# VISION THROUGH FIELDS: LEARNING REAL-SIM-REAL REASONING FOR MLLM UNDERSTANDING OF MULTIPHYSICS COMBUSTION DYNAMICS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal large language models (MLLMs) have made notable progress in image and video understanding, yet their ability to reason about physical processes remains limited. To systematically study this problem, we design a comprehensive multiphysics physical reasoning benchmark grounded in complex engineering scenarios, covering combustion dynamics across diverse environments. The benchmark spans a hierarchy from intuitive physical perception to engineering level reasoning. Evaluations on this benchmark reveal a substantial gap between visual perception and physical state understanding in current MLLMs. To bridge this gap, we propose Vision Through Fields (VTF), a framework for learning real-sim-real physical reasoning, where MLLMs first ground visual observations using aligned simulation fields provided at observed timestamps, reason over field states, field evolution, and cross field relations, and finally transfer the resulting understanding back to visual decisions. To make this reasoning process learnable, we decompose it into real-to-sim grounding, simulation field reasoning, and sim-to-real transfer, and train these capabilities with structured supervision derived from physical fields. This learned real-sim-real reasoning raises mean benchmark accuracy from 42.50% to 54.85% with an MLLM backbone. More broadly, our results highlight the value of simulation derived physical fields as reusable supervision for field guided MLLM physical reasoning. All datasets and code will be made publicly available.

## 1 INTRODUCTION

Multimodal large language models (MLLMs) have made substantial progress in image and video understanding (Bai et al., 2025; Zhu et al., 2025b), yet intuitive physics understanding (IPU) from visual observations remains challenging (Zheng et al., 2024; Chow et al., 2025; Li et al., 2026). Studies of human intuitive physics link scene understanding to inferring latent physical properties and predicting subsequent dynamics, beyond recognizing visible objects and events (Battaglia et al., 2013; Hamrick et al., 2016).

Existing IPU benchmarks assess physical properties, future states, and counterfactual outcomes, often combining perception, language comprehension, and physical reasoning (Zheng et al., 2024; Chow et al., 2025). Beyond Static Vision isolates dynamic perception through next frame selection and temporal coherence judgment (Li et al., 2026). Yet visual prediction need not establish latent property inference: Physion++ tests predictions that specifically require estimating mass, friction, or elasticity from observed dynamics (Tung et al., 2023). ContPhy examines fluid, rope, cloth, and ball systems, while Beyond Static Vision focuses primarily on fluid dynamics, including pouring and stirring (Zheng et al., 2024; Li et al., 2026). These evaluations leave open how IPU extends from localized interactions to large engineering environments with coupled heat release, flow, and species transport across a range of geometries, ventilation settings, and source configurations.

To address these gaps, we study combustion dynamics in engineering environments. A localized fire releases heat and generates buoyant flow, while geometry and ventilation govern the transport of heat and combustion products throughout the surrounding space (McGrattan et al., 2000). Combustion therefore provides a representative setting for examining how models connect local physical states



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.

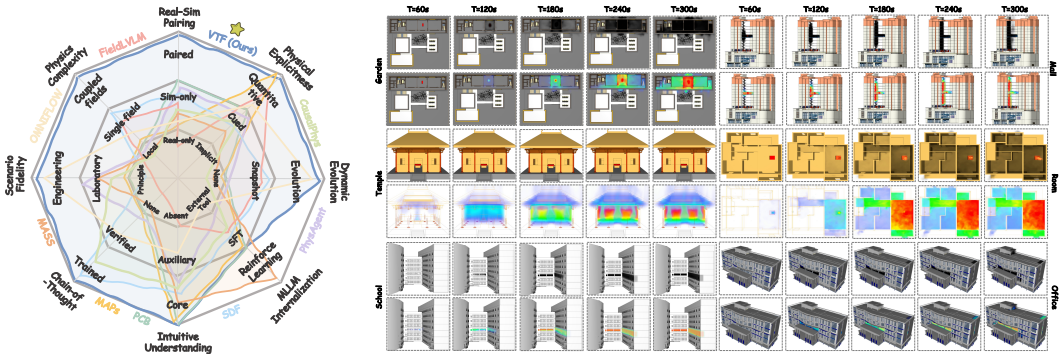



Figure 1: **Left:** Comparison between our benchmark and representative benchmarks and methods for intuitive physics understanding (IPU) in MLLMs across key evaluation dimensions. **Right:** Overview of the combustion scenarios evaluated in our benchmark.

to evolving environmental conditions. Our benchmark (Figure 1) spans seven scenarios and 228 simulation cases, with 67,716 paired scene and physical field samples. Its six tasks distinguish three levels: temporal consistency, physical state perception, and physical environment assessment. These evaluate whether models can recognize coherent evolution, identify fire and ventilation conditions, and assess smoke control outcomes. Evaluations reveal a gap across these levels: the strongest overall baseline, Claude Opus 5.5, achieves 64.15% on temporal coherence judgment but only 51.88% on environmental judgment, close to the latter task's 50% chance level.

Scientific simulation offers a rich resource for developing these capabilities. Established well before contemporary LLMs, it has accumulated extensive records of physical processes, including spatially and temporally resolved temperature, gas concentration, and velocity fields (Li et al., 2008; Ohana et al., 2024). These fields encode numerical solutions to governing physical equations under specified geometry and boundary conditions, exposing states and interactions that are difficult to infer from appearance alone. Velocity fields expose flow direction and magnitude, while temperature and concentration fields describe the distribution of heat and combustion products. Together, these quantities provide physical context for interpreting the movement and accumulation of visible smoke. Although simulation data already support scientific analysis and machine learning, their potential as reusable supervision for connecting visual observations to coupled physical dynamics in MLLM IPU remains underexplored. This raises a question: how can accumulated scientific simulation resources be repurposed to teach MLLMs to interpret physical processes from visual observations (Chen et al., 2026)?

We propose Vision Through Fields (VTF), a framework for learning real-sim-real (RSR) reasoning with simulation derived physical fields. VTF grounds observations in field evidence, reasons about field evolution and relationships, and uses this evidence to answer the original visual question. Precisely aligned real recordings and simulation fields are scarce; we therefore use scene renderings and fields produced from the same simulations to obtain consistent training pairs. Sharing geometry, timestamps, and underlying states ties each rendered observation to its physical context, avoiding the need to align independently recorded videos with a separately constructed simulation. We organize these pairs into four tasks that train scene to field grounding, field evolution prediction, coherence judgment, and transfer back to visual decisions. These tasks encourage the model to connect visual observations to field states and their evolution, making the fields intermediate evidence for the final decision. At inference, VTF receives visual observations together with corresponding field visualizations at the observed timestamps. No future field states are supplied.

Experiments show complementary gains from field prompting and fine tuning. Across six unseen scenarios, VTF improves mean accuracy from 41.88% under zero-shot inference to 53.79%, while shuffling field order reduces performance. We also apply the checkpoint trained on tunnel cases directly to real camera images from the NIST Station Nightclub experiment (Madrzykowski et al., 2006; Grosshandler et al., 2005), without real data fine tuning. With auxiliary simulation fields, VTF improves accuracy from 38.26% to 49.04%, providing evidence of transfer from rendered training observations to real imagery.

Our contributions are threefold:

- We introduce an engineering combustion benchmark that distinguishes three levels of IPU, revealing limitations beyond recognizing plausible temporal evolution.
- We propose VTF, which converts paired scene renderings and physical fields into structured supervision for grounding observations, reasoning over fields, and answering visual questions.
- We demonstrate generalization to unseen scenarios and real fire imagery with auxiliary simulation fields, and examine how field alignment and rendering affect performance.

## 2 Related Work

**Physics Augmentation.** Physical knowledge enters MLLMs through external evidence or learned structure. Physics Context Builders translate simulations into physical descriptions, while MAPS and PhysAgent combine foundation models with simulation traces or specialist vision modules Balazadeh et al. (2025); Zhu et al. (2025a); Chow et al. (2025). MASS, Scene Dynamic Field, and CausalPhys internalize depth, motion, field cues, or causal graphs through targeted tuning Wu et al. (2026b); Li et al. (2026); Tang et al. (2026). Most of these methods represent physics as text, auxiliary visual cues, or abstract structure.

**Video and Multimodal Reasoning.** Video reasoning studies latent prediction, adaptive sampling, visualized reasoning, and reinforcement learning Bardes et al. (2024); Buch et al. (2025); Li et al. (2025); Feng et al. (2025); Tan et al. (2025). Methods address long contexts through modular fusion, object cues, reflection, memory, and extended training, while VTimeCoT encodes temporal position explicitly Yu et al. (2025b); Ranasinghe et al. (2025); Song et al. (2025); Diko et al. (2025); Lin et al. (2025); Zhang et al. (2025a). Broad benchmarks evaluate expert knowledge, aggregation, view transfer, online perception, and narrative reasoning Zhao et al. (2025); Peng et al. (2025); He et al. (2025); Xun et al. (2025); Yu et al. (2025a), but also show that more frames or longer reasoning do not ensure grounded inference Saravanan et al. (2025); Yi et al. (2025); Jung et al. (2025); Ye et al. (2025); Guo et al. (2025); Xu et al. (2025); Luo et al. (2025).

**Explicit Physical Field Reasoning.** FieldLVLM compresses flow attributes into structured descriptions, while OMNIFLOW maps flow tensors to topological language constrained by physical laws Zhang et al. (2025b); Wu et al. (2026a). Both demonstrate the value of explicit fields but focus on scientific field data. VTF instead aligns visual observations with evolving simulation fields, learns their relations, and transfers field regularities to visual combustion reasoning.

## 3 BENCHMARK CONSTRUCTION

### 3.1 Motivation

Most video understanding benchmarks emphasize actions, events, and semantic relationships, making IPU difficult to isolate. Recent IPU work has used next frame selection and temporal coherence verification (Li et al., 2026). These tasks probe temporal consistency, but not whether models infer the physical states behind visible motion or assess its effects on the surrounding environment. This distinction matters for combustion, where heat transfer, species transport, and flow couple local states to environmental outcomes. Such geometry dependent, engineering scale behavior remains underrepresented in current IPU benchmarks. We therefore evaluate temporal consistency (I), physical state perception (P), and physical environment assessment (E) separately.

Meanwhile, scientific simulations describe these coupled processes through spatially and temporally resolved fields, providing reusable evaluation resources (Li et al., 2008; Ohana et al., 2024). Yet existing datasets are primarily designed for numerical modeling and field prediction; paired scene videos, aligned physical fields, and diagnostic IPU tasks remain scarce. We therefore construct paired scene renderings and simulation fields from representative engineering fire scenarios, with questions and answers grounded in their shared states. This benchmark connects simulation resources to layered IPU assessments.

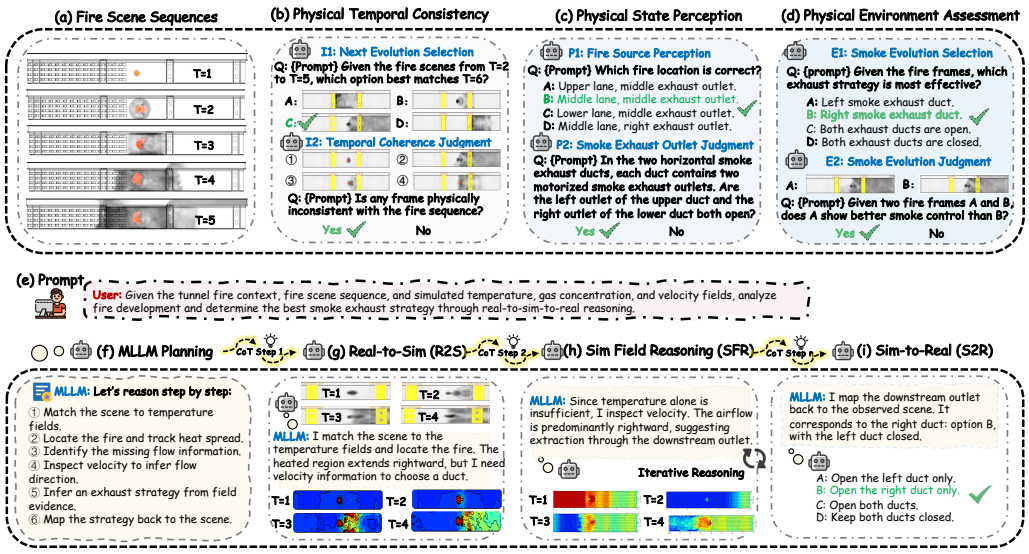



Figure 2: Overview of benchmark tasks and VTF reasoning. Panels (a)–(d) show scene sequences and the three evaluation levels. Panels (e)–(i) illustrate field guided RSR reasoning: temperature fields localize the fire, velocity fields reveal flow direction, and their combined evidence guides exhaust strategy selection and transfer back to the scene.

### 3.2 Evaluation System

Figure 2 organizes IPU into three levels: temporal consistency (I), physical state perception (P), and physical environment assessment (E). Each level includes selection among four candidates and binary judgment, denoted by the suffixes 1 and 2, respectively.

**Evaluation Protocols.** Throughout, $k$ identifies a task, $i$ indexes its samples, and $N_k$ denotes its sample count. For sample $i$, $\mathcal{C}_i^k$ contains the visual evidence, question, and candidate references where applicable. For model parameters $\theta$, $p_\theta(a \mid \mathcal{C}_i^k)$ denotes the probability of answer $a$. For selection tasks $k \in \{\mathrm{I1}, \mathrm{P1}, \mathrm{E1}\}$, $A_i^k$ contains one correct answer $a_i^{k,+}$ and three distractors. Assuming a unique highest scoring answer, selection accuracy is

$$\mathrm{Acc}_k^{\mathrm{sel}} = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{I}\left[ p_\theta(a_i^{k,+} \mid \mathcal{C}_i^k) > \max_{a \in A_i^k \setminus \{a_i^{k,+}\}} p_\theta(a \mid \mathcal{C}_i^k) \right]. \tag{1}$$

For judgment tasks $k \in \{\mathrm{I2}, \mathrm{P2}, \mathrm{E2}\}$, the model returns YES or NO. Let $a_i^{k,+} \in \{\mathrm{YES}, \mathrm{NO}\}$ denote the correct label. Judgment accuracy is

$$\mathrm{Acc}_k^{\mathrm{jud}} = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{I}\left[ \arg\max_{a \in \{\mathrm{YES}, \mathrm{NO}\}} p_\theta(a \mid \mathcal{C}_i^k) = a_i^{k,+} \right]. \tag{2}$$

Here, $\mathbb{I}[\cdot]$ is the indicator function. The task level expressions follow Li et al. (2026); in implementation, correctness is determined from generated answers using the same parsing rule for all models. We report accuracy separately for each task.

**Level I: Temporal Consistency.** This level tests whether a model can distinguish plausible temporal progression from disrupted evolution. In Next Evolution Selection (I1), the model selects the correct continuation of an observed clip from four candidate frames. Temporal Coherence Judgment (I2) presents either an intact clip or one with two randomly swapped frames; the model answers YES if the clip is inconsistent and NO otherwise. Together, these tasks test anticipation of subsequent dynamics and sensitivity to temporal order.

**Level P: Physical State Perception.** This level asks models to infer the physical configuration behind observed flame and smoke behavior. In Physical Perception Selection (P1), the model selects the matching configuration from four candidates. Smoke Exhaust Outlet Judgment (P2) asks

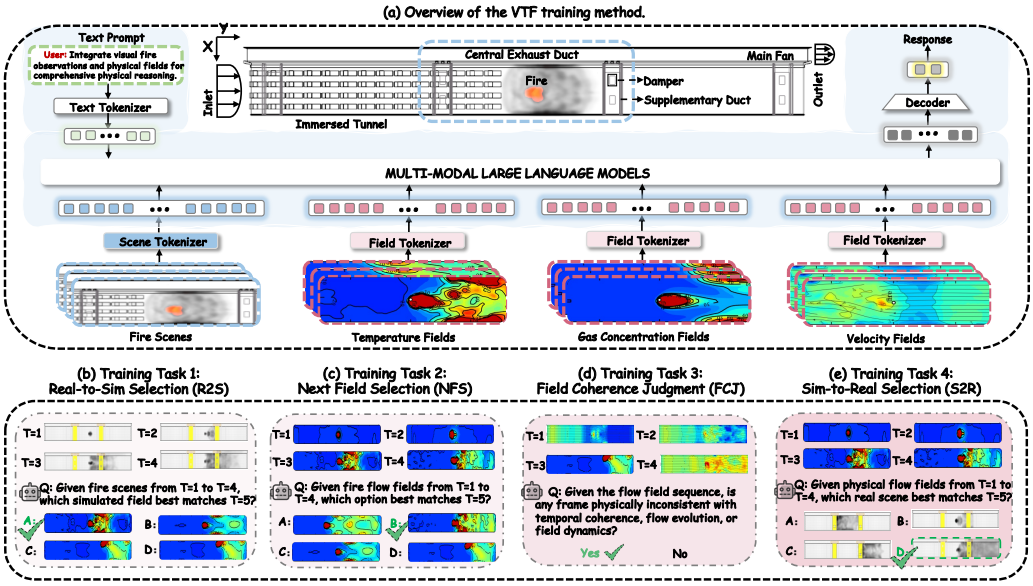



Figure 3: Overview of the VTF training framework and training tasks. Panel (a) shows the overall VTF training framework, where training is conditioned on reasoning paths and field relations. Panels (b)–(e) show the four VTF training tasks: real to sim selection, next field selection, field coherence selection, and sim to real selection.

whether the exhaust outlets specified in the question are open; the model answers YES if they are and NO otherwise. Questions concern fire source locations and exhaust outlet states. Both tasks require locating the relevant physical objects and inferring their states or spatial configuration from observed dynamics.

**Level E: Physical Environment Assessment.** This level asks models to evaluate how coupled fire and flow dynamics affect the surrounding environment. In Smoke Evolution Selection (E1), the model selects the most effective smoke exhaust strategy from four candidates based on the observed fire conditions. In Smoke Evolution Judgment (E2), it compares two clips with similar fire conditions and answers YES if clip A better satisfies the criterion than clip B, and NO otherwise. For simulated cases, the criterion is the fraction of 3D temperature monitoring points above $50\,^{\circ}\mathrm{C}$ at steady state; lower values indicate less thermal impact. Unlike Level P, these tasks require assessing the environmental consequences of physical states rather than identifying the states alone.

