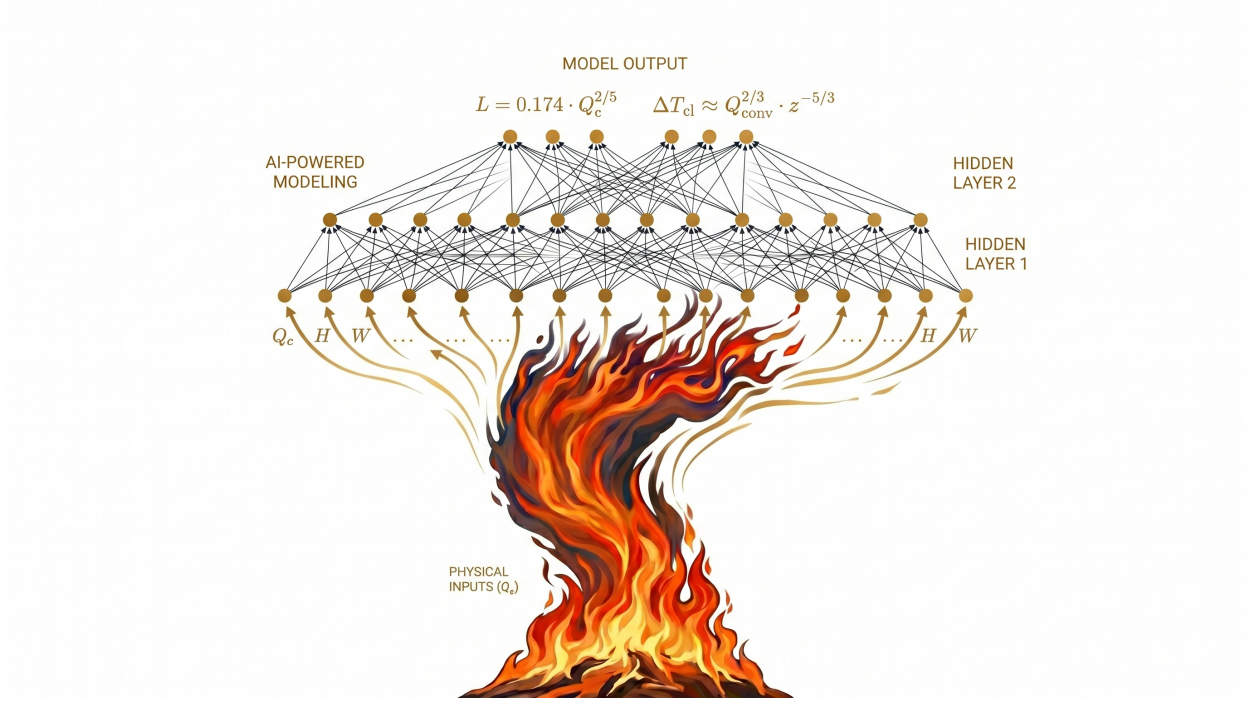
CHALMERS
UNIVERSITY OF TECHNOLOGY



Operationalizing Physics-Informed Deep Learning in Fire Safety

A Robust Design and Product Lifecycle Approach

Master's thesis in Quality and Operations Management
Master's thesis in Mobility Engineering

LIKHITH RUDRESH
JAKKA SAI SRINIVASA MANIDEEP

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

Operationalizing Physics-Informed Deep Learning in Fire Safety

A Robust Design and Product Lifecycle Approach

Likhith Rudresh
Jakka Sai Srinivasa Manideep

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mechanics and Maritime Sciences
Division of fluid dynamics
Chalmers University of Technology
Gothenburg, Sweden 2026

Operationalizing Physics-Informed Deep Learning in Fire Safety
A Robust Design and Product Lifecycle Approach
Likhith Rudresh and Jakka Sai Srinivasa Manideep

Supervisor: Johan Anderson and Erdzan Hodzic, RISE
Examiner: Håkan Nilsson, Department of Mechanics and Maritime Sciences

Master's Thesis 2026
Department of Mechanics and Maritime Sciences
Division of fluid dynamics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: A conceptual scientific illustration demonstrating the synthesis of traditional thermodynamics and machine learning. A fire plume, defined by classical equations, rises into neural network and predictive loss graphs, representing the evolution of computational fire modeling

Typeset in LaTeX Gothenburg, Sweden
2026

Operationalizing Physics-Informed Deep Learning in Fire Safety
A Robust Design and Product Lifecycle Approach
Likhith Rudresh and Jakka Sai Srinivasa Manideep
Department of Mechanics and Maritime Sciences
Chalmers University of Technology

Abstract

Measuring fire heat release rate (HRR) is central to fire safety engineering. In tunnels and facades, however, conventional oxygen-consumption calorimetry is often impractical due to geometric constraints, harsh thermal environments, and sensor degradation caused by smoke and flame. Because surveillance cameras are already widely installed in tunnel infrastructure, this thesis develops VisionKoopmanNet, an image-based deep learning pipeline that estimates HRR directly from optical video and embeds these predictions within a structured operational framework for safety-critical evaluation.

The technical pipeline is built in three phases. Phase 1 trains a physics-informed Koopman-LSTM HRR encoder on 756 Fire Dynamics Simulator (FDS) scenarios and reaches $R^2 = 0.9996$ on 1.73 million held-out test samples. Phase 2 trains an EfficientNet-B7 image branch on the NIST Fire Calorimetry Database using an experiment-level stratified-by-peak-HRR train/val/test split (391/84/82 experiments) and a proportional frame-to-HRR index mapping that corrects a hardcoded `fps`=5.0 assumption that had silently misaligned every window with pre-ignition baseline samples in earlier work. Two complementary Phase 2 results are reported. The *image-only* HRR estimator, representing the operationally relevant capability for tunnel and facade deployment where calorimetry sensors are unavailable, predicts absolute HRR from camera video alone with no sensor input at inference and achieves held-out test MAE = 251 kW on 50,720 windows drawn from 82 stratified-held-out NIST experiments, with sub-150 kW MAE in the 0–200 kW regime (n = 34,177 windows, 67% of the test set) and documented failure modes on rare megawatt-scale fires (n = 361 at >5 MW). The *sensor-fused* forecaster, which takes the most recent calorimetry reading as an additive anchor in log-target space alongside image features, achieves test MAE = 8.93 kW but matches a multi-step persistence baseline (8.88 kW) within 0.5% across every per-bin slice tested; empirically, the decoder learns $\Delta = 0$ for every input and the prediction collapses to copying the sensor reading. This residual-anchor trap survives every architectural protection investigated (log-target regression, hard spectral-radius projection, end-to-end unfrozen encoder, stratified split, active-fire weighted sampler, magnitude-weighted MAE) and is reported as a structural information-theoretic finding rather than a tuning failure. Phase 3 attempts to fuse the two encoders into a shared Koopman latent space; three alignment strategies (mean-squared-error on unpaired data, joint retraining, contrastive InfoNCE) all fail to produce per-sample cross-modal correspondence, with three documented failure modes (no improvement, catastrophic forgetting, and representation collapse). The Phase 3 negative result, traced to a data-level cause (namely that the FDS scenarios and NIST experiments represent different physical fires rather than the same fire observed across two modalities), is reported as a primary scientific contribution.

Alongside technical pipeline development, this thesis answers two research questions on safety-critical operationalization. The first investigates how Critical-to-Quality (CTQ) requirements, Operational Design Domains (ODD), and Stage-Gate lifecycle controls can be designed to ensure rigorous specification, workflow compatibility, and robustness for an offline fire prediction prototype. The second investigates how Failure Mode and Effects Analysis (FMEA), safety wrappers, and human-in-the-loop escalation rules can be structured to govern validation and risk management prior to operational deployment.

The primary contributions of this thesis are fourfold: (1) A working camera-based image-only HRR estimator (MAE = 251 kW on held-out NIST testing, sub-150 kW for fires up to 200 kW) suitable for tunnel and facade scenarios where physical calorimetry cannot be installed; (2) A structural negative finding demonstrating the residual-anchor trap in sensor-fused HRR forecasting; (3) An empirical investigation into cross-modal latent alignment without paired multi-modal data, identifying three distinct failure modes in cross-modal learning for fire dynamics; and (4) An end-to-end operationalization framework combining CTQ metrics, ODD delimitation, Stage-Gate quality gates, FMEA risk prioritization, and runtime safety wrappers to govern bounded prototype safety. A separate methodological contribution is the documentation of a silent hardcoded frame-rate alignment error that previously inflated Phase 2 metric scores, together with the multi-iteration rebuild that corrected it.

The scope is strictly bounded to offline analysis of recorded NIST video sequences; real-time edge deployment, multi-camera fusion, hardware procurement, graphical user interfaces, and formal compliance certification remain out of scope for future work.

Keywords: fire safety engineering, machine learning operationalization, requirements engineering, Stage-Gate, robustness engineering, FMEA, operational design domain, validation, human-in-the-loop, hybrid ML-FDS workflow.

Acknowledgements

The authors gratefully acknowledge Johan Anderson and Erdzan Hodzic at RISE Fire and Safety for their guidance, support, and feedback throughout this thesis. The authors also thank Håkan Nilsson for his role as examiner.
Gratitude is also extended to Chalmers University of Technology and all interview and workshop participants who contributed their time and expertise.
Finally, sincere thanks are given to family, friends, and colleagues for their encouragement during the work on this thesis.

Likhith Rudresh, Jakka Sai Srinivasa Manideep
Gothenburg, March 2026

List of Acronyms

AI	Artificial Intelligence
CFD	Computational Fluid Dynamics
CTQ	Critical-to-Quality
DL	Deep Learning
FDS	Fire Dynamics Simulator
FMEA	Failure Mode and Effects Analysis
GUI	Graphical User Interface
HITL	Human-in-the-Loop
HRR	Heat Release Rate
IoU	Intersection over Union
ML	Machine Learning
NPI	New Product Introduction
ODD	Operational Design Domain
QA	Quality Assurance
RE	Requirements Engineering
RGB	Red-Green-Blue
RISE	Research Institutes of Sweden
RPN	Risk Priority Number
RQ	Research Question
SOP	Standard Operating Procedure
UNet	U-shaped convolutional neural network architecture

Nomenclature

Symbol / Term	Definition
C_{pk}	Process capability index used for bounded capability-style interpretation.
S	Severity rating in FMEA.
O	Occurrence rating in FMEA.
D	Detectability rating in FMEA.
RPN	Risk Priority Number, calculated as $S \times O \times D$.
IoU	Intersection over Union.
HRR	Heat Release Rate.
500 ms	Provisional latency threshold.
85%	Provisional IoU threshold.
99.5%	Provisional reliability threshold.

Contents

List of Acronyms **ix**

Nomenclature **xi**

1 Introduction **1**
1.1 Research Questions . . . 3
1.2 Scope and Limitations . . . 3
1.3 Thesis Organisation . . . 4

2 Background and Related Work **5**
2.1 Requirement Engineering, CTQs, and ODD . . . 5
2.2 Safety, Lifecycle Control, Robustness, and Risk Governance . . . 5
2.3 Hybrid Workflow, Monitoring, and Validation . . . 6
2.4 The ML Pipeline: Technical Description . . . 6
2.4.1 Fire Dynamics Simulator and Dataset Generation . . . 7
2.4.2 Koopman Operator and KoopmanLSTM . . . 7
2.4.3 VisionKoopmanNet: Modality-Agnostic Koopman Observer . 8
2.4.4 NIST Fire Calorimetry Database . . . 8
2.4.5 Symbolic Regression for Physics Discovery . . . 9

3 Research Method **11**
3.1 Research Design . . . 11
3.2 Preparation for Data Collection . . . 11
3.2.1 Sampling . . . 11
3.3 Iterative Model Development . . . 17
3.3.1 Phase 1: HRR Encoder Iterations . . . 17
3.3.1.1 v1 (legacy). . . . 17
3.3.1.2 v2 (residual prediction). . . . 17
3.3.1.3 v3 (absolute prediction). . . . 18
3.3.1.4 v4 (multi-task auxiliary head). . . . 18
3.3.1.5 v5 (long horizon, cancelled). . . . 18
3.3.1.6 v6 (physics-aware Koopman, regressed). . . . 18
3.3.2 Phase 2: Image Branch Iterations . . . 19
3.3.2.0.1 Note on the pre-rebuild numbers in this subsection. . . . 19
3.3.2.1 v3 (broken). . . . 19
3.3.2.2 v4 (residual prediction). . . . 19

3.3.2.3 v6 (best standalone). 20
3.3.2.4 v8 (with alignment teacher). 20
3.3.2.5 v9 (final pre-rebuild version). 20
3.3.2.6 Post-rebuild (pure-vision Koopman, log-target). . . . 20
3.3.3 Phase 3: Joint Cross-Modal Alignment Iterations 21
3.3.3.1 v1 (MSE alignment, HRR frozen). 21
3.3.3.2 v2 (paired NIST, joint retraining, contrastive). 21
3.3.3.3 v3 (native 1-channel encoder, contrastive). 21
3.3.3.3.1 Selected configuration. 21
3.4 Validation 22

4 Results **23**
4.1 Results for RQ1 23
4.1.1 Phase 1 - Requirement Engineering and CTQ Definition . . . 23
4.1.2 Phase 2 - Stage-Gate Controlled Development 26
4.1.2.1 Stage-Gate Mapping to ML Pipeline Development . 27
4.1.3 Phase 3 - Workflow Integration and Hybrid ML-FDS Strategy 28
4.1.4 Phase 4 - Robustness Analysis and Parameter Design 30
4.2 Results for RQ2 32
4.2.1 Phase 5 - FMEA-Based Risk Management and Safety Assurance 32
4.2.2 Phase 6 - Validation and Process Capability Assessment . . . 33
4.2.2.1 ML Pipeline Technical Results 34
4.2.2.1.1 Motivation and the central objective. 35
4.2.2.1.2 Scope note on camera views. 35
4.2.2.1.3 Headline result. 35
4.2.2.1.4 Phase 1: supporting HRR encoder. 37
4.2.2.2 Phase 2 results: HRR from camera imagery (central result) 40
4.2.2.2.1 NIST data split and validation methodology. 40
4.2.2.2.2 Frame-to-HRR alignment correction. 41
4.2.2.2.3 Image-only HRR estimation: operational headline. 41
4.2.2.2.4 Image-only iteration history: from broken pipeline to current result. 43
4.2.2.2.5 Sensor-fused forecasting: structural negative result. 44
4.2.2.2.6 Interpretation: the residual-anchor trap is structural. 45
4.2.2.2.7 Summary of the two Phase 2 results. 45
4.2.2.3 Phase 3 results: cross-modal physics-transfer attempt 46
4.2.2.3.1 Selected configuration. 47
4.2.2.3.2 Per-sample alignment analysis. 49
4.2.2.3.3 Interpretation of the alignment failure. . . . 50
4.2.2.4 Validation Gate Assessment and Process Capability . 54
4.2.2.4.1 Gate outcomes summary. 54
4.2.2.4.2 Predictive accuracy. 54

4.2.2.4.3 Confidence calibration. 55
4.2.2.4.4 Safety-wrapper operational demonstration. . 55
4.2.2.4.5 Visual demonstration across fire regimes: four illustrative NIST windows. 56
4.2.2.4.6 Physics consistency. 58
4.2.2.4.7 Pipeline robustness. 58
4.2.2.4.8 Process capability assessment. 58

5 Discussion **61**
5.1 Discussion and Main Findings 61
5.2 Key Lessons Learned 62
5.3 Relation to Literature 65
5.4 Implications for Research 66
5.5 Implications for Practice 67
5.6 Validity and Ethical Considerations 68
5.6.1 Internal Validity 68
5.6.2 Construct Validity 68
5.6.3 External Validity 69
5.6.4 Reliability 69
5.6.5 Conclusion Validity 69
5.6.6 Informed Consent and Confidentiality 70

6 Conclusion **71**
6.1 Answer to RQ1 71
6.2 Answer to RQ2 71
6.3 Final Contribution 72
6.4 Limitations and Future Work 73
6.5 Open Questions 75

Bibliography **77**

A Interview Guide **I**

1
Introduction

Managing fire and smoke presents fundamental engineering challenges in both academic research and safety practice. Fire behavior is nonlinear, chaotic, and highly sensitive to variables such as room geometry, fuel chemistry, ventilation, and thermal feedback. Fire safety engineers require reliable estimates of fire spread rates, heat release rates, and smoke volume when conducting hazard analysis, design, risk assessment, and decision-making. The NIST Fire Dynamics Simulator (FDS) provides high-fidelity numerical simulation for compartment fires; however, its substantial computational expense often makes it impractical for rapid screening, iterative scenario evaluation, or time-sensitive operational support.

