: the latter provides controlled physical coverage, while the former tests whether the same capability boundaries remain observable under more realistic evidence conditions.

Fire scenario task decomposition. The fire scenario decomposition reveals substantial heterogeneity across Localized Onset, Coupled Propagation, and Critical Transition. In particular, strong performance at one stage of fire evolution is not a reliable indicator of strong performance at the remaining stages. This is visible for both structured and multimodal observations and is consistent with the stage-specific trends observed on the controlled simulation subset.

Table 12: **Choice-format evaluation along the fire scenario task axis.** Models are evaluated on the real-world-aligned subset under the structured-observation track (S-track) and multimodal image-observation track (I-track). Each dimension reports accuracy (Acc), macro-F1 (F1), and Brier score; Average is the equal-weight mean across T1–T3.

Model	T1			T2			T3			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records												
Open-weight models: fewer than 20B total parameters												
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	24.49	34.01	0.60	27.80	39.92	0.53	24.70	35.44	0.58	25.66	36.46	0.57
Qwen3-8B (Yang et al., 2025)	32.70	42.44	0.53	32.72	44.39	0.47	27.35	38.89	0.57	30.92	41.91	0.52
Gemma-3-12B-IT (Gemma Team et al., 2025)	43.37	54.98	0.39	40.45	52.80	0.40	42.37	52.84	0.41	42.06	53.54	0.40
Open-weight models: 20–50B total parameters												
Gemma-3-27B-IT (Gemma Team et al., 2025)	43.75	54.38	0.38	39.43	48.29	0.46	44.12	55.20	0.37	42.43	52.62	0.40
Qwen3-32B (Yang et al., 2025)	40.86	49.07	0.40	49.19	60.34	0.27	49.73	58.91	0.28	46.59	56.11	0.32
Open-weight models: 50–200B total parameters												
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	43.32	54.98	0.27	43.25	52.52	0.27	53.74	64.01	0.19	46.77	57.17	0.24
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	50.38	61.26	0.30	40.69	52.20	0.38	40.49	50.52	0.39	43.85	54.66	0.36
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	51.46	61.00	0.25	49.76	58.36	0.24	54.95	64.46	0.18	52.06	61.27	0.22
Open-weight models: more than 200B total parameters												
Qwen3-235B-A22B (Yang et al., 2025)	40.12	48.56	0.43	44.31	54.88	0.33	47.98	58.26	0.28	44.14	53.90	0.34
DeepSeek-V3.2 (Liu et al., 2025)	44.70	53.19	0.29	50.89	59.27	**0.19**	56.99	65.42	0.14	50.86	59.29	0.21
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	51.18	58.61	0.29	48.62	57.17	0.26	50.73	57.11	0.18	50.18	57.63	0.25
Closed-source models												
GPT-5.6-Luna (OpenAI, 2026)	**60.38**	66.34	0.23	**62.52**	**68.29**	0.20	**69.73**	**75.33**	**0.12**	**64.21**	**69.99**	**0.18**
Gemini-2.5-Flash (Comanici et al., 2025)	52.69	61.18	0.32	53.66	62.76	0.30	65.60	72.13	0.17	57.32	65.36	0.26
Claude-Haiku-4.5 (Anthropic, 2025)	60.17	**66.71**	**0.20**	56.54	64.96	0.22	55.59	64.62	0.17	57.43	65.43	0.20
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	54.86	62.67	0.30	56.71	66.59	0.24	55.71	64.91	0.26	55.76	64.72	0.27
I-track: multimodal physics-rendered visual evidence												
Open-weight multimodal models: 7–8B												
InternVL3-8B (Zhu et al., 2025)	26.54	35.95	**0.15**	34.66	47.34	0.24	26.39	32.30	0.41	29.20	38.53	0.27
Qwen3-VL-8B-Instruct (Bai et al., 2025)	35.13	43.54	0.52	24.92	32.81	0.61	29.64	36.34	0.51	29.90	37.56	0.55
Open-weight multimodal models: 20–50B												
InternVL3-38B (Zhu et al., 2025)	42.69	48.77	0.48	43.31	53.44	0.40	39.62	45.79	0.41	41.87	49.33	0.43
Qwen3-VL-32B-Instruct (Bai et al., 2025)	53.03	59.13	0.34	42.42	50.52	0.36	39.81	44.70	0.35	45.09	51.45	0.35
Open-weight multimodal models: more than 50B total parameters												
InternVL3-78B (Zhu et al., 2025)	37.54	46.31	0.50	42.89	52.60	0.42	25.71	30.82	0.62	35.38	43.24	0.51
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	44.13	48.97	0.44	43.26	51.98	0.37	50.57	56.17	0.31	45.99	52.37	0.37
Closed-source multimodal models												
GPT-5.6-Luna (OpenAI, 2026)	**61.77**	**66.26**	0.25	**52.71**	**59.38**	**0.23**	**58.80**	62.73	0.19	**57.76**	**62.79**	0.22
Gemini-2.5-Flash (Comanici et al., 2025)	53.21	58.56	0.39	38.75	48.33	0.44	58.58	**63.61**	0.28	50.18	56.83	0.37
Claude-Haiku-4.5 (Anthropic, 2025)	58.03	62.97	0.23	42.89	52.01	0.23	56.91	62.51	**0.17**	52.61	59.16	**0.21**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	48.44	54.77	0.38	37.58	45.73	0.43	47.24	53.11	0.35	44.42	51.20	0.39

Under the choice interface, several models attain high scores at individual fire-evolution stages while exhibiting much weaker performance elsewhere. The relative ranking also changes across T1–T3, which indicates that model scale or overall benchmark accuracy alone does not determine competence at a specific fire stage. In the reported I-track results, Coupled Propagation has a lower model-level mean Acc. than Localized Onset and Critical Transition, while the S-track is more even across the three categories.

Table 13: **Open-report evaluation along the fire scenario task axis.** Models are evaluated on the real-world-aligned subset under the structured-observation track (S-track) and multimodal image-observation track (I-track). Each dimension reports accuracy (Acc), macro-F1 (F1), and Brier score; Average is the equal-weight mean across T1–T3.

Model	T1			T2			T3			Average		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
S-track: structured sensor and physical records												
Open-weight models: fewer than 20B total parameters												
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	43.18	34.09	0.28	23.58	20.16	0.47	45.22	40.07	0.28	37.33	31.44	0.34
Qwen3-8B (Yang et al., 2025)	45.54	37.86	0.36	47.15	39.07	0.32	46.50	43.15	0.36	46.40	40.03	0.35
Gemma-3-12B-IT (Gemma Team et al., 2025)	48.26	37.67	0.29	30.89	26.61	0.46	41.83	34.20	0.32	40.33	32.83	0.36
Open-weight models: 20–50B total parameters												
Gemma-3-27B-IT (Gemma Team et al., 2025)	44.15	33.71	0.33	46.34	39.31	0.34	57.57	50.75	0.25	49.35	41.26	0.31
Qwen3-32B (Yang et al., 2025)	43.03	36.64	0.40	47.97	39.50	0.33	61.49	54.51	0.21	50.83	43.55	0.31
Open-weight models: 50–200B total parameters												
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	56.21	48.50	**0.20**	49.59	41.03	0.22	**66.05**	59.36	**0.16**	57.28	49.63	**0.19**
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	51.95	45.28	0.35	57.72	51.23	0.26	59.62	52.78	0.27	56.43	49.76	0.29
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	46.87	40.72	0.35	52.85	47.43	0.27	60.17	53.93	0.22	53.30	47.36	0.28
Open-weight models: more than 200B total parameters												
Qwen3-235B-A22B (Yang et al., 2025)	50.72	41.20	0.27	44.72	38.62	0.33	47.33	40.64	0.28	47.59	40.15	0.29
DeepSeek-V3.2 (Liu et al., 2025)	**57.90**	50.83	0.25	43.90	37.05	0.34	53.34	47.94	0.27	51.71	45.27	0.29
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	52.77	47.37	0.30	46.34	40.00	0.28	58.29	51.80	0.20	52.47	46.39	0.26
Closed-source models												
GPT-5.6-Luna (OpenAI, 2026)	52.10	46.54	0.35	**66.67**	**58.14**	0.17	65.90	**60.47**	0.18	**61.56**	**55.05**	0.24
Gemini-2.5-Flash (Comanici et al., 2025)	52.26	44.74	0.31	50.41	42.45	0.25	60.71	53.44	0.22	54.46	46.88	0.26
Claude-Haiku-4.5 (Anthropic, 2025)	57.13	**51.10**	0.27	59.35	50.16	**0.16**	58.43	53.95	0.22	58.30	51.74	0.22
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	49.33	42.53	0.35	60.16	53.41	0.23	65.11	59.67	0.22	58.20	51.87	0.27
I-track: multimodal physics-rendered visual evidence												
Open-weight multimodal models: 7–8B												
InternVL3-8B (Zhu et al., 2025)	29.74	43.44	0.45	22.66	33.92	0.52	19.36	24.40	0.67	23.92	33.92	0.55
Qwen3-VL-8B-Instruct (Bai et al., 2025)	39.49	35.69	0.51	36.20	29.40	0.48	**53.39**	51.05	0.39	43.03	38.71	0.46
Open-weight multimodal models: 20–50B												
InternVL3-38B (Zhu et al., 2025)	27.69	42.53	0.53	**48.18**	**59.06**	0.29	22.53	31.93	0.51	32.80	44.51	0.44
Qwen3-VL-32B-Instruct (Bai et al., 2025)	**53.08**	46.63	**0.31**	44.01	35.97	0.35	52.20	47.64	0.30	**49.76**	43.41	0.32
Open-weight multimodal models: more than 50B total parameters												
InternVL3-78B (Zhu et al., 2025)	37.44	**53.60**	0.45	40.36	52.27	0.36	41.92	**53.70**	0.39	39.91	**53.19**	0.40
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	33.33	28.67	0.53	35.94	29.13	0.42	48.60	45.50	0.33	39.29	34.43	0.43
Closed-source multimodal models												
GPT-5.6-Luna (OpenAI, 2026)	49.23	44.46	0.31	44.79	38.54	0.29	49.16	44.22	0.25	47.73	42.41	**0.28**
Gemini-2.5-Flash (Comanici et al., 2025)	45.64	41.18	0.41	42.97	32.93	0.34	49.89	45.31	0.35	46.17	39.81	0.37
Claude-Haiku-4.5 (Anthropic, 2025)	37.69	33.54	0.44	44.01	35.77	**0.28**	44.42	39.32	**0.24**	42.04	36.21	0.32
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	51.03	45.54	0.37	40.89	33.03	0.40	45.54	43.02	0.38	45.82	40.53	0.38

Open reporting makes these stage-specific differences more pronounced. The strongest systems can still achieve high performance at particular fire-evolution stages, yet the reported model-level means differ across Localized Onset, Coupled Propagation, and Critical Transition. Critical Transition attains the highest mean Acc. on both observation tracks, while Coupled Propagation remains lower on the I-track. The multimodal track further changes the relative ranking of several models, showing that visual grounding and stage-level fire reasoning remain only partially coupled in current systems.

Overall, the real-world-aligned evaluation reaches the same qualitative conclusion as the controlled simulation study from a complementary source of evidence. Current foundation models exhibit localized strengths rather than a single general fire-world reasoning capability. Performance depends jointly on the physical capability being tested, the fire scenario task, the observation modality, and whether the answer is candidate-conditioned or generated openly.

C.4 Open-Report Explanation Metrics

Answer correctness alone does not determine whether an open report is physically well supported. We therefore provide the three explanation-oriented metrics that are defined only for the open-report interface: Evidence-F1 (Evi-F1), Mechanism Alignment (Mech), and Gold-Linked Support (GLS). Evi-F1 measures whether the generated report recovers the evidence elements required by the Gold annotation, Mech evaluates agreement with the target physical mechanism, and GLS measures whether the final report is supported by Gold-linked evidence. Higher values are better for all three metrics.