### 3.3 Dataset Construction

We use high fidelity videos rendered by simulation postprocessors as proxies for real observations for two reasons. First, modeling errors create discrepancies between recorded scenes and their simulated counterparts, making precise alignment between real videos and simulation fields difficult. Rendered videos share the same model and simulation states as the fields, avoiding this mismatch. Second, postprocessing tools in mature simulation software can generate these videos directly from simulation outputs, making paired data easier to obtain and the construction procedure easier to apply across scenarios.

**Scenarios and Simulation Cases.** We simulate 228 cases across 7 fire scenarios: 300 s per tunnel case and over 300 s per case in the other scenarios. All domain dimensions below follow $L \times W \times H$. The immersed tunnel contributes 222 cases covering 32 smoke exhaust outlet configurations, 6 fire locations, and heat release rates (HRRs) ranging from 10 to 50 MW, with specific combinations detailed in Appendix C. The reference configuration has a $200 \times 24 \times 8$ m domain bounding box, 341,760 cells, and transverse extraction ($3.44\,\mathrm{m/s}$ inlet; $250\,\mathrm{m^3/s}$ exhaust activated at 20 s). The other six scenarios are reserved for testing: garden (174,720 cells, $52 \times 41 \times 10$ m), shopping mall (169,360 cells, $46 \times 23 \times 58$ m), office (177,660 cells, $54 \times 28 \times 25$ m), residential room

Table 1: Zero-shot benchmark accuracy of existing MLLMs across six tasks. Bold and underlined values mark the best and second highest baseline scores in each column, respectively.

| Setting | Model | Param | Temporal Acc. | | State Acc. | | Environment Acc. | | Avg. (%) |
|---|---|---|---|---|---|---|---|---|---|
| | | | I1 (%) | I2 (%) | P1 (%) | P2 (%) | E1 (%) | E2 (%) | |
| Open Source | InternVL2.5 | 8B | 27.12 | 50.84 | 26.38 | 46.27 | 25.41 | 41.64 | 36.28 |
| | InternVL3.5 | 8B | 30.71 | 57.69 | 29.70 | 50.97 | 25.45 | 45.94 | 40.08 |
| | GLM-4.1V | 9B | 29.05 | 54.37 | 27.84 | 49.92 | 26.51 | 44.68 | 38.73 |
| | GLM-4.6V-Flash | 9B | 33.28 | 54.04 | 30.85 | 55.64 | 29.33 | 49.63 | 42.13 |
| | mPLUG-Owl3 | 7B | 31.02 | 58.36 | 29.88 | 52.47 | 27.94 | 46.79 | 41.08 |
| | Qwen2-VL | 7B | 30.14 | 56.42 | 28.73 | 50.31 | 27.02 | 45.83 | 39.74 |
| | Qwen2.5-VL | 7B | 32.32 | 60.20 | 30.58 | 54.35 | 29.13 | 48.41 | 42.50 |
| Closed Source | GPT-5.5 | – | **33.68** | <u>63.74</u> | <u>33.29</u> | <u>57.92</u> | <u>30.42</u> | 51.43 | <u>45.08</u> |
| | Gemini-3.5 | – | 33.49 | 62.41 | 32.67 | 57.23 | 29.28 | <u>51.44</u> | 44.42 |
| | Claude Opus 5.5 | – | <u>33.61</u> | **64.15** | **33.68** | **59.49** | **30.65** | **51.88** | **45.58** |

(594,000 cells, $11 \times 12 \times 4.5$ m), school (328,812 cells, $74 \times 46 \times 33$ m), and temple (112,360 cells, $16 \times 16 \times 12$ m). Each case provides temperature, gas concentration (CO), and velocity fields.

**Dataset Pipeline and Tasks.** Each clip contains synchronized scene and field frames at physical times $b$, $b + s$, $b + 2s$, and $b + 3s$. The temporal stride $s$ is measured in seconds (default 1 s); prediction tasks target $b + (3 + h)s$ (default $h = 1$). Zero-shot and SFT evaluation use scene frames only. Level I tests temporal consistency through next frame selection (I1) and judgment of clips with two randomly swapped frames (I2). Level P uses simulation settings to label configuration selection (P1) and the opening states of specified exhaust outlets (P2). For Level E, each original video carries a case level label based on the steady state fraction of temperature monitoring points above $50\,^{\circ}\mathrm{C}$. CoT and full VTF add the three aligned fields at observed timestamps, using the fixed field guided prompt in Figure 2. R2S, NFS, FCJ, and S2R supervise field grounding, evolution, coherence, and transfer to visual decisions. Across seven scenarios, we obtain 67,716 paired scene and field clips. Each 300 s tunnel case yields 297 clips with next frame targets, totaling 65,934 across 222 cases, split by case approximately 7:2:1 for training, validation, and test. The other six scenarios contribute 1,782 clips, all reserved for test.

# 4 Vision Through Fields

## 4.1 Multiphysics Field Guided RSR Reasoning

VTF uses multiphysics simulation fields as intermediate evidence for physical reasoning. Given visual observations $\mathcal{X}^{\mathrm{r}}$, aligned field visualizations $\mathcal{F}^{\mathrm{s}}_{\mathrm{obs}}$, and a question $q$, it grounds observations in fields, reasons about field evolution, and answers $q$:

$$(\mathcal{X}^{\mathrm{r}}, \mathcal{F}^{\mathrm{s}}_{\mathrm{obs}}, q) \xrightarrow{\mathcal{R}_{\mathrm{r2s},\phi}} \mathcal{Z}^{\mathrm{s}} \xrightarrow{\mathcal{R}_{\mathrm{s2s},\phi}} \mathcal{H}^{\mathrm{s}} \xrightarrow{\mathcal{R}_{\mathrm{s2r},q}} \hat{a}. \tag{3}$$

Here, $\mathcal{Z}^{\mathrm{s}}$ denotes the field representation associated with the observations, $\mathcal{H}^{\mathrm{s}}$ the resulting field reasoning state, and $\hat{a}$ the predicted answer. The operators $\mathcal{R}$ denote prompted reasoning stages within the MLLM; $\phi$ specifies the physical quantity or relation of interest. Superscripts $\mathrm{r}$ and $\mathrm{s}$ identify the visual and simulation domains. Figure 2 illustrates this process. The real domain uses scene renderings as proxies for recorded observations.

**Evolution of Field Representations.** The fields used by VTF originate from a shared simulation state. Let $\boldsymbol{U} = (\rho, \rho\boldsymbol{u}^{\mathrm{T}}, \rho E, \rho\boldsymbol{Y}^{\mathrm{T}})^{\mathrm{T}}$ collect mass, momentum, total energy, and species densities. Their normal flux separates advective transport, pressure contributions, and viscous, thermal, and species transport:

$$\boldsymbol{F}_n = u_n \begin{bmatrix} \rho \\ \rho\boldsymbol{u} \\ \rho E \\ \rho\boldsymbol{Y} \end{bmatrix} + \begin{bmatrix} 0 \\ p\boldsymbol{n} \\ pu_n \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} 0 \\ -\boldsymbol{\tau}\boldsymbol{n} \\ (\boldsymbol{q} - \boldsymbol{\tau}\boldsymbol{u}) \cdot \boldsymbol{n} \\ \boldsymbol{J}\boldsymbol{n} \end{bmatrix}. \tag{4}$$

Here, $\boldsymbol{n}$ is the outward unit normal and $u_n = \boldsymbol{u} \cdot \boldsymbol{n}$. Density, velocity, pressure, specific total energy, and species mass fractions are denoted by $\rho$, $\boldsymbol{u}$, $p$, $E$, and $\boldsymbol{Y}$. The terms $\boldsymbol{\tau}$, $\boldsymbol{q}$, and $\boldsymbol{J}$ represent viscous stress, nonadvective energy flux, and species diffusion, respectively.

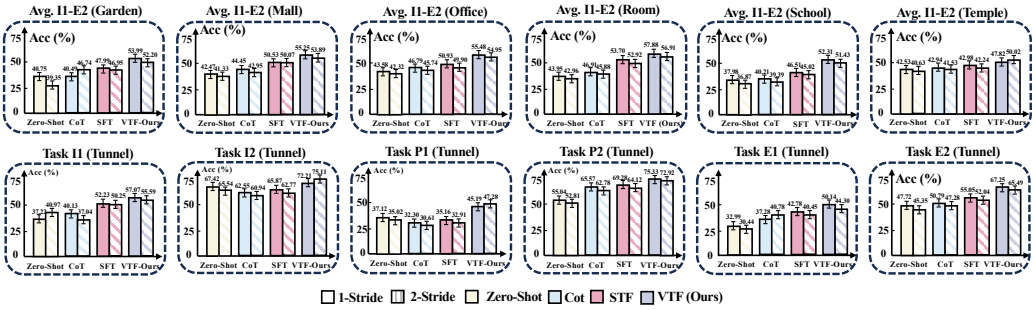



Figure 4: Qwen2.5-VL-7B accuracy; stride 2 uses sparser observations than default stride 1.

The transformation $\boldsymbol{\Psi}$ extracts temperature, gas concentration, and velocity. For image location $\boldsymbol{p}$, the kernel $\boldsymbol{W}(\boldsymbol{p}, \boldsymbol{x})$ weights spatial positions $\boldsymbol{x}$ in the simulation domain $\Omega$. At physical time $t$, the sampled representation and weighted Jacobian are

$$\boldsymbol{g}(\boldsymbol{p}, t) = \int_{\Omega} \boldsymbol{W}(\boldsymbol{p}, \boldsymbol{x}) \boldsymbol{\Psi}(\boldsymbol{U}(\boldsymbol{x}, t)) \, \mathrm{d}V, \qquad \boldsymbol{B} = \boldsymbol{W} \frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{U}}. \tag{5}$$

Let $\boldsymbol{F}$ denote the flux matrix, with $\boldsymbol{F}_n = \boldsymbol{F}\boldsymbol{n}$. The chain rule and integration by parts yield

$$\partial_t \boldsymbol{g} + \int_{\partial\Omega} \boldsymbol{B}\boldsymbol{F}_n \, \mathrm{d}A = \int_{\Omega} \left[ \boldsymbol{B}\boldsymbol{S} + (\nabla \boldsymbol{B}) : \boldsymbol{F} \right] \mathrm{d}V. \tag{6}$$

Here, $\boldsymbol{S}$ collects volume sources, and : denotes contraction over the conserved components and spatial directions. The identity assumes a fixed domain, sampling weights independent of time, and sufficiently smooth fields and variable transformations. Appendix A gives the component definition, derivation, and energy conventions.

Equation (6) relates changes in the sampled fields to boundary transport, physical sources, and a correction induced by spatial sampling and nonlinear variable transformations. It characterizes the physical basis of our representations before color mapping, $\boldsymbol{z} = \boldsymbol{C}(\boldsymbol{g})$, where $\boldsymbol{C}$ maps sampled values to colors. Numerical simulation outputs approximate this continuous description through discretization and model closures, including the low Mach formulation used for fire dynamics (McGrattan et al., 2000).

Scene renderings and field visualizations share the same simulated states, geometry, and boundary conditions. We reuse these paired trajectories to supervise visual grounding, field evolution prediction, field coherence judgment, and transfer to visual decisions. The evolution identity motivates this supervision; it is not imposed as a residual loss or evaluated by a solver during MLLM training. The following section defines the four unit tasks that train these abilities.

### 4.2 Supervised Fine Tuning of VTF

We train VTF through four unit tasks: Real to Sim Selection (R2S), Next Field Selection (NFS), Field Coherence Judgment (FCJ), and Sim to Real Selection (S2R). Figure 3 summarizes these tasks. Each prompt specifies the evidence, reasoning stage, and target physical quantity or relation.

**Task Formulation.** Each training example consists of an input context $\mathcal{C}_i^k$, a candidate answer set $A_i^k$, and a reference answer $a_i^{k,+}$, using the notation in Section 3.2. R2S, NFS, and S2R select one answer from a candidate set containing the reference answer and distractors:

$$\hat{a}_i^k = \underset{a \in A_i^k}{\arg\max} \, p_\theta(a \mid \mathcal{C}_i^k), \qquad k \in \{\mathrm{R2S}, \mathrm{NFS}, \mathrm{S2R}\}. \tag{7}$$

FCJ predicts whether the field sequence is inconsistent, with YES denoting inconsistency and NO denoting coherence:

$$\hat{a}_i^{\mathrm{FCJ}} = \underset{a \in \{\mathrm{YES}, \mathrm{NO}\}}{\arg\max} \, p_\theta(a \mid \mathcal{C}_i^{\mathrm{FCJ}}). \tag{8}$$

These expressions define the task level predictions; the reference answers $a_i^{k,+}$ provide supervision for fine tuning. Performance is measured by the selection and judgment accuracies in Section 3.2.

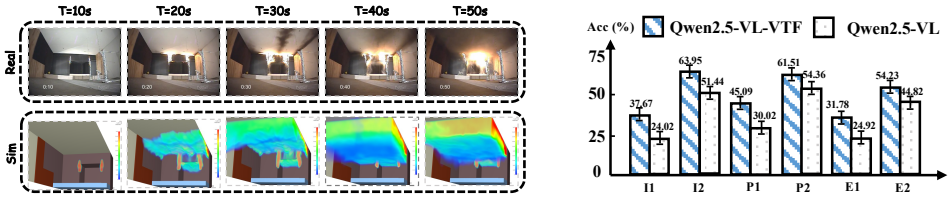



Figure 5: Real fire evaluation: camera images and simulated fields (left), and task accuracy (right).

**Real to Sim Selection (R2S).** R2S trains visual grounding in simulation fields. The input contains visual observations and a query about the relevant physical quantity. The model selects the correct simulation field from candidate fields, linking visible phenomena to underlying physical states.

**Next Field Selection (NFS).** NFS trains field evolution prediction. Given an observed field sequence and a next state query, the model selects the ground truth next field from candidate continuations.

**Field Coherence Judgment (FCJ).** FCJ trains temporal and physical consistency judgment. The input is a field sequence that may contain a temporally inconsistent frame. The correct answer is YES for physical inconsistency and NO for coherent field evolution.

**Sim to Real Selection (S2R).** S2R trains transfer from field reasoning to visual decisions. Given simulation field evidence, the physical reasoning context, and a question about the observed scene, the model selects the answer supported by the fields from candidate real world answers.

Table 2: Parameter scaling. Bold and underlined values mark the best and second best results.

| | | VTF SFT Accuracy (%) | | | | | Benchmark Evaluation Accuracy (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Param** | **Model** | **R2S** | **FCJ** | **NFS** | **S2R** | **Avg.** | **I1** | **I2** | **P1** | **P2** | **E1** | **E2** | **Avg.** |
| 2B | InternVL2.5 | 42.51 | 51.87 | 35.23 | 39.54 | 42.29 | 31.09 | 56.67 | 29.16 | 51.02 | 27.39 | 47.63 | 40.49 |
| 3B | Qwen2.5-VL | 49.90 | 56.88 | 42.27 | 45.77 | 48.71 | 37.18 | 65.94 | 36.05 | 62.15 | 32.62 | 57.12 | 48.51 |
| 4B | InternVL2.5 | 45.97 | 53.49 | 37.95 | 41.57 | 44.75 | 33.45 | 58.47 | 31.24 | 54.14 | 28.96 | 50.13 | 42.73 |
| 7B | Qwen2.5-VL | 55.10 | 61.81 | 47.68 | 50.06 | 53.66 | 44.13 | 72.37 | **43.00** | 67.49 | **39.88** | **62.26** | **54.85** |
| 8B | InternVL2.5 | 51.29 | 58.42 | 43.47 | 46.47 | 49.91 | 38.36 | 63.33 | 36.48 | 60.72 | 32.13 | 55.82 | 47.81 |
| 26B | InternVL2.5 | 51.13 | 59.87 | 43.92 | 47.75 | 50.67 | 39.42 | 64.95 | 35.61 | 62.93 | 31.29 | 56.26 | 48.41 |
| 72B | Qwen2.5-VL | **56.65** | **62.66** | **48.13** | **50.98** | **54.60** | **45.13** | **73.73** | 41.61 | **68.14** | 38.53 | 61.60 | 54.79 |

## 5 EXPERIMENTS

**Experimental Setup.** VTF fine tunes Qwen2.5-VL-7B-Instruct with four bit LoRA for three epochs on two NVIDIA H800 80GB GPUs; inference uses one. We select the checkpoint with the lowest Tunnel validation loss. All models receive the same questions and answer parsing. Results average five runs, each task across seven scenes and then across six tasks. Appendix E.1 gives the training and inference settings.

**Benchmark Findings.** Table 1 reveals a gap between temporal recognition and physical interpretation. Qwen2.5-VL leads open source models at 42.50%, and Claude Opus 5.5 leads overall at 45.58%. The latter's I2, P2, and E2 accuracy falls from 64.15% to 59.49% and 51.88%, approaching the 50% chance level. The best selection scores are 33.68% on both I1 and P1 and 30.65% on E1, versus 25% chance. Plausible visual evolution thus does not ensure reliable inference of physical states or consequences.

**Effectiveness of VTF.** Zero-shot uses the base model with scene images only. CoT adds aligned fields through the fixed prompt in Figure 2. SFT uses the tuned checkpoint with scene images only; full VTF combines this checkpoint with the same field guided CoT. Their tunnel accuracies are 46.25%, 48.10%, 53.40%, and 61.20%, respectively (Figure 4). Field guided CoT gains 7.80 points after tuning versus 1.85 before tuning, indicating complementary benefits. Across six unseen environments, VTF improves the mean from 41.88% to 53.79%.

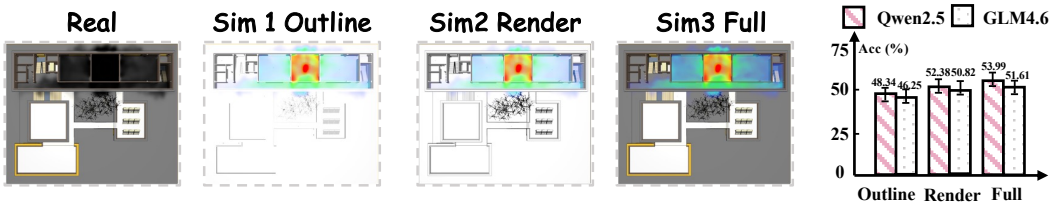



Figure 7: Garden field rendering variants (left) and accuracy (right).

**Parameter Scaling.** Table 2 shows diminishing returns from model size. Qwen2.5-VL benchmark accuracy rises from 48.51% at 3B to 54.85% at 7B; scaling to 72B yields 54.79%, providing no further gain despite more than tenfold growth in parameters. Although 72B improves the unit task average by 0.94 points, this does not translate into higher benchmark accuracy. We therefore use 7B as the VTF backbone.