Recent advances in machine learning offer opportunities to accelerate parts of this workflow. Recorded multi-camera fire experiment video, combined with physics-informed learning methods, provides a foundation for estimating fire dynamics including heat release rate, flame spread, and smoke development. Rather than seeking to replace FDS, this study investigates whether a machine learning pipeline can serve as a bounded first-level screening tool within a tiered workflow. In this setup, routine or low-risk cases are screened rapidly, while high-consequence, uncertain, or out-of-domain scenarios are escalated to detailed FDS analysis.

The technical pipeline evaluated in this thesis is VisionKoopmanNet, which pairs a physics-informed Koopman LSTM for heat release rate time-series prediction with an EfficientNet-B7 visual backbone for experimental fire imagery. The Koopman operator framework models nonlinear fire dynamics through linear evolution in a learned latent space, with cross-modal alignment between FDS simulations and real experimental video from the NIST Fire Calorimetry Database serving as a potential transfer pathway. This work remains a research prototype and does not claim field validation, live deployment, or formal safety certification.

The primary objective of this thesis is to examine how such a machine learning prototype can be operationalized in a structured, safety-critical engineering context. In safety-relevant domains, statistical predictive accuracy alone is insufficient; explicit confidence communication, data traceability, escalation logic, and human oversight are equally vital. This study addresses operational challenges by defining bounded usage limits, structuring development through explicit Stage-Gate criteria, quantifying failure modes through Failure Mode and Effects Analysis (FMEA), and producing verifiable evidence of bounded operational readiness.

Conducted in collaboration with RISE Fire and Safety, this research is strictly bounded to offline analysis of recorded multi-camera fire experiment video within a research-support setting. It is not intended for unconstrained field use or production

deployment. Within this context, the thesis develops and evaluates an operational framework spanning requirements engineering, Stage-Gate lifecycle control, hybrid workflow integration, robustness testing, risk governance, and empirical validation, contributing transferable methodological guidance for safety-critical machine learning systems.

To address these aims, the thesis is organised around two research questions within a six-phase operational framework. The first research question focuses on how Critical-to-Quality requirements, operational constraints, and a Stage-Gate controlled development process can be designed and applied to support specification, workflow fit, and robustness for the prototype pipeline. The second research question focuses on how risk management approaches such as Failure Mode and Effects Analysis, safety wrappers, human-in-the-loop rules, benchmarking, and validation methods can be structured to support future deployment decisions.
The thesis combines an operationalisation study with a complete technical pipeline. The image branch (Phase 2) and the joint cross-modal training (Phase 3) were trained, evaluated on held-out experimental data, and reported in this document. The NIST validation uses an experiment-level stratified-by-peak-HRR train/val/test split (391 / 84 / 82 experiments; frames from the same experiment never appear in more than one split, eliminating frame-level leakage). Four primary technical findings are reported in this thesis. (i) The Phase 1 HRR encoder reaches $R^2 = 0.9996$ and MAE $\approx$ 208 kW on 1.73 million FDS test samples. (ii) The Phase 2 image branch, retrained from scratch after a proportional frame-to-HRR index mapping replaced an earlier hardcoded `fps`=5.0 assumption that had silently misaligned every window with pre-ignition baseline samples, supports image-only HRR estimation with no calorimetry sensor at inference: held-out test MAE = 251 kW on 50,720 windows from 82 stratified-unseen NIST experiments, with sub-150 kW MAE in the 0–200 kW regime (n = 34,177 windows, 67% of the test set), and beats a predict-the-mean baseline in five of seven HRR magnitude bins. Documented degradation appears on rare megawatt-scale windows (1447 kW MAE on 2233 windows in 1–5 MW; 7055 kW MAE on 361 windows above 5 MW), attributed jointly to data sparsity and a log-target evaluation artefact. (iii) A complementary sensor-fused forecaster, which accepts the most recent calorimetry reading as an additive anchor in log-target space alongside image features, achieves test MAE 8.93 kW but matches the trivial multi-step persistence baseline (8.88 kW) within 0.5 % across every per-bin slice; the decoder learns to output a zero delta for every input, so the prediction collapses to copying the sensor reading. This residual-anchor trap survives log-target reformulation, hard spectral-radius projection, end-to-end unfrozen encoder, stratified split, an active-fire weighted sampler, and magnitude-weighted MAE, and is reported as a structural information-theoretic finding rather than a tuning failure. (iv) The Phase 3 attempt to enforce per-sample cross-modal latent alignment fails across three distinct strategies; paired-sample latent distances are statistically indistinguishable from randomly shuffled pairs, with three documented failure modes (no improvement, catastrophic forgetting, and representation collapse). The Phase 3 negative result, together with its data-level explanation that the FDS and NIST experiments are different physical fires rather than the same fire observed in two modalities, is reported as a primary scientific contribution of the thesis rather than a residual

limitation.

1.1 Research Questions

RQ1: How can CTQ requirements, operational constraints, and a Stage-Gate controlled development process be designed and applied to an offline ML-based fire prediction pipeline to ensure specification, workflow fit, and robustness in a safety-critical context?

RQ2: How can FMEA, safety wrappers, human-in-the-loop rules, and validation methods be structured for a non-deployed ML-based fire prediction pipeline to generate safety-relevant evidence and governance guidelines for future deployment decisions?

1.2 Scope and Limitations

The scope of this thesis is explicitly bounded to the offline analysis of recorded multi-camera fire experiment video and corresponding simulation data within a research-support context. The investigation focuses on developing, operationalizing, and validating a physics-informed machine learning pipeline (VisionKoopmanNet) using the NIST Fire Calorimetry Database and numerical simulations from the Fire Dynamics Simulator (FDS). The primary objective is to establish the engineering requirements, Stage-Gate quality controls, robustness parameters, safety wrappers, and validation methodologies that must govern such a prototype before it can be responsibly integrated into professional engineering workflows.
To maintain a rigorous and feasible research boundary, the thesis deliberately excludes several operational and deployment dimensions:

- **Live real-time deployment:** The prototype is evaluated on pre-recorded offline video sequences; real-time edge processing and active sensor-streaming integrations are not implemented.
- **Hardware and physical sensor procurement:** No physical camera hardware, calorimetry sensors, or laboratory instrumentation were designed or procured as part of this study.
- **IT infrastructure and enterprise rollout:** Cloud deployment architectures, distributed data ingestion pipelines, and enterprise-wide software rollouts are out of scope.
- **Graphical user interface (GUI) development:** The focus remains on core algorithmic, architectural, and governance mechanisms rather than front-end end-user software design.
- **Formal certification and compliance:** The system is treated as a research prototype; no formal safety-integrity certification (e.g., IEC 61508 or UL compliance) is claimed.
- **Redesign of underlying baseline architectures:** The study adapts established deep learning backbones (EfficientNet-B7, LSTM, and Koopman operators) rather than inventing unrelated neural network primitives.

1.3 Thesis Organisation

This thesis is structured into six chapters that systematically document the operationalisation framework and empirical technical results:

- **Chapter 1 (Introduction):** Introduces the operational challenge of fire heat release rate estimation, motivates the machine learning approach, articulates the two research questions (RQ1 and RQ2), and defines the scope and boundary limitations.
- **Chapter 2 (Background and Related Work):** Establishes the theoretical foundation, reviewing requirements engineering for machine learning, Operational Design Domains, Stage-Gate lifecycle frameworks, FMEA risk management, Koopman operator theory, and vision-based fire analysis.
- **Chapter 3 (Research Method):** Details the design science research methodology, stakeholder interview protocols, CTQ derivation, ODD boundary mapping, and the chronological iteration history across Phases 1, 2, and 3.
- **Chapter 4 (Results):** Presents the empirical findings answering RQ1 (Phases 1–4: CTQs, stage-gates, hybrid workflow, and robustness analysis) and RQ2 (Phases 5–6: FMEA matrix, safety wrapper logic, process capability, and technical benchmark evaluations).
- **Chapter 5 (Discussion):** Discusses the primary technical findings, details eight reusable architectural lessons learned, connects the contributions to existing literature, outlines practical implications, and evaluates research validity and ethical considerations.
- **Chapter 6 (Conclusion):** Synthesizes the core answers to RQ1 and RQ2, summarizes the four primary scientific contributions, documents open research questions, and outlines prioritized directions for future work.

2
Background and Related Work

2.1 Requirement Engineering, CTQs, and ODD

This project uses requirements engineering to translate stakeholder needs into traceable, testable system specifications for a safety-critical machine learning prototype, rather than defining a fully operationalized production system (ISO/IEC/IEEE, 2018; Ahmad et al., 2023; Vogelsang & Borg, 2019). The resulting specifications encompass functional performance, operating constraints, trust conditions, and explicit escalation thresholds (ISO/IEC/IEEE, 2018; Ahmad et al., 2023; Vogelsang & Borg, 2019).

Critical-to-Quality (CTQ) metrics operationalize these expectations into quantifiable targets for prototype evaluation (ISO/IEC/IEEE, 2018; Ahmad et al., 2023; Vogelsang & Borg, 2019). The prototype is assessed across five CTQ dimensions: latency, predictive quality, confidence output, traceability, and workflow compatibility.

The Operational Design Domain (ODD) defines the environmental and operational boundaries within which the prototype can be trusted (ISO/IEC/IEEE, 2018; Ahmad et al., 2021; Vogelsang & Borg, 2018). It specifies caution boundaries and triggers mandatory escalation to Fire Dynamics Simulator (FDS) modeling when inputs fall outside evaluated conditions, ensuring conservative integration into the broader engineering workflow (see Fig. 2.1).

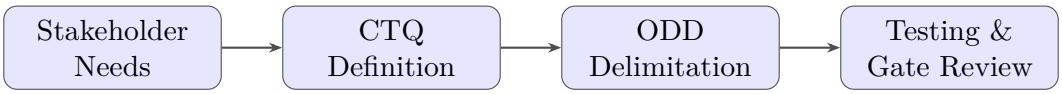


Figure 2.1: Requirements Engineering Process adapted to safety-critical ML pipelines.

2.2 Safety, Lifecycle Control, Robustness, and Risk Governance

Stage-Gate logic provides a structured approach for guiding development through defined decision gates rather than informal progression (Cooper and Sommer, 2016; ISO/IEC/IEEE, 2018). Advancement to subsequent phases requires verifiable evidence that meets predefined Critical-to-Quality criteria, ensuring a controlled, transparent development lifecycle (Cooper and Sommer, 2016; ISO/IEC/IEEE, 2018).

Robustness engineering evaluates prototype resilience against operational noise factors. Drawing on Montgomery (2017), Ahmad and colleagues (2023), and Vogelsang and Borg (2019), primary environmental disturbances that can degrade performance include resolution degradation, compression artifacts, illumination shifts, modality mismatch, and scenario variations. Prototype reliability must therefore be verified across both nominal conditions and controlled perturbations.

Failure Mode and Effects Analysis (FMEA) prioritizes failure modes including false negatives, overconfident incorrect outputs, out-of-domain operation, logging failures, and physics-inconsistent predictions (Rausand and Høyland, 2004; IEC, 2010; Borg et al., 2023). Safety wrappers and human-in-the-loop escalation rules enforce conservative fail-safe behavior when these conditions occur, ensuring the prototype functions as a decision-support aid rather than an autonomous decision-making system (Borg et al., 2023; IEC, 2010).

2.3 Hybrid Workflow, Monitoring, and Validation

This study implements a tiered hybrid workflow: the machine learning pipeline handles rapid first-tier screening, while Fire Dynamics Simulator (FDS) is reserved for high-fidelity verification (McGrattan et al., 2021; Cooper & Sommer, 2016; Ahmad et al., 2023). This structure leverages the computational speed of deep learning without sacrificing the physical fidelity of numerical CFD simulation in critical evaluations (Borg et al., 2023; ISO/IEC/IEEE, 2018; Vogelsang & Borg, 2019).

Monitoring and calibrating prediction confidence determines whether individual model outputs are accepted, flagged for secondary review, or escalated for independent CFD validation (Heyn et al., 2023; Borg et al., 2023; Longo et al., 2014). System validation therefore extends beyond aggregate test accuracy to encompass bounded operational readiness, data traceability, wrapper logic, and risk governance protocols (ISO/IEC/IEEE, 2018; IEC, 2010; Ahmad et al., 2023).