Table 14: **Open-report explanation metrics along the physical capability axis.** Models are evaluated on the controlled simulation subset. Each dimension reports Evidence-F1 (Evi-F1), Mechanism Alignment (Mech), and Gold-Linked Support (GLS); Average is the equal-weight mean across P1–P5. Higher is better for all three metrics.

Model	P1			P2			P3			P4			P5			Average		
	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS
S-track: structured sensor and physical records																		
Open-weight models: fewer than 20B total parameters																		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	57.16	32.34	16.73	52.59	13.55	27.53	64.15	31.07	24.54	31.04	11.76	30.00	12.03	15.59	20.35	43.39	20.86	23.83
Qwen3-8B (Yang et al., 2025)	**68.70**	39.75	20.06	49.63	19.74	28.70	65.96	29.61	35.09	35.95	10.34	**41.88**	10.50	16.08	23.64	46.15	23.10	29.87
Gemma-3-12B-IT (Gemma Team et al., 2025)	64.09	33.68	18.38	51.43	16.68	36.79	69.85	32.71	26.86	43.84	13.66	26.00	14.40	15.88	19.57	48.72	22.52	25.52
Open-weight models: 20–50B total parameters																		
Gemma-3-27B-IT (Gemma Team et al., 2025)	65.27	35.23	21.23	51.40	16.26	36.30	65.23	26.04	29.94	39.84	12.93	29.39	12.96	15.46	24.42	46.94	21.18	28.26
Qwen3-32B (Yang et al., 2025)	61.59	36.69	21.21	51.65	23.52	35.00	61.75	31.70	32.53	39.81	16.81	39.09	13.11	18.86	28.29	45.58	25.52	31.22
Open-weight models: 50–200B total parameters																		
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	62.77	36.26	16.12	55.68	13.50	29.94	68.46	33.64	34.17	39.26	14.53	31.21	13.30	18.91	21.71	47.89	23.37	26.63
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	61.89	40.85	20.61	49.12	21.17	31.42	60.05	29.82	30.30	41.00	14.47	30.45	12.20	20.80	23.06	44.85	25.42	27.17
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	62.63	40.16	18.57	50.53	22.66	30.56	61.02	32.29	31.89	40.28	19.61	38.24	9.55	19.12	25.39	44.80	26.77	28.93
Open-weight models: more than 200B total parameters																		
Qwen3-235B-A22B (Yang et al., 2025)	63.70	38.33	21.47	50.49	20.81	29.57	61.15	29.96	32.89	33.68	15.05	38.67	12.02	18.40	31.20	44.21	24.51	30.76
DeepSeek-V3.2 (Liu et al., 2025)	63.81	42.73	19.66	**55.78**	26.60	38.02	60.59	35.60	36.13	42.24	20.41	40.48	15.16	20.18	**34.88**	47.52	29.10	**33.83**
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	57.37	36.98	20.38	49.28	24.79	39.44	53.53	30.54	29.62	39.68	19.75	29.18	8.03	21.18	27.33	41.58	26.65	29.19
Closed-source models																		
GPT-5.6-Luna (OpenAI, 2026)	65.26	32.93	23.79	52.28	22.63	38.46	**71.96**	32.79	29.62	41.60	21.19	37.24	**19.27**	17.43	33.14	50.07	25.39	32.45
Gemini-2.5-Flash (Comanici et al., 2025)	65.48	**45.20**	21.17	55.37	26.75	38.40	70.37	**39.71**	**36.57**	**45.34**	20.79	27.52	16.16	21.46	29.07	**50.54**	30.78	30.55
Claude-Haiku-4.5 (Anthropic, 2025)	58.53	40.34	**25.66**	48.90	**30.58**	**40.74**	60.83	37.28	28.74	40.22	**26.35**	31.55	11.36	21.45	23.06	43.97	**31.20**	29.95
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	61.83	40.18	20.48	51.10	23.79	31.73	62.04	33.42	35.69	41.31	18.60	37.94	17.02	**21.79**	27.91	46.66	27.56	30.75
I-track: multimodal physics-rendered visual evidence																		
Open-weight multimodal models: 7–8B																		
InternVL3-8B (Zhu et al., 2025)	47.11	29.19	13.85	23.02	14.55	15.19	43.44	19.45	31.90	15.25	1.88	17.52	0.00	13.15	12.50	25.76	15.64	18.19
Qwen3-VL-8B-Instruct (Bai et al., 2025)	53.36	31.23	20.87	24.31	19.52	12.16	41.77	16.78	34.01	15.30	4.44	15.11	0.00	15.99	34.17	26.95	17.59	23.26
Open-weight multimodal models: 20–50B																		
InternVL3-38B (Zhu et al., 2025)	56.51	37.03	27.96	23.77	22.37	18.08	43.58	19.69	33.41	15.68	13.71	**33.72**	0.00	20.97	19.58	27.91	22.75	26.55
Qwen3-VL-32B-Instruct (Bai et al., 2025)	54.77	**42.98**	48.02	27.13	**31.71**	**29.65**	51.86	28.84	36.27	14.96	**22.16**	27.10	0.09	20.55	33.75	29.76	**29.25**	**34.96**
Open-weight multimodal models: more than 50B total parameters																		
InternVL3-78B (Zhu et al., 2025)	**59.65**	35.65	25.40	23.44	21.11	26.04	43.57	17.11	25.94	16.26	13.03	23.36	0.00	18.62	38.33	28.58	21.10	27.81
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	31.03	20.48	13.16	18.50	18.36	18.47	36.20	16.38	22.02	9.30	9.44	15.03	0.00	13.79	12.50	19.01	15.69	16.24
Closed-source multimodal models																		
GPT-5.6-Luna (OpenAI, 2026)	57.97	35.13	34.10	24.62	22.83	27.35	49.84	20.20	36.27	**20.28**	16.87	20.40	7.04	14.77	17.50	31.95	21.96	27.12
Gemini-2.5-Flash (Comanici et al., 2025)	48.65	23.50	27.30	26.57	21.46	25.71	36.92	20.14	35.52	19.45	13.32	20.40	10.17	19.21	37.50	28.35	19.53	29.29
Claude-Haiku-4.5 (Anthropic, 2025)	51.01	38.37	**49.45**	**29.06**	29.15	24.52	**53.06**	**31.15**	35.67	17.97	22.11	25.23	**14.41**	18.96	33.75	**33.10**	27.95	33.72
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	51.25	35.95	27.59	26.06	25.59	27.81	51.13	28.07	**37.33**	18.62	11.95	26.95	11.92	**21.83**	**45.42**	31.80	24.68	33.02

Tables 14 and 15 report these metrics for controlled simulations, while Tables 16 and 17 provide the corresponding real-world-aligned results. These diagnostics are reported separately from Acc., F1, and Brier because they characterize different properties of an open physical-world report rather than another form of answer accuracy.

Controlled-simulation explanation quality. The physical-axis results reveal a substantial separation between evidence recovery and mechanism reconstruction. For many models and capability levels, Evi-F1 is considerably higher than Mech, indicating that identifying relevant physical observations is easier than correctly connecting them through the underlying causal mechanism. This gap is visible even for strong models and becomes particularly important at the higher-level capability dimensions.

GLS provides a complementary view. A model may recover individual Gold evidence items without consistently using them to support the final physical commitment, and conversely a report with reasonable answer accuracy may still have weak explicit evidence linkage. The three explanation metrics should therefore not be interpreted as redundant measures of the same property.

Table 15: **Open-report explanation metrics along the fire scenario task axis.** Models are evaluated on the controlled simulation subset. Each dimension reports Evidence-F1 (Evi-F1), Mechanism Alignment (Mech), and Gold-Linked Support (GLS); Average is the equal-weight mean across T1–T3. Higher is better for all three metrics.

Model	T1			T2			T3			Average		
	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS
S-track: structured sensor and physical records												
Open-weight models: fewer than 20B total parameters												
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	46.56	12.70	19.10	47.66	27.58	25.80	40.98	17.85	14.30	45.07	19.38	19.73
Qwen3-8B (Yang et al., 2025)	46.53	17.17	22.60	51.38	23.35	35.00	50.69	24.86	19.50	49.53	21.79	25.70
Gemma-3-12B-IT (Gemma Team et al., 2025)	52.96	14.91	23.80	54.63	26.63	22.00	49.54	23.67	18.20	52.38	21.74	21.33
Open-weight models: 20–50B total parameters												
Gemma-3-27B-IT (Gemma Team et al., 2025)	49.68	14.29	24.00	51.71	23.09	26.30	49.82	22.02	21.40	50.40	19.80	23.90
Qwen3-32B (Yang et al., 2025)	48.90	21.68	26.00	50.71	28.47	30.50	46.69	27.02	21.40	48.77	25.72	25.97
Open-weight models: 50–200B total parameters												
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	54.24	14.90	20.50	51.26	30.58	32.30	47.72	23.60	14.40	51.07	23.03	22.40
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	48.17	18.71	23.50	49.79	25.74	23.50	46.13	28.06	20.60	48.03	24.17	22.53
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	49.54	20.00	23.90	49.74	32.81	28.50	45.92	25.27	20.20	48.40	26.03	24.20
Open-weight models: more than 200B total parameters												
Qwen3-235B-A22B (Yang et al., 2025)	44.67	19.77	23.10	49.47	26.44	29.60	47.62	25.42	23.10	47.25	23.88	25.27
DeepSeek-V3.2 (Liu et al., 2025)	52.46	23.76	**27.00**	50.99	32.32	**36.90**	49.22	29.16	19.80	50.89	28.41	**27.90**
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	47.65	20.27	26.90	45.75	28.36	26.80	42.65	27.62	21.70	45.35	25.42	25.13
Closed-source models												
GPT-5.6-Luna (OpenAI, 2026)	50.66	18.18	26.00	**57.21**	29.95	33.80	**54.11**	28.06	17.20	53.99	25.40	25.67
Gemini-2.5-Flash (Comanici et al., 2025)	**54.74**	22.98	25.70	56.98	34.45	32.50	51.36	**32.88**	22.90	**54.36**	30.10	27.03
Claude-Haiku-4.5 (Anthropic, 2025)	47.45	**27.13**	25.90	50.55	**35.89**	32.30	43.92	29.40	**24.50**	47.31	**30.81**	27.57
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	49.27	21.68	24.90	51.25	29.91	32.40	48.34	30.98	22.20	49.62	27.52	26.50
I-track: multimodal physics-rendered visual evidence												
Open-weight multimodal models: 7–8B												
InternVL3-8B (Zhu et al., 2025)	21.22	11.28	17.20	32.97	13.07	25.20	33.94	23.62	13.40	29.38	15.99	18.60
Qwen3-VL-8B-Instruct (Bai et al., 2025)	23.84	16.70	9.50	31.36	11.34	26.90	34.29	21.23	25.40	29.83	16.42	20.60
Open-weight multimodal models: 20–50B												
InternVL3-38B (Zhu et al., 2025)	22.33	17.17	23.30	32.78	20.44	**32.40**	37.68	31.45	25.10	30.93	23.02	26.93
Qwen3-VL-32B-Instruct (Bai et al., 2025)	25.67	**28.11**	23.30	37.73	29.96	30.80	37.45	29.00	43.10	33.62	**29.02**	32.40
Open-weight multimodal models: more than 50B total parameters												
InternVL3-78B (Zhu et al., 2025)	22.27	16.51	26.30	32.96	18.10	24.10	38.81	29.73	29.90	31.35	21.45	26.77
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	18.47	17.17	14.80	24.88	14.49	16.60	22.03	15.42	12.90	21.79	15.69	14.77
Closed-source multimodal models												
GPT-5.6-Luna (OpenAI, 2026)	25.98	20.01	26.60	36.64	20.25	29.50	**40.28**	28.06	28.30	34.30	22.77	28.13
Gemini-2.5-Flash (Comanici et al., 2025)	27.08	17.74	24.90	28.27	19.25	29.50	35.29	22.01	30.80	30.21	19.67	28.40
Claude-Haiku-4.5 (Anthropic, 2025)	**28.12**	24.81	26.60	**38.40**	**30.34**	30.00	38.30	**31.63**	**44.00**	**34.94**	28.93	**33.53**
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	26.19	21.60	**32.60**	37.28	22.68	28.70	37.59	31.05	33.80	33.69	25.11	31.70

The fire-task decomposition exhibits the same general phenomenon. Evidence grounding, mechanism alignment, and Gold-linked support vary differently across Localized Onset, Coupled Propagation, and Critical Transition. In particular, mechanism alignment remains substantially more difficult than recovering surface-level evidence for many model–stage combinations. This helps explain why improvements in open-report answer accuracy do not always imply an equivalent improvement in physically accountable reasoning.