**Generalization to Unseen Scenarios.** Trained on tunnel cases, VTF leads in all six unseen environments at stride 1 (Figure 4, upper panels), reaching 53.79% mean accuracy versus 41.88% for zero-shot and 48.77% for SFT. The gains over zero-shot are largest in the school and room, at 14.33 and 13.93 percentage points, respectively. The temple remains the hardest scenario at 47.82%, with a smaller gain of 5.29 points, showing that transfer varies across environments. At stride 2, mean accuracy remains 53.23%, only 0.56 points below stride 1. Together, these results indicate that field guided reasoning transfers beyond the training geometry and remains effective with sparser observations, although performance still depends on the target scenario.

**Robustness on Real Fire Images.** We apply the tunnel checkpoint to real camera images from NIST's full scale Station Nightclub mockup experiment (Madrzykowski et al., 2006; Grosshandler et al., 2005), without fine tuning on real data (Figure 5). The zero-shot baseline receives scene images and questions; VTF additionally receives simulation fields at the observed timestamps, with no future fields. E1/E2 labels derive from experimental measurements in the NIST report. VTF improves all six tasks, raising mean accuracy from 38.26% to 49.04%. Gains range from 6.86 points on E1 to 15.07 on P1. E1 remains the weakest selection task (31.78%), indicating persistent difficulty in environmental assessment. Appendix E.2 details the protocol.

**Field Alignment.** Holding the model, prompt, scenes, and questions fixed, shuffling field order reduces tunnel CoT accuracy from 48.10% to 40.53% (Figure 6), below zero-shot accuracy of 46.25%. I1 and I2 show the largest declines, of 10.06 and 9.43 points, respectively. Degradation across all six tasks suggests that disrupting temporal correspondence also compromises state and environmental judgments. Misaligned fields can therefore be less useful than scene images alone.

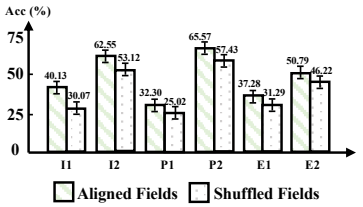



Figure 6: Field alignment in tunnels.

**Field Rendering.** We compare three rendering variants of the same physical fields, using Garden as an example. Only the architectural background varies from Outline to Render to Full (Figure 7, left). Full performs best for both models (Figure 7, right), exceeding Outline by 5.65 and 5.36 percentage points for Qwen2.5 and GLM4.6, respectively, but Render by only 1.61 and 0.79 points. The consistent ranking across both backbones suggests that architectural context facilitates scene and field correspondence. Most of the improvement occurs between Outline and Render, indicating that additional background detail yields diminishing gains when the physical fields are unchanged.

## 6 Conclusion

We introduced a combustion benchmark with three levels and VTF for reasoning with fields. Field prompting and fine tuning provide complementary gains. Transfer to unseen scenarios and real fire imagery, together with sensitivity to field alignment, supports simulation fields as reusable evidence for IPU beyond rendered training scenes when auxiliary simulation fields are available. Further validation is needed across broader real fire conditions and other physical processes.

## REFERENCES


Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G. Krishnan. Physics context builders: A modular framework for physical reasoning in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7318–7328, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, and Mahmoud Assran. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. doi: 10.1073/pnas.1306572110.

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29071–29082, 2025.

Qiang Chen, Hao Guo, Huatai Zhu, Tairan Huang, Yichao Cao, Hongyan Xu, Keke Huang, Haifeng Li, Yi Chen, and Xiu Su. FireWorldBench: Benchmarking complex physical world intelligence through coupled-field fire dynamics. *arXiv preprint arXiv:2609.23064*, 2026.

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.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. ReWind: Understanding long videos with instructed learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13734–13743, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, 2025.

William Grosshandler, Nelson Bryner, Daniel Madrzykowski, and Kenneth Kuntz. Report of the technical investigation of The Station Nightclub Fire: Appendices. Technical Report NIST NCSTAR 2v2, National Institute of Standards and Technology, Gaithersburg, MD, June 2005. URL `https://doi.org/10.6028/NIST.NCSTAR.2v2`.

Meng-Hao Guo, Xuanyu Chu, Qianrui Yang, Zhe-Han Mo, Yiqing Shen, Pei-Lin Li, et al. RBench-V: A primary assessment for visual reasoning models with multimodal outputs. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Jessica B. Hamrick, Peter W. Battaglia, Thomas L. Griffiths, and Joshua B. Tenenbaum. Inferring mass in complex scenes by mental simulation. *Cognition*, 157:61–76, 2016. doi: 10.1016/j.cognition.2016.08.012.

Yuping He, Yifei Huang, Guo Chen, Baoqi Pei, Jilan Xu, Tong Lu, and Jiangmiao Pang. EgoExoBench: A benchmark for first- and third-person view video understanding in MLLMs. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. On the consistency of video large language models in temporal comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722, 2025.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Proceedings of the International Conference on Machine Learning*, pp. 36340–36364, 2025.

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 multi-modal large language models. In *International Conference on Learning Representations*, 2026.

Yi Li, Eric Perlman, Minping Wan, Yunke Yang, Charles Meneveau, Randal Burns, Shiyi Chen, Alexander Szalay, and Gregory Eyink. A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence. *Journal of Turbulence*, 9:N31, 2008. doi: 10.1080/14685240802376389.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, et al. Unleashing hour-scale video training for long video-language understanding. In *Advances in Neural Information Processing Systems*, 2025.

Romy Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Understanding and correcting reasoning failures in multimodal models. In *Advances in Neural Information Processing Systems*, 2025.

Daniel M. Madrzykowski, Nelson P. Bryner, and Stephen I. Kerber. NIST station nightclub fire investigation: Physical simulation of the fire. *Fire Protection Engineering*, (31), 2006. URL `https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=910278`.

Kevin B. McGrattan, Howard R. Baum, Ronald G. Rehm, Anthony P. Hamins, and Glenn P. Forney. Fire Dynamics Simulator: Technical reference guide. NIST Interagency/Internal Report 6467, National Institute of Standards and Technology, 2000.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The Well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tianhao Peng, Haochen Wang, Yuanxing Zhang, Noah Wang, Zili Wang, Ge Zhang, et al. MVU-Eval: Towards multi-video understanding evaluation for multimodal LLMs. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael Ryoo. Understanding long videos with multimodal language models. In *International Conference on Learning Representations*, 2025.

Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. VELOCITI: Benchmarking video-language compositional reasoning with strict entailment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18914–18924, 2025.

Zihan Song, Xin Wang, Zi Qian, Hong Chen, Longtao Huang, Hui Xue, and Wenwu Zhu. Modularized self-reflected video reasoner for multimodal LLM with application to video question answering. In *Proceedings of the International Conference on Machine Learning*, pp. 56389–56413, 2025.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning. In *Advances in Neural Information Processing Systems*, 2025.

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 *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel M. Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel L. K. Yamins, Judith Fan, and Kevin A. Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

Hao Wu, Yongheng Zhang, Yuan Gao, Fan Xu, Fan Zhang, Ruobing Xie, Ruijian Gou, Yuxuan Liang, Xiaomeng Huang, and Xian Wu. OMNIFLOW: A physics-grounded multimodal agent for generalized scientific reasoning. *arXiv preprint arXiv:2603.15797*, 2026a.

Xiyang Wu, Zongxia Li, Jihui Jin, Guangyao Shi, Gouthaman KV, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. MASS: Motion-aware spatial-temporal grounding for physics reasoning and comprehension in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Y. Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In *Advances in Neural Information Processing Systems*, 2025.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, et al. RTV-Bench: Benchmarking MLLM continuous perception, understanding and reasoning through real-time video. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Junyan Ye, Dongzhi Jiang, Jun He, Baichuan Zhou, Zilong Huang, Zhiyuan Yan, et al. BLINK-Twice: You see, but do you observe? a reasoning benchmark on visual perception. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Han Yi, Yulu Pan, Feihong He, Xinyu Liu, Benjamin J. Zhang, Oluwatumininu Oguntola, and Gedas Bertasius. ExAct: An executable action benchmark for multimodal physical reasoning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, et al. VRBench: A benchmark for multi-step reasoning in long narrative videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21655–21666, 2025a.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. CREMA: Generalizable and efficient video-language reasoning via multimodal modular fusion. In *International Conference on Learning Representations*, 2025b.

Jinglei Zhang, Yuanfan Guo, Rolandos Alexandros Potamias, Jiankang Deng, Hang Xu, and Chao Ma. VTimeCoT: Thinking by drawing for video temporal grounding and reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Xiaomei Zhang, Hanyu Zheng, Xiangyu Zhu, Jinghuan Wei, Junhong Zou, Zhen Lei, and Zhaoxiang Zhang. Improving large vision-language models' understanding for flow field data. *arXiv preprint arXiv:2507.18311*, 2025b.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, et al. MMVU: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8475–8489, 2025.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B. Tenenbaum, and Chuang Gan. ContPhy: Continuum physical concept learning and reasoning from videos. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61526–61558. PMLR, 2024. URL `https://proceedings.mlr.press/v235/zheng24l.html`.

Erle Zhu, Yadi Liu, Zhe Zhang, Xujun Li, Jin Zhou, Xinjie Yu, Minlie Huang, and Hongning Wang. MAPS: Advancing multi-modal reasoning in expert-level physical science. *arXiv preprint arXiv:2501.10768*, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

# APPENDIX

## AI USE STATEMENT

We used generative AI tools to help refine the conceptual framing and benchmark methodology, assist with mathematical derivations and their exposition, interpret reported results, and translate and edit the manuscript. We also used them to identify relevant literature and format the manuscript. The authors reviewed the content developed with AI assistance and take responsibility for the final paper.

## A FIELD REPRESENTATION EVOLUTION DERIVATION

We derive the field representation evolution identity from continuous balance laws and spatial sampling.

**Physical State and Balance Laws.** Consider a fixed domain $\Omega \subset \mathbb{R}^d$ with a sufficiently regular boundary and outward unit normal $\boldsymbol{n}$. A multicomponent viscous fluid has conserved state $\boldsymbol{U} = (\rho, \rho\boldsymbol{u}^{\mathrm{T}}, \rho E, \rho\boldsymbol{Y}^{\mathrm{T}})^{\mathrm{T}}$ and flux matrix

$$\boldsymbol{F} = \begin{bmatrix} \rho\boldsymbol{u}^{\mathrm{T}} \\ \rho\boldsymbol{u}\otimes\boldsymbol{u} + p\boldsymbol{I} - \boldsymbol{\tau} \\ \left[(\rho E + p)\boldsymbol{u} - \boldsymbol{\tau}\cdot\boldsymbol{u} + \boldsymbol{q}\right]^{\mathrm{T}} \\ \rho\boldsymbol{Y}\otimes\boldsymbol{u} + \boldsymbol{J} \end{bmatrix}, \qquad \partial_t \boldsymbol{U} + \sum_{\ell=1}^{d} \partial_{x_\ell} \boldsymbol{F}_\ell = \boldsymbol{S}. \tag{A.1}$$

Here, $d$ is the spatial dimension, $\boldsymbol{I}$ the identity matrix, and $\boldsymbol{F}_\ell$ column $\ell$ of $\boldsymbol{F}$. For symmetric viscous stress, $\boldsymbol{F}_n = \boldsymbol{F}\boldsymbol{n}$. The rows of $\boldsymbol{J}$ are species diffusion fluxes. We use $E = e + \|\boldsymbol{u}\|^2/2$, including formation energy in $e$ and species diffusion energy in $\boldsymbol{q}$. The source $\boldsymbol{S}$ collects body forces, their work, external heating, and species production, without counting reaction energy twice. Its mass component vanishes without volumetric injection. Solvers approximate these continuum balances using numerical schemes and closures (McGrattan et al., 2000).

**Variable Transformation and Sampling.** Write $\boldsymbol{U} = (\rho, \boldsymbol{m}^{\mathrm{T}}, \varepsilon, \boldsymbol{r}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{m} = \rho\boldsymbol{u}$, $\varepsilon = \rho E$, and $\boldsymbol{r} = \rho\boldsymbol{Y}$ are momentum, total energy, and species densities. The primitive variables satisfy

$$\boldsymbol{u} = \frac{\boldsymbol{m}}{\rho}, \qquad e = \frac{\varepsilon}{\rho} - \frac{\|\boldsymbol{m}\|^2}{2\rho^2}, \qquad \boldsymbol{Y} = \frac{\boldsymbol{r}}{\rho}, \qquad T = \mathcal{T}(\rho, e, \boldsymbol{Y}). \tag{A.2}$$

The thermodynamic relation $\mathcal{T}$ and concentration conversions define $\boldsymbol{\Psi}$, assuming $\rho > 0$ and sufficient differentiability. For species $j$, use $c_j = \rho Y_j$ for mass concentration or $Y_j$ for mass fraction, matching the exported units. The $d+2$ field components are sampled by a diagonal or block diagonal matrix $\boldsymbol{W}$. At image location $\boldsymbol{p}$, define

$$\boldsymbol{g}(\boldsymbol{p}, t) = \int_\Omega \boldsymbol{W}(\boldsymbol{p}, \boldsymbol{x})\boldsymbol{\Psi}(\boldsymbol{U}(\boldsymbol{x}, t))\,\mathrm{d}V, \qquad \boldsymbol{B} = \boldsymbol{W}\frac{\partial\boldsymbol{\Psi}}{\partial\boldsymbol{U}}. \tag{A.3}$$

The Jacobian is evaluated at $\boldsymbol{U}(\boldsymbol{x}, t)$, and $\mathrm{d}V$ is the volume element. This sampling precedes color mapping.

**Chain Rule and Integration by Parts.** For a fixed domain and weights independent of time, Equation (A.1) gives

$$\partial_t \boldsymbol{g} = \int_\Omega \boldsymbol{B}\,\partial_t \boldsymbol{U}\,\mathrm{d}V = \int_\Omega \boldsymbol{B}\boldsymbol{S}\,\mathrm{d}V - \sum_{\ell=1}^{d}\int_\Omega \boldsymbol{B}\,\partial_{x_\ell}\boldsymbol{F}_\ell\,\mathrm{d}V. \tag{A.4}$$

The product rule yields $\boldsymbol{B}\partial_{x_\ell}\boldsymbol{F}_\ell = \partial_{x_\ell}(\boldsymbol{B}\boldsymbol{F}_\ell) - (\partial_{x_\ell}\boldsymbol{B})\boldsymbol{F}_\ell$. The divergence theorem gives

$$\partial_t \boldsymbol{g} + \int_{\partial\Omega} \boldsymbol{B}\boldsymbol{F}\boldsymbol{n}\,\mathrm{d}A = \int_\Omega \left[\boldsymbol{B}\boldsymbol{S} + \sum_{\ell=1}^{d}(\partial_{x_\ell}\boldsymbol{B})\boldsymbol{F}_\ell\right]\,\mathrm{d}V. \tag{A.5}$$

Contracting over conserved components and spatial directions, define

$$(\nabla\boldsymbol{B}) : \boldsymbol{F} \equiv \sum_{\ell=1}^{d}(\partial_{x_\ell}\boldsymbol{B})\boldsymbol{F}_\ell.$$

With $\boldsymbol{F}_n = \boldsymbol{F}\boldsymbol{n}$, this yields the identity in the main paper. Its correction follows from

$$\partial_{x_\ell}\boldsymbol{B} = (\partial_{x_\ell}\boldsymbol{W})\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{U}} + \boldsymbol{W}\partial_{x_\ell}(\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{U}}). \tag{A.6}$$

The terms arise from spatially varying sampling and nonlinear variable conversion. Setting $\boldsymbol{\Psi}(\boldsymbol{U}) = \boldsymbol{U}$ and $\boldsymbol{W} = \boldsymbol{I}$ recovers the original integral balance.

**Temporal Intervals and Scope.** Integrating Equation (A.5) over an interval gives

$$\boldsymbol{g}(\boldsymbol{p}, t+\Delta t) - \boldsymbol{g}(\boldsymbol{p}, t) = \int_t^{t+\Delta t} \left\{ \int_\Omega [\boldsymbol{B}\boldsymbol{S} + (\nabla \boldsymbol{B}) : \boldsymbol{F}]\, \mathrm{d}V - \int_{\partial\Omega} \boldsymbol{B}\boldsymbol{F}_n\, \mathrm{d}A \right\} \mathrm{d}\xi, \tag{A.7}$$

All integrands are evaluated at $\xi$. NFS and FCJ use labels derived from simulations, not this identity as a loss. Weights that vary with time add $\int_\Omega (\partial_t \boldsymbol{W})\boldsymbol{\Psi}(\boldsymbol{U})\, \mathrm{d}V$ to Equation (A.5). The derivation assumes smooth fields; discretization and rendering need not preserve the identity, and no conservation guarantee applies to RGB images or model predictions.

# B Case Study

## B.1 Garden Fire

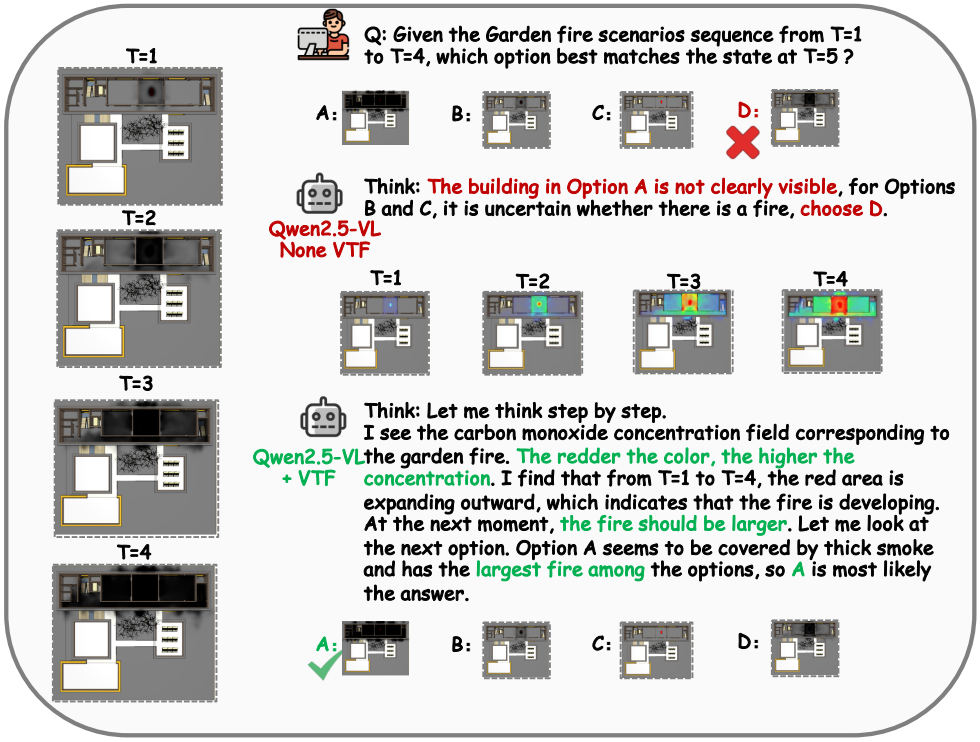



Figure B.1: Garden next state selection without and with VTF.