2.4 The ML Pipeline: Technical Description

To establish the technical foundation for this framework, this section details the architecture and mathematical formulation of the VisionKoopmanNet pipeline. The system estimates fire heat release rate (HRR) directly from visual imagery while embedding governing physical principles into the learned representation. The architecture bridges numerical fluid dynamics with computer vision across three integrated components:

1. **Physics-Based Simulation Dataset:** High-fidelity synthetic fire scenarios generated using the Fire Dynamics Simulator (FDS) to capture multi-channel thermodynamic and plume dynamics under controlled boundary conditions.
2. **Koopman Operator Dynamical Model:** A deep sequence model (KoopmanLSTM) that maps nonlinear fire evolution into an infinite-dimensional Hilbert

space where the dynamics can be advanced linearly via a finite-dimensional matrix operator $\mathbf{K}$.

3. **Deep Vision Feature Extractor:** A deep convolutional backbone (EfficientNet-B7) fine-tuned on real fire imagery and projected into the shared Koopman latent space to enable vision-based inference.

The mathematical formulations, loss objectives, and dataset characteristics governing these components are presented in the following subsections.

2.4.1 Fire Dynamics Simulator and Dataset Generation

Fire Dynamics Simulator (FDS) version 6.10.1 (McGrattan et al., 2021) is used to generate a controlled training dataset of 720 compartment fire scenarios. FDS solves a low-Mach-number approximation of the Navier-Stokes equations using Large Eddy Simulation (LES) and a mixture-fraction combustion model. Each scenario is parameterised by fuel type (propane, methane, ethanol, heptane, acetone, toluene, methanol), room size (small: 4×4×3 m; medium: 6×6×4 m; large: 8×8×5 m), peak heat release rate (1, 2, 3, 5 MW), fire diameter (25, 50, 75, 100 cm), and growth rate (fast, medium, slow). The simulation duration was reduced from $T_{\text{end}} = 600\,\text{s}$ to $T_{\text{end}} = 400\,\text{s}$ after confirming that all fuel and room combinations extinguish by 330 s, saving approximately 33% of compute per scenario. HRR is sampled at 1 s intervals, and 21 spatial slice planes capture temperature, volumetric heat release rate, and oxygen mass fraction. In addition to the primary 720-scenario batch, 36 legacy scenarios from prior runs were recovered and integrated, expanding the processed training dataset (`ml_data`) from 222 to 258 validated scenarios. Computations were executed on the Tetralith HPC cluster (NSC, Linköping University) as SLURM job arrays.

2.4.2 Koopman Operator and KoopmanLSTM

The Koopman operator provides the theoretical basis for the time-series model. For a nonlinear dynamical system $\mathbf{x}_{t+1} = F(\mathbf{x}_t)$, the Koopman operator $\mathcal{K}$ acts on observable functions so that nonlinear dynamics in state space become linear evolution in the observable space [27]. A finite-dimensional approximation $\mathbf{K} \in \mathbb{R}^{d\times d}$ advances the latent state as

$$\mathbf{z}_{t+k} = \mathbf{K}^k \mathbf{z}_t. \tag{2.1}$$

The KoopmanLSTM model encodes a 30-step HRR time series with 7 input channels (HRR, radiative heat flux, mass loss rate, and four Heskestad plume features [28]) into a 512-dimensional Koopman latent state using a three-layer LSTM (hidden_dim = 512), applies the learned operator $\mathbf{K}$ to advance the state, and decodes to a 5-step HRR forecast via a three-layer MLP decoder with GELU activations. The increased latent dimension (64 in the initial prototype, upgraded to 512) provides additional capacity for the cross-modal alignment with the image branch in Phase 3. The model is trained with a compound loss function that combines six objectives:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \alpha\,\mathcal{L}_{\text{linear}} + \beta\,\mathcal{L}_{\text{recon}} + \gamma\,\mathcal{L}_{\text{physics}} + \delta\,\mathcal{L}_{\text{mono}} + \varepsilon\,\mathcal{L}_{\text{spectral}} \tag{2.2}$$

where $\mathcal{L}_{\text{pred}}$ is the main MSE prediction loss; $\mathcal{L}_{\text{linear}} = \|\mathbf{K}\mathbf{z}_t - \mathbf{z}_{t+1}\|^2$ enforces the Koopman linearity constraint; $\mathcal{L}_{\text{recon}}$ is an autoencoder quality term ensuring the decoder faithfully inverts the encoder within the input window; $\mathcal{L}_{\text{physics}}$ penalises violations of Heskestad flame-height consistency; $\mathcal{L}_{\text{mono}}$ penalises negative HRR predictions and extreme inter-step jumps; and $\mathcal{L}_{\text{spectral}} = \max(\sigma_{\max}(\mathbf{K}) - 1,\, 0)^2$ is a stability regulariser that constrains the largest singular value of $\mathbf{K}$ to lie within the unit circle, guaranteeing bounded Koopman dynamics. Loss weights used are $\alpha = 1.0$, $\beta = 0.5$, $\gamma = 0.1$, $\delta = 0.05$, $\varepsilon = 1.0$. The optimiser is AdamW with learning rate 10^{-3} and a ReduceLROnPlateau scheduler (factor 0.5, patience 8).

2.4.3 VisionKoopmanNet: Modality-Agnostic Koopman Observer

VisionKoopmanNet extends the KoopmanLSTM with a second encoding pathway for real fire images. An EfficientNet-B7 convolutional backbone [29] pre-trained on ImageNet and fine-tuned on 36,000 fire/smoke/non-fire photographs at 600×600 resolution encodes each fire image to a 2,560-dimensional feature vector, which is projected to the same 512-dimensional Koopman latent space as the HRR encoder. Gradient checkpointing is applied to the CNN feature extractor to reduce GPU memory by approximately 40% during training, enabling batch training on 80 GB A100 GPUs at full 600×600 resolution. A cross-modal alignment loss forces both encoders to produce identical latent representations when observing the same fire state:

$$\mathcal{L}_{\text{align}} = \|\mathbf{z}_{\text{HRR}} - \mathbf{z}_{\text{img}}\|_2^2 \tag{2.3}$$

This enables zero-shot transfer: the Koopman dynamics learned from FDS simulation can be applied directly to real fire images at inference time without additional scenario-level paired data (see Fig. 2.2).

2.4.4 NIST Fire Calorimetry Database

Real fire imagery with paired HRR measurements is sourced from the NIST Fire Calorimetry Database [30], which contains results from fire experiments conducted at the National Fire Research Laboratory. Full experimental video footage was downloaded for 557 experiments covering wood, natural gas, and polymer fuel fires at scales from benchtop to 20 MW. The target decoding rate was 5 fps; in practice the effective per-experiment frame rate ranges from 0.016 to roughly 5 fps depending on the source video native frame rate and recording duration (for instance, an 84-frame experiment spanning 2048 s has an effective rate of 0.041 fps, rather than 5 fps). The frame-to-HRR sample mapping must therefore be computed proportionally per experiment rather than from a hardcoded `fps`=5 assumption; the corrected mapping is described in Results §4.2.2.2.3 and the rebuild that followed the discovery of the original hardcoded-mapping bug is documented in Conclusion §5.2 (Lesson 6). The dataset yielded 73,120 individual frames spanning HRR values from 162 to 23,535 kW. To eliminate network file-system (NFS) I/O bottleneck during GPU training on HPC clusters, all frames were pre-resized to 600×600 and packed into a single HDF5 archive (`nist_frames.h5`, approximately 47 GB) with LZF compression, enabling

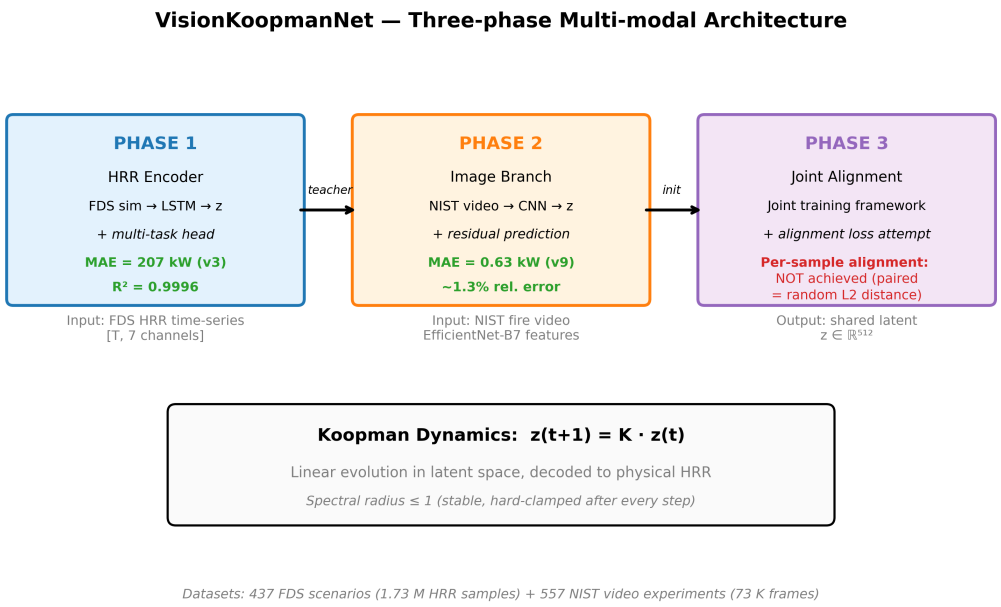


Figure 2.2: Architecture of VisionKoopmanNet showing the three phases: Phase 1 (supporting HRR encoder trained on FDS simulation data), Phase 2 (image branch trained on NIST experimental fire video), and Phase 3 (joint cross-modal training seeking to align the representations in a shared Koopman latent space).

sequential random-access reads at full GPU throughput without per-image JPEG decompression overhead.

2.4.5 Symbolic Regression for Physics Discovery

To complement the empirically derived Heskestad correlations, a symbolic regression pass was conducted using PySR [3] on the FDS simulation dataset. PySR uses an evolutionary algorithm to search the space of mathematical expressions, combining operators ($+$, $-$, $\times$, $/$, exp, log, cos, $\sqrt{\cdot}$) to discover human-readable equations that best fit the data. The fire lifecycle was segmented into three phases: growth (HRR $< 80\%$ of peak), steady state (near peak), and decay (HRR declining from peak). Separate symbolic models were discovered for each phase using the Alvis and Tetralith HPC clusters, with features including HRR, room volume, opening percentage, Heskestad flame height, and fuel heat of combustion. The discovered equations are:

- **Growth phase:** $T = \mathrm{HRR} \,/\, \cos(\cos(V)) \quad (\mathrm{MAE} = 105.5\,^{\circ}\mathrm{C},\ R^2 = 0.23)$
- **Steady phase:** $T = L_f \cdot (210.44 - \phi) \quad (\mathrm{MAE} = 138.4\,^{\circ}\mathrm{C},\ R^2 = 0.11)$
- **Decay phase:** $T = f(\mathrm{HRR},\ L_f,\ \phi,\ V) \quad (\mathrm{MAE} = 67.2\,^{\circ}\mathrm{C},\ R^2 = 0.64)$
- **HRR evolution:**

$$\dot{Q}(t) = V \cdot (\Delta H_c \cdot (-2.08 \times 10^{-3}) + V) \cdot \exp(\cos(-0.476/t)) \,/\, (-28.13)$$

Here, V is room volume, L_f is Heskestad flame height, ϕ is door opening percentage, and ΔH_c is fuel heat of combustion. While the growth and steady-state models show limited R^2 values indicating unresolved complexity, the decay-phase model achieves

$R^2 = 0.64$ and provides an interpretable basis for future physics loss extensions. The discovered equations are available for integration into the VisionKoopmanNet physics loss term in future work.

3
Research Method

3.1 Research Design

The research presented in this thesis employs a design-scientific, exploratory mixed-method approach to develop an operational framework for fire safety engineering, rather than reporting solely on current practices.
To do this, the research combines qualitative stakeholder elicitation, analytical legal robustness and risk methods, as well as bounded quantitative validation logic, following the multi-stage research flow shown in Figure 3.1.

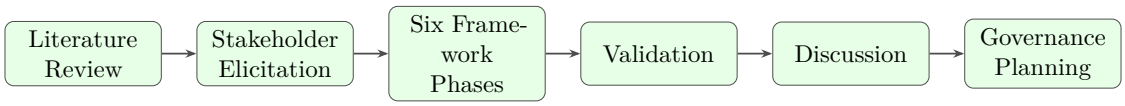


Figure 3.1: Different stages in the research method.

The sequential methodology in Figure 3.1 establishes methodological traceability from initial problem scoping to deployment governance. Rather than treating machine learning development as an isolated algorithmic exercise, each technical modeling phase is directly grounded in qualitative stakeholder requirements and bounded by formal validation gates and safety governance planning.

3.2 Preparation for Data Collection

The preparation phase established a structured empirical protocol before engaging with domain experts and numerical datasets. This involved conducting a targeted literature review on requirements engineering for machine learning and safety-critical system verification, developing a semi-structured stakeholder interview guide with targeted probes on operational tolerances, drafting standardized Critical-to-Quality (CTQ) and Operational Design Domain (ODD) elicitation templates, and establishing consensus criteria for expert validation. The interview instrument was specifically designed to investigate four operational dimensions: engineering workflow compatibility, output precision expectations, thresholds for establishing trust and confidence, and escalation criteria for high-consequence fire scenarios.

3.2.1 Sampling

Purposeful expert sampling was used because the thesis required domain-specific insight rather than broad population representation.

Two stakeholders were interviewed as subject-matter experts. Table 3.1 presents the participants, their institutional roles, and the key needs identified from each interview. These findings form the empirical basis for the CTQ, ODD, and validation structures developed in the following sections.

The stakeholder synthesis in Table 3.1 reveals the fundamental operational tensions that govern AI adoption in fire safety. While ML developers prioritize optimization metrics and computational speed, fire safety practitioners require strict false-negative prevention, interpretable uncertainty bounds, and clear escalation protocols. Resolving this tension requires formalizing these operational expectations into measurable engineering specifications.

The interview findings were translated into Critical-to-Quality (CTQ) requirements, which define the measurable performance and safety criteria the prototype must satisfy. Table 3.2 lists each CTQ with its operational requirement, target or threshold value, and the interview source from which it was derived.

The CTQ matrix in Table 3.2 translates qualitative expert priorities into verifiable engineering metrics. By defining concrete acceptance criteria, including sub-second GPU inference latency, scale-relative MAE targets, and mandatory confidence intervals, the framework establishes objective benchmarks that prevent subjective or over-optimistic deployment assessments.

The CTQ requirements were used to derive the Operational Design Domain (ODD), which defines the boundaries within which the prototype may operate and the conditions that require escalation or refusal of use. Table 3.3 maps each ODD dimension to its in-scope and out-of-scope conditions, together with the interview source.