These results complement the main benchmark metrics by showing where an apparently correct report obtains its support. FireWorldBench therefore distinguishes between producing the right final commitment, citing relevant evidence, recovering the correct mechanism, and linking the final claim to evidence sanctioned by the benchmark annotation.

Real-world-aligned explanation quality. The same diagnostics are evaluated on the real-world-aligned subset to test whether explanation failures persist outside the controlled simulation setting. The results remain highly heterogeneous across models and capability levels, but the distinction between evidence identification and mechanism understanding is again clear.

Table 16: **Open-report explanation metrics along the physical capability axis.** Models are evaluated on the real-world-aligned subset. Each dimension reports Evidence-F1 (Evi-F1), Mechanism Alignment (Mech), and Gold-Linked Support (GLS); Average is the equal-weight mean across P1–P5. Higher is better for all three metrics.

Model	P1			P2			P3			P4			P5			Average		
	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS
S-track: structured sensor and physical records																		
Open-weight models: fewer than 20B total parameters																		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	50.70	34.30	23.26	54.36	13.86	28.64	85.66	30.91	27.27	41.83	4.78	16.67	18.19	15.37	16.67	50.15	19.84	22.50
Qwen3-8B (Yang et al., 2025)	65.34	43.53	10.85	56.10	21.61	28.64	86.51	36.36	**60.61**	45.97	8.75	38.33	25.94	14.86	22.22	55.97	25.02	32.13
Gemma-3-12B-IT (Gemma Team et al., 2025)	63.52	38.90	28.68	57.30	18.58	32.05	84.87	33.03	36.36	48.31	13.49	23.33	26.63	16.77	20.37	56.13	24.15	28.16
Open-weight models: 20–50B total parameters																		
Gemma-3-27B-IT (Gemma Team et al., 2025)	**68.03**	43.88	26.87	60.78	18.32	34.32	79.35	32.12	45.45	46.02	11.78	37.50	27.83	15.17	16.67	56.40	24.25	32.16
Qwen3-32B (Yang et al., 2025)	63.76	36.37	28.42	54.44	25.18	31.59	81.03	**41.52**	56.06	49.14	12.19	30.83	24.88	17.87	20.37	54.65	26.63	33.45
Open-weight models: 50–200B total parameters																		
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	63.19	38.58	23.77	**65.35**	15.84	36.36	**90.35**	35.45	39.39	53.30	8.67	42.50	26.80	17.12	25.93	**59.80**	23.13	33.59
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	63.00	41.28	20.93	52.14	27.46	43.41	71.79	32.73	50.00	47.09	17.14	38.33	17.59	20.62	25.93	50.32	27.85	35.72
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	59.66	44.80	27.13	56.24	23.44	33.64	77.96	33.03	39.39	51.59	22.45	35.00	**29.71**	19.78	31.48	55.03	28.70	33.33
Open-weight models: more than 200B total parameters																		
Qwen3-235B-A22B (Yang et al., 2025)	58.36	38.95	29.20	60.57	24.03	37.05	75.69	32.12	46.97	35.21	12.43	41.67	16.81	19.20	29.63	49.33	25.35	36.90
DeepSeek-V3.2 (Liu et al., 2025)	53.15	45.57	14.73	56.05	23.73	44.09	64.32	38.48	42.42	50.60	21.65	30.83	20.74	22.61	22.22	48.97	30.41	30.86
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	52.33	37.18	19.64	52.92	26.66	44.32	57.50	28.48	39.39	46.18	27.15	24.17	16.87	**28.38**	25.93	45.16	29.57	30.69
Closed-source models																		
GPT-5.6-Luna (OpenAI, 2026)	56.95	36.51	29.46	59.68	26.81	**45.91**	83.02	29.39	50.00	**60.82**	24.12	41.67	20.10	18.31	**42.59**	56.11	27.03	**41.93**
Gemini-2.5-Flash (Comanici et al., 2025)	61.14	**47.38**	29.97	62.23	25.08	42.50	79.34	37.27	46.97	53.77	16.49	35.83	22.62	18.82	27.78	55.82	29.01	36.61
Claude-Haiku-4.5 (Anthropic, 2025)	53.45	42.87	**34.11**	48.70	**30.22**	45.45	65.17	37.27	50.00	45.00	**27.71**	41.67	18.04	23.42	22.22	46.07	**32.30**	38.69
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	61.50	40.34	28.68	56.49	27.47	39.55	72.64	32.12	50.00	47.46	18.43	**45.00**	19.78	19.89	31.48	51.57	27.65	38.94
I-track: multimodal physics-rendered visual evidence																		
Open-weight multimodal models: 7–8B																		
InternVL3-8B (Zhu et al., 2025)	47.28	26.93	7.32	19.56	15.37	20.00	43.57	18.78	23.33	19.78	3.48	18.18	0.00	13.89	15.00	26.04	15.69	16.77
Qwen3-VL-8B-Instruct (Bai et al., 2025)	54.34	28.73	16.53	25.73	17.29	**28.15**	43.33	22.37	33.33	25.62	13.15	17.42	0.00	19.13	25.00	29.80	20.13	24.09
Open-weight multimodal models: 20–50B																		
InternVL3-38B (Zhu et al., 2025)	55.42	30.18	11.11	27.34	22.84	24.81	45.66	23.58	35.83	19.38	16.82	40.15	0.00	23.37	15.00	29.56	23.36	25.38
Qwen3-VL-32B-Instruct (Bai et al., 2025)	55.37	**37.92**	**28.18**	**32.92**	31.45	26.30	53.96	30.47	40.42	25.15	21.45	**44.70**	4.76	24.49	35.00	34.43	29.16	**34.92**
Open-weight multimodal models: more than 50B total parameters																		
InternVL3-78B (Zhu et al., 2025)	**61.84**	33.48	20.87	28.59	15.36	24.44	48.80	20.10	25.83	21.65	13.24	40.15	3.33	22.83	**40.00**	32.84	21.00	30.26
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	47.82	33.11	22.49	23.25	21.78	21.11	34.00	18.66	16.25	23.86	16.66	21.21	3.10	22.04	25.00	26.41	22.45	21.21
Closed-source multimodal models																		
GPT-5.6-Luna (OpenAI, 2026)	53.86	35.14	20.05	23.66	21.60	26.30	47.88	20.60	31.67	25.52	15.59	27.27	3.43	19.46	30.00	30.87	22.48	27.06
Gemini-2.5-Flash (Comanici et al., 2025)	53.06	34.52	15.45	25.01	25.09	25.93	40.75	24.73	36.25	**27.31**	15.89	40.15	7.60	**26.99**	35.00	30.75	25.44	30.56
Claude-Haiku-4.5 (Anthropic, 2025)	52.99	35.13	23.04	30.58	**31.67**	22.59	**54.84**	**35.27**	**42.92**	22.83	**22.21**	31.06	**16.17**	24.22	20.00	**35.48**	**29.70**	27.92
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	55.43	32.44	18.43	28.24	25.94	27.04	49.46	27.47	37.50	24.94	12.64	32.58	5.67	25.08	20.00	32.75	24.71	27.11

For the physical capability axis, several models achieve relatively strong Evi-F1 while obtaining much lower Mechanism Alignment. The gap is particularly informative at the more demanding capability levels: a model can identify observations associated with the target phenomenon without recovering the causal interpretation required to explain that phenomenon correctly. GLS further shows that correct evidence extraction does not guarantee that the generated answer is explicitly grounded in the appropriate Gold-linked support.

Table 17: **Open-report explanation metrics along the fire scenario task axis.** Models are evaluated on the real-world-aligned subset. Each dimension reports Evidence-F1 (Evi-F1), Mechanism Alignment (Mech), and Gold-Linked Support (GLS); Average is the equal-weight mean across T1–T3. Higher is better for all three metrics.

Model	T1			T2			T3			Average		
	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS	Evi-F1	Mech	GLS
S-track: structured sensor and physical records												
Open-weight models: fewer than 20B total parameters												
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	49.24	9.93	27.60	67.04	20.38	17.90	41.11	28.72	21.30	52.46	19.68	22.27
Qwen3-8B (Yang et al., 2025)	53.64	17.45	33.70	66.44	23.58	47.20	53.71	35.07	14.20	57.93	25.37	31.70
Gemma-3-12B-IT (Gemma Team et al., 2025)	56.03	16.09	31.90	65.34	25.31	26.00	52.64	32.37	26.20	58.00	24.59	28.03
Open-weight models: 20–50B total parameters												
Gemma-3-27B-IT (Gemma Team et al., 2025)	54.78	15.04	36.10	65.86	24.55	40.70	**56.17**	35.41	23.90	58.94	25.00	33.57
Qwen3-32B (Yang et al., 2025)	52.59	19.76	29.10	66.47	29.86	48.00	52.29	30.91	26.00	57.12	26.84	34.37
Open-weight models: 50–200B total parameters												
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	**62.82**	13.15	36.40	71.01	23.63	43.90	52.46	32.25	24.40	**62.10**	23.01	34.90
Qwen3-Next-80B-A3B-Instruct (Qwen Team, 2025)	50.54	23.49	37.60	60.28	26.50	51.20	49.60	35.18	22.40	53.47	28.39	37.07
GLM-4.5-Air (106B-A12B) (Zeng et al., 2025)	54.31	21.13	29.40	66.43	31.27	44.70	50.82	37.42	28.40	57.19	29.94	34.17
Open-weight models: more than 200B total parameters												
Qwen3-235B-A22B (Yang et al., 2025)	50.05	19.64	**40.50**	60.61	24.01	41.50	46.10	33.12	29.30	52.25	25.59	37.10
DeepSeek-V3.2 (Liu et al., 2025)	56.03	21.01	37.00	55.21	33.93	41.50	43.59	38.80	16.90	51.61	31.25	31.80
Kimi-K2-Instruct (1T-A32B) (Team et al., 2025)	51.93	22.79	34.10	50.38	34.25	38.20	41.86	34.58	21.50	48.06	30.54	31.27
Closed-source models												
GPT-5.6-Luna (OpenAI, 2026)	57.48	24.44	37.70	**76.80**	29.32	**56.90**	46.08	31.14	**33.30**	60.12	28.30	**42.63**
Gemini-2.5-Flash (Comanici et al., 2025)	60.70	22.68	35.90	65.58	27.05	48.80	49.78	**38.95**	29.30	58.69	29.56	38.00
Claude-Haiku-4.5 (Anthropic, 2025)	48.30	**26.74**	37.40	54.57	**37.07**	56.90	43.00	37.13	30.60	48.62	**33.65**	41.63
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	53.26	22.58	37.00	61.46	28.89	54.50	49.19	34.31	29.50	54.64	28.59	40.33
I-track: multimodal physics-rendered visual evidence												
Open-weight multimodal models: 7–8B												
InternVL3-8B (Zhu et al., 2025)	16.96	11.45	19.50	36.96	13.30	21.40	32.88	22.65	9.80	28.93	15.80	16.90
Qwen3-VL-8B-Instruct (Bai et al., 2025)	25.94	14.03	22.10	36.44	20.93	30.20	36.52	25.58	19.30	32.97	20.18	23.87
Open-weight multimodal models: 20–50B												
InternVL3-38B (Zhu et al., 2025)	22.81	18.30	19.70	36.91	23.76	**47.40**	38.63	27.94	12.40	32.78	23.33	26.50
Qwen3-VL-32B-Instruct (Bai et al., 2025)	**30.87**	24.84	**31.00**	42.81	30.67	43.00	38.78	**33.52**	**30.40**	**37.49**	29.68	**34.80**
Open-weight multimodal models: more than 50B total parameters												
InternVL3-78B (Zhu et al., 2025)	24.14	11.90	23.60	40.50	20.37	37.00	**41.17**	29.99	27.10	35.27	20.75	29.23
Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025)	23.21	18.15	18.70	30.43	20.00	20.60	33.15	29.48	23.30	28.93	22.54	20.87
Closed-source multimodal models												
GPT-5.6-Luna (OpenAI, 2026)	24.99	17.81	24.40	38.73	20.69	32.30	37.32	30.00	23.30	33.68	22.83	26.67
Gemini-2.5-Flash (Comanici et al., 2025)	26.91	21.29	25.60	34.49	22.40	42.40	38.15	32.05	21.90	33.18	25.25	29.97
Claude-Haiku-4.5 (Anthropic, 2025)	28.10	**25.34**	20.30	**42.94**	**33.85**	43.50	40.92	31.55	22.00	37.32	**30.25**	28.60
Qwen3.5-Plus-02-15 (Qwen Team, 2026)	27.40	21.04	30.50	40.09	22.73	33.90	39.11	30.03	18.90	35.53	24.60	27.77