**Reasoning under Smoke Occlusion.** Figure B.1 compares Qwen2.5-VL responses in the unseen garden environment. Without VTF, the model rejects option A because the building is obscured, expresses uncertainty about the fire in B and C, and incorrectly selects D. With VTF, the response refers to the aligned CO concentration fields and tracks the expanding region of high concentration over the observed sequence. It then selects A, the correct continuation with more extensive smoke coverage. This case illustrates how field representations make evolving physical conditions explicit when smoke obscures scene details, supporting the transfer from field interpretation to a visual prediction in an unseen environment.

### B.2 SCHOOL FIRE

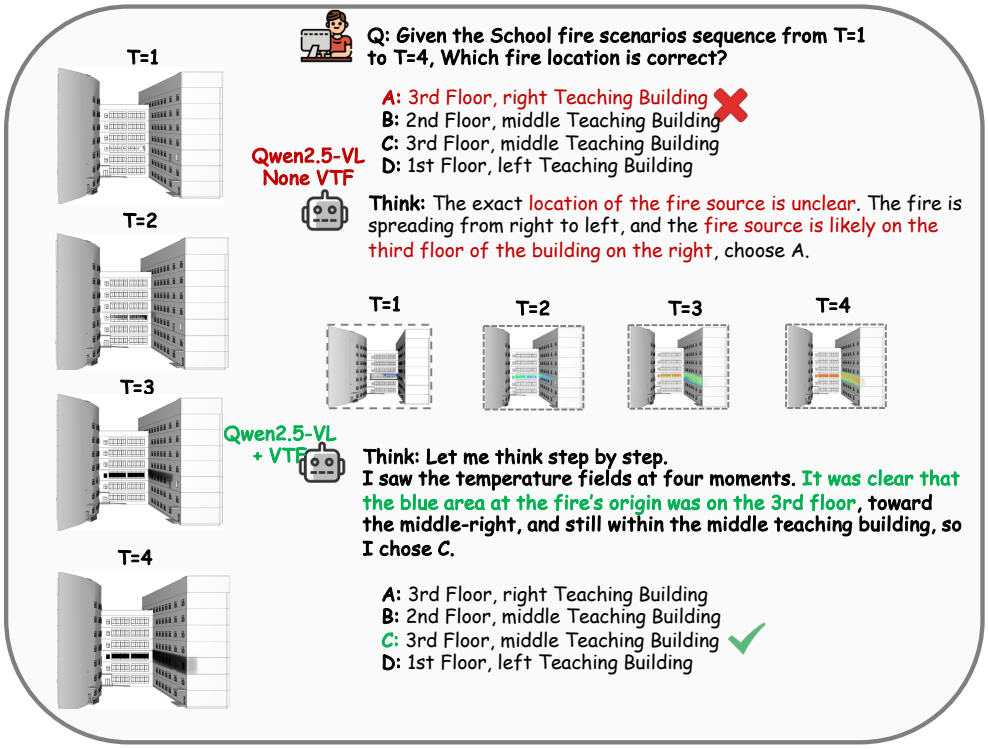



Figure B.2: School fire source localization without and with VTF.

**Localizing the Fire Source with Field Evidence.** Figure B.2 illustrates physical state perception in the unseen school environment. Without VTF, Qwen2.5-VL infers the source location from the apparent direction of fire spread and incorrectly selects the third floor of the right teaching building (A). With VTF, the response uses four aligned views of temperature fields to locate the source on the third floor of the middle teaching building and selects the correct answer (C). This example illustrates how field evidence can help distinguish the source location from the visible spread of combustion products, supporting spatial interpretation of physical states in an unseen building layout.

### B.3 TUNNEL FIRE

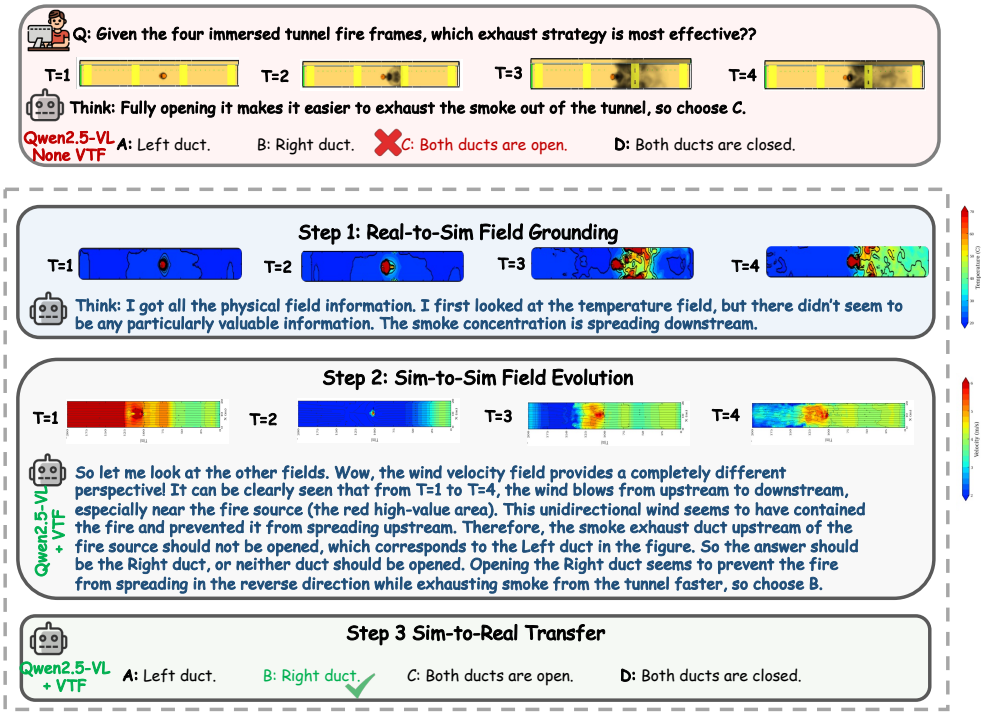



Figure B.3: Tunnel exhaust strategy selection without and with VTF.

**Assessing Exhaust Strategies with Flow Evidence.** Figure B.3 illustrates environmental assessment in the immersed tunnel. Without VTF, Qwen2.5-VL assumes that opening both ducts improves smoke removal and incorrectly selects C. With VTF, the response examines temperature and velocity fields, identifies downstream flow near the fire source, and uses this directional evidence to select the right duct (B), the correct answer in this case. The comparison illustrates how VTF can support an engineering judgment for the given configuration by connecting field patterns to an exhaust decision. It complements the garden and school cases by extending interpretation guided by fields from temporal evolution and source localization to environmental consequences.

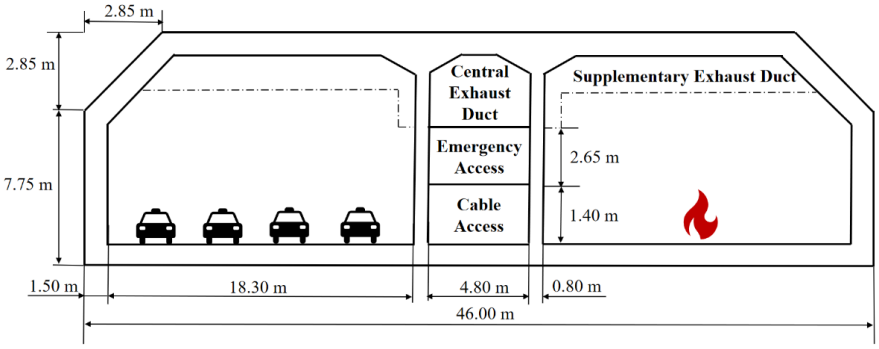
# C EXTRAWIDE IMMERSED TUNNEL FIRE

This appendix describes the immersed tunnel fire case used to construct simulation fields for training and evaluating multimodal large language models (MLLMs). It summarizes the geometry, ventilation, numerical settings, monitoring layout, and scenario matrix needed to interpret the benchmark.

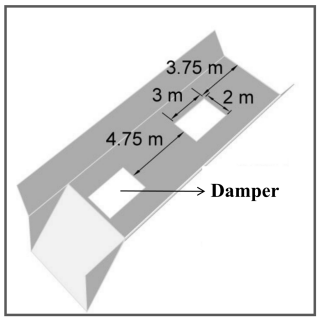
## C.1 TUNNEL GEOMETRY

The Shenzhen–Zhongshan immersed tunnel is part of the Shenzhen–Zhongshan river crossing channel and has a total length of 6.8 km. Its standard cross section is 46.0 m wide and 10.6 m high and contains two traffic tubes with four lanes each. The shared emergency access space has a clear height of 2.65 m, fire doors are arranged at intervals of 82.5 m, and the central exhaust duct area exceeds 15 $m^2$.

A supplementary exhaust duct system is used to reduce plug-holing and excessive fresh air entrainment. Each supplementary duct is 7.2 m wide and 18.3 m long and opens downward through two electrically controlled dampers. Adjacent ducts are spaced 60 m apart. Figure C.1 summarizes the tunnel cross section and the local damper geometry.




**(a)** Tunnel cross section




**(b)** Supplementary exhaust duct

Figure C.1: Engineering geometry of the extrawide immersed tunnel case. The dual traffic tubes, central service spaces, supplementary exhaust duct, and damper dimensions are shown at their corresponding scales.

## C.2 SIMULATION SETUP

The fire scenarios were modeled with Fire Dynamics Simulator (FDS) 6.7. The large-eddy simulation (LES) mode was adopted to resolve the turbulent structures carrying most of the energy while modeling the unresolved subgrid scales. The baseline heat release rate (HRR) was 50 MW, representing a severe design fire, with an initial soot production rate of 16.47 g/s. A longitudinal velocity of 3.44 m/s was imposed to suppress upstream smoke rollback. The reference setup used ignition at $t = 0$ s and extraction at $t = 20$ s.

Table C.1: Reference settings for the immersed tunnel fire simulations.

| **Parameter** | **Setting** |
|---|---|
| Ambient temperature | 20 °C |
| Ambient pressure | 101 kPa |
| Baseline HRR | 50 MW |
| Initial soot production rate | 16.47 g/s |
| Longitudinal wind speed | 3.44 m/s |
| Concrete density | 2280 kg/m$^3$ |
| Concrete specific heat | 1.04 kJ/(kg·K) |
| Concrete conductivity | 1.8 W/(m·K) |
| Main fan exhaust flow rate | 250 m$^3$/s |
| Mechanical exhaust start time | 20 s |
| Mesh size near the fire | $0.4 \times 0.4 \times 0.4$ m$^3$ |
| Simulation duration | 300 s |

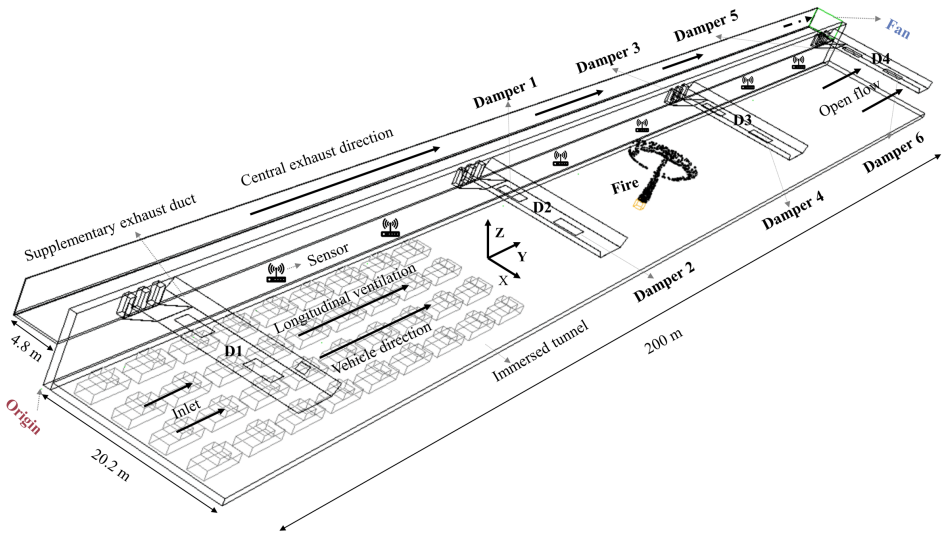



Figure C.2: Global FDS geometry of the extrawide immersed tunnel fire case.

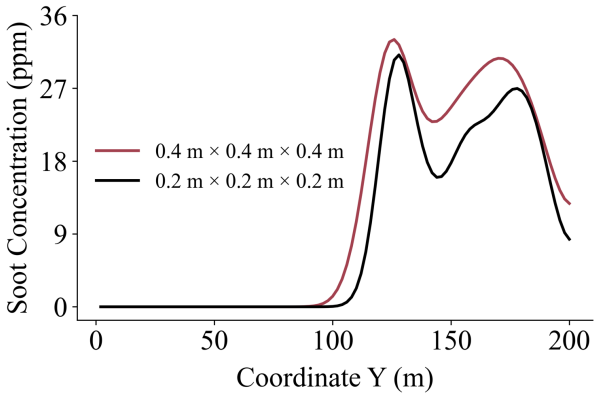
### C.3 Grid Resolution

The characteristic fire diameter was used to determine the required local mesh resolution:

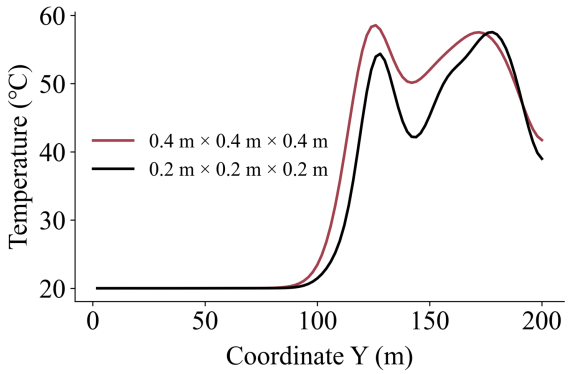
$$D^* = \left( \frac{\dot{Q}}{\rho_\infty c_p T_\infty \sqrt{g}} \right)^{2/5}, \tag{C.1}$$

where $\dot{Q}$ is the HRR, $\rho_\infty$ is the ambient density, $c_p$ is the specific heat, $T_\infty$ is the ambient temperature, and $g$ is gravitational acceleration. For the 50 MW fire, $D^* = 4.4$ m, giving a recommended cell size near the source of 0.22–0.44 m.

Grid sensitivity tests compared local cell sizes of 0.2 and 0.4 m using the longitudinal soot concentration and temperature profiles. The finer mesh produced only a limited change relative to its substantially higher computational cost. Accordingly, 0.4 m cells were used near the fire and the supplementary exhaust ducts, while 0.8 m cells were used in the far field.




**(a)** Soot concentration




**(b)** Temperature

Figure C.3: Grid sensitivity comparison for 0.2 and 0.4 m local meshes.

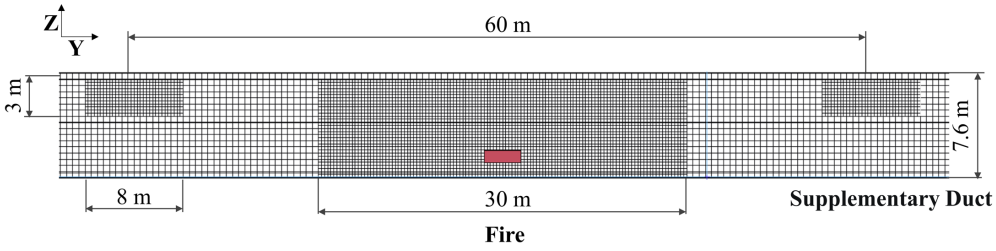
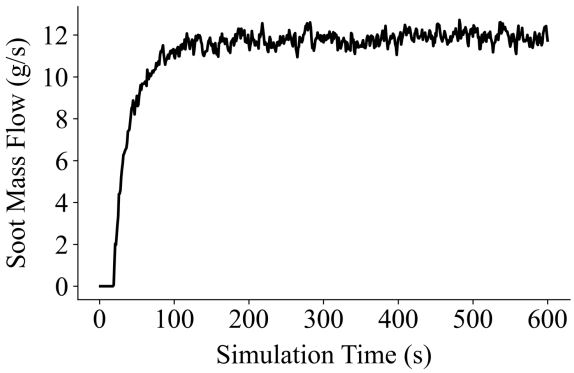


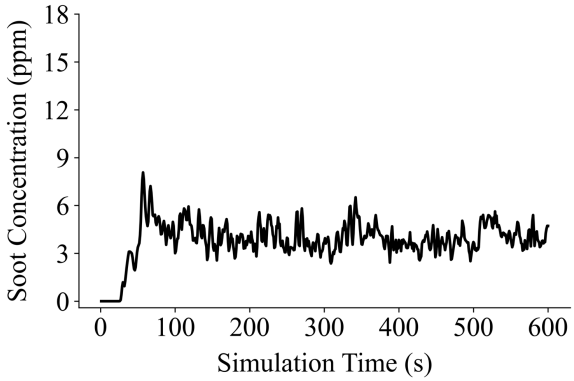

Figure C.4: Mesh arrangement near the fire source and supplementary exhaust ducts.