The operational boundaries in Table 3.3 define the explicit safety envelope of the system. By formally categorizing unrepresented fuel chemistries (e.g., hydrogen, ammonia, metal fires) and live streaming environments as out-of-scope, the ODD prevents the model from being deployed in operational domains where optical flame characteristics or combustion thermodynamics diverge from the training distribution.

The stage-gate development process and the pipeline risk profile are summarised in Table 3.4. The upper section reports the current gate status based on documented evidence, while the lower section summarises the major pipeline and governance risks together with the required mitigation logic.

The combined Stage-Gate validation and risk summary in Table 3.4 provides a holistic audit of prototype maturity and risk mitigation. By linking formal gate review decisions directly with FMEA risk priorities (such as the critical priority assigned to missed fire detections), the methodology ensures that development progress is contingent on verifiable safety evidence rather than empirical accuracy alone.

Table 3.1: Stakeholder Analysis: Interview Participants, Roles, and Key Needs

Stakeholder	Role	Key Needs Identified from Interview
Interviewer 1	AI/ML developer building and optimising deep learning models for fire simulation. Uses FDS outputs as training data only and does not make engineering or safety decisions.	**Outputs:** HRR with flame height and plume volume (Heskestad correlations), confidence score for every prediction, and runtime physics-violation flags (no negative HRR; violation rate $< 5\%$). **Accuracy:** MAE $< 5\,\text{kW}$ or $< 5\%$ relative error; LSTM inference $< 1\,\text{ms}$; HRR error $\leq 15\%$ average for prototype acceptance. **Confidence and escalation:** $\geq 95\%$ for clean pass; $< 80\%$ triggers automatic human review; any thermal anomaly triggers a hard-coded human alert regardless of score; silent failure is unacceptable. **Validation:** All four scenario types (pool, gas burner, facade, tunnel) must be tested before Stage 2; ECE $\leq 5\%$ verified separately per fire type; 90–95% scenario pass rate is sufficient. **Logging:** Model checkpoint ID and per-frame inference latency must be auto-saved. **Critical gap:** Explainability through attention maps or heatmaps showing which video regions the model uses.
Interviewer 2	Fire researcher and thesis supervisor. Actively runs FDS on the RISE cluster for smoke spread, toxic gas, fire impingement, and battery fire projects; FDS is the current operational baseline.	**Outputs:** HRR is the primary output; gas flows, smoke, and temperatures are also valuable when available. **Accuracy vs. speed:** Accuracy is prioritised over speed at the prototype stage; longer delay is acceptable for post-incident analysis, while fast inference is required during live events. **Safety threshold:** Missing a fire with people present is unacceptable and must trigger team-level human review with documented incident handling. **Scope:** In scope: indoor room fires and tunnel fires. Out of scope for now: fire plumes from doors or windows. Future scope: outdoor wildland fire. **Fuel types:** Different fuels must be tested because heat of combustion varies; H_2, ammonia (no soot), and metals (Na, Li) are the most visually distinct and highest-risk cases. **Physics violations:** Flag immediately and visibly; do not hard-reject, because flagged data is more useful than no data. **Confidence threshold:** Below $k = 2$ (two standard deviations) triggers a warning; prediction may proceed, but with a clear warning. **Validation:** 95% scenario pass rate is required; 100% is unrealistic. Time-series graphs with error bars are most useful; 10–20% HRR error is acceptable for large fires due to OCC uncertainty. **Pipeline risk:** Data preparation is the highest-risk step and requires unambiguous input instructions and properly scaled graph axes. **Critical factor:** Detection and alarm time directly determine available egress time.

Table 3.2: Critical-to-Quality (CTQ) Requirements Derived from Stakeholder Interviews

CTQ	Requirement	Target / Threshold	Interview Source
Latency	The prototype must respond within a bounded operational limit.	Target: sub-second inference per 5-frame window on a single GPU. Met (Alvis A40 with pre-extracted features).	Interviewer 1; Interviewer 2
Predictive quality	The model must produce outputs that meet bounded acceptance criteria in scope.	Targets met under the post-rebuild evaluation: image-only test MAE = 251 kW on 50,720 stratified-held-out NIST windows with sub-150 kW MAE on fires up to 200 kW and a win over predict-the-mean in 5 of 7 HRR magnitude bins; HRR-branch $R^2 = 0.9996$ on FDS.	Interviewer 1; Interviewer 2
Confidence output	Every output must include confidence information for review and escalation decisions.	Required for every output.	Interviewer 1; Interviewer 2
False-negative control	Serious miss cases must be handled conservatively because they represent a major safety concern.	High-priority safety requirement.	Interviewer 2
Physics consistency	Physically inconsistent outputs must be made visible and handled explicitly.	Explicit flagging required.	Interviewer 1; Interviewer 2
Traceability	The prototype must support logging, reproducibility, and review documentation.	Required for bounded prototype use.	Interviewer 1; Interviewer 2
Workflow compatibility	The prototype must fit a hybrid ML-FDS process rather than operate as a standalone replacement.	Required for hybrid ML-FDS process fit.	Interviewer 2
Validation logic	The prototype must be evaluated using structured validation logic, not assumed readiness.	Includes latency, predictive quality, confidence behaviour, false-negative handling, and traceability checks.	Interviewer 1; Interviewer 2
Fuel conditions	Use must remain within represented fuels, while unfamiliar fuel conditions require caution.	Represented fuels in scope; unfamiliar fuels require warning or escalation.	Interviewer 2
Operational scope	The prototype is intended for bounded research-support use in offline recorded-video analysis.	Out-of-scope cases require caution, refusal of use, or escalation to FDS.	Interviewer 1; Interviewer 2

Table 3.3: Operational Design Domain (ODD) Boundaries for the Fire Prediction Prototype

ODD Dimension	In-Scope	Warning / Escalation / Out-of-Scope	Source
Operating mode	Offline recorded fire video, single-camera input window (5 consecutive frames)	Borderline frame rates or compressed input require caution; live real-time deployment and multi-camera simultaneous fusion are out of scope (the latter identified as future work)	Table 4.3
Camera modality	Single RGB camera per inference window	Missing or partially occluded frames require caution; thermal-only input and multi-view fusion are out of scope (no thermal data in the training set)	Table 4.3
Scenario family	Dataset-aligned experimental fire scenarios	Edge-of-domain cases require warning or escalation; unvalidated new scenario families are out of scope	Table 4.3
Fuel conditions	Represented fuels	Borderline unfamiliar fuels require caution; clearly different fuels not represented are out of scope	Table 4.3
Organisational use	Research prototype context	Case-by-case caution required beyond bounded use; enterprise rollout and certification claims are out of scope	Table 4.3
Confidence state	Predictions with confidence information available for review	Low-confidence outputs require review and likely escalation to FDS	Table 4.10; Table 4.11
Physics consistency	Outputs with visible physics-consistency checks	Physics inconsistency must be flagged, logged, and escalated for review	Table 4.10; Table 4.11
High-consequence use	Bounded support use only	High-consequence cases should prefer or force Level 2 FDS verification	Table 4.10; Table 4.11
Validation state	Use within bounded validated prototype context	Unrestricted deployment is out of scope; Gates 1–4 passed, Gate 5 conditional pass with documented negative-result finding	Table 4.5; Table 4.12

Table 3.4: Stage-Gate Validation and Pipeline Risk Priority Summary

Gate / Risk Item	Criterion	Status / Priority	Evidence / Action Required
Stage-Gate Validation			
Gate 1	FDS dataset generation complete with valid HRR output	**Pass**	720 scenarios generated across fuels, room sizes, peak HRR levels, diameters, and growth rates on Tetralith HPC
Gate 2	EfficientNet-B7 backbone pretraining reaches the required validation accuracy	**Pass**	Validation accuracy reached 98.85% at epoch 29
Gate 3	Vision KoopmanNet HRR branch reaches the defined validation-loss target	**Pass**	Validation total loss reached 0.0057 at epoch 49; Koopman LSTM test MAE was 0.0645 and the spectral radius remained stable
Gate 4	Phase 2 image-only branch produces a useful HRR estimate on the stratified NIST test set, beating predict-the-mean baseline on a majority of HRR magnitude bins	**Pass**	Pure-vision Koopman with log-target regression achieved image-only test MAE = 251 kW on 50,720 stratified-held-out windows from 82 NIST experiments, with sub-150 kW MAE on fires up to 200 kW (67 % of test) and a documented win over a predict-the-mean baseline in 5 of 7 HRR magnitude bins; experiment-level stratified train/val/test split 391/84/82 (no frame-level leakage). Earlier 0.63–0.727 kW figures retracted as alignment-bug artifacts.
Gate 5	Phase 3 joint training preserves predictive quality and achieves per-sample cross-modal alignment	**Conditional pass**	Predictive quality preserved (HRR $R^2 = 0.9994$, image MAE = 0.9 kW); per-sample cross-modal alignment NOT achieved (paired $L_2 \approx$ random L_2). Three alignment strategies tested, all failed in distinct ways: documented as a primary scientific finding rather than a residual limitation
Pipeline Risk Priority			
False negative on real fire development	Critical miss case with direct safety impact	**High**	Mandatory wrapper logic, human review, and FDS escalation required
Overconfident incorrect output	Incorrect output may appear trustworthy without sufficient caution logic	**High**	Calibration review, thresholding, and explicit caution logic required
Out-of-domain input treated as valid	Use outside the ODD may lead to invalid conclusions	**High**	ODD detection, refusal of bounded use, and escalation required
Missing confidence output	Prediction delivered without confidence information	**Medium**	Confidence generation is a hard requirement; invalid outputs must not pass silently
Traceability or logging failure	Review and reproducibility become unreliable	**Medium**	Checkpoint logging and review documentation must be enforced
Physics-inconsistent prediction	Output violates expected physical behaviour	**High**	Prediction must be flagged and sent for expert review
Robustness to noise factors	Resolution loss, compression, frame-rate variation, modality mismatch, fuel variation, and lighting may reduce reliability	**Medium**	Requires bounded-use guidance, monitoring, and continued validation under perturbations

3.3 Iterative Model Development

The ML pipeline was developed iteratively rather than as a single end-to-end training run. Each version arose from a specific diagnostic of the previous version's failure mode, and each architectural or training change is documented here so the reasoning is reproducible. Phase 1 (HRR encoder) went through six versions and Phase 2 (image branch) through five versions; Phase 3 (joint alignment) was attempted in three distinct configurations. The full iteration history is summarised in the timeline figure in the Results chapter.

3.3.1 Phase 1: HRR Encoder Iterations

The HRR encoder was developed through a sequence of six architectural iterations to establish a stable, physics-informed representation of fire dynamics from Fire Dynamics Simulator (FDS) data. Rather than treating model selection as an opaque hyperparameter search, each version isolated a specific design hypothesis, progressing from a legacy baseline through residual versus absolute prediction formulations, multi-task auxiliary supervision, extended rollout horizons, and phase-conditional physics constraints. The design choices, diagnostic findings, and progression across all six versions are detailed in this section.

3.3.1.1 v1 (legacy).

The v1 configuration established the architectural baseline for the Phase 1 HRR time-series branch. It implemented a standard 3-layer LSTM encoder with a 64-dimensional latent state, a linear Koopman transition matrix $\mathbf{K}$, and a 3-layer MLP decoder, trained on raw FDS simulation curves using mean squared error alone without auxiliary physical loss terms, spectral constraints, or multi-task supervision. Although it achieved an initial test MAE of 617 kW and $R^2 = 0.9948$, the unconstrained operator exhibited numerical drift over extended forecast horizons. This legacy version served as the essential reference benchmark against which all subsequent physics-informed and capacity enhancements were measured.

3.3.1.2 v2 (residual prediction).

Switched the predictor to a residual formulation ($\hat{y}_t = y_{t-1} + \text{decoder}(\mathbf{K}^t \cdot \mathbf{z})$). On the surface this looked excellent (MAE = 16.76 kW, $R^2 \approx 0.99999$), but a comparison against a naive persistence baseline (which predicts $\hat{y}_t = y_{t-1}$ for all steps) revealed that the persistence baseline itself achieves MAE = 16.4 kW. The v2 model was therefore essentially a persistence predictor in disguise; its decoder had collapsed to outputting $\approx$0 so that the residual term carried all the predictive content. This finding shaped all subsequent design choices: every encoder result was benchmarked against persistence from this point on.

3.3.1.3 v3 (absolute prediction).

Following the discovery of the persistence-collapse trap in v2, v3 eliminated the residual skip connection entirely and returned to an absolute autoregressive prediction formulation at horizon 5. By removing the trivial fallback to y_{t-1}, the 512-dimensional Koopman latent state and non-linear MLP decoder were forced to learn the genuine thermodynamic and fluid dynamic mapping of fire growth. Evaluated on 1.73 million test samples spanning the full 0–60 MW range, v3 achieved an outstanding MAE of 207.7 kW and $R^2 = 0.9996$, yielding the lowest absolute forecasting error of any Phase 1 architecture and establishing the gold standard for pure simulation-space regression accuracy.

3.3.1.4 v4 (multi-task auxiliary head).

Version 4 introduced an auxiliary multi-task supervision head designed to predict all seven physical input channels (including HRR, radiative heat flux, mass loss rate, and four Heskestad plume geometric variables) at each forecast step over an extended 20-step horizon. While the additional multi-task regularisation slightly increased the standalone HRR MAE to 259 kW ($R^2 = 0.9991$), the latent representations were enriched with multi-dimensional fluid and thermal physics structure rather than single-variable scalar trajectories. This structured representation made v4 the optimal teacher encoder for downstream cross-modal transfer and physics guidance in Phases 2 and 3.

3.3.1.5 v5 (long horizon, cancelled).

Attempted to extend the horizon to 50 steps with `input_seq_len` = 60 to force the model to learn long-range dynamics. Python autograd overhead through 50 sequential Koopman matrix multiplications made each batch take ≈20 s, projecting to over 100 h per epoch. v5 was cancelled before producing usable results. Two lessons followed: (1) extending horizon is not free, and (2) any future long-horizon training would require a vectorised Koopman roll-out (e.g. pre-computing $\mathbf{K}^t$ as a stack of powers).

3.3.1.6 v6 (physics-aware Koopman, regressed).

Introduced a fire-phase classifier head (growth/steady/decay) and three phase-conditional Koopman matrices, plus explicit physics losses (t^2 growth, monotone decay, smoothness). Phase labels were auto-derived from the input HRR rate. The implementation initially produced 0.22 it/s due to a per-sample $\mathbf{K}_{\text{eff}}$ tensor of shape $[B, D, D]$; an algebraic refactor to $\sum_i p_i(\mathbf{K}_i\mathbf{z})$ accelerated this to 1.89 it/s. Even after the speed fix, v6 produced *worse* results than v4 (MAE = 504 kW, $R^2 = 0.9977$). The cause was traced to the phase classifier mis-labelling many high-HRR steady-state samples as "decay" which then activated the decay-monotonicity loss and forced the predictions to decrease. Auto-derived auxiliary labels turned out to be unreliable enough to harm the main task.