The fire scenario task axis leads to the same conclusion from a fire-evolution perspective. Explanation quality varies substantially across Localized Onset, Coupled Propagation, and Critical Transition, and the model ranking under Evi-F1 does not necessarily coincide with the ranking under Mechanism Alignment or GLS. This divergence is important because a fire-world model that reaches a plausible stage-specific prediction for the wrong physical reason may behave unreliably when the environment, observation modality, or physical state changes.

Taken together, the additional results clarify the main findings of FireWorldBench. First, response format matters: multiple-choice competence systematically overstates some abilities that become substantially harder under open reporting. Second, structured and multimodal observations expose different model strengths, and strong performance on sensor records does not guarantee robust reasoning from physics-rendered visual evidence. Third, both the physical capability axis and the fire scenario task axis reveal pronounced internal heterogeneity that is obscured by a single aggregate score. Finally, open-report diagnostics show that answer correctness, evidence recovery, causal mechanism understanding, and evidence-grounded support are related but distinct dimensions of physical-world intelligence. These observations motivate evaluating fire-world models through the complete two-axis, multi-interface protocol rather than through final-answer accuracy alone.

C.5 Environment-family analysis

Beyond the capability- and task-level decompositions, we further examine whether open-report performance varies across the seven environment families represented in the controlled simulation subset. Table 18 reports the six evaluation metrics after grouping the results by their source environment family. This analysis provides a complementary view of FireWorldBench by revealing whether answer quality, calibration, and physical explanation remain stable across different classes of fire environments.

Table 18: **Open-report performance across environment families on the controlled simulation subset.** Results are aggregated separately over the seven environment families defined in FireWorld-Bench. Acc., F1, Evi-F1, Mech, and GLS are reported as percentages and are higher-is-better, whereas Brier is reported on the $[0, 1]$ scale and lower is better. The best result in each metric is shown in bold.

Environment Family	Acc.	F1	Brier	Evi-F1	Mech	GLS
Transportation and Infrastructure	41.6	**51.0**	0.42	36.4	**45.9**	24.6
Office and Digital Facilities	41.0	45.8	0.37	41.6	23.9	23.7
Healthcare and Education	**46.8**	49.5	0.33	**46.9**	24.9	28.3
Commercial and Public Spaces	44.4	49.1	0.35	42.0	24.5	27.1
Residential and Care Settings	45.5	48.0	**0.32**	39.3	24.0	**28.8**
Industry, Energy, and Logistics	40.5	48.1	0.37	44.1	24.2	23.6
Wildland and Wildland–Urban Interface (WUI)	41.8	41.1	0.37	37.6	21.6	21.1

Table 19: **FireWorldGPT results under multiple-choice evaluation.** Rows compare Vanilla, physics-aware prompting, chain-of-thought reasoning, and supervised fine-tuning across four foundation models. Colored annotations report changes relative to Vanilla: green denotes performance improvement and red denotes degradation, while arrows indicate the numerical direction of change. Acc. and F1 are higher-is-better, whereas lower Brier is better.

Method	Llama-3.1-8B-Instruct			Qwen3-8B			InternVL3-8B			Qwen3-VL-8B		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
Vanilla	24.4	34.9	0.90	30.5	40.9	0.86	22.9	31.8	0.52	28.2	37.5	0.84
+Prompt	28.2 ↑3.8	39.0 ↑4.1	0.97 ↑0.07	31.3 ↑0.8	41.3 ↑0.4	0.89 ↑0.03	23.8 ↑0.9	34.1 ↑2.3	0.48 ↓0.04	33.3 ↑5.1	43.1 ↑5.6	0.89 ↑0.05
+CoT	24.0 ↓0.4	34.2 ↓0.7	**0.67** ↓0.23	32.7 ↑2.2	43.2 ↑2.3	0.82 ↓0.04	26.4 ↑3.5	35.2 ↑3.4	**0.21** ↓0.31	30.3 ↑2.1	40.5 ↑3.0	0.87 ↑0.03
+SFT	**42.6** ↑18.2	**55.9** ↑21.0	0.95 ↑0.05	**42.6** ↑12.1	**52.2** ↑11.3	**0.81** ↓0.05	**40.2** ↑17.3	**50.4** ↑18.6	0.87 ↑0.35	**41.3** ↑13.1	**51.2** ↑13.7	**0.83** ↓0.01

The results reveal clear environment-dependent differences that are not fully captured by answer accuracy alone. Healthcare and Education achieves the highest Acc. (46.8) and Evi-F1 (46.9), indicating comparatively strong answer completion and evidence recovery. Transportation and Infrastructure obtains the highest F1 (51.0) and, more notably, the highest Mechanism Alignment (45.9), suggesting substantially stronger recovery of the underlying physical mechanisms. Residential and Care Settings yields the best calibration (Brier $= 0.32$) and the highest GLS (28.8). In contrast, Wildland and Wildland–Urban Interface (WUI) records the lowest F1 (41.1), Mech (21.6), and GLS (21.1), while Industry, Energy, and Logistics has the lowest Acc. (40.5).

A particularly notable observation is that cross-family variation is much larger for mechanism-level understanding than for final-answer accuracy. Acc. varies within a relatively narrow range of 40.5–46.8, whereas Mech ranges from 21.6 to 45.9. This gap suggests that models can maintain superficially similar answer-level performance across different fire environments while differing substantially in whether they recover the correct underlying physical mechanism. Together with the preceding capability- and task-level results, this environment-family analysis further supports evaluating physical-world understanding through answer quality, calibration, evidence grounding, and mechanism recovery jointly rather than relying on accuracy alone.

D FireWorldGPT Additional Results

We provide the complete controlled-simulation results of the FireWorldGPT method study. While the main paper reports the equal average of the multiple-choice and open-report interfaces, we separate the two response formats here to expose their interface-specific behavior. All methods are evaluated on the same frozen test instances as their corresponding Vanilla configuration. The Vanilla rows in Tables 19 and 20 serve as the paired reference runs for the FireWorldGPT intervention study rather than as reproductions of the benchmark-wide baseline summaries in Appendix C. Accordingly, all reported method differences are interpreted relative to the matched Vanilla run within the same intervention-study configuration. For Prompt, CoT, and SFT, the colored annotation following each absolute score reports its signed difference from Vanilla. Green upward arrows denote numerical increases and red downward arrows denote numerical decreases. Acc. and F1 are reported in percentage points, whereas Brier and its change are reported directly on the $[0, 1]$ scale. Higher Acc. and F1 are better, while lower Brier is better.

Table 20: **FireWorldGPT results under open-report evaluation.** Rows compare Vanilla, physics-aware prompting, chain-of-thought reasoning, and supervised fine-tuning across four foundation models. Colored annotations report changes relative to Vanilla: green denotes performance improvement and red denotes degradation, while arrows indicate the numerical direction of change. Acc. and F1 are higher-is-better, whereas lower Brier is better.

Method	Llama-3.1-8B-Instruct			Qwen3-8B			InternVL3-8B			Qwen3-VL-8B		
	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier	Acc	F1	Brier
Vanilla	37.8	30.8	0.63	36.2	30.5	**0.74**	27.2	21.0	0.65	30.2	21.5	0.74
+Prompt	38.5 ↑0.7	29.9 ↓0.9	**0.62** ↓0.01	**39.7** ↑3.5	**32.7** ↑2.2	0.75 ↑0.01	**35.0** ↑7.8	24.7 ↑3.7	**0.64** ↓0.01	35.7 ↑5.5	24.2 ↑2.7	**0.66** ↓0.08
+CoT	40.3 ↑2.5	32.9 ↑2.1	0.72 ↑0.09	38.6 ↑2.4	30.9 ↑0.4	0.75 ↑0.01	29.9 ↑2.7	**24.8** ↑3.8	0.69 ↑0.04	33.2 ↑3.0	27.2 ↑5.7	0.76 ↑0.02
+SFT	**43.5** ↑5.7	**33.2** ↑2.4	0.83 ↑0.20	38.9 ↑2.7	29.4 ↓1.1	0.87 ↑0.13	28.0 ↑0.8	24.2 ↑3.2	0.86 ↑0.21	**40.6** ↑10.4	**27.4** ↑5.9	0.81 ↑0.07

D.1 Multiple-Choice Results

Multiple-choice evaluation shows a consistent advantage for SFT in answer quality across all four backbones. For Llama-3.1-8B-Instruct, SFT raises Acc. from 24.4 to 42.6 and F1 from 34.9 to 55.9. InternVL3-8B exhibits similarly large gains, increasing from 22.9 to 40.2 in Acc. and from 31.8 to 50.4 in F1. Qwen3-VL-8B follows the same trend, with SFT improving Acc. from 28.2 to 41.3 and F1 from 37.5 to 51.2. Thus, under candidate-conditioned evaluation, supervised adaptation produces the strongest answer-quality gains for every evaluated backbone.

Confidence calibration exhibits a less uniform pattern. CoT achieves the lowest Brier for Llama-3.1-8B-Instruct and InternVL3-8B, reducing the score from 0.90 to 0.67 and from 0.52 to 0.21, respectively. SFT instead achieves the lowest Brier for Qwen3-8B and Qwen3-VL-8B, at 0.81 and 0.83. In contrast, the strong Acc./F1 gains of InternVL3-8B under SFT are accompanied by a substantial Brier increase from 0.52 to 0.87. These results indicate that improvements in candidate-conditioned answer quality do not necessarily translate into uniformly better calibration.

D.2 Open-Report Results

Open-report evaluation reveals a more heterogeneous interaction between adaptation strategy and model backbone. SFT remains highly effective for several models even after candidate answers are removed. For Llama-3.1-8B-Instruct, SFT improves Acc. from 37.8 to 43.5 and F1 from 30.8 to 33.2. The strongest open-report SFT effect appears on Qwen3-VL-8B, where Acc. increases from 30.2 to 40.6 and F1 from 21.5 to 27.4. These gains show that parameter-level adaptation can improve open-ended physical-state reconstruction rather than only candidate-conditioned answer selection.