### C.4 Duration and Monitoring

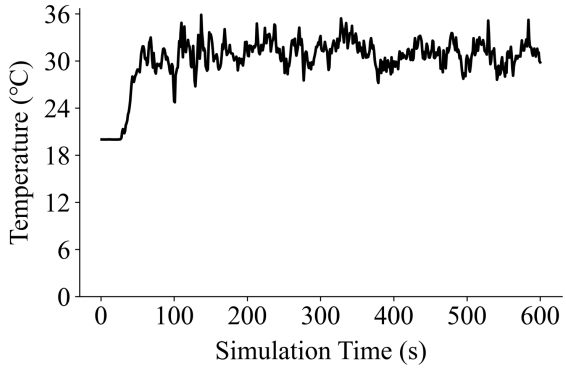
The required simulation duration was evaluated by opening all six dampers and extending the calculation to 600 s. The soot mass flow at the central exhaust outlet stabilized near 11.8 g/s after approximately 200 s. At the tunnel outlet ($Y = 200$ m), soot concentration and temperature similarly fluctuated around 5 ppm and 30 °C. A duration of 300 s was therefore adopted for the complete case matrix so that both transient development and the developed regime were retained.




**(a)** Exhaust soot mass flow




**(b)** Outlet soot concentration




**(c)** Outlet temperature

Figure C.5: Stability assessment in the extended run used to select the 300 s simulation horizon.

Soot concentration and temperature sensors were placed every 1 m along the longitudinal domain of 200 m at a height of 1.5 m. Each sensor returned one sample per second. Figure C.6 shows the fire source, six controllable dampers, and longitudinal monitoring positions.

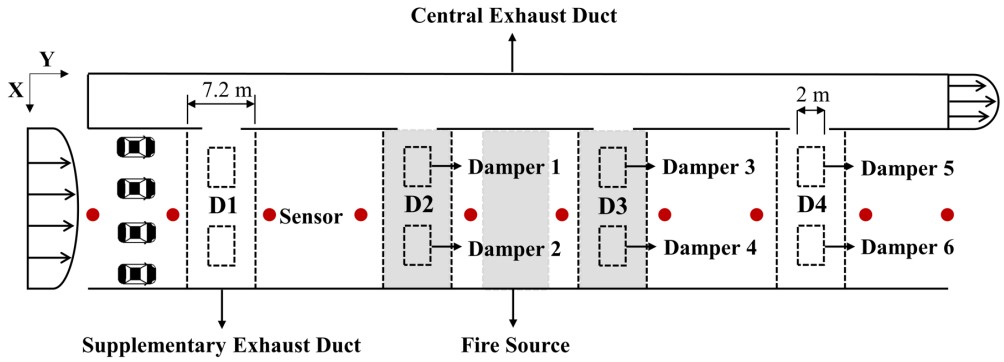



Figure C.6: Monitoring layout in plan view with the fire source, supplementary exhaust ducts, six dampers, and longitudinal sensor positions.

### C.5 Scenario Matrix

The primary dataset varies two factors: fire source location and exhaust damper combination. Six source locations cover fires in the middle and side lanes at three longitudinal positions:

$$(9.15, 70),\ (9.15, 100),\ (9.15, 130),$$
$$(2.15, 70),\ (2.15, 100),\ (2.15, 130) \quad \text{m}.$$

The 32 damper vectors at each location give $6 \times 32 = 192$ baseline cases (Table C.3). The 30 additional cases use optimal damper settings per location, with exhaust activation timing differing from the baselines at 50 MW and HRRs of $\{10, 20, 30, 40, 50\}$ MW.

Table C.2: Summary of the two groups of simulation cases.

| Case group | Fire locations | Damper setting | HRR | Cases |
|---|---|---|---|---|
| Primary matrix | Six coordinates | 32 six-bit vectors | 50 MW | 192 |
| HRR extension | Six coordinates | Optimal per location | 10–50 MW | 30 |

Table C.3: Six-bit damper vectors used in the primary matrix of 192 cases. Bits $d_1 \ldots d_6$ indicate closed (0) or open (1) dampers.

| No. | Vector | No. | Vector | No. | Vector | No. | Vector |
|---|---|---|---|---|---|---|---|
| 32 | 000000 | 24 | 010001 | 16 | 100001 | 8 | 110000 |
| 31 | 000011 | 23 | 010010 | 15 | 100010 | 7 | 110011 |
| 30 | 000101 | 22 | 010100 | 14 | 100100 | 6 | 110101 |
| 29 | 000111 | 21 | 010111 | 13 | 100111 | 5 | 110110 |
| 28 | 001001 | 20 | 011000 | 12 | 101000 | 4 | 111001 |
| 27 | 001010 | 19 | 011011 | 11 | 101011 | 3 | 111010 |
| 26 | 001100 | 18 | 011101 | 10 | 101101 | 2 | 111100 |
| 25 | 001111 | 17 | 011110 | 9 | 101110 | 1 | 111111 |

### C.6 Spatial Physical Fields

Figures C.7–C.10 show the simulated temperature and soot concentration fields for all 32 damper configurations at the six fire source locations. Within each panel, the narrow field maps follow the six-bit damper ordering in Table C.3; the common color scale and the corresponding vectors

are shown above each group of field maps. The panels are ordered by fire source position as (a) $(9.15, 70)$, (b) $(9.15, 100)$, (c) $(9.15, 130)$, (d) $(2.15, 70)$, (e) $(2.15, 100)$, and (f) $(2.15, 130)$ m.

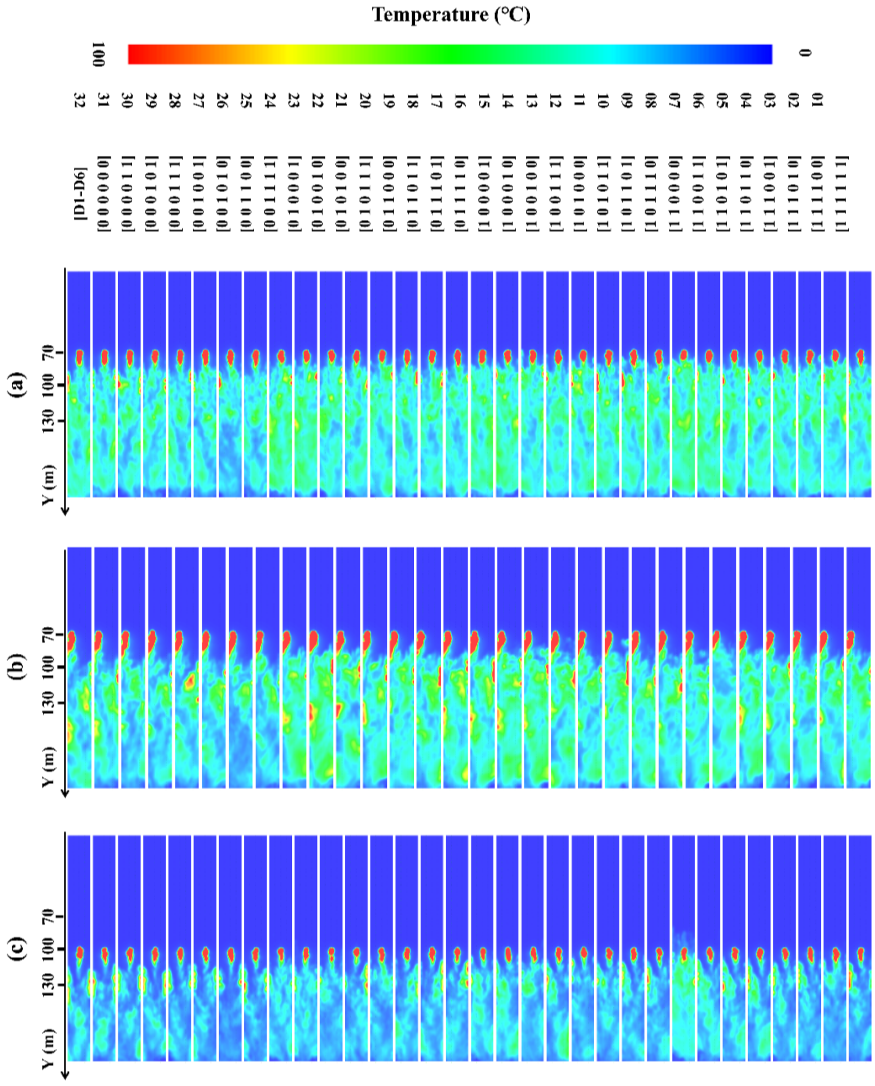



Figure C.7: Spatial temperature fields at $(9.15, 70)$, $(9.15, 100)$, and $(9.15, 130)$ m for all 32 damper configurations.

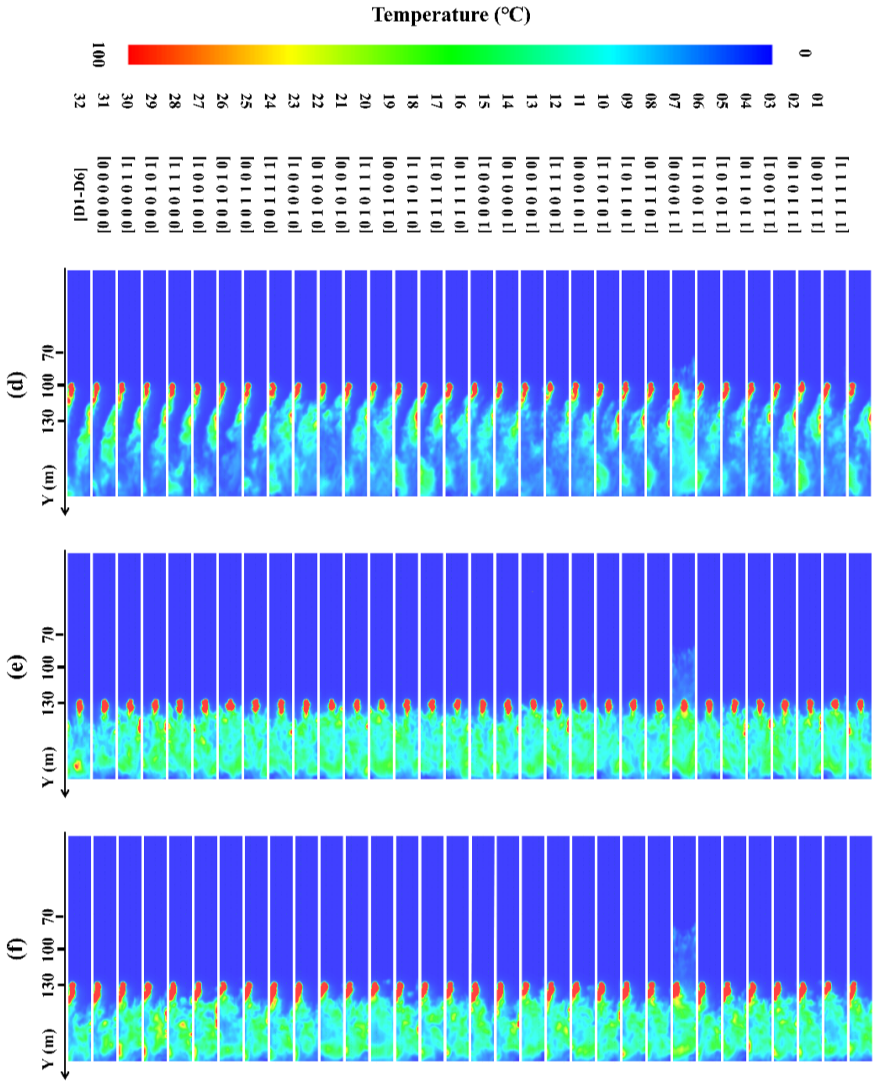



Figure C.8: Spatial temperature fields at $(2.15, 70)$, $(2.15, 100)$, and $(2.15, 130)$ m for all 32 damper configurations.

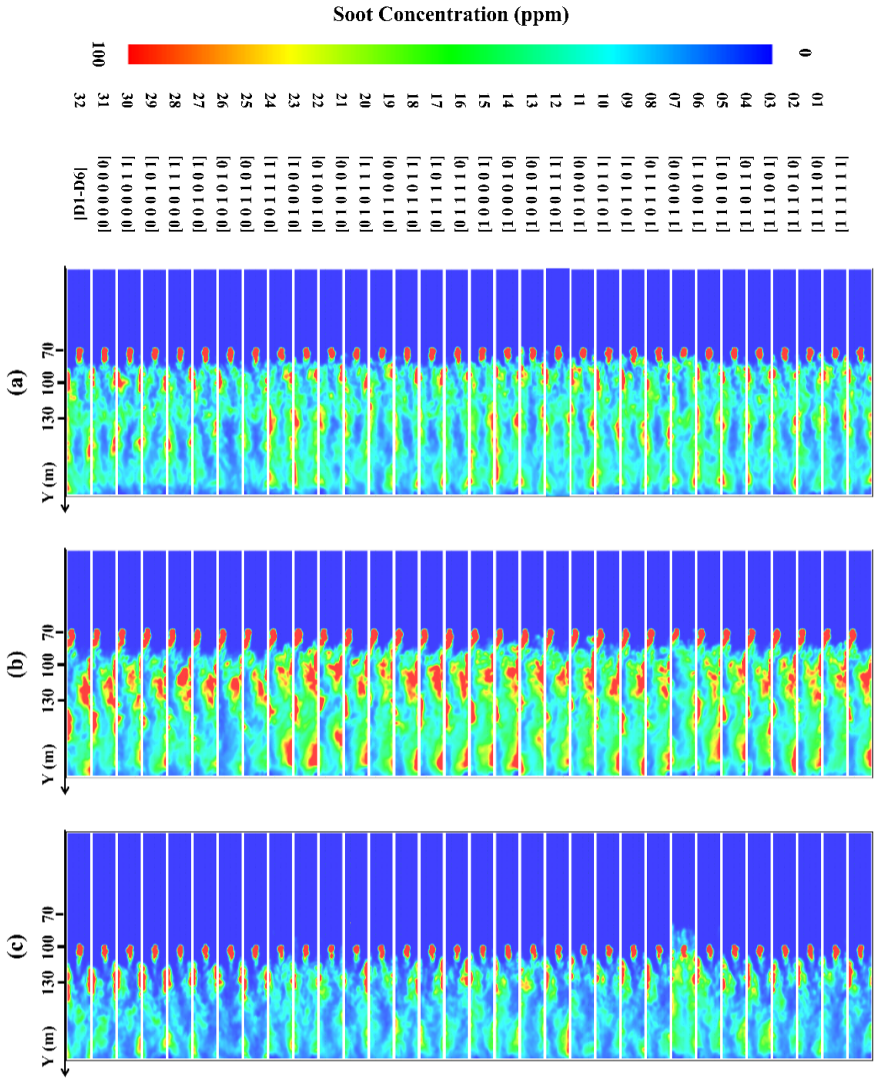



Figure C.9: Spatial soot concentration fields at $(9.15, 70)$, $(9.15, 100)$, and $(9.15, 130)$ m for all 32 damper configurations.

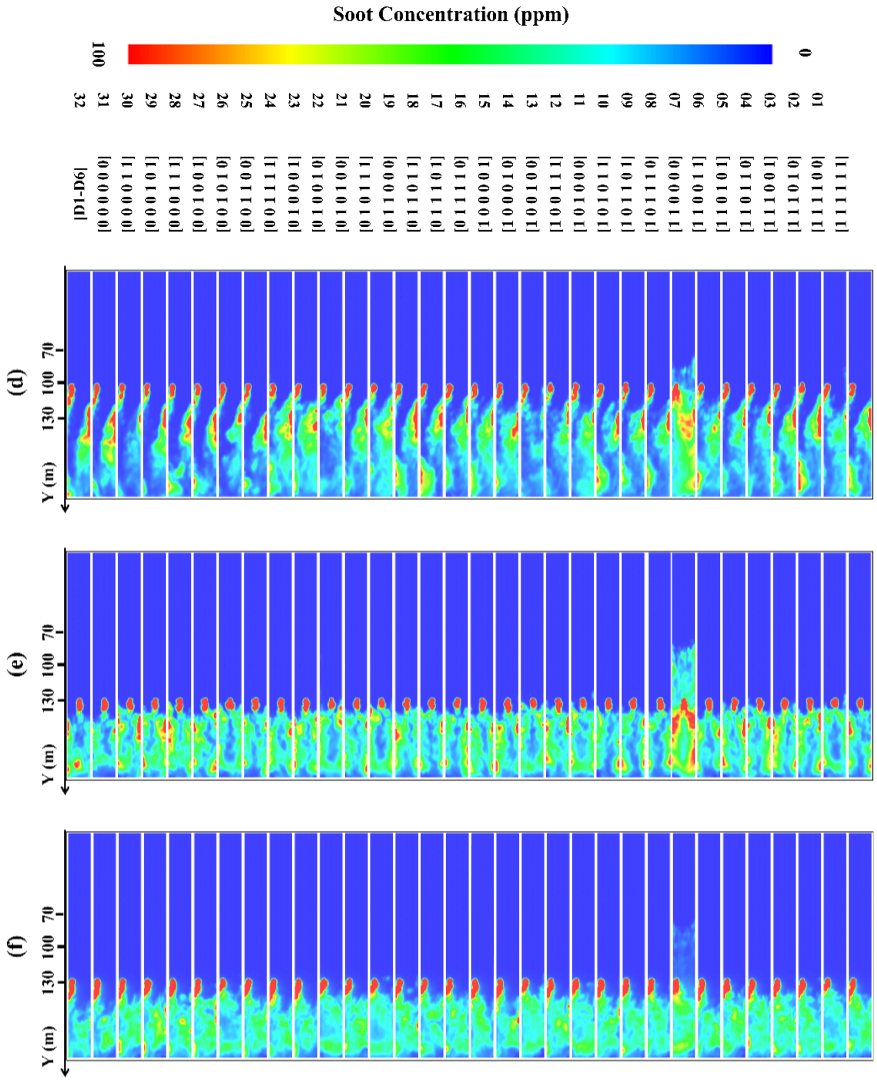



Figure C.10: Spatial soot concentration fields at $(2.15, 70)$, $(2.15, 100)$, and $(2.15, 130)$ m for all 32 damper configurations.

# D FIRE SCENARIOS FOR ADDITIONAL EVALUATION

## D.1 GARDEN FIRE

The garden scenario is reserved for evaluating model generalization. Its computational domain measures $52 \times 41 \times 10$ m $(L \times W \times H)$ and contains 174,720 cells. The following views document the building layout and courtyard geometry; dimension lines indicate the computational domain.

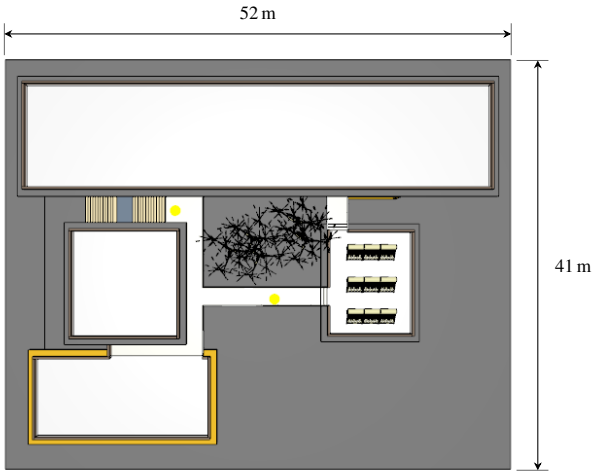



Figure D.1: Garden model: top view.