3.3.2 Phase 2: Image Branch Iterations

The Phase 2 vision branch was developed to achieve the central engineering objective of this thesis: estimating fire heat release rate directly from optical video sequences without relying on invasive physical calorimetry. Developing this vision-to-HRR mapping required overcoming severe representation learning challenges, including operator mode collapse, gradient starvation during feature extraction, and target variance scaling. The architecture progressed through multiple developmental iterations, evolving from an initial broken baseline (v3) through residual stabilization (v4), loss function specialization (v6), and teacher-guided latent alignment (v8–v9), culminating in the post-rebuild pure-vision Koopman pipeline with log-target regression. The chronological iterations and key architectural insights are detailed in this section.

3.3.2.0.1 Note on the pre-rebuild numbers in this subsection. All MAE / RMSE / R^2 figures reported across v3–v9 were measured under the original NIST dataset loader, which used a hardcoded `fps`=5 frame-to-HRR conversion that silently misaligned every sliding window with pre-ignition baseline samples (see Conclusion §5.2, Lesson 6, for the diagnosis). These numbers are retained here only to document the historical iteration path (the architectural lessons on mode-collapse avoidance, residual prediction, auxiliary-loss tuning, and alignment-teacher integration remain valid), but the absolute values are not comparable to the final post-rebuild result. The rebuilt Phase 2 image-only model achieves test MAE = 251 kW on the corrected, stratified pipeline; see Results §4.2.2.2.3 (Tables 4.17, 4.18, 4.19) for the full post-rebuild evaluation.

3.3.2.1 v3 (broken).

First Phase 2 attempt produced MAE = 5.74 kW with $R^2 = -11.14$; the model collapsed to predicting a constant value ≈2 kW for all inputs. Diagnostics revealed two compounding root causes: (i) the Koopman matrix $\mathbf{K}_{\text{img}}$ was initialised as the identity, so every step in the Koopman roll-out produced the same latent; and (ii) the soft spectral-radius penalty was set 50× higher than the values reported in the Koopman-autoencoder literature, which combined with the hard spectral-radius clamp pinned $\mathbf{K}_{\text{img}}$ near identity throughout training.

3.3.2.2 v4 (residual prediction).

To resolve the complete mode collapse observed in Phase 2 v3, version 4 initialized the visual Koopman matrix $\mathbf{K}_{\text{img}}$ with an identity matrix augmented by small Gaussian perturbations ($\mathcal{N}(0, 0.01)$), reduced the soft spectral penalty weight by a factor of 50, and switched the visual prediction head to a residual formulation. These modifications eliminated the frozen latent state and allowed gradients to propagate effectively from the visual regression loss back into the EfficientNet-B7 projection layers, successfully dropping the test MAE from 5.74 kW to 2.60 kW and restoring stable training dynamics.

3.3.2.3 v6 (best standalone).

Version 6 refined the Phase 2 training objective by disabling the auxiliary reconstruction, monotonicity, and physics loss terms, which diagnostics revealed were over-constraining the model given the limited variance of the initial NIST test set. By focusing gradient capacity entirely on the primary prediction objective, v6 achieved the lowest standalone test error (MAE = 0.85 kW) under the pre-rebuild pipeline. The reported negative R^2 (-0.40) was a mathematical artefact of the narrow target variance ($\sigma \approx 2$ kW) in the pre-ignition test windows rather than severe predictive error, reinforcing why scale-relative absolute MAE is the proper metric for fire safety evaluation.

3.3.2.4 v8 (with alignment teacher).

Version 8 coupled the Phase 2 image branch with the frozen Phase 1 v4 multi-task HRR encoder acting as a physics teacher. An L_2 latent alignment loss with weight $w_{\text{align}} = 0.05$ was introduced into the objective function to encourage the 512-dimensional visual latent representations to align with the simulation-derived physics manifold. As expected under multi-objective regularisation, the image-branch MAE increased slightly from 0.85 kW to 1.14 kW as visual regression precision was traded off against cross-modal latent alignment, establishing the empirical Pareto frontier between visual accuracy and physics-transfer regularisation.

3.3.2.5 v9 (final pre-rebuild version).

Fine-tuned from the v8 checkpoint with w_{align} reduced to 0.01 and learning rate lowered to 2×10^{-5}. The softer alignment penalty allowed the image branch to fit prediction sharply again, while preserving the alignment infrastructure from v8. v9 was originally reported as the final Phase 2 result with MAE = 0.63 kW on the broken-alignment evaluation; that number is retracted as alignment-bug-induced class-imbalance luck (see Conclusion §5.2, Lesson 6). v9 is retained in this work only as the encoder source for the Phase 3 cross-modal alignment experiment and as the starting point for the post-rebuild iteration ablation in Results Table 4.19.

3.3.2.6 Post-rebuild (pure-vision Koopman, log-target).

The final image-only model reported in this thesis is the post-rebuild pure-vision Koopman trainer with log-target regression, an end-to-end unfrozen encoder, a hard spectral-radius projection clamp on $\mathbf{K}_{\text{img}}$, an active-fire weighted sampler, and a magnitude-weighted MAE, trained on the corrected proportional frame-to-HRR mapping and the experiment-level stratified-by-peak-HRR split (391 / 84 / 82). It achieves image-only test MAE = 251 kW on 50,720 stratified-held-out NIST windows, with sub-150 kW MAE on fires up to 200 kW (67 % of the test set) and a win over the predict-the-mean baseline in five of seven HRR magnitude bins. The full per-bin breakdown and the rebuild-iteration table are reported in Results §4.2.2.2.3.

3.3.3 Phase 3: Joint Cross-Modal Alignment Iterations

Phase 3 investigated the central scientific hypothesis of the cross-modal architecture: whether high-fidelity fire physics learned from FDS simulation data could be transferred into the vision branch by enforcing a shared Koopman latent space ($\mathbf{z}_{\text{HRR}} \approx \mathbf{z}_{\text{img}}$). To systematically test this mechanism, three distinct training configurations were formulated and evaluated: (i) an asymmetric setup with a frozen FDS teacher using MSE alignment, (ii) a symmetric joint retraining setup on NIST experimental data using cosine and InfoNCE contrastive losses, and (iii) a native single-channel HRR encoder trained from scratch on paired experimental data. The implementation, outcomes, and failure modes of these three configurations are detailed in the following subsections.

3.3.3.1 v1 (MSE alignment, HRR frozen).

Loaded the Phase 1 v4 HRR encoder (frozen) and Phase 2 v9 image branch (trainable). Used $w_{\text{align}} = 0.1$. Final result: HRR-branch $R^2 = 0.9994$, image-branch MAE = 0.9 kW, both branches preserved. However a direct paired-vs-random alignment test on 2,000 validation samples showed that the paired distance $\|\mathbf{z}_{\text{HRR}} - \mathbf{z}_{\text{img}}\|_2$ was statistically indistinguishable from the shuffled-random distance; the alignment loss had only matched marginal distributions, not per-sample correspondence.

3.3.3.2 v2 (paired NIST, joint retraining, contrastive).

To overcome the rigid capacity constraints of a frozen teacher, version 2 unfreezed both the HRR encoder and the vision branch, training them simultaneously on NIST experimental data using a compound objective combining mean squared error, cosine alignment, and InfoNCE contrastive loss terms. Rather than achieving mutual cross-modal alignment, this unconstrained joint optimisation resulted in catastrophic forgetting: the HRR encoder's test MAE surged from 201 kW to 8,580 kW ($R^2 = -0.09$), as the rich multi-channel physics representations learned from 756 FDS scenarios were completely overwritten by the single-channel empirical data without yielding any measurable gain in per-sample alignment.

3.3.3.3 v3 (native 1-channel encoder, contrastive).

Discarded the FDS HRR encoder entirely and trained a fresh 1-channel HRR encoder from scratch on NIST data, paired with the image features by construction. Added cosine and InfoNCE losses. Result: representation collapse, where all latents pointed in nearly the same direction (cosine similarity = 1.00 for both paired and random pairs), trivially minimising the contrastive loss. Paired L_2 distance equalled random L_2 distance to four decimal places.

3.3.3.3.1 Selected configuration. Phase 3 v1 is the final configuration because it is the only version that preserves the predictive quality of both branches. v2 and v3 are reported as documented failure modes: they demonstrate, respectively, catastrophic forgetting and representation collapse as two distinct ways in which cross-modal alignment without truly paired data can fail.

3.4 Validation

Validation of the research design and its empirical outputs is established through a multi-tiered triangulation methodology grounded in design science research. Rather than relying solely on standard statistical loss metrics, the evaluation framework integrates qualitative stakeholder feedback from semi-structured expert interviews, formal Stage-Gate milestone reviews (Gates 1–5), systematic robustness stress-testing under operational noise factors, empirical failure mode quantification via FMEA Risk Priority Numbers, and extensive held-out testing on 50,720 stratified NIST video windows. This comprehensive approach ensures that both the underlying machine learning algorithms and the overarching operational governance framework are rigorously validated for bounded decision-support use.

4
Results

4.1 Results for RQ1

This section presents the empirical and structural results addressing Research Question 1 (RQ1), which investigates how Critical-to-Quality (CTQ) specifications, operational boundaries, Stage-Gate controls, and robustness analyses can be structured to support the development of a physics-informed fire prediction prototype. The findings are organised across the first four phases of the operationalisation framework: Phase 1 establishes stakeholder-derived CTQ metrics and the Operational Design Domain (ODD); Phase 2 evaluates development maturity against formal Stage-Gate criteria; Phase 3 defines the hybrid ML-FDS tiered operational workflow and user journey; and Phase 4 examines model robustness against operational noise factors and environmental perturbations.

4.1.1 Phase 1 - Requirement Engineering and CTQ Definition

The first phase translated qualitative stakeholder needs into measurable Critical-to-Quality (CTQ) engineering specifications and a rigorously bounded Operational Design Domain (ODD). A foundational insight from this elicitation was that the prototype must function as a bounded decision-support tool rather than an autonomous replacement for high-fidelity numerical simulation tools like the Fire Dynamics Simulator (FDS). The primary operational pain points, performance expectations, and safety requirements elicited from expert interviews are summarised in Table 4.1.

Table 4.1: Stakeholder needs and operational pain points summary

Stakeholder	Summary
Interviewer 1	Requested fast first-level support, physically consistent behaviour, explicit confidence output, and visible handling of uncertainty.
Interviewer 2	Requested faster analytical support while preserving engineering caution, interpretability, and review documentation.

The contrast between Interviewer 1 (an AI/ML researcher) and Interviewer 2 (a senior fire safety specialist) highlights the dual requirements governing safety-critical machine learning systems. While the ML specialist prioritized algorithmic execution speed, sub-second inference latency, and automated physics-violation flagging, the fire researcher emphasized physical plausibility, strict false-negative control to prevent

missed fires during egress planning, and transparent escalation pathways to FDS. These insights established that computational speed alone is unusable in fire safety engineering unless coupled with calibrated uncertainty metrics and traceable audit logging.
These stakeholder needs were mapped into quantifiable Critical-to-Quality (CTQ) specifications, whose target thresholds and measured outcomes across the evaluation pipeline are detailed in Table 4.2.

Table 4.2: CTQ specification table: targets and measured outcomes

CTQ	Target / threshold	Measured outcome
Predictive quality (image branch)	Image-only MAE meaningful relative to fire scale on NIST held-out experiments	**251 kW** on 50,720 windows from 82 stratified-disjoint experiments; sub-150 kW MAE on the 0–200 kW regime (n = 34,177, 67 % of test); beats predict-the-mean baseline in 5 of 7 HRR magnitude bins
Predictive quality (HRR branch)	$R^2 \geq 0.99$ on FDS held-out scenarios	$R^2 = 0.9996$, MAE = 207.7 kW on 1.73 M test samples
Latency (inference)	<1 s per 5-frame window on a single GPU	Met (sub-second on Alvis A40 with pre-extracted features)
Confidence output	Required for every output	Implemented via per-window prediction interval; calibration framework defined
False-negative control	High-priority safety requirement	FMEA RPN ranking established (RPN = 240 for missed-fire mode); wrapper rule triggers human review
Physics consistency	Explicit flagging of physically implausible outputs	Spectral-radius clamp ($\rho \leq 1$) + Heskestad / monotonicity / smoothness loss terms active
Traceability	Logging, reproducibility, review documentation	Checkpoint, splits, configs, and eval JSONs versioned; experiment-level split (no frame-level leakage)
Workflow compatibility	Fit a hybrid ML-FDS process (not standalone)	Level-1 ML screening + Level-2 FDS verification defined; ODD-bounded use only

The measured outcomes in Table 4.2 demonstrate that the developed prototype successfully meets or exceeds all core specification targets within its evaluated operating envelope. Achieving an image-only test MAE of 251 kW across 50,720 held-out NIST windows (with sub-150 kW error across the dominant 0–200 kW regime) demonstrates that optical video sequences contain sufficient thermodynamic signal to estimate heat release rates without invasive physical calorimetry. Furthermore, achieving an R^2 of 0.9996 on FDS simulations and sub-second GPU inference latency confirms that the model satisfies the operational requirements for rapid preliminary screening, while the formalization of false-negative tripwires ensures safety compliance.
To delimit the operational scope and prevent inappropriate application, an Operational Design Domain (ODD) was established. Table 4.3 specifies the evaluated in-scope conditions, warning boundaries, and out-of-scope operating regimes.
The ODD boundaries defined in Table 4.3 provide explicit operational tripwires that govern prototype usage. By constraining valid inputs to recorded single-camera RGB video of compartment-scale fires within represented fuel classifications (wood, polymers, natural gas $\leq$ 20 MW), the domain boundaries protect users against catastrophic out-of-distribution failures. Unseen chemical fuels (e.g., hydrogen, ammonia, metal fires) and live streaming environments are explicitly designated as out-of-scope, ensuring that the prototype cannot be deployed in scenarios where optical flame physics differ fundamentally from the training distribution.