Prompting is more competitive for the other backbones. It achieves the highest Acc. and F1 for Qwen3-8B, and the highest Acc. for InternVL3-8B, while CoT obtains the highest InternVL3-8B F1. Calibration again behaves differently from answer quality. Prompting produces the lowest Brier for Llama-3.1-8B-Instruct, InternVL3-8B, and Qwen3-VL-8B, including a reduction from 0.74 to 0.66 for Qwen3-VL-8B. For Qwen3-8B, the Vanilla configuration retains the lowest Brier at 0.74. Overall, no single intervention dominates every open-report metric: supervised adaptation provides the strongest open-ended gains for some backbones, whereas prompting and CoT remain competitive on others, and confidence calibration can move independently of answer accuracy.

E FireWorldGPT Implementation Details

E.1 Controlled Enhancement Protocol

The FireWorldGPT study compares three complementary enhancement strategies against a shared Vanilla baseline: physics-aware prompting, chain-of-thought reasoning, and supervised fine-tuning. Vanilla uses the standard benchmark instruction together with the task-admissible observations and the prescribed response schema. The three enhancements are evaluated as mutually exclusive conditions rather than being composed: +Prompt modifies the presentation and organization of the available physical evidence, +CoT modifies the inference process while keeping the evidence fixed, and +SFT updates the model parameters using training-split supervision. FireWorldGPT specifically denotes the model obtained through the +SFT condition.

Within each model comparison, the benchmark split, logical items, observation track, temporal cutoff, response schema, decoding configuration, and deterministic evaluator are held fixed. The methods therefore differ only in the enhancement being tested. Prompting and CoT do not update the model parameters, whereas SFT uses no test-time reasoning scaffold beyond the standard Vanilla task instruction. No condition is allowed to

access hidden simulator states, post-cutoff observations, Gold evidence, or fields excluded by the task-specific observation contract.

E.2 Physics-Aware Prompting

Physics-aware prompting provides an explicit protocol for reading task-admissible fire-world evidence. The prompt directs the model to identify the relevant physical variable, spatial region, observation time, legend or measurement scale, and task-specific output fields before producing its prediction. For the structured-observation track, this protocol is applied to the supplied sensor traces and physical records. For the multimodal observation track, the input is supplemented with available observation-aligned visualizations of temperature, soot density, visibility, or velocity, restricted to timestamps permitted by the task cutoff. The model is explicitly instructed not to infer numerical values that cannot be reliably resolved from the input.

The augmentation is constructed only from public, task-admissible observations. It does not introduce derived evidence tables, region rankings, risk summaries, propagation annotations, answer-conditioned statistics, hidden thresholds, or future frames. Each added visualization retains its source world, physical variable, timestamp, observation cutoff, rendering configuration, and artifact hash. If no nonredundant and admissible field representation can be constructed for an item, the augmentation reduces to a no-op and the item retains its Vanilla evidence. Consequently, +Prompt changes neither the physical target nor the model parameters; it tests whether an explicit field-reading interface is sufficient to improve physical grounding.

E.3 Chain-of-Thought Reasoning

The +CoT condition uses exactly the same observations as Vanilla and adds a single instruction requesting an explicit four-stage inference process (Wei et al., 2022). The model is asked to first identify the relevant observations, then infer the latent physical state, connect that state to a coupled-field mechanism, and finally derive the task-specific consequence. This ordering follows the observation–state–mechanism–consequence formulation introduced in Section 4 and is applied consistently to grounding, forecasting, diagnosis, and intervention questions.

The complete intermediate rationale is retained for qualitative auditing, but only the final answer is scored. A deterministic extractor projects the final JSON object into the same choice-format or open-report schema used by Vanilla. The method does not employ an external physical rule base, retrieval system, simulator call, verifier, correction stage, self-consistency voting, multi-turn refinement, or LLM judge. It therefore measures the effect of exposing a model's internal inference path while holding its parameters and observable evidence fixed.

E.4 Supervised Fine-Tuning

The +SFT condition adapts the base model using publicly available examples from the controlled-simulation training split of FireWorldBench (Ouyang et al., 2022). Training examples are deduplicated at the logical-item level before serialization, ensuring that paired response formats or repeated renderings do not unintentionally multiply the supervision associated with a single physical target. All temporal variants and related instances from the same source event remain subject to the event-group split, and no evaluation or real-world-aligned test item is used for optimization.

Each training input contains the standard task instruction, question, and task-admissible observations from the corresponding observation track. Choice-format targets contain the correct option set. Open-report targets contain the structured `prediction`, a concise `conclusion`, Gold-linked `evidence`, and the corresponding physical `mechanism`. Visual-model training uses the authorized multimodal observations associated with the item, whereas the structured-observation counterpart uses only the corresponding textual sensor and physical records. Evidence is never transferred across tracks when it is unavailable under the original task contract.

Let $\mathcal{D}_{\text{train}}$ denote the resulting training set, x an input sequence, and $y = (y_1, \ldots, y_{|y|})$ its serialized target. The model is optimized using the standard autoregressive next-token objective:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{(x,y)\in\mathcal{D}_{\text{train}}} \sum_{t=1}^{|y|} m_t \log p_\theta(y_t \mid x, y_{<t}). \tag{34}$$

Here, θ denotes the trainable model parameters and m_t is a token-level supervision mask. Because the released training examples do not provide authoritative confidence annotations, tokens corresponding to the confidence value are excluded from the supervised objective:

$$m_t = \begin{cases} 0, & t \in \mathcal{T}_{\text{conf}}, \\ 1, & t \notin \mathcal{T}_{\text{conf}}, \end{cases} \tag{35}$$

where $\mathcal{T}_{\text{conf}}$ is the set of target positions occupied by the confidence value. The remaining response structure is fully supervised. We introduce no auxiliary classification objective, contrastive loss, pairwise preference loss, reward model, synthetic negative examples, or additional self-training stage. Thus, improvements under +SFT can be attributed to learning the mapping from admissible observations to structured physical predictions, evidence, and mechanisms rather than to an auxiliary training signal.

At inference time, FireWorldGPT receives the same standard task prompt and response schema as Vanilla. Confidence is generated by the adapted model but is not learned from an authoritative confidence target. Brier scores can therefore still be computed by the common evaluator, but they should not be interpreted as evidence that SFT directly optimizes calibration.

E.5 Paired Evaluation and Reproducibility

All four conditions—Vanilla, +Prompt, +CoT, and +SFT—are evaluated on matched test instances. For a metric M, method a, and a fixed evaluation cell $\mathcal{D}_k$, the reported improvement is computed relative to the corresponding Vanilla run:

$$\Delta M_{a,k} = M(f_a; \mathcal{D}_k) - M(f_{\text{Vanilla}}; \mathcal{D}_k). \tag{36}$$

Here, f_a denotes the prediction function induced by enhancement a, while $\mathcal{D}_k$ fixes the model, benchmark subset, capability group, observation track, and response interface. Deltas are computed from the unrounded scores before being displayed in Tables 19 and 20. This paired construction prevents an apparent improvement from being caused by a different test subset, missing predictions, or a changed observation policy.

Every run records the base checkpoint revision, enhancement condition, benchmark split, prompt and template hashes, observation configuration, temporal cutoff policy, decoding parameters, parser version, and evaluator version. Prompt runs additionally record the provenance of every attached field visualization; CoT runs retain the unmodified rationale and the deterministically extracted final response; SFT runs record the training-split manifest and target-serialization version. All predictions are scored using the deterministic protocol in Appendix B, and invalid responses remain in the applicable metric denominators. These controls make the three enhancement strategies directly comparable while preserving the distinction between input-level evidence structuring, inference-time reasoning, and parameter-level domain adaptation.

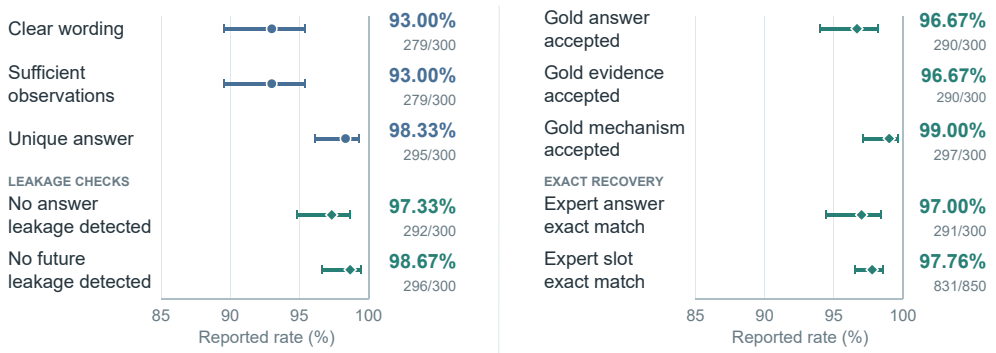
F Expert-Validated Questions and Answers.

Figure 8 provides evidence of task answerability and reference reliability within the audited subset. Question clarity and observation sufficiency each receive 93.0% fully positive ratings, although rechecks flag answer leakage in eight items and future-information leakage in four. Gold-answer and mechanism acceptance reach 96.67% and 99.0%, respectively, while reported expert exact agreement reaches 97.0% across answers and 97.76% across output slots. These findings indicate that most audited questions are clearly formulated, provide sufficient observations, and exhibit no detected answer or future-information leakage, while strong agreement with expert responses supports the validity and accuracy of the Gold answers.

G Qualitative Benchmark Cases and Multimodal Fire-World Visualizations

Figures 9–13 present one representative benchmark case for each dimension of the physical capability axis, ordered from Temporal Evolution Forecasting (P1) to Counterfactual Intervention Reasoning (P5). Each case pairs a choice-format question with its matched open-report counterpart under the same physical target and multimodal evidence, and displays the Gold response alongside a representative answer from GPT-5.6-Luna. These examples illustrate how the benchmark progresses from identifying temporal order and grounding localized field observations to integrating coupled fields, attributing causal mechanisms, and predicting the consequences of interventions. They also expose the distinction between recognizing a plausible answer and producing a structured report whose conclusion is explicitly connected to physical evidence and mechanism.

Figures 14–20 complement the question-level cases with one controlled fire world from each of the seven environment families, following the ordering used in the main paper. For every world, time-indexed three-dimensional scene renderings are co-registered with temperature, soot-density, velocity, visibility, and streamwise-velocity fields, making both the semantic environment and its latent physical evolution inspectable. Figure 21 then



(a) Question quality

(b) Expert-Gold agreement

Figure 8: **Question quality and expert–Gold agreement.** Expert audit results on question quality, leakage, Gold acceptance, and agreement with Gold answers and output slots.

presents a real-world-aligned heptane pool-fire event: recorded video frames establish the observed chronology, whereas the aligned three-dimensional reconstruction and FDS-derived fields provide the simulation-completed latent state. The two provenance sources remain visually separated to avoid presenting reconstructed physical fields as direct measurements.

P1.Temporal Evolution Forecasting:

Multiple Choice:

Question: Between t=8 s and t=16 s, the soot plume reaches the ceiling and the low-visibility region expands laterally. If conditions remain unchanged, what is likely next?

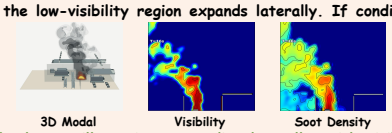
Options:

A. Further upper-layer spreading with lower nearby visibility

B. Upper-layer contraction with recovering visibility

C. Nearly unchanged smoke distribution

D. Direct floor-level soot migration



3D Modal | Visibility | Soot Density

GPT-5.6 Luna: A — The observed progression indicates that the upper smoke layer will continue spreading laterally, with visibility further decreasing in adjacent ceiling regions.

Open Report:

Question: Using the observed soot and visibility changes, predict the next short-time state and explain the temporal evolution.