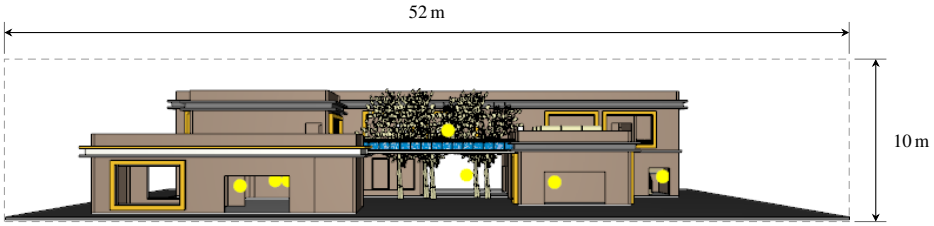



Figure D.2: Garden model: front view.

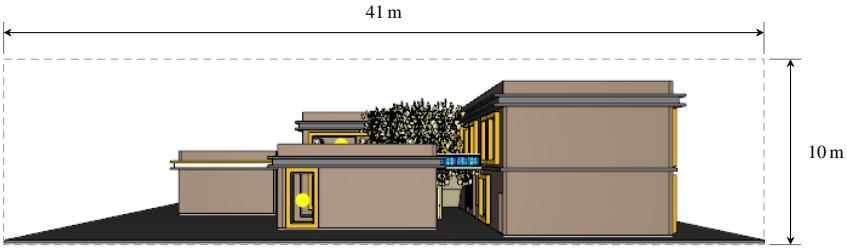



Figure D.3: Garden model: right view.

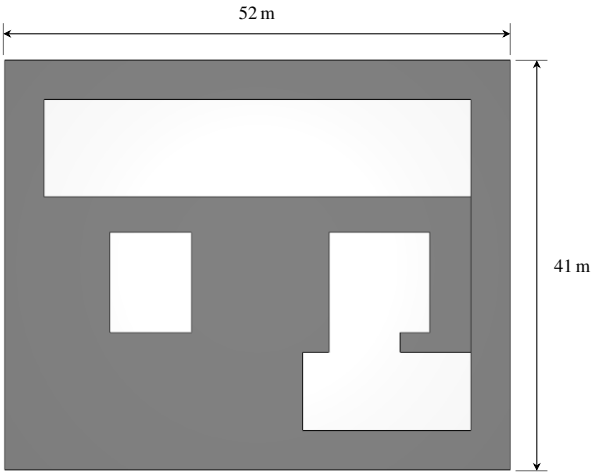



Figure D.4: Garden model: bottom view.

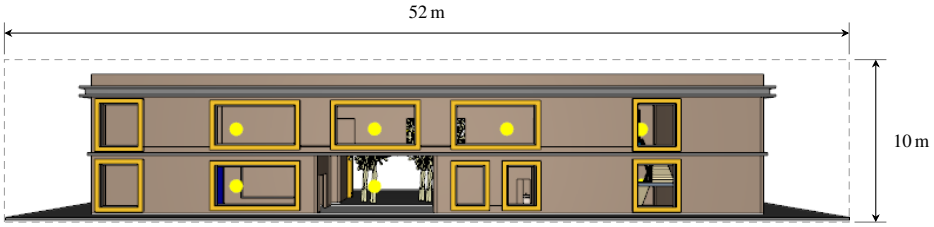



Figure D.5: Garden model: rear view.

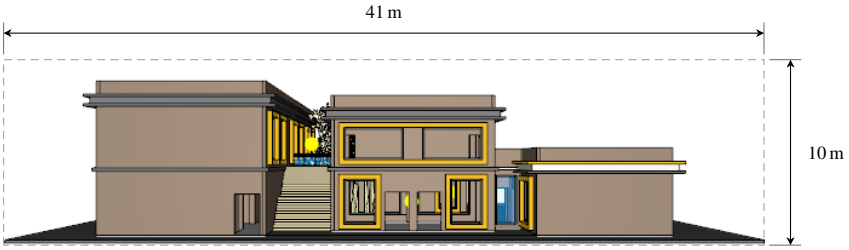



Figure D.6: Garden model: left view.

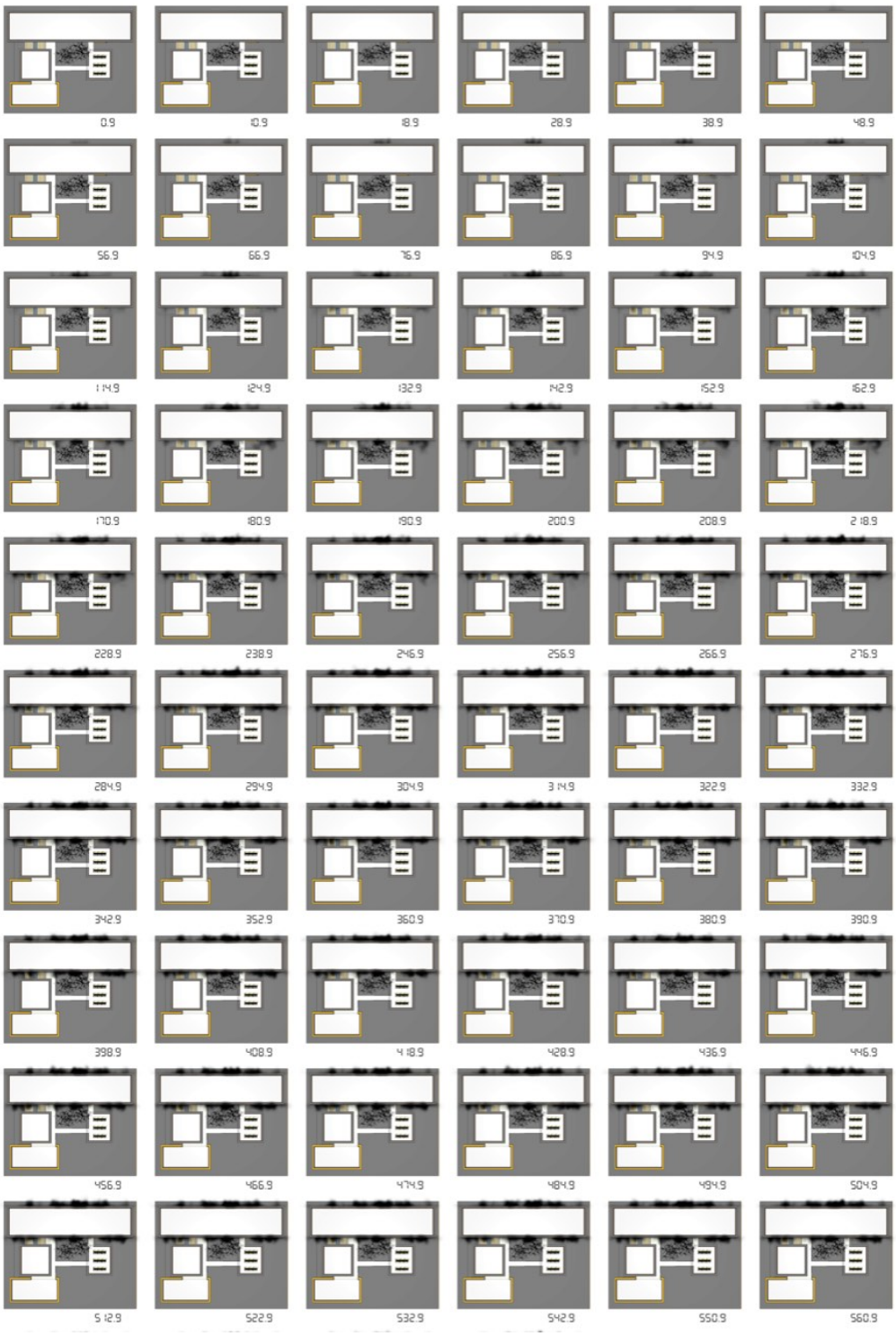


Figure D.7: Garden scene renderings. Frames sampled across the sequence.

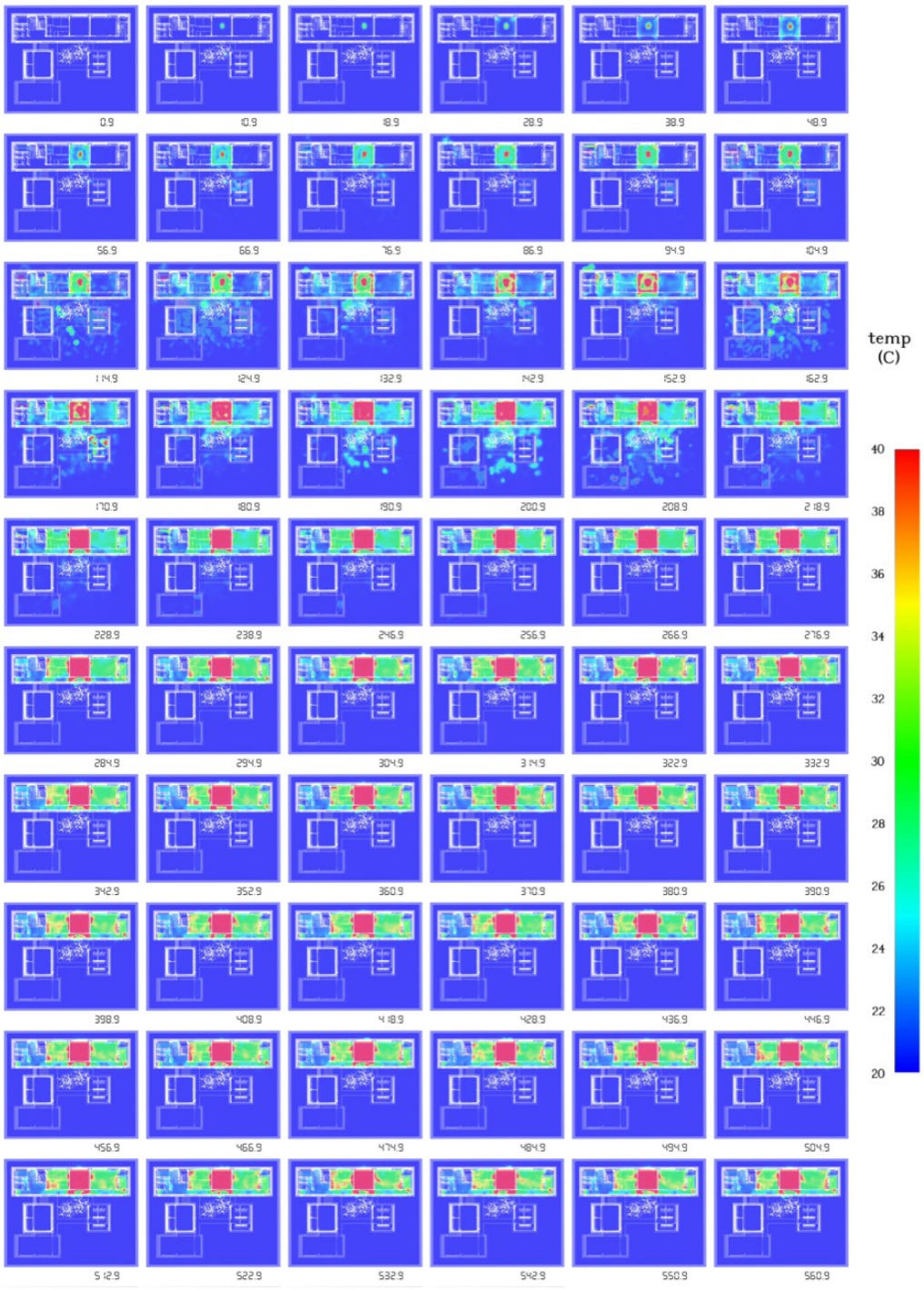



Figure D.8: Garden physical field visualizations. Frames sampled across the sequence.

### D.2 SCHOOL FIRE

The school scenario is reserved for evaluating model generalization. Its computational domain measures $74 \times 46 \times 33$ m $(L \times W \times H)$ and contains 328,812 cells. The views below show the building layout and connecting structures; domain dimensions are indicated schematically by dashed frames and dimension lines.

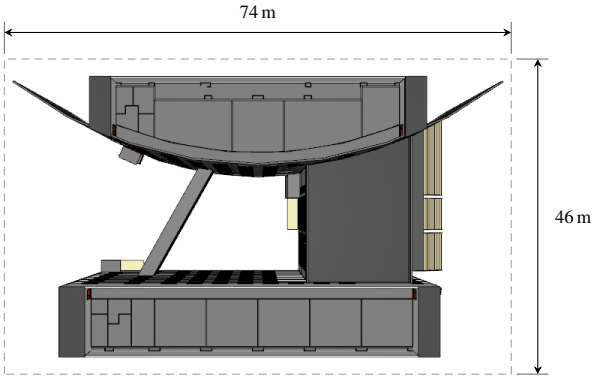



Figure D.9: School model: top view.

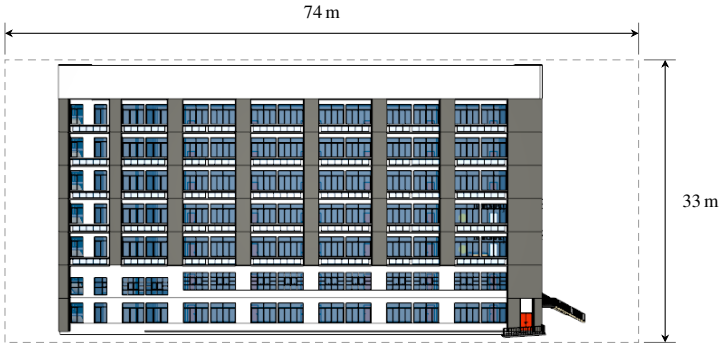



Figure D.10: School model: front view.

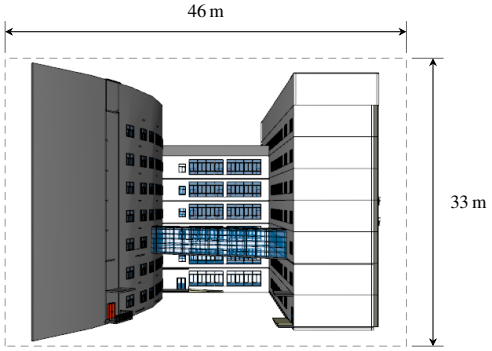



Figure D.11: School model: side view.

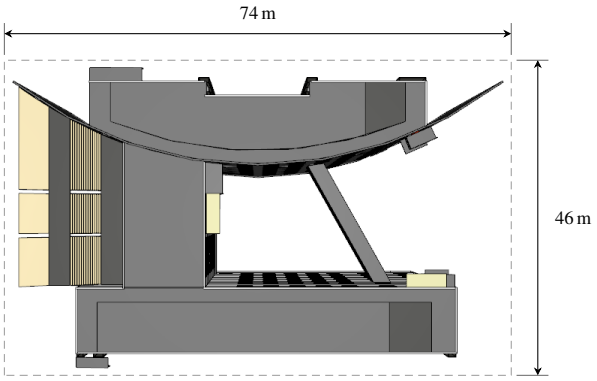



Figure D.12: School model: bottom view.




Figure D.13: School model: rear view.



Figure D.14: School scene renderings. Frames sampled across the sequence.

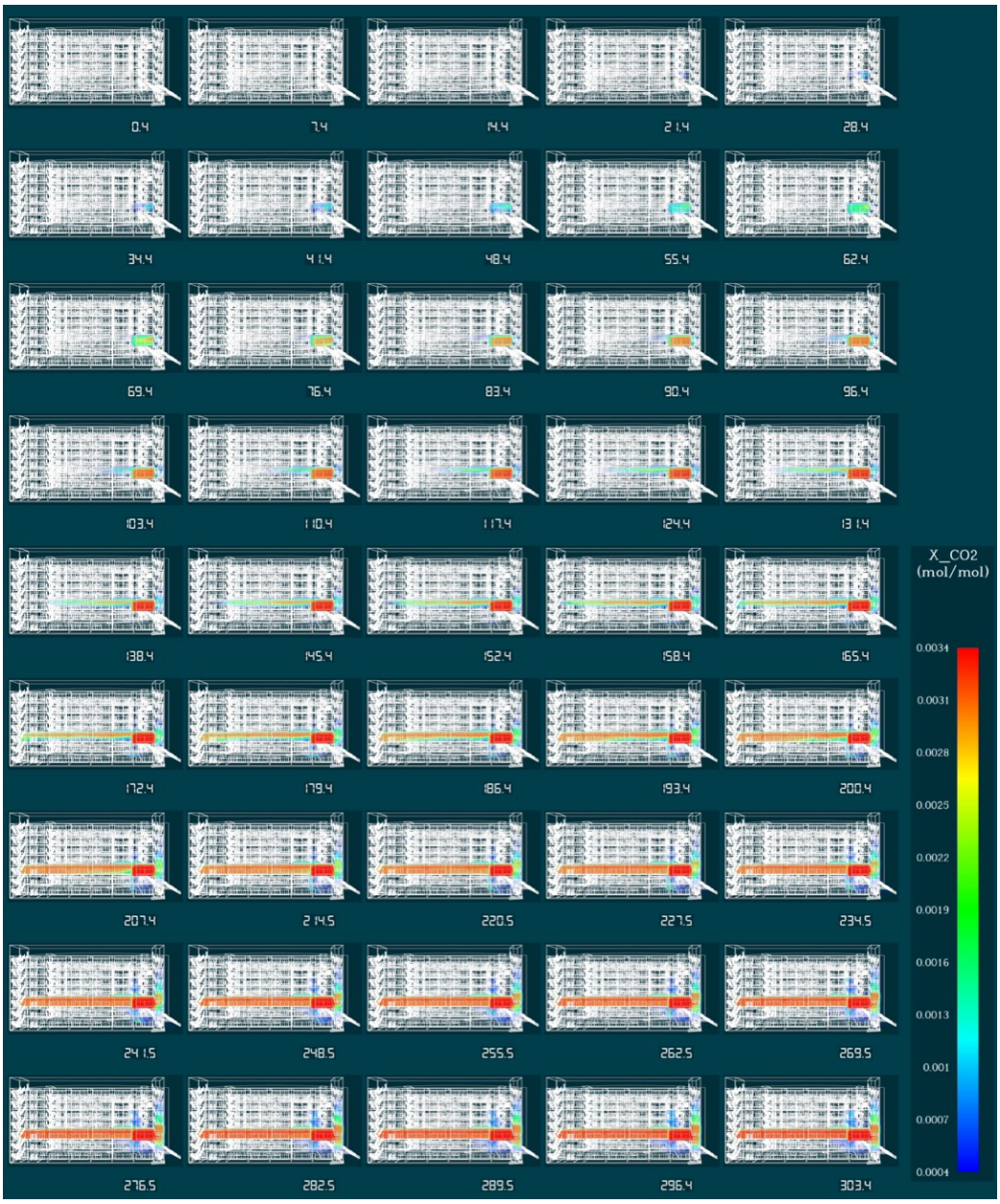



Figure D.15: School physical field visualizations. Frames sampled across the sequence.