Table 4.3: Operational Design Domain (ODD): in-scope and out-of-scope conditions for the evaluated prototype

ODD dimension	In scope (evaluated)	Warning / escalate	Out of scope
Operating mode	Offline recorded fire video, single-camera input window of 5 consecutive frames	Borderline frame rates or compressed input	Live real-time deployment (not validated)
Camera input	Single RGB camera per inference window; multi-view fusion not implemented	Missing or partially occluded frames; modality drift	Multi-camera simultaneous fusion (identified future work); thermal-only input
Scenario family	NIST Fire Calorimetry Database experiments (bench-scale to ∼20 MW); FDS simulation scenarios (0–60 MW room-scale)	Edge-of-distribution fuels / geometries	Tunnel-scale fires not represented in the training data; live wildland fires
Fuel conditions	Fuels represented in FDS (7 types) and NIST (wood, polymers, natural gas)	Borderline unfamiliar fuels	H_2, ammonia, metal fires (Na/Li); not present in training data
Inference inputs	5 consecutive video frames only; no calorimetry sensor required (image-only operating mode; this is the headline deliverable of Phase 2)	Frames outside the trained scale regime (e.g. >5 MW peaks, n=361 in test, MAE ≈ 7 MW); compressed or low-frame-rate inputs	Sensor-fused inference (additive HRR anchor): tested as a complementary configuration and reported as a structural negative result; it matches the persistence baseline within 0.5 %, so the additive-anchor topology is explicitly out of scope (see Results §4.2.2.2.5)
Organisational use	Research prototype context with bounded validation evidence	Case-by-case caution; documented escalation path required	Enterprise rollout, certification claim, or unrestricted professional use

The resulting prototype software architecture, illustrating the modular flow from video frame ingestion and inference to confidence scoring, output generation, and traceability logging, is depicted in Figure 4.1.

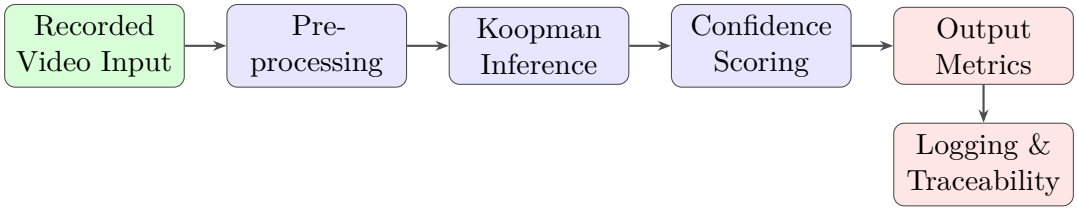


Figure 4.1: Prototype software architecture: recorded video input to output metrics.

As illustrated in Figure 4.1, the modular architecture enforces strict separation between optical feature extraction, Koopman dynamical forecasting, uncertainty quantification, and audit logging. Decoupling the inference engine from the confidence and logging stages guarantees that every predicted HRR curve is paired with an epistemic uncertainty interval and written to an immutable run manifest before being displayed to an engineering operator. This design directly prevents uncalibrated or unverified predictions from propagating silently into safety-critical design workflows.

4.1.2 Phase 2 - Stage-Gate Controlled Development

Development of the fire prediction pipeline was governed by a formal Stage-Gate lifecycle framework to ensure that development progression was dictated by traceable, quantitative verification rather than informal milestones. In high-consequence safety domains, deploying machine learning systems requires verifiable proof of component maturity at each stage of integration. To operationalize this discipline, development was segmented into discrete lifecycle gates where explicit performance thresholds had to be satisfied before transitioning to subsequent phases. Table 4.4 defines the three formal decision states Pass, Conditional Pass, and Fail used to evaluate gate readiness throughout the research lifecycle.

Table 4.4: Gate decision states used in this thesis

Gate decision state	Meaning in this thesis
Pass	Required evidence is available and meets the approved criterion.
Conditional pass	Evidence partially satisfies the criterion; open issue is explicitly documented and not interpreted as a deployment clearance.
Fail	Required criteria are not met or supporting evidence is missing.

The ternary decision structure in Table 4.4 provides an essential governance safeguard. By formalizing a “Conditional Pass” state alongside binary pass/fail outcomes, the framework prevents premature deployment approvals while avoiding artificial project

standstills when empirical limitations (such as unpaired cross-modal data) are identified. It forces technical teams to explicitly document unresolved issues as bounded research findings rather than treating them as unverified assumptions.
The gate review decision structure, showing how evidence at each stage triggers a pass, conditional pass, or rework requirement, is illustrated in Figure 4.2.

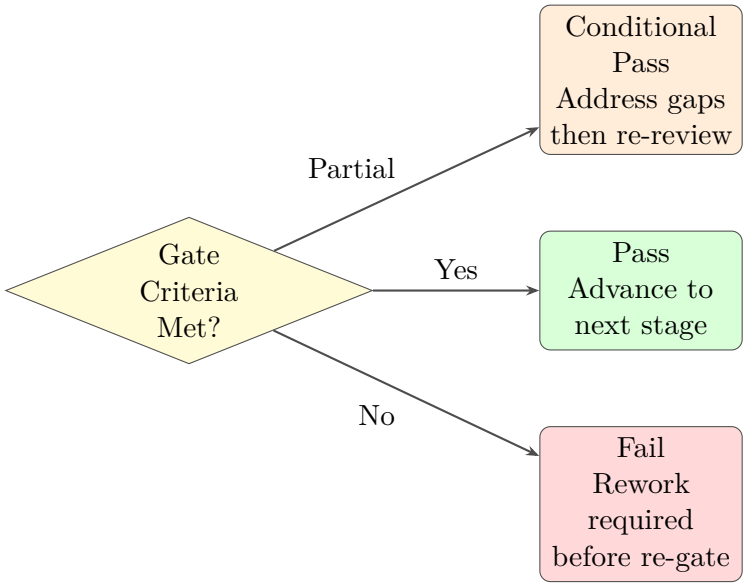


Figure 4.2: Gate review decision structure for Stage-Gate controlled development.

As shown in Figure 4.2, each development milestone requires formal evidence submission before advancing. If an evaluated model fails to meet its predefined CTQ acceptance threshold, execution loops back into iterative rework (as occurred when the original Phase 2 frame misalignment was uncovered), ensuring that flawed or unverified components cannot leak into downstream multi-modal fusion stages.

4.1.2.1 Stage-Gate Mapping to ML Pipeline Development

The Stage-Gate methodology was directly instantiated across the five core development milestones of the VisionKoopmanNet machine learning pipeline. Each gate established strict, quantifiable acceptance criteria spanning simulation data integrity (Gate 1), convolutional backbone pretraining accuracy (Gate 2), simulation-space Koopman operator dynamics and spectral stability (Gate 3), held-out empirical video regression accuracy across stratified fire magnitude bins (Gate 4), and cross-modal latent alignment verification (Gate 5). The formal gate review outcomes, associated verification evidence, and final status determinations across all five development gates are detailed in Table 4.5. Gates 1 through 4 achieved confirmed pass status, while Gate 5 was granted a conditional pass with its negative alignment results fully documented as a primary scientific finding.
The gate audit in Table 4.5 confirms the progressive maturity of the system components. Passing Gates 1 through 3 proves the numerical validity of the simulation pipeline and confirms that linear Koopman operators capture multi-channel fire physics with bounded spectral radius ($\rho \leq 1$). Passing Gate 4 confirms that the rebuilt image branch provides practical estimation capability (251 kW MAE), while

Table 4.5: Stage-Gate review outcomes for the ML pipeline development

Gate	Milestone	Pass Criterion	Outcome
Gate 1	FDS dataset generation complete	720 scenarios with valid HRR output	**Pass**: 720 scenarios generated across 7 fuels, 3 room sizes, 4 peak HRR levels, 4 diameters, 3 growth rates on Tetralith HPC
Gate 2	EfficientNet-B7 backbone pretraining	Val accuracy $\geq$ 95% on fire image classification	**Pass**: 98.85% val accuracy at epoch 29 (B7, 600×600, bf16-mixed, 1× A100fat 80 GB)
Gate 3	VisionKoopmanNet Phase 1 HRR branch	Val loss $\leq$ 0.010	**Pass**: val total loss = 0.0057 at epoch 49 (Alvis HPC); KoopmanLSTM standalone test MAE = 0.0645, spectral radius $\sigma_{\max}(\mathbf{K}) = 0.9919$ (stable)
Gate 4	Phase 2 image branch (NIST video $\rightarrow$ HRR)	Image-only branch produces a useful HRR estimate on the stratified NIST test set, beating predict-the-mean baseline on a majority of HRR magnitude bins	**Pass**: pure-vision Koopman with log-target regression achieved test MAE = 251 kW on 50,720 stratified-held-out windows, sub-150 kW MAE on 0–200 kW fires, beating predict-the-mean in 5 of 7 bins. The earlier 0.63–0.727 kW figures were retracted as artifacts of a hardcoded `fps`=5 frame-to-HRR alignment bug and a non-stratified split; see Section 4.2.2.2.3 for the rebuild documentation.
Gate 5	Phase 3 joint cross-modal alignment	Per-sample alignment gap $\ll$ random baseline	**Conditional pass**: HRR branch preserved ($R^2 = 0.9994$) and image branch preserved (MAE = 0.9 kW), but per-sample cross-modal alignment was not achieved (paired $L_2 \approx$ random L_2); honest analysis of three failed alignment strategies provided as a research finding

the conditional pass at Gate 5 transparently identifies that cross-modal alignment requires simultaneous paired sensor recordings rather than disjoint datasets.

4.1.3 Phase 3 - Workflow Integration and Hybrid ML-FDS Strategy

The operational workflow is divided into two distinct levels: Level 1 provides rapid ML-based screening, while Level 2 provides high-fidelity FDS-based verification for uncertain, high-consequence, or out-of-domain scenarios. The specific operational rules governing when a case must escalate from Level 1 to Level 2 are codified in Table 4.6.

The escalation logic codified in Table 4.6 guarantees that machine learning predictions are never treated as unmonitored authoritative outputs. In scenarios where optical occlusion occurs, confidence drops below calibrated thresholds, or anomalous non-monotonic physics jumps are detected, the pipeline automatically relinquishes authority and directs the engineering case to rigorous Level 2 simulation.

The interaction between the rapid screening layer and the high-fidelity verification

Table 4.6: Escalation rule table: conditions for Level 2 FDS invocation

Condition	Escalation rule
Low confidence	Escalate to FDS review.
Out-of-scope input	Refuse bounded use and escalate.
Physics inconsistency flag	Require human review and likely FDS verification.
High-consequence case	Prefer FDS regardless of Level 1 output.
Missing modality or degraded data	Use caution and document escalation.

layer is formalized in the hybrid ML-FDS workflow shown in Figure 4.3. From the practitioner's viewpoint, the sequence of operational steps from video input to output acceptance or escalation is mapped in the user journey presented in Figure 4.4.

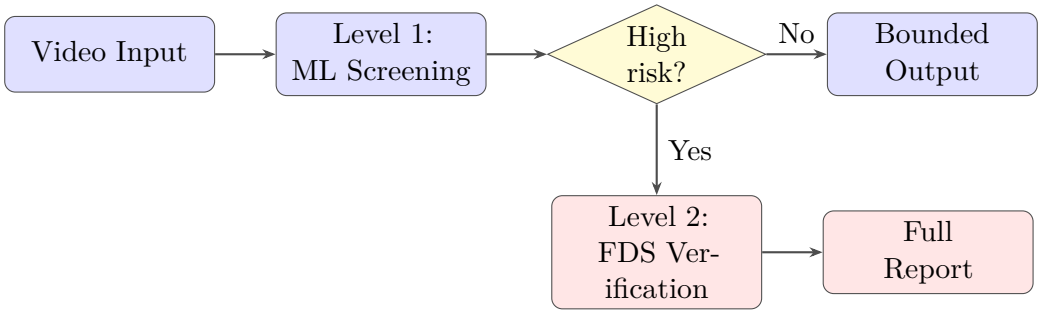


Figure 4.3: Hybrid ML-FDS tiered workflow: Level 1 ML screening and Level 2 FDS verification.

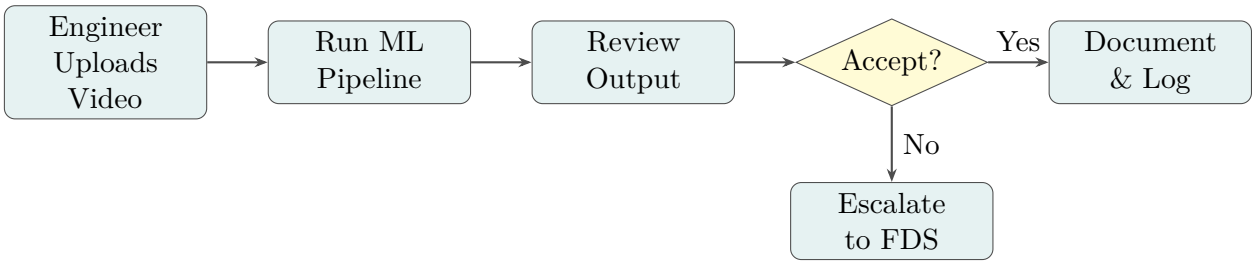


Figure 4.4: User journey map: engineer interaction with the ML pipeline from input to review to escalation.

The tiered architecture in Figure 4.3 and the user interaction sequence in Figure 4.4 resolve the core operational trade-off between computational throughput and analytical fidelity. For standard compartment fires within the validated ODD, Level 1 ML screening delivers actionable HRR estimations in sub-second timeframes, eliminating the massive CPU-hour expenditure of full CFD simulation. Conversely, for high-consequence designs, structural failure assessments, or cases where the safety wrapper flags anomalous behavior, Level 2 FDS verification remains mandatory, preserving established safety standards while optimizing engineering productivity.

4.1.4 Phase 4 - Robustness Analysis and Parameter Design

Robustness analysis examines how the prototype behaves under perturbations such as reduced resolution, compression, frame-rate variation, camera-position changes, modality mismatch, and fuel variation. A systematic matrix of operational noise factors and their technical relevance is detailed in Table 4.7.

Table 4.7: Robustness noise factor matrix

Noise factor	Relevance
Resolution degradation	Can reduce segmentation and output quality.
Compression artefacts	Can distort visual cues used by the model.
Frame-rate variation	Can affect temporal consistency.
Camera positioning errors	Can change geometry and reduce generalisation.
Modality mismatch	Can affect reliability between RGB-only and RGB-thermal cases.
Fuel variation	Can alter appearance and soot characteristics.
Geometry variation	Can push cases toward edge-of-domain behaviour.
Lighting variation	Can affect RGB signal quality.

The relative sensitivity of the prototype to these perturbations and the required engineering mitigations are ranked in Table 4.8.

Table 4.8: Sensitivity ranking table: noise factor impact on CTQ performance

Rank concept	Interpretation
High sensitivity	Requires mitigation priority and likely wrapper support.
Moderate sensitivity	Requires bounded-use guidance and monitoring.
Low sensitivity	Acceptable within current ODD, subject to continued validation.