Correct Answer:


```
{
  "conclusion": "The upper smoke layer will continue spreading, with visibility deteriorating in adjacent ceiling regions.",
  "prediction": {
    "smoke_evolution": "continued_upper_layer_spread"
  },
  "evidence": "From t=8 s to t=16 s, soot evolves from a narrow rising plume into a ceiling-reaching cloud, while the low-visibility region expands across the same upper area and the source and ventilation conditions remain unchanged throughout this interval.",
  "mechanism": "Continued combustion sustains buoyant transport toward the ceiling. Once vertical motion is constrained by the upper boundary, smoke spreads laterally and accumulates, increasing soot coverage and progressively reducing visibility across adjacent ceiling regions.",
  "confidence": 0.95
}
```


GPT-5.6 Luna:


```
{
  "conclusion": "The upper smoke layer is likely to continue expanding laterally, with further visibility loss near the ceiling.",
  "prediction": {
    "smoke_evolution": "continued_upper_layer_spread"
  },
  "evidence": "Between t=8 s and t=16 s, the soot plume grows from a narrow column into a broader ceiling-level cloud, while the low-visibility region expands over the same upper area.",
  "mechanism": "Sustained buoyant transport carries smoke upward, and once it reaches the ceiling the flow spreads laterally, increasing soot accumulation and reducing visibility.",
  "confidence": 0.88
}
```


Figure 9: **P1: Temporal Evolution Forecasting.** A matched choice-format and open-report case asks how the smoke state will evolve after the soot plume reaches the ceiling and the low-visibility region expands laterally from $t = 8\,\mathrm{s}$ to $t = 16\,\mathrm{s}$. The Gold response predicts continued upper-layer spreading with further visibility degradation in adjacent ceiling regions. The representative GPT-5.6-Luna response reaches the same forecast and grounds it in the observed ceiling interaction, lateral soot expansion, and progressive visibility loss.

P2.Physical Field Perception and Grounding:

Multiple Choice:

Question: At t=16 s, the co-registered temperature and soot-density maps show nonuniform patterns across the ward. Which region is most strongly supported as the common location of the dominant physical anomaly?

Options:

A. The west-corridor end beside the local alcove
B. The central nurse-station and nearby rooms
C. The east patient-room cluster
D. The south entrance and lobby area



GPT-5.6 Luna: A — The west-corridor end is the only region where the strongest temperature and soot-density anomalies are spatially co-located.

Open Report:

Question: Using the 3D layout and both physical-field maps, identify the region containing the dominant anomaly and explain how their spatial distributions support that grounding over the alternatives.

Correct Answer:


```
{
  "conclusion": "The dominant thermal and soot anomaly is grounded to the west-corridor end rather than another ward region.",
  "prediction": {
    "anomaly_region": "west_corridor_end"
  },
  "evidence": "At t=16 s, the temperature maximum and highest soot-density region are co-located at the west-corridor end; both fields decrease with distance from that location, while the central, eastern, and entrance regions remain comparatively close to background and show no matching joint hotspot.",
  "mechanism": "A local combustion region produces heat and soot in the same neighborhood, so co-location of the two field maxima provides consistent spatial evidence for the west-corridor end rather than another ward region lacking a corresponding multi-field anomaly.",
  "confidence": 0.95
}
```


GPT-5.6 Luna:

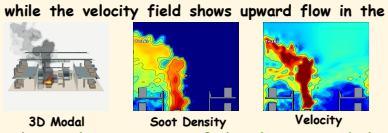

```
{
  "conclusion": "The dominant anomaly is localized at the west-corridor end.",
  "prediction": {
    "anomaly_region": "west_corridor_end"
  },
  "evidence": "The temperature and soot-density maps both peak near the west-corridor end, whereas the central, eastern, and entrance regions remain much closer to background conditions.",
  "mechanism": "The co-location of elevated temperature and soot indicates a common localized combustion-related anomaly rather than independent changes elsewhere in the ward.",
  "confidence": 0.92
}
```


Figure 10: **P2: Physical Field Perception and Grounding.** A localized anomaly in a care-ward corridor is presented through a 3D scene together with temperature and soot-density fields. The task requires distinguishing a nearby linen-cart or combustible-material source from spatially broader alternatives such as ventilation or system-level failure. The representative GPT-5.6-Luna response localizes the source to the west corridor and supports this attribution using the spatial co-occurrence and rapid decay of the thermal and soot anomalies.

P3.Cross-Field Coupling Understanding:

Multiple Choice:

Question: The soot plume rises above the kiosk and bends toward the upper-left, while the velocity field shows upward flow in the plume core and leftward flow aloft. Which interpretation best explains both fields?



Options:

A. Upper geometry redirects smoke despite the measured local airflow
B. Buoyancy dominates while horizontal flow contributes negligibly to plume shape
C. Buoyancy lifts soot upward while leftward airflow advects the upper plume
D. Soot transport and airflow evolve independently in the observed region

GPT-5.6 Luna: C — Buoyancy carries the soot upward, while the leftward airflow advects the upper part of the plume toward the upper-left.

Open Report:

Question: Using soot morphology, velocity, and the 3D structure, explain how the two physical fields jointly determine the observed smoke transport pattern.

Correct Answer:


```
{
  "conclusion": "The plume shape reflects coupled buoyant rise and leftward advection through the upper atrium rather than independent field behavior.",
  "prediction": {
    "field_coupling": "buoyancy_plus_leftward_advection"
  },
  "evidence": "The soot field forms a connected rising column bending upper-left, while the velocity field shows upward motion in the plume core and leftward flow across the same upper region.",
  "mechanism": "Heating generates buoyant ascent that carries soot upward; the horizontal airflow then advects soot-bearing gases leftward, while the atrium geometry constrains the combined transport path through the upper space.",
  "confidence": 0.90
}
```


GPT-5.6 Luna:


```
{
  "conclusion": "The plume reflects coupled buoyant rise and leftward advection in the upper atrium.",
  "prediction": {
    "field_coupling": "buoyancy_plus_leftward_advection"
  },
  "evidence": "The soot column rises and bends toward the upper-left, while the velocity field shows upward motion near the plume core and leftward motion in the corresponding upper region.",
  "mechanism": "Buoyancy transports soot vertically from the source, after which the horizontal airflow advects the soot-bearing gases leftward along the available upper-space pathway.",
  "confidence": 0.90
}
```


Figure 11: **P3: Cross-Field Coupling Understanding.** A kiosk fire in a public atrium is queried through its 3D structure, soot-density distribution, and velocity field. The task requires combining plume morphology with the flow state to forecast smoke transport under an unchanged source and an open upper exhaust route. The representative GPT-5.6-Luna response predicts continued plume rise and flow-guided lateral transport, linking buoyant ascent, ambient advection, and geometric confinement in a single coupled-field explanation.

P4.Causal Mechanism Attribution:

Multiple Choice:

Question: The charging-bay plume rises and then bends left along the service corridor, where the u-velocity field is persistently leftward. What mechanism best explains the deflection?

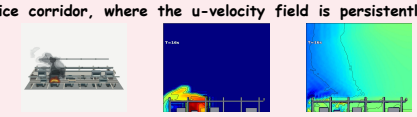
Options:

A. Corridor crossflow advects the buoyant plume leftward

B. Buoyancy alone redirects the plume horizontally without crossflow

C. Radiative heating transports soot laterally along the corridor

D. A secondary local source produces the observed leftward extension



GPT-5.6 Luna: A — The plume first rises through buoyancy and is then deflected leftward by the persistent corridor crossflow.

Open Report:

Question: Describe the smoke direction, identify the dominant mechanism causing its lateral deflection, and distinguish this mechanism from the initial vertical rise.

Correct Answer:


```
{
  "conclusion": "The plume initially rises by buoyancy but is deflected leftward mainly by longitudinal corridor crossflow acting on the soot-bearing gases.",
  "prediction": {
    "dominant_mechanism": "corridor_crossflow",
    "smoke_direction": "upward_then_leftward"
  },
  "evidence": "The soot-density field extends upward and leftward from the source, while the u-velocity field shows persistent leftward flow across the same service-corridor region.",
  "mechanism": "Fire heating produces buoyant vertical ascent, whereas the corridor flow supplies horizontal advection that bends and transports the plume leftward; surrounding geometry constrains the resulting smoke pathway.",
  "confidence": 0.90
}
```


GPT-5.6 Luna:


```
{
  "conclusion": "The lateral plume deflection is mainly caused by longitudinal corridor crossflow rather than buoyancy alone.",
  "prediction": {
    "dominant_mechanism": "corridor_crossflow",
    "smoke_direction": "upward_then_leftward"
  },
  "evidence": "The soot plume extends upward and then leftward, while the u-velocity field shows sustained leftward motion through the same service-corridor region.",
  "mechanism": "Buoyancy explains the initial vertical rise, whereas horizontal momentum from the corridor airflow advects the plume leftward and determines its sustained lateral displacement.",
  "confidence": 0.89
}
```


Figure 12: **P4: Causal Mechanism Attribution.** A charging-bay fire produces a buoyant smoke plume that subsequently bends leftward along a service corridor under a persistent leftward airflow. The Gold response attributes the lateral deflection primarily to corridor crossflow, while distinguishing this mechanism from the initial buoyant rise. The representative GPT-5.6-Luna response reaches the same causal attribution, linking the upward plume development to buoyancy and the sustained lateral transport to corridor crossflow.

P5.Counterfactual Intervention Reasoning:

Multiple Choice:

Question: Two refinery-fire runs are identical until t=20 s. The counterfactual then isolates the leak-feeding pump while all other conditions remain unchanged. What later difference is expected?

Options:

A. The isolated case develops a weaker plume and lower nearby equipment exposure

B. Both cases retain similar plumes because ignition has already occurred

C. Isolation redirects the plume while preserving the same source strength

D. Isolation increases later heat release despite stopping continued fuel delivery

GPT-5.6 Luna: A — Isolating the leak-feeding pump reduces continued fuel supply, producing a weaker later plume and lower exposure of nearby equipment.

Open Report:

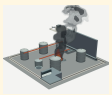
Question: Compare the continuing-feed case with the isolated case and decompose the expected effects on source persistence, plume extent, and nearby equipment exposure.

Correct Answer:

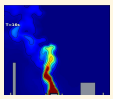
{ "conclusion": "The isolated case develops a less persistent source, a smaller plume, and lower exposure of nearby equipment.",
"prediction": {
"source_persistence": "isolated_lower",
"plume_extent": "isolated_smaller",
"equipment_exposure": "isolated_lower" },
"evidence": "Before isolation, both runs share the same geometry, ignition state, and plume development; afterward, only continued pump-fed fuel delivery differs, while ventilation, equipment layout, and other boundary conditions remain unchanged.",
"mechanism": "Stopping continued fuel delivery weakens the sustained heat-release source, reducing buoyant plume generation and soot production; the smaller thermal and soot plume consequently lowers heating and contamination exposure to adjacent equipment.",
"confidence": 0.90 }

GPT-5.6 Luna:

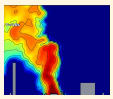
{
"conclusion": "The isolated case develops a weaker and less persistent fire, with reduced plume extent and equipment exposure.",
"prediction": {
"source_persistence": "isolated_lower",
"plume_extent": "isolated_smaller",
"equipment_exposure": "isolated_lower"
},
"evidence": "The two runs share the same pre-intervention state and differ only in continued pump-fed fuel delivery after t=20 s.",
"mechanism": "Pump isolation removes the sustained fuel supply, lowering subsequent heat release and soot production and therefore reducing plume development and nearby thermal exposure.",
"confidence": 0.91
}



3D Modal



Temperature



Soot Density

Figure 13: **P5: Counterfactual Intervention Reasoning.** A refinery pipe-rack fire is evaluated under the counterfactual intervention of earlier pump isolation. The task requires decomposing the intervention effect into changes in source strength, plume spread, and equipment exposure. Using the 3D scene, temperature field, and soot-density field, the representative GPT-5.6-Luna response predicts that earlier isolation would reduce continued fuel delivery, suppress subsequent plume development, and lower the thermal and contamination risks to adjacent equipment.