### D.3 Office Fire

The office scenario is reserved for evaluating model generalization. Its computational domain measures $54 \times 28 \times 25$ m $(L \times W \times H)$ and contains 177,660 cells. The following views document the building geometry; domain dimensions are indicated schematically by dashed frames and dimension lines.

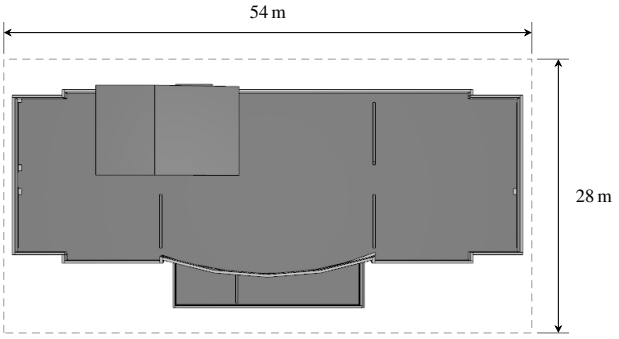



Figure D.16: Office model: top view.

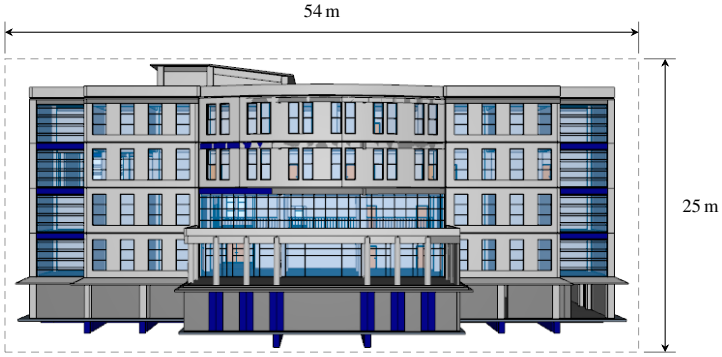



Figure D.17: Office model: front view.

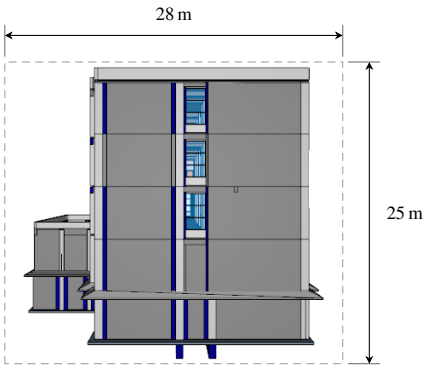



Figure D.18: Office model: right view.

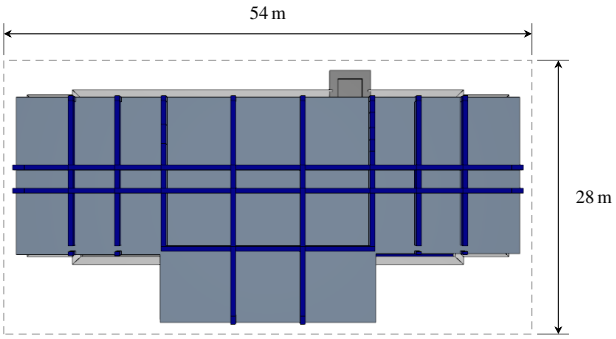



Figure D.19: Office model: bottom view.




Figure D.20: Office model: rear view.

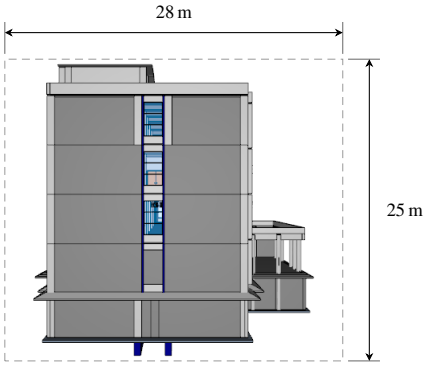



Figure D.21: Office model: left view.



Figure D.22: Office scene renderings. Frames sampled across the sequence.

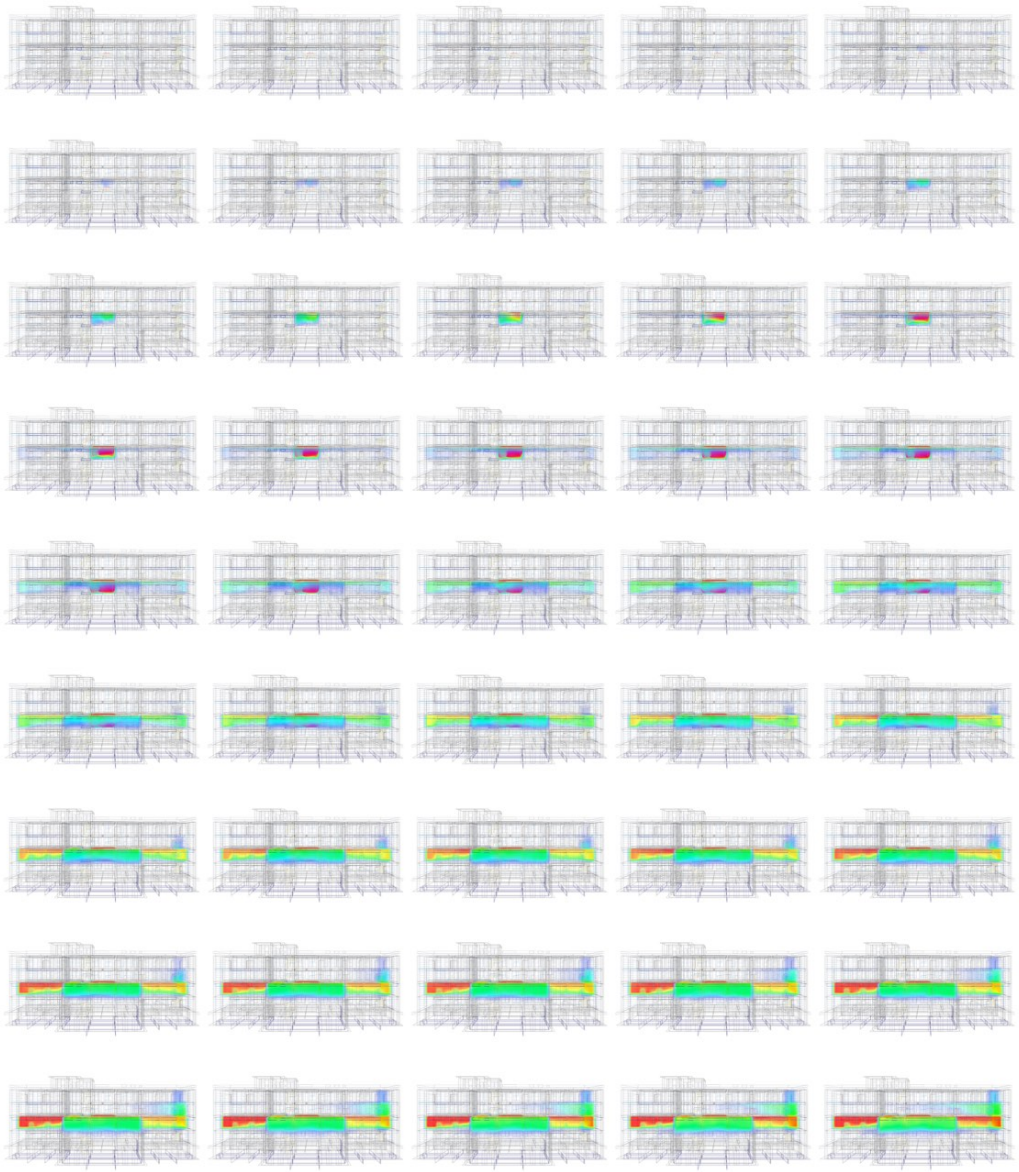


Figure D.23: Office physical field visualizations. Frames sampled across the sequence.

### D.4 Residential Room Fire

The residential room scenario is reserved for evaluating model generalization. Its computational domain measures $11 \times 12 \times 4.5$ m $(L \times W \times H)$ and contains 594,000 cells.

The following views show the room layout and exterior geometry; domain dimensions are indicated schematically by dashed frames and dimension lines.

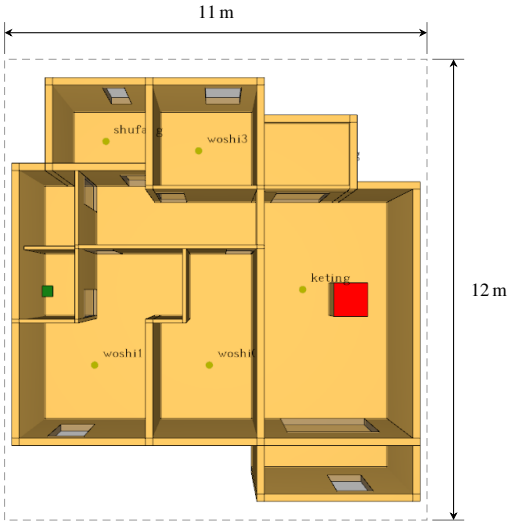



Figure D.24: Residential room model: top view.

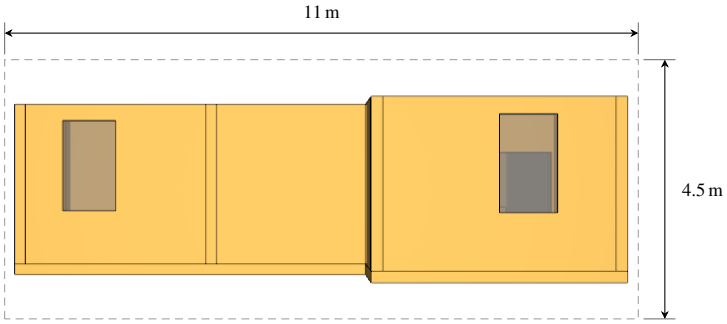



Figure D.25: Residential room model: front view.

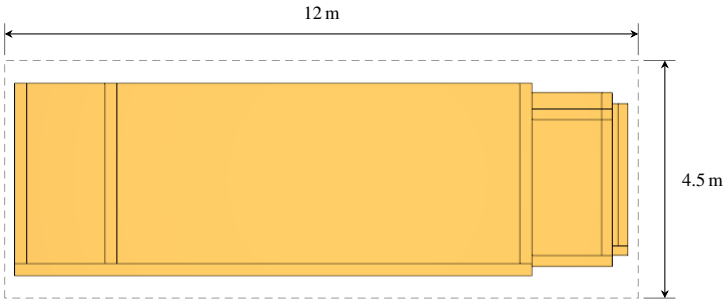



Figure D.26: Residential room model: right view.

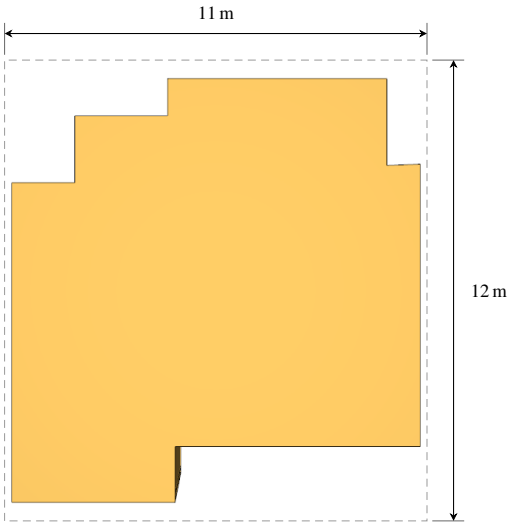



Figure D.27: Residential room model: bottom view.

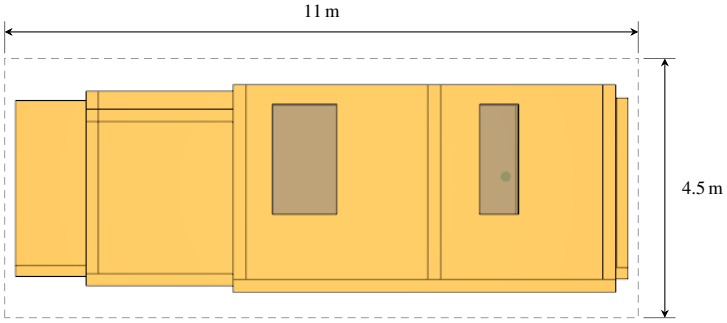



Figure D.28: Residential room model: rear view.

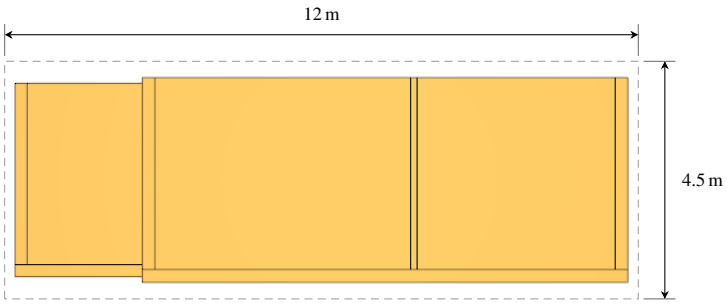



Figure D.29: Residential room model: left view.

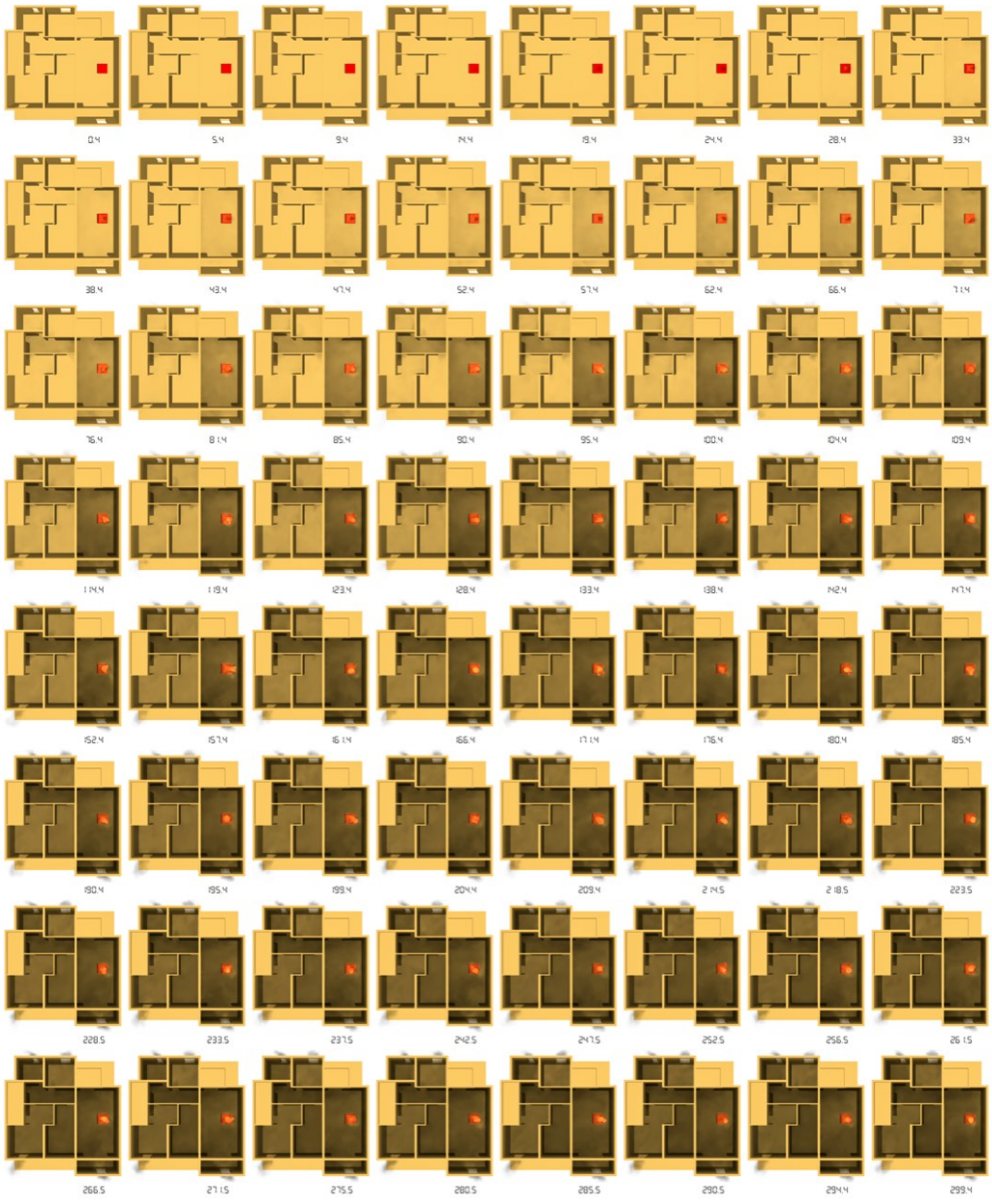


Figure D.30: Residential room scene renderings. Frames sampled across the sequence.