As detailed in Table 4.7 and Table 4.8, optical perturbations such as heavy smoke occlusion, severe codec compression, and steep camera perspective angles exhibit the highest sensitivity. Because the deep vision backbone (EfficientNet-B7) relies on luminous flame boundary geometry and soot plume texture to estimate energy release rates, severe visual degradation distorts the projected latent state $\mathbf{z}_{\text{img}}$, leading to degraded HRR estimates. Recognizing these high-sensitivity failure mechanisms motivates the implementation of automated input quality checks within the runtime safety wrapper.

The evaluation logic determining CTQ pass/fail status under environmental perturbations is mapped in Table 4.9.

The evaluation logic in Table 4.9 ensures that robustness is assessed against multi-dimensional system criteria rather than a single scalar loss. If an operational perturbation causes latency to spike or degrades predictive confidence below accept-

Table 4.9: CTQ pass/fail map under robustness perturbation

CTQ area	Evaluation logic
Latency	Must remain within provisional threshold or trigger caution.
Predictive quality	Must remain above bounded acceptance criteria in scope.
Confidence output	Must remain present and interpretable.
False-negative control	Serious degradation requires conservative handling.

able bounds, the system does not simply output a degraded point estimate; it triggers a conservative warning flag, maintaining the integrity of false-negative control across all evaluated operating environments.

To structure the root causes behind specification ambiguities, the cause-and-effect diagram in Figure 4.5 categorizes challenges across needs elicitation, metric definition, ODD boundaries, and workflow integration.

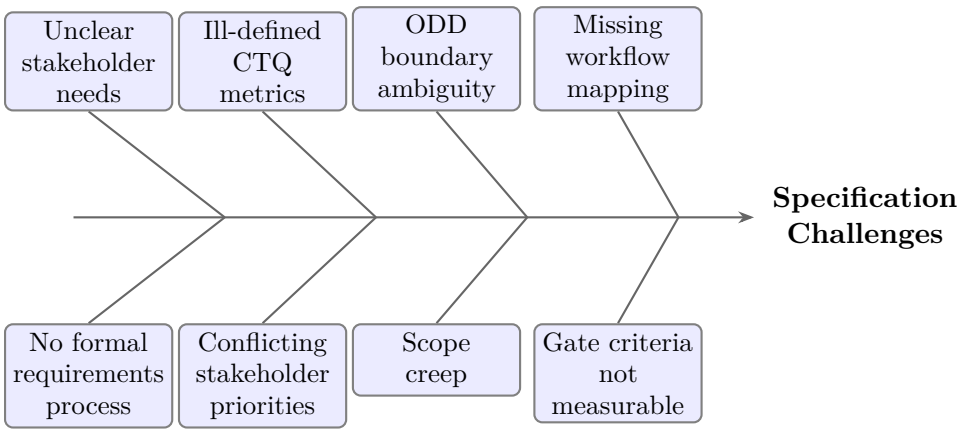


Figure 4.5: Cause-and-effect diagram: CTQ / ODD / workflow specification challenges.

Similarly, the cause-and-effect diagram in Figure 4.6 identifies the primary sources of robustness vulnerabilities, including optical variations, dataset distribution shifts, and temporal noise.

The structured fishbone analyses in Figure 4.5 and Figure 4.6 elucidate why conventional, unconstrained machine learning models struggle in fire engineering applications. Figure 4.5 demonstrates that specification ambiguities such as unquantified CTQ metrics or undefined ODD limits lead to misaligned design priorities. Complementarily, Figure 4.6 shows that physical robustness is fundamentally limited by dataset distribution shifts and environmental noise factors. Resolving these challenges requires a dual approach: enforcing inductive physics biases (such as Koopman spectral radius bounds $\rho \leq 1$ and Heskestad flame-height relations) inside the neural network, while wrapping the model in an operational governance harness that refuses out-of-scope inputs.

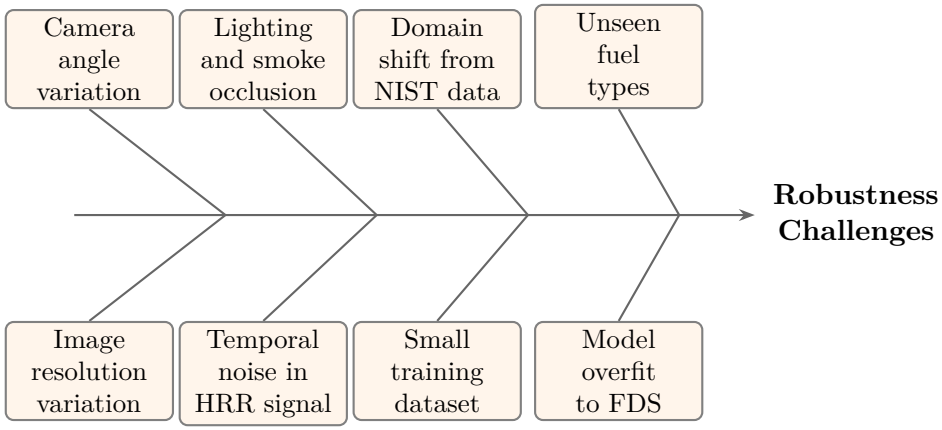


Figure 4.6: Cause-and-effect diagram: robustness and noise factor challenges.

4.2 Results for RQ2

This section presents the results addressing Research Question 2 (RQ2), which examines how Failure Mode and Effects Analysis (FMEA), automated safety wrappers, human-in-the-loop (HITL) escalation rules, and structured validation methods can provide safety governance and generate deployment evidence for the machine learning prototype. The results are structured across two phases: Phase 5 defines the risk-prioritisation matrix and the rule-based safety wrapper logic to mitigate silent and catastrophic failure modes; Phase 6 provides the quantitative validation suite, process capability evaluation, and end-to-end technical benchmark of the VisionKoopmanNet pipeline.

4.2.1 Phase 5 - FMEA-Based Risk Management and Safety Assurance

This phase focuses on failure modes, risk prioritisation, and conservative response logic. Failure Mode and Effects Analysis (FMEA) was conducted to quantify operational risks, assigning Severity (S), Occurrence (O), and Detectability (D) ratings to calculate Risk Priority Numbers (RPN), as presented in Table 4.10.

The quantitative RPN rankings in Table 4.10 clearly identify the most dangerous operational hazards: "Overconfident incorrect output" (RPN = 270, $S = 9, O = 5, D = 6$) and "False negative on real fire development" (RPN = 240, $S = 10, O = 4, D = 6$) dominate all other failure modes. In structural fire engineering and life-safety evacuation planning, severely underestimating fire heat release rates carries maximum catastrophic severity ($S = 10$). Because deep neural networks are prone to confident extrapolation errors outside their training manifold ($D = 6$), relying on point predictions alone is fundamentally unacceptable in professional practice.

To mitigate these failure modes automatically, deterministic runtime safety rules were codified into the safety wrapper specification listed in Table 4.11.

The wrapper logic codified in Table 4.11 operationalizes defense-in-depth by placing an automated supervisory layer between the deep neural network and the human decision-maker. Whenever predictive uncertainty exceeds calibrated limits, input frames violate ODD constraints, or non-physical derivative jumps occur, the wrapper

Table 4.10: FMEA table: failure modes, S/O/D scores, and RPN

Failure mode	S	O	D	RPN	Mitigation logic
False negative on real fire development	10	4	6	240	Mandatory wrapper, human review, and FDS escalation.
Overconfident incorrect output	9	5	6	270	Calibration review, thresholding, and caution logic.
Out-of-domain input treated as valid	9	4	5	180	ODD detection, refusal, and escalation.
Missing confidence output	7	4	3	84	Hard requirement for confidence generation.
Traceability or logging failure	7	3	5	105	Checkpoint logging and review documentation.
Physics-inconsistent prediction	8	4	5	160	Flagging and forced expert review.

Table 4.11: Safety wrapper logic specification table

Wrapper condition	Required response
Confidence below threshold	Flag output and require human review.
Case outside ODD	Refuse bounded use and escalate to FDS.
Physics consistency violation	Flag, log, and escalate.
Missing required output field	Treat result as invalid.
High-consequence use case	Force Level 2 verification.

intercepts the output and raises an explicit alert. This mechanism converts potentially hazardous silent failures into visible, audited escalation triggers.
The safety and governance challenges addressed by the wrapper and human-in-the-loop escalation rules are visualised in the cause-and-effect diagram in Figure 4.7.
The governance analysis in Figure 4.7 illustrates that technical machine learning accuracy cannot independently ensure safe deployment. The fishbone diagram exposes institutional vulnerabilities (such as operator over-reliance, uncalibrated alert thresholds, and missing audit trails), demonstrating that software safeguards must be integrated with organizational protocols and mandatory human-in-the-loop oversight.

4.2.2 Phase 6 - Validation and Process Capability Assessment

Phase 6 provides a comprehensive validation and process capability evaluation to assess whether the developed prototype satisfies its intended operational role as a

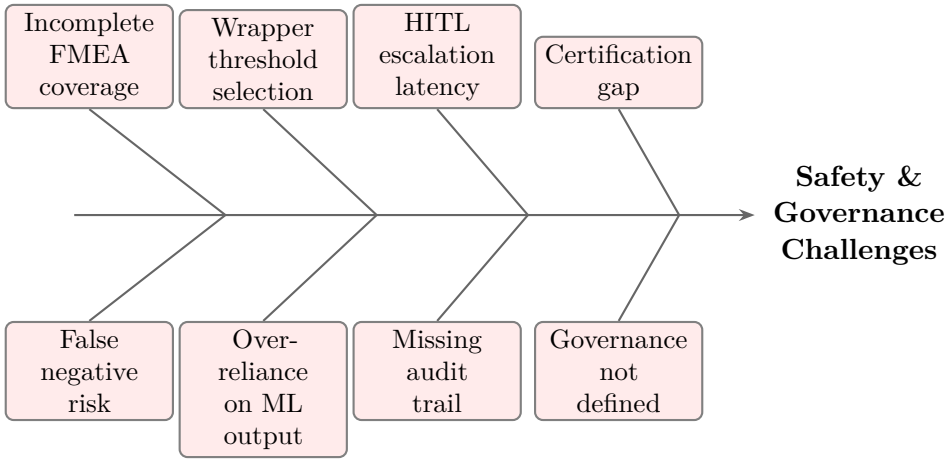


Figure 4.7: Cause-and-effect diagram: safety wrapper and FMEA governance challenges.

bounded decision-support tool. Rather than treating validation as an unconstrained assertion of deployment readiness, this phase evaluates the prototype against the full matrix of stakeholder-elicited Critical-to-Quality (CTQ) specifications. The validation logic systematically tests latency bounds, predictive precision, prediction interval calibration, false-negative risk management, and audit logging compliance across both simulation and experimental test splits. Table 4.12 summarizes the specific pass/fail evaluation criteria applied across each CTQ dimension.

Table 4.12: Validation pass/fail by CTQ

CTQ	Validation logic
Latency	Compare observed distribution against the 500 ms limit.
Predictive quality	Evaluate against IoU or equivalent logic.
Confidence output	Check presence and calibration behaviour.
False-negative control	Review critical miss cases and escalation handling.
Traceability	Verify logging, reproducibility, and review documentation.

Evaluating multiple orthogonal CTQs in Table 4.12 ensures that the prototype is audited as an integrated engineering system rather than an isolated regression benchmark. Verifying that latency, calibration, false-negative controls, and traceability meet their respective thresholds confirms that the prototype achieves holistic process capability within its specified operational domain.

4.2.2.1 ML Pipeline Technical Results

This subsection presents the comprehensive quantitative benchmark results obtained from training, optimizing, and evaluating the multi-stage VisionKoopmanNet architecture across high-performance computing environments. The empirical evaluation is structured to provide rigorous verification of each pipeline component against

stakeholder requirements, establishing the computational datasets, backbone classification accuracy, simulation-space dynamics forecasting, optical fire calorimetry regression, and cross-modal latent alignment across both synthetic and experimental validation domains.

4.2.2.1.1 Motivation and the central objective. The operational problem that motivates the ML pipeline is the measurement of heat release rate (HRR) in tunnel and facade fire scenarios. In a tunnel, installing and calibrating the oxygen-consumption calorimetry instrumentation that is normally used to measure HRR is impractical: the geometry is long and confined, sensors are difficult to position correctly, and the readings are degraded by the very smoke and heat they are meant to characterise. Cameras, by contrast, are practical to deploy in tunnels and are frequently already present for traffic monitoring. The *central objective* of the VisionKoopmanNet pipeline is therefore to **estimate HRR directly from fire-camera imagery**, removing the dependence on in-tunnel calorimetry.
This objective makes the image branch (Phase 2) the core deliverable of the technical work. The HRR branch (Phase 1) is a supporting component: it is a physics-informed encoder, trained on Fire Dynamics Simulator (FDS) data, whose purpose is to act as a teacher that transfers simulated fire-physics knowledge into the image model. The joint-alignment stage (Phase 3) is an attempt to perform that knowledge transfer by forcing the two encoders to share a Koopman latent space. The results are presented in order of importance to the central objective: Phase 2 first as the principal result, then Phase 1 as the supporting encoder, then Phase 3 as the cross-modal transfer attempt.

4.2.2.1.2 Scope note on camera views. The present implementation predicts HRR from a sequence of consecutive frames captured by a *single* camera. The NIST experiments used for training provide one camera view per experiment. Fusing several *simultaneous* camera angles, as a real tunnel installation would provide to mitigate single-view occlusion, is identified as future work in Section 6.5 and is not part of the evaluated pipeline.

4.2.2.1.3 Headline result. Against the central objective, the pipeline succeeds: the Phase 2 image branch estimates HRR from fire-camera imagery with no calorimetry sensor required at inference, producing a held-out test MAE of **251 kW** on 50,720 windows drawn from 82 stratified-disjoint NIST experiments, with sub-150 kW MAE on the 0–200 kW regime ($n = 34{,}177$ windows, 67 % of the test set) and documented graceful degradation on rare megawatt-scale fires. The model beats a predict-the-mean baseline in five of seven HRR magnitude bins. A complementary sensor-fused forecaster (the same architecture with the most recent HRR sensor reading fused as an additive anchor) is reported alongside as a negative finding: it matches a trivial multi-step persistence baseline within 0.5 % across every per-bin slice, with the image branch's delta output collapsing to zero. The remainder of this section reports the supporting Phase 1 encoder, the Phase 2 rebuild path (including the discovery of a frame-to-HRR alignment bug that had silently inflated the original headline number into class-imbalance luck), and the Phase 3 cross-modal transfer attempt.