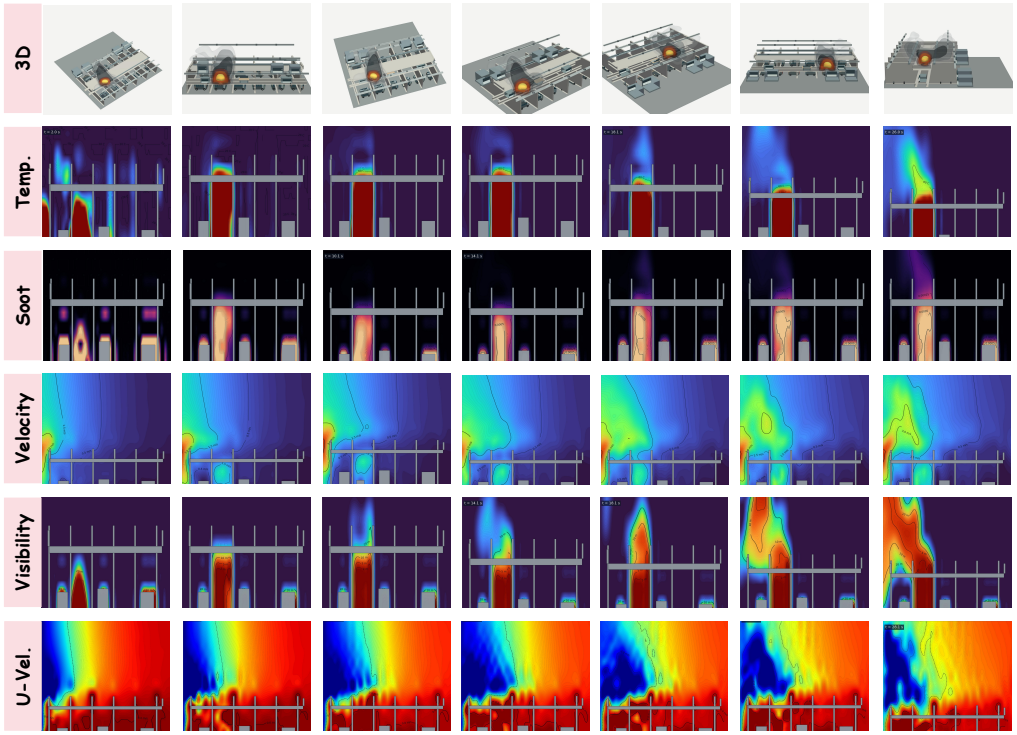


Figure 14: **Transportation and Infrastructure: Electric Bus Terminal Charging-bay Fire.** This controlled fire world shows the time-resolved development of a charging-bay fire through co-registered 3D scenes and temperature, soot-density, velocity, visibility, and streamwise-velocity fields. The terminal geometry makes the interaction between buoyant plume rise, bay-scale confinement, and longitudinal smoke transport directly observable.

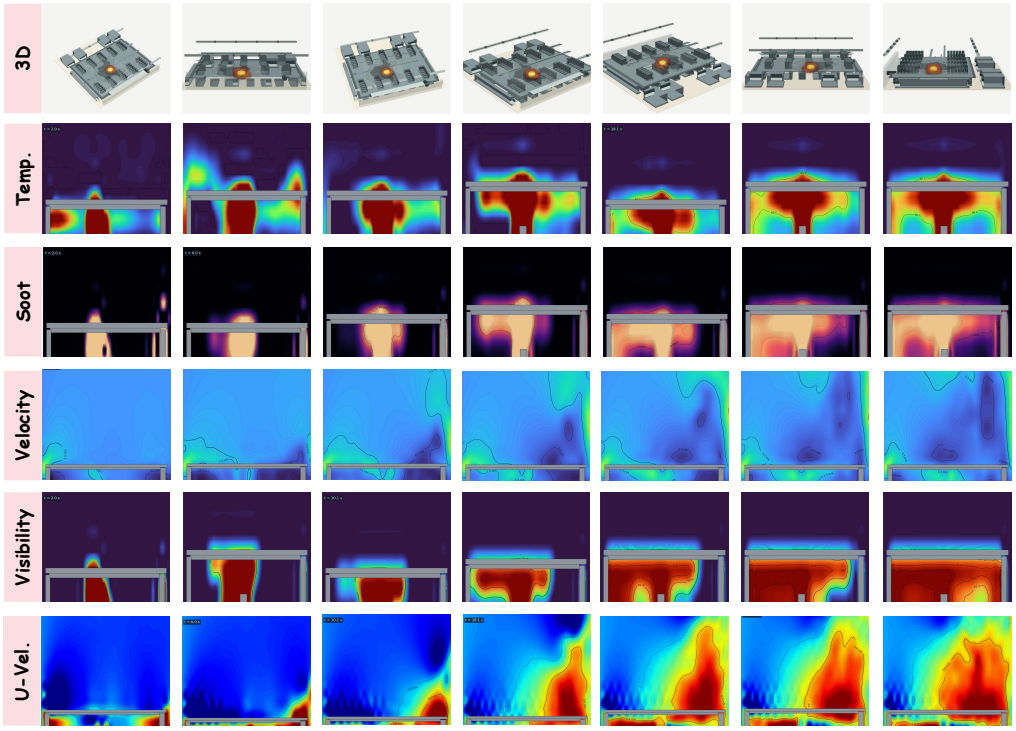


Figure 15: **Office and Digital Facilities: Data-centre Cold-aisle Cable-tray Fire.** This controlled sequence depicts the growth and transport of heat and smoke from a cable-tray fire within a confined cold-aisle layout. The co-registered fields reveal how overhead structures and aisle geometry constrain the thermal plume, redistribute soot, modify local flow, and progressively degrade visibility.

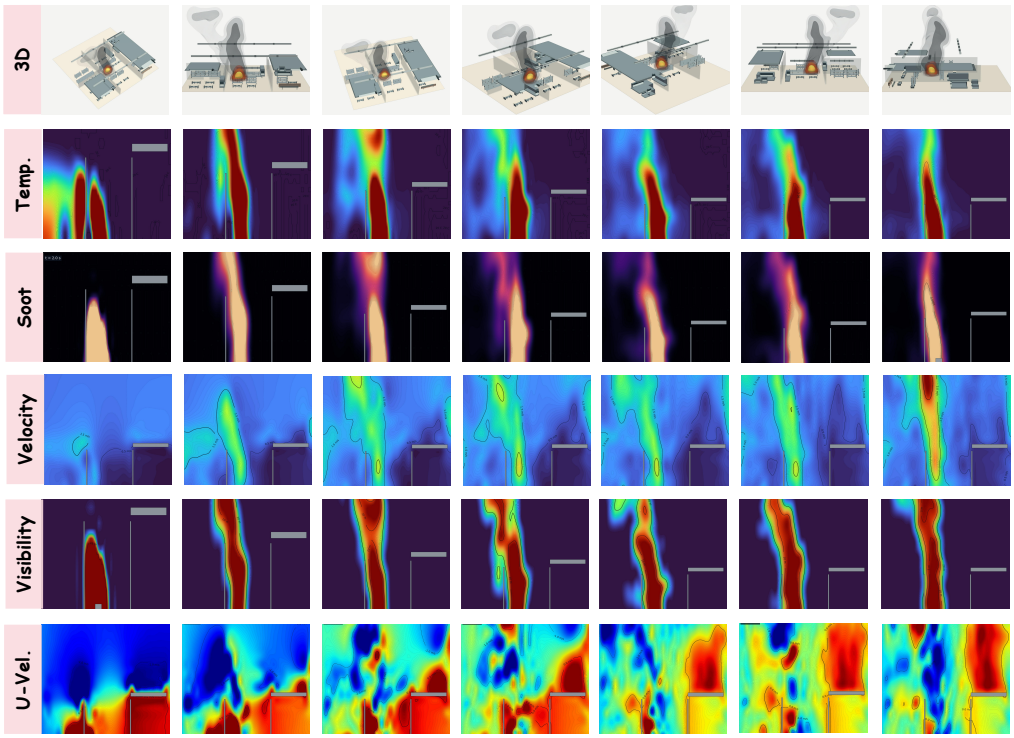


Figure 16: **Healthcare and Education: Teaching Medical Laboratory Solvent-bench Fire.** A localized solvent-bench source develops into a vertically extended thermal and soot plume inside a laboratory compartment. The synchronized physical fields expose the resulting changes in plume velocity, streamwise transport, and visibility, while the 3D views preserve the spatial relationship between the source, laboratory fixtures, and surrounding enclosure.

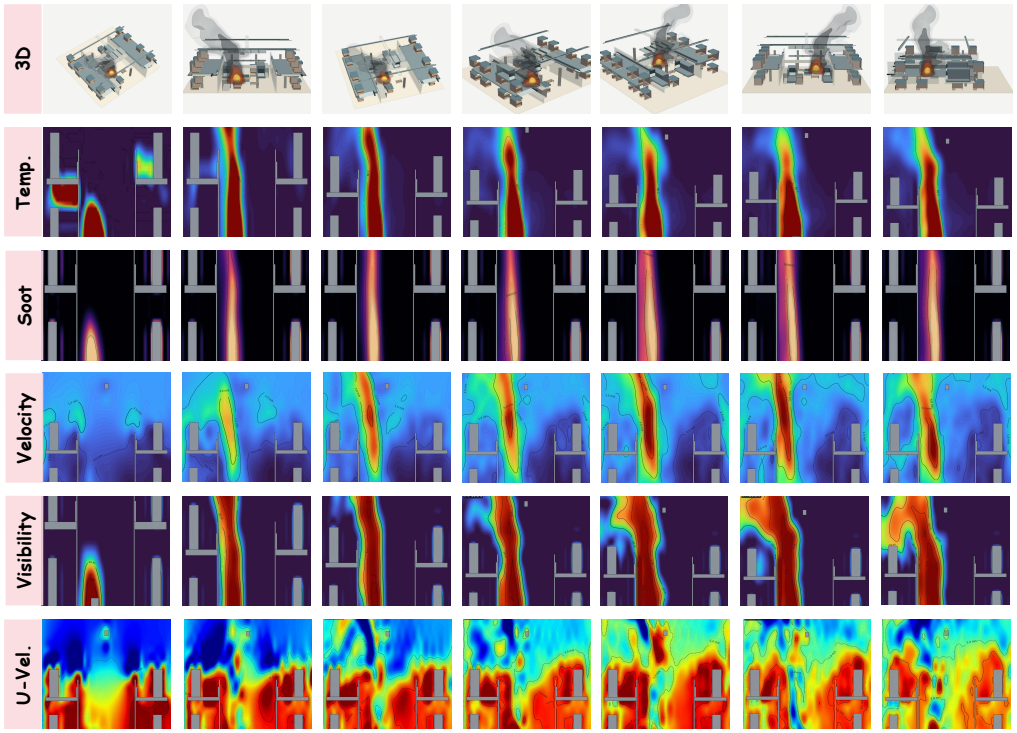


Figure 17: **Commercial and Public Spaces: Public Atrium Kiosk Fire.** This controlled fire world captures plume growth from a kiosk source within a tall, structurally open atrium. The temporal sequence illustrates how buoyancy, upper-level geometry, and the surrounding flow field jointly determine plume rise, lateral smoke spreading, and the spatial progression of visibility loss.