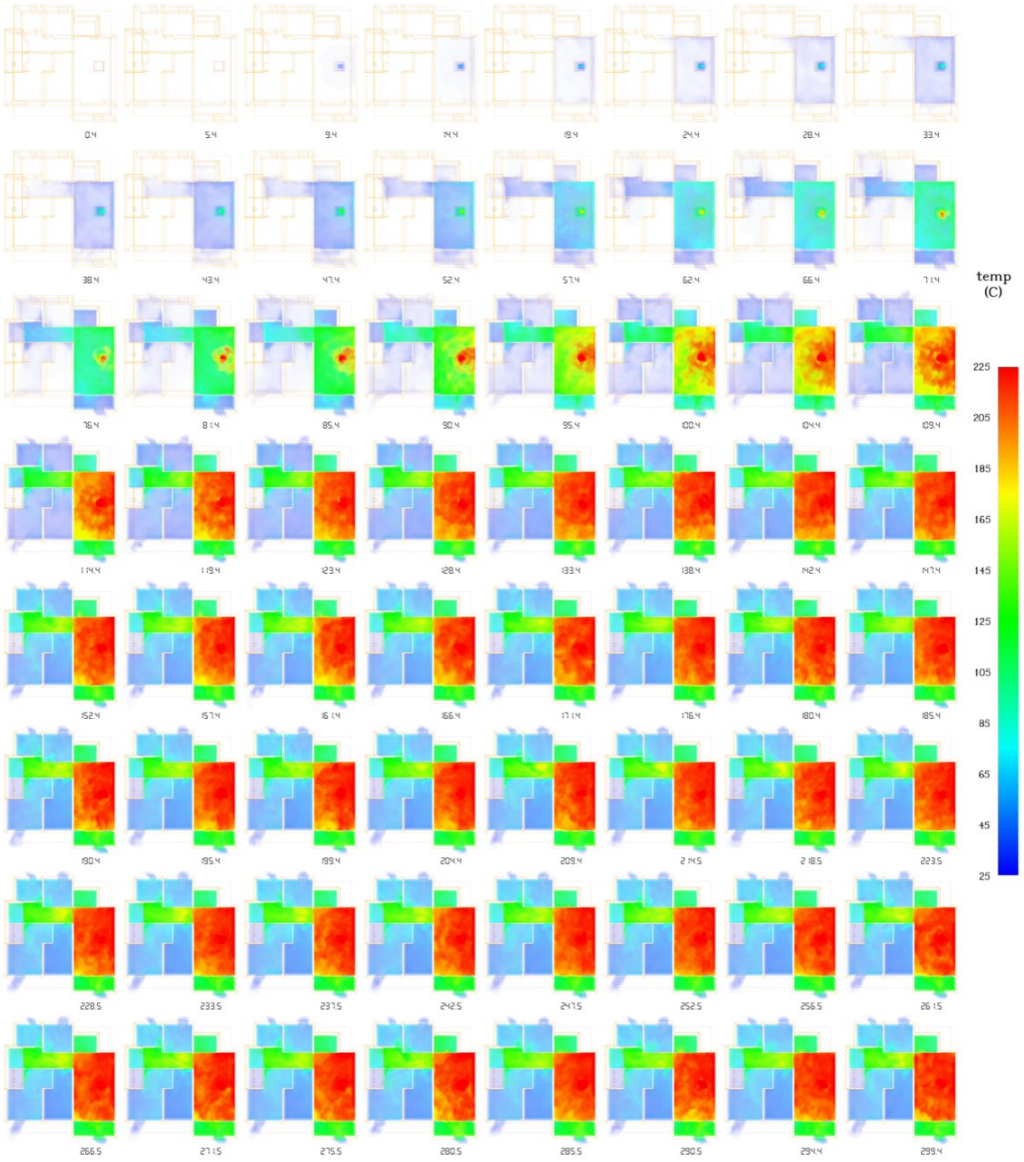



Figure D.31: Residential room physical field visualizations. Frames sampled across the sequence.

### D.5 SHOPPING MALL FIRE

The shopping mall scenario is reserved for evaluating model generalization. Its computational domain measures $46 \times 23 \times 58$ m $(L \times W \times H)$ and contains 169,360 cells.

The following views document the building geometry; domain dimensions are indicated schematically by dashed frames and dimension lines.

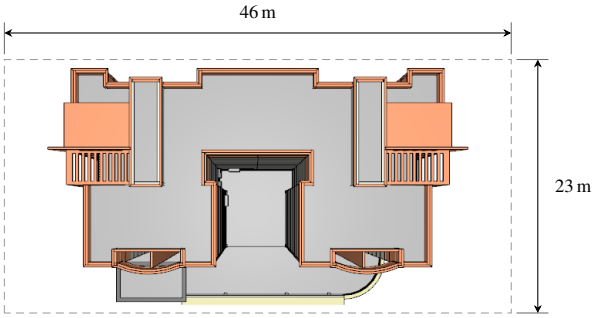



Figure D.32: Shopping mall model: top view.

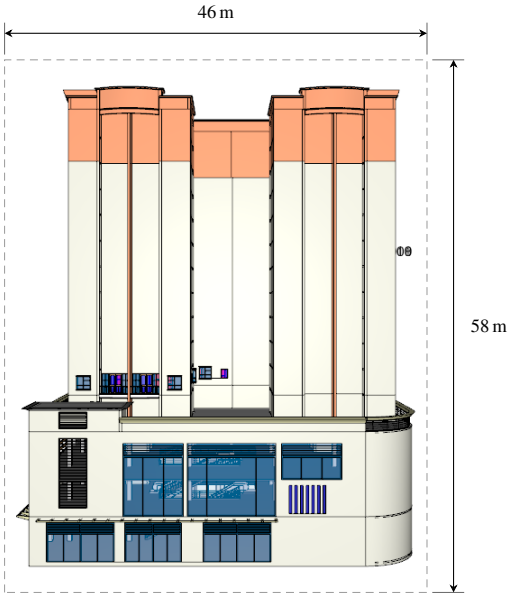



Figure D.33: Shopping mall model: front view.

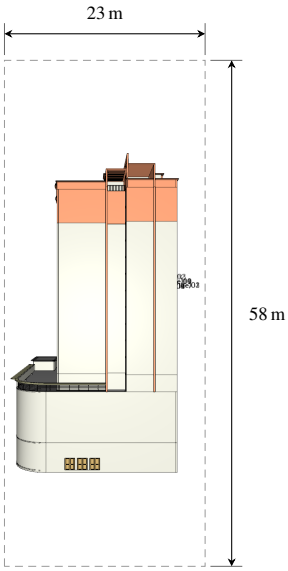



Figure D.34: Shopping mall model: right view.

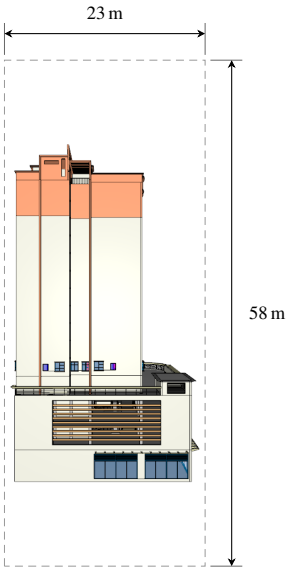



Figure D.35: Shopping mall model: left view.

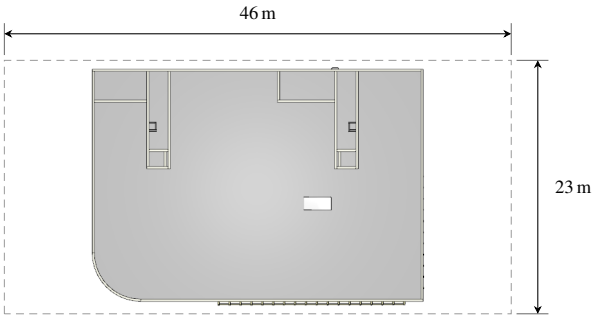



Figure D.36: Shopping mall model: bottom view.

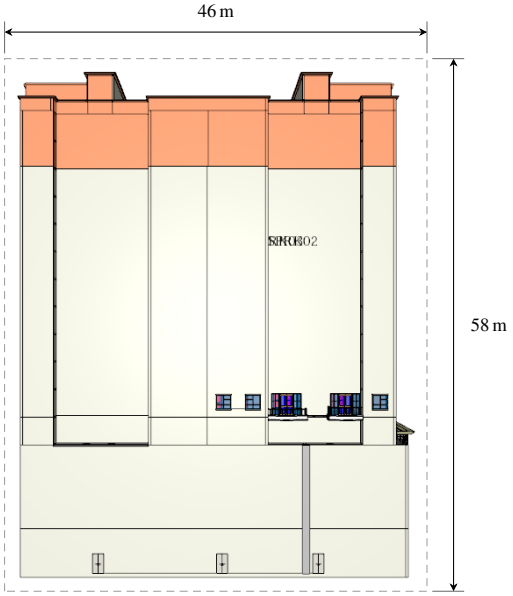



Figure D.37: Shopping mall model: rear view.



Figure D.38: Shopping mall scene renderings. Frames sampled across the sequence.




Figure D.39: Shopping mall physical field visualizations. Frames sampled across the sequence.

### D.6 TEMPLE FIRE

The temple scenario is reserved for evaluating model generalization. Its computational domain measures $16 \times 16 \times 12$ m $(L \times W \times H)$ and contains 112,360 cells.

The following views document the building geometry; domain dimensions are indicated schematically by dashed frames and dimension lines.

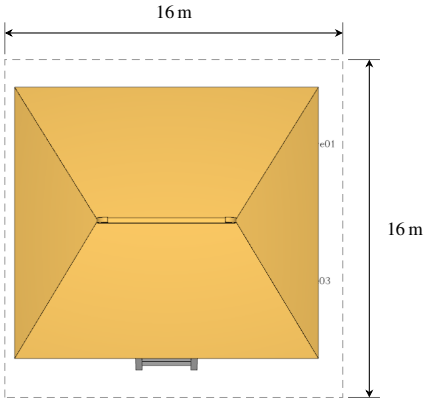



Figure D.40: Temple model: top view.

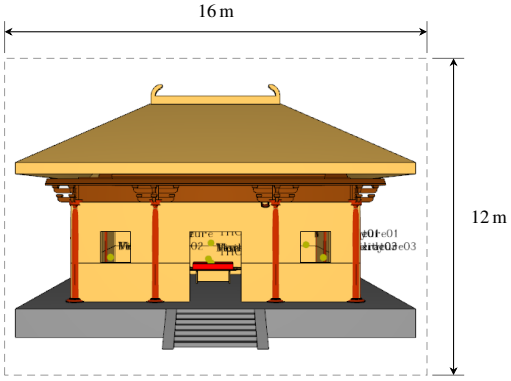



Figure D.41: Temple model: front view.

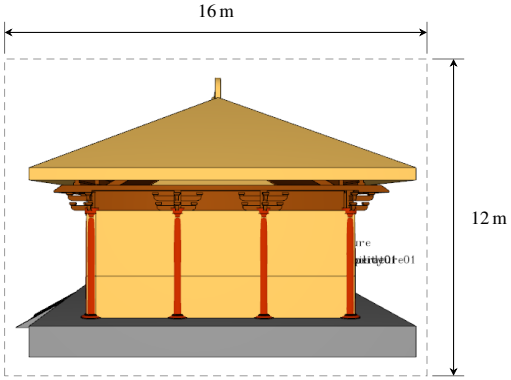



Figure D.42: Temple model: right view.

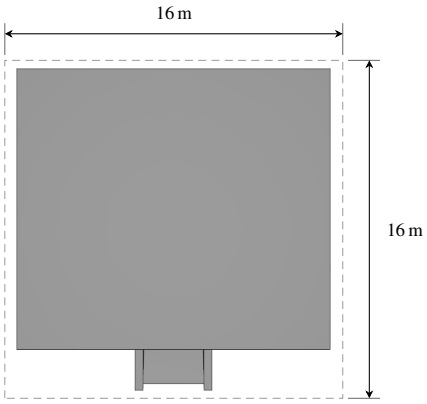



Figure D.43: Temple model: bottom view.

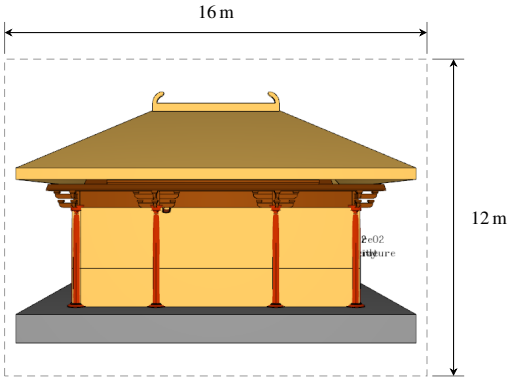



Figure D.44: Temple model: rear view.

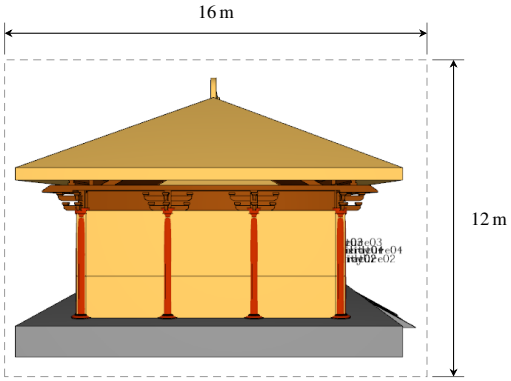



Figure D.45: Temple model: left view.

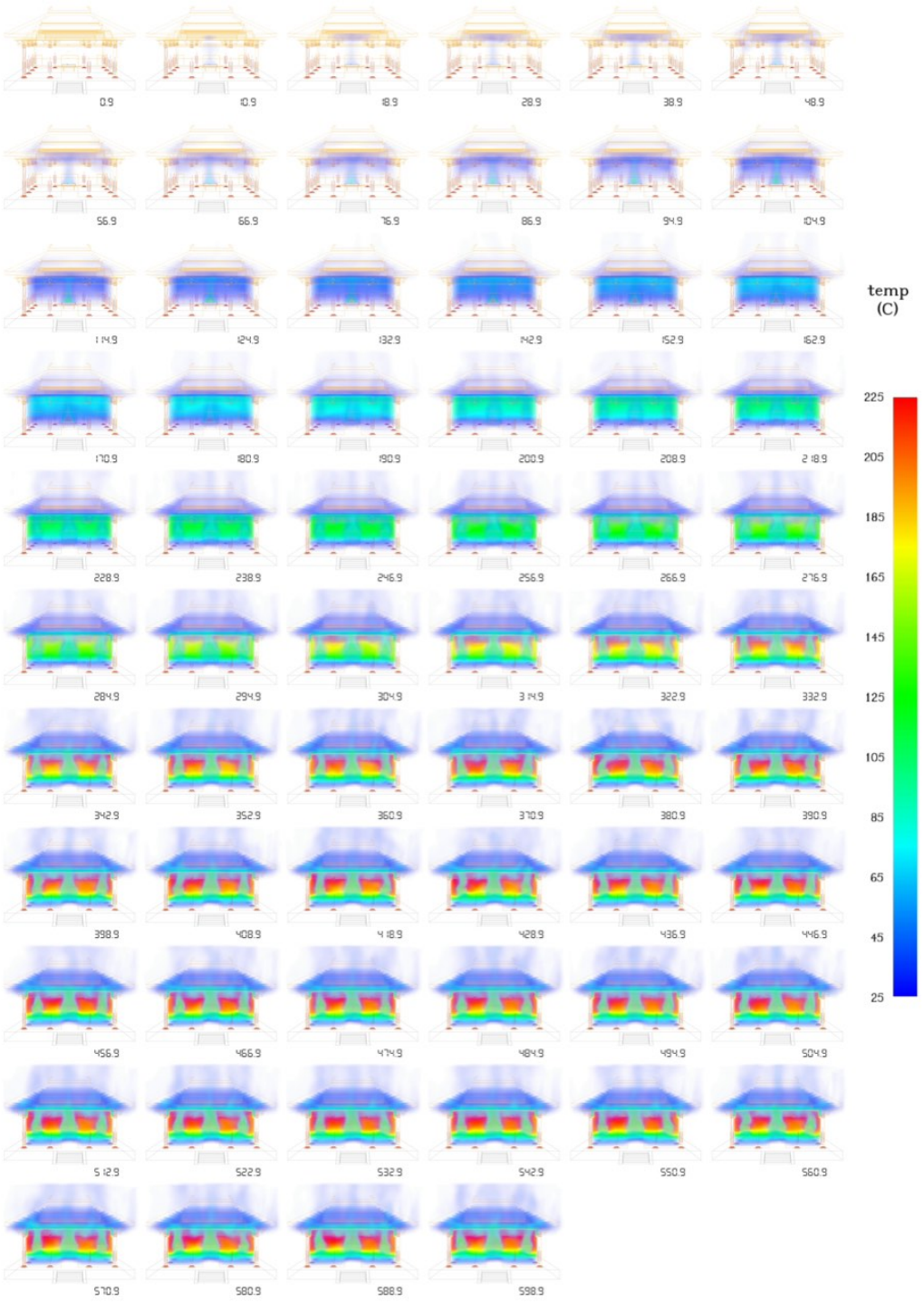



Figure D.46: Temple physical field visualizations. Frames sampled across the sequence.

# E Experimental Details

## E.1 Training and Inference Settings

VTF fine tunes Qwen2.5-VL-7B-Instruct with a four bit LoRA adapter of rank 16, alpha 32, and dropout 0.05 for three epochs on two NVIDIA H800 80GB GPUs. AdamW uses a $1 \times 10^{-5}$ learning rate, cosine decay, 0.03 warmup, and 0.01 weight decay. One sample per device and four accumulation steps give a global batch size of 8. Training uses FP16, gradient checkpointing, a sequence limit of 2,048 tokens, an image limit of 50,176 pixels, and seed 20270720. Inference uses one GPU, FP16, and greedy decoding. The checkpoint has the lowest Tunnel validation loss.

## E.2 Real Fire Evaluation Protocol

**Data and Checkpoint.** We evaluate real camera images from NIST's full scale Station Nightclub mockup experiment (Madrzykowski et al., 2006; Grosshandler et al., 2005). VTF uses the checkpoint trained on tunnel simulation data directly, without fine tuning on real fire data.

**Inputs and Temporal Access.** The zero-shot Qwen2.5-VL baseline receives the question and real scene images. VTF additionally receives field visualizations derived from simulations and aligned with the timestamps of the observed images. No field from a future timestamp is supplied. The fields are auxiliary simulation evidence, rather than measured physical fields recovered from the camera images.

**Tasks and Label Sources.** P1/P2 evaluate the observed physical configuration against distractor alternatives, including incorrect fire source locations. These distractors are answer alternatives, not additional independently conducted fire experiments. E1/E2 labels are derived from real experimental measurements documented in the official NIST technical report (Grosshandler et al., 2005), rather than from the simulation fields supplied to VTF. This experiment evaluates transfer to real camera observations with auxiliary simulation evidence.