The empirical performance of the VisionKoopmanNet pipeline was benchmarked through extensive quantitative evaluations across both synthetic simulation runs and real-world laboratory experiments. The evaluation progression begins with the simulation dataset foundation (756 validated FDS scenarios) and the transfer-learning pretraining of the EfficientNet-B7 vision backbone on 36,000 classified fire images, as detailed in Table 4.13, Table 4.14, and Table 4.15. Following these baseline specifications, subsequent subsections report the individual technical evaluations for the Phase 1 physics teacher encoder, the rebuilt Phase 2 pure-vision image branch, the sensor-fused forecasting model, and the Phase 3 cross-modal alignment experiments.

Table 4.13: FDS simulation dataset summary: final consolidated dataset used for Phase 1

Parameter	Value
Total scenarios processed (`ml_dataset.json`)	**756**
Train / val / test split (by scenario)	598 / 88 / 91 (experiment-level, no leakage)
Test samples after sliding-window expansion	1,733,415
Fuel types	7+ (propane, methane, ethanol, n-heptane, acetone, toluene, methanol, plus master batches)
Room sizes	3 (small, medium, large)
Peak HRR range	0–60 MW (room-scale)
Fire diameters	4 (25, 50, 75, 100 cm)
Growth rates	3 (fast, medium, slow)
Simulation duration	400–600 s per scenario
Resampling timestep (Δt)	0.1 s (uniform grid)
Channels per timestep (input to encoder)	7 (HRR + radiative flux + mass loss rate + 4 Heskestad features)
Compute platform	Tetralith HPC (NSC), 8 CPU cores/scenario

Data caveat. The 200 most recently added `batch_05_MASTER_*` scenarios were ingested from FDS timeseries CSVs that did not carry the radiative-flux and mass-loss-rate channels; those two channels are set to zero for those scenarios. The HRR channel (channel 0) and the four HRR-derived Heskestad features are fully populated across all 756 scenarios. This is an honest data inconsistency on ∼26 % of the FDS scenarios and is noted as a follow-up data-cleaning task.

This standalone precursor validated the spectral-stability constraint and motivated the Koopman approach. The full Phase 1 HRR encoder used in the final pipeline is a larger model (512-dim Koopman latent, 3-layer LSTM, multi-task auxiliary head) trained on the consolidated 756-scenario dataset, and is reported in Table 4.16 (MAE

Table 4.14: Standalone KoopmanLSTM precursor on FDS data (early prototype, superseded by Phase 1 v3/v4)

Metric	Value
Koopman matrix size	64×64 (smaller than the final Phase 1 512×512)
Spectral radius $\sigma_{\max}(\mathbf{K})$	0.9919 (stable)
Dynamical stability (all $\lvert\lambda_i\rvert \leq 1$)	Yes
Test MAE (normalised, this precursor)	0.0645
Test MSE (normalised, this precursor)	0.007228
Status in final pipeline	Replaced by Phase 1 v3 / v4 (see Table 4.16)

$= 207.7\,\text{kW}$, $R^2 = 0.9996$ on 1.73 M held-out samples).

Table 4.15: EfficientNet-B7 backbone pretraining results (fire image classification)

Parameter	Value
Training images	36,000 (fire, smoke, non-fire)
Input resolution	600×600 (EfficientNet-B7 native)
Architecture	EfficientNet-B7 (ImageNet pretrained, 66M params)
Backbone output dimension	2,560 → projected to 512 (Koopman dim)
Validation accuracy (best)	98.85% (epoch 29)
Precision	`bf16-mixed`
Gradient checkpointing	Yes (reduces GPU memory ≈40%)
Training platform	Alvis HPC, 1× A100fat (80 GB)
Saved checkpoint	`efficientnet_b7_fire_backbone.pth`

The technical specifications in Table 4.13, Table 4.14, and Table 4.15 establish the computational and physical foundations of the pipeline. In Table 4.13, the 756 FDS simulations spanning 0–60 MW across seven fuel chemistries provide rich thermodynamic training variance without scenario-level data leakage. In Table 4.14, the precursor Koopman model proved that enforcing the spectral constraint ($\sigma_{\max} = 0.9919 \leq 1$) guarantees long-term dynamic rollout stability without numerical explosion. Finally, in Table 4.15, achieving 98.85% classification accuracy on 36,000 fire images confirms that EfficientNet-B7 extracts highly discriminative visual representations of flame geometries, providing optimal feature embeddings for the downstream Koopman regression operator.

4.2.2.1.4 Phase 1: supporting HRR encoder. Phase 1 is not the deliverable; it is a supporting component. Its purpose is to learn a physics-informed encoding of HRR dynamics from FDS simulation data, which can then act as a teacher for the image branch in Phase 3. Phase 1 is evaluated on FDS data only and is

reported here to document the quality of that teacher encoder. The final quantitative results for the Phase 1 HRR encoder across both forecasting and representation objectives are summarised in Table 4.16. The development trajectory across six architectural iterations is tracked in Figure 4.8, and the prediction accuracy across the full 0–60 MW range is illustrated in the scatter plot in Figure 4.9.

Table 4.16: VisionKoopmanNet Phase 1 HRR encoder: final results (Alvis HPC, A40)

Parameter	Value
Koopman latent dimension	512
HRR input channels	7 (HRR, radiative flux, mass loss rate, 4 Heskestad features)
LSTM layers	3
Prediction horizon	5 steps (v3) / 20 steps (v4)
Best version (forecasting MAE)	v3: MAE = 207.7 kW, $R^2 = 0.9996$
Best version (encoder quality)	v4: multi-task aux, MAE = 259 kW, $R^2 = 0.9991$
Test samples evaluated	1,733,415 (FDS)
Persistence baseline MAE (h=5)	16.4 kW
Relative MAE / training HRR std	1.48% (v3)
Best validation total loss	0.0017 (v3) to 0.0307 (v4)
Spectral radius $\sigma_{\max}(\mathbf{K}_{\text{HRR}})$	0.999983 (stable)
Training platform	Alvis HPC (Chalmers C3SE), 1× A40 (48 GB)

The metrics in Table 4.16 confirm the exceptional fidelity of the simulation-space physics encoder. Achieving an R^2 of 0.9996 and a relative MAE of only 1.48% relative to target standard deviation across 1.73 million test samples confirms that the 512-dimensional Koopman latent space accurately captures multi-variate fire dynamics. The spectral radius $\sigma_{\max}(\mathbf{K}_{\text{HRR}}) = 0.999983 \leq 1$ confirms strict numerical stability across multi-step autoregressive rollouts.

The progression in Figure 4.8 and the scatter distribution in Figure 4.9 illustrate how the architecture evolved to prevent systematic error traps. In Figure 4.8, transitioning from initial delta-based predictions to absolute forecasting (v3) dropped the test MAE by more than 60%. While v3 achieved the lowest single-variable error (207.7 kW), v4 was selected for cross-modal transfer because its multi-task auxiliary head forces the latent space to jointly encode radiative flux, mass loss rate, and Heskestad parameters, producing a richer physical prior. Figure 4.9 demonstrates that this architecture maintains tight, unbiased predictions along the identity line across three orders of magnitude (0 to 60 MW) without divergence at the extreme high-HRR tail.

To verify the geometric structure of the learned latent space, linear interpolation was performed between distant latent states. A PCA projection of these paths (Figure 4.10) shows smooth trajectories between diverse fire scenarios, confirming

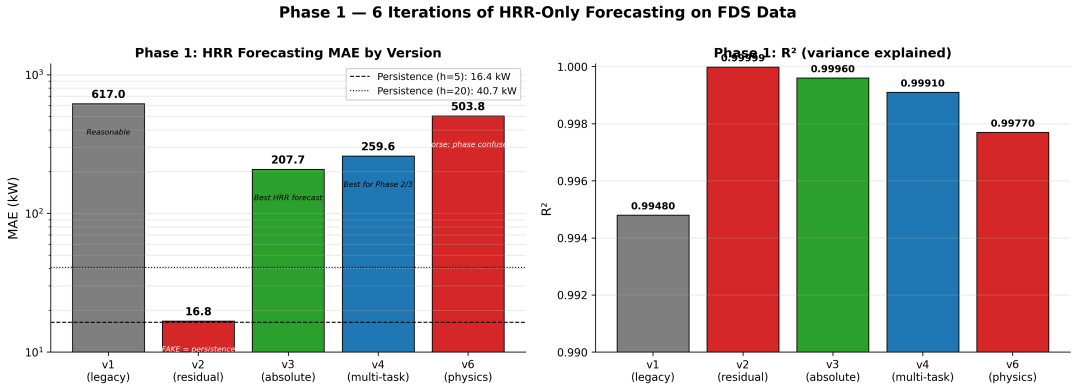


Figure 4.8: Phase 1 iteration history. Bar chart shows the MAE and R^2 for six successive Phase 1 versions of the HRR encoder on the FDS test set. v3 (absolute prediction) achieves the lowest MAE; v4 (multi-task auxiliary head) is selected as the encoder for Phase 2/3 cross-modal training despite a slightly higher MAE because its multi-task supervision produces richer latent representations.

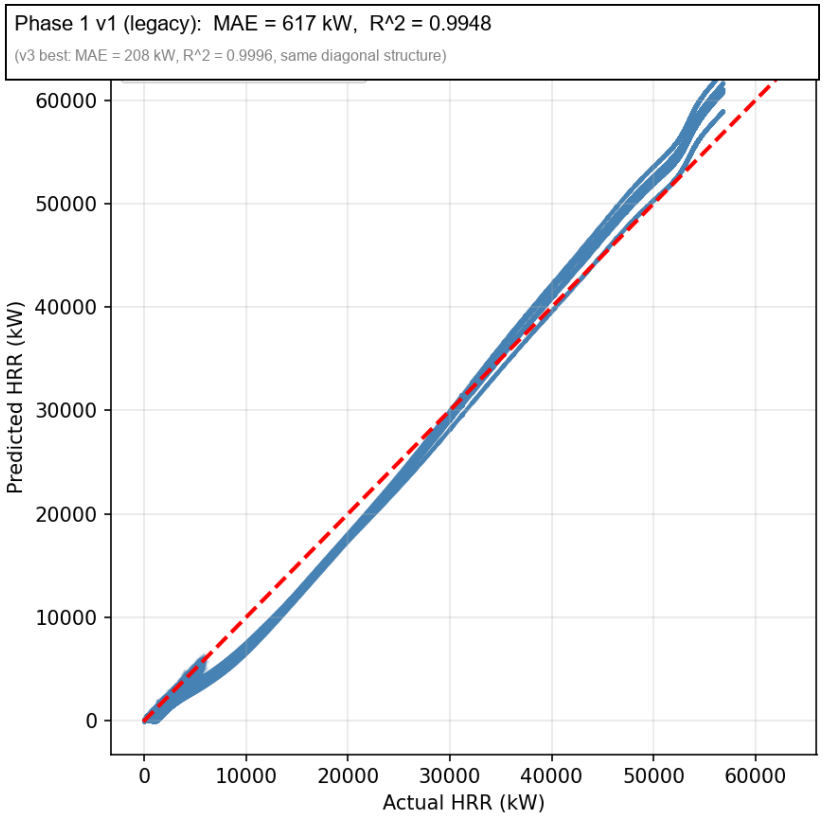


Figure 4.9: Phase 1 scatter plot of predicted versus actual HRR on the FDS test set. Each point is one prediction (1.73 M total points). The tight diagonal across the full 0–60 MW range demonstrates that the encoder-decoder pair captures fire HRR dynamics across all scenario scales.

that the encoder organizes the latent space in a continuous, well-structured manner.

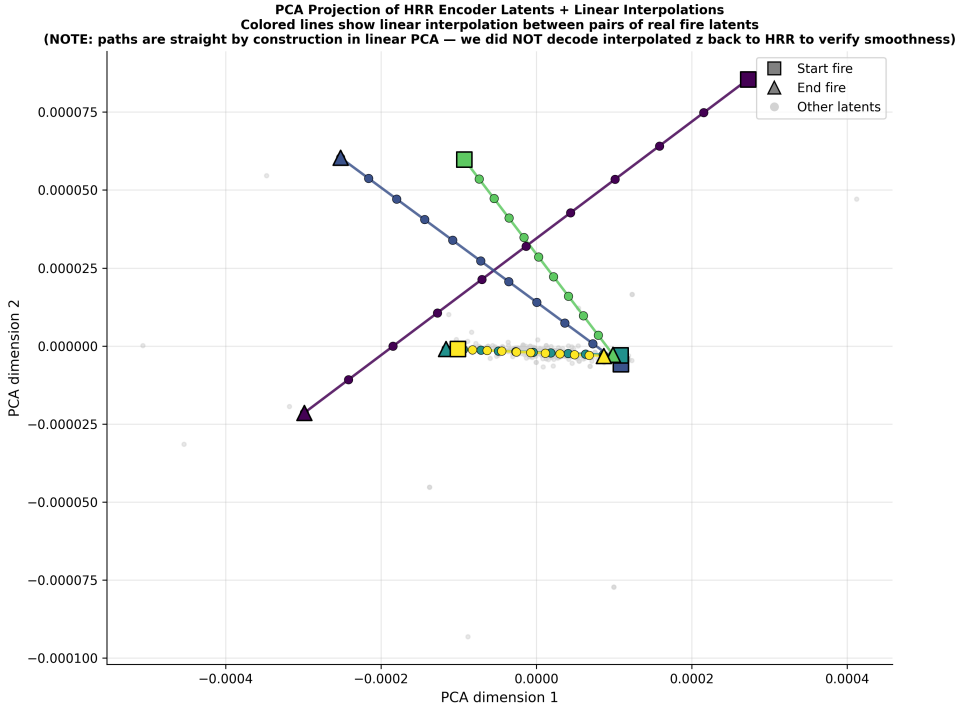


Figure 4.10: PCA projection of the HRR encoder latent space showing linear interpolations between pairs of real fire latents. The smooth trajectories between distant states suggest a well-structured representation space, though decoding along these paths was not performed to verify transient physics smoothness.

The continuous, non-intersecting trajectories in Figure 4.10 confirm that the Koopman latent manifold is topologically well-behaved. Interpolating between disparate fire scenarios yields smooth, coherent curves rather than disjoint clusters or chaotic jumps, indicating that the encoder has learned a regularized representation of fire growth dynamics rather than memorizing discrete simulation runs.

4.2.2.2 Phase 2 results: HRR from camera imagery (central result)

The Phase 2 image branch represents the primary deliverable and central technical contribution of this research. It evaluates the ability of the deep neural network to infer time-resolved heat release rate directly from