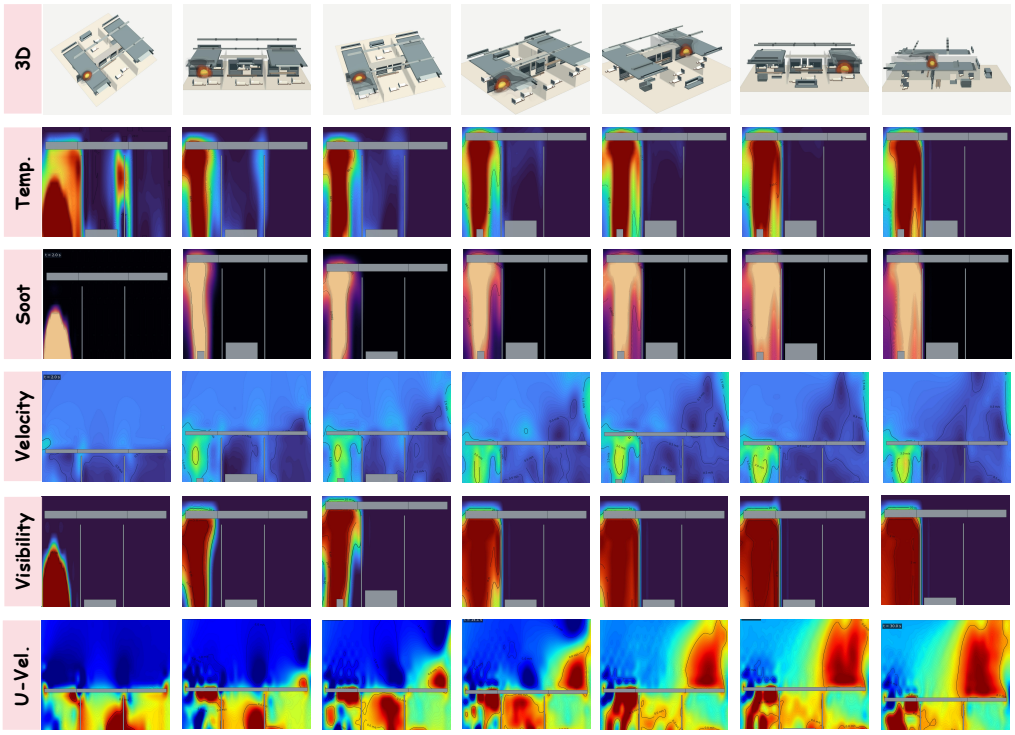


Figure 18: **Residential and Care Settings: Night-shift Care Ward Linen-cart Fire.** A corridor-adjacent linen-cart fire is represented through its evolving 3D scene and five co-registered physical fields. The sequence highlights the localized origin of the thermal and soot anomalies, their subsequent extension through the connected ward geometry, and the associated deterioration of visibility along the transport path.

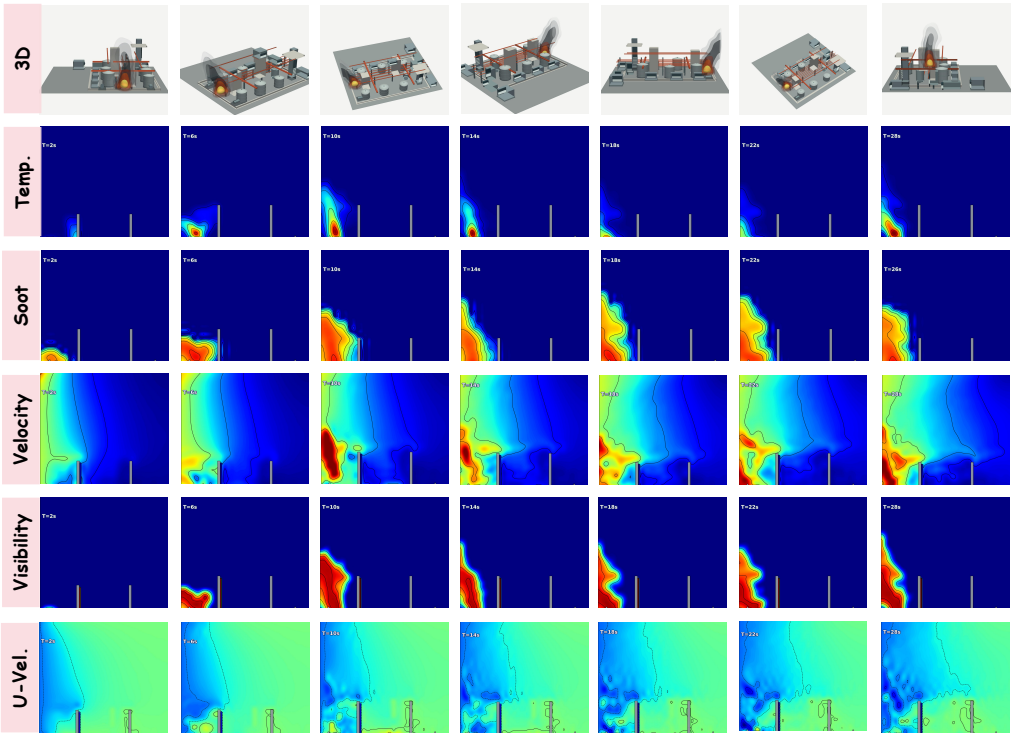


Figure 19: **Industry, Energy, and Logistics: Integrated Refinery Pipe-rack Energy Fire.** This controlled outdoor fire world places the source among pipe racks, processing units, and adjacent equipment. The multimodal sequence exposes the growth of the buoyant plume, the transport of heat and soot through the equipment layout, and the spatially varying velocity and visibility conditions that determine downstream exposure.

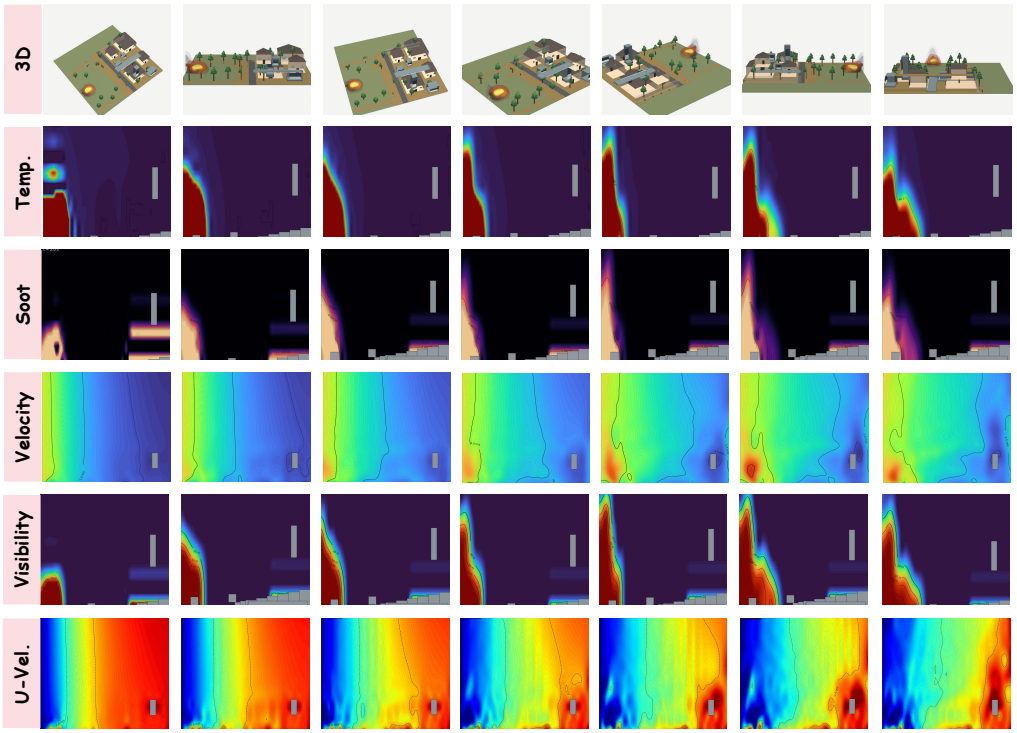


Figure 20: **Wildland and Wildland–Urban Interface: Hillside WUI Evacuation-lane Fire.** This controlled scene represents fire development near vegetation, structures, and a hillside evacuation route. The co-registered field sequence illustrates how terrain, built obstacles, and ambient flow shape the propagation of heat and smoke, producing spatially heterogeneous visibility and transport conditions along the wildland–urban interface.

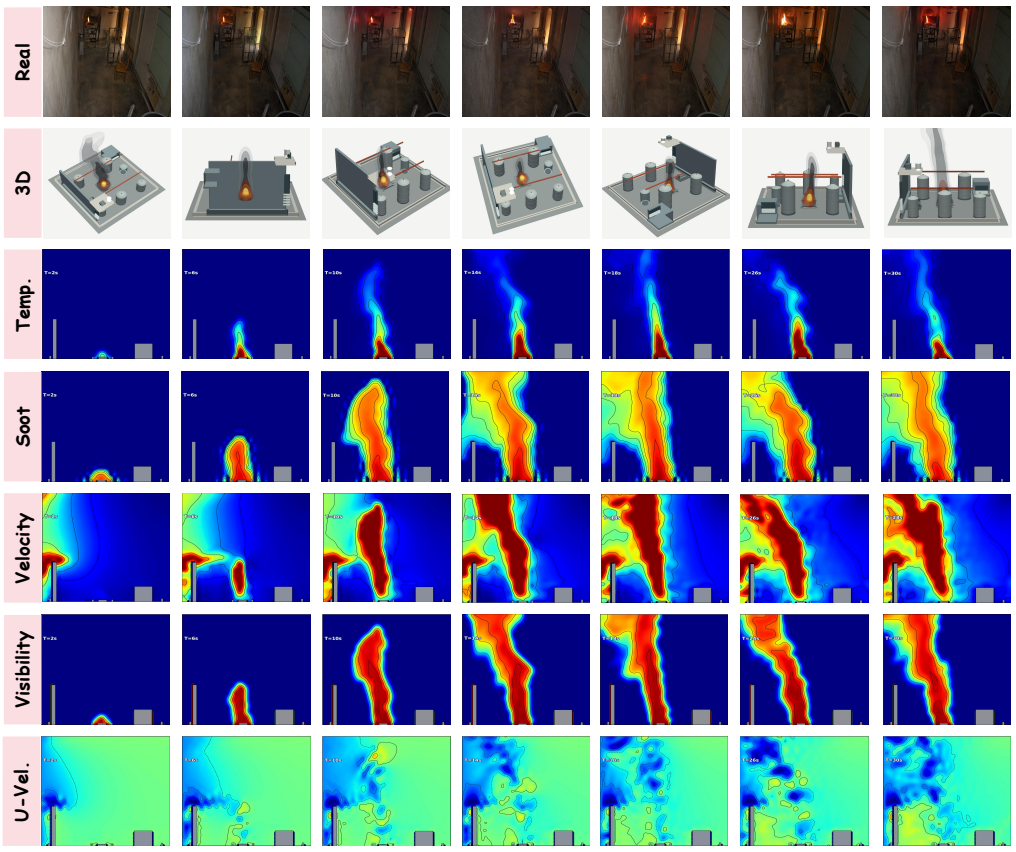


Figure 21: **Real-world-aligned event: Heptane Pool-fire Experiment.** Unlike the preceding controlled worlds, this example begins with an observed experimental sequence. Recorded video frames anchor the event chronology and visible fire development, while the aligned 3D reconstruction and FDS realization provide simulation-completed temperature, soot-density, velocity, visibility, and streamwise-velocity fields at matched time points. The separate rows preserve the distinction between observed evidence and reconstructed latent physical state.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, code implementation and debugging, and the detection of potentially fabricated references. We also used AI tools to search for prior work related to our research ideas to reduce the risk of overlooking similar studies. No AI-generated text was directly incorporated into the manuscript. All AI-assisted work was reviewed by the authors, and AI-generated code was verified and tested for correctness. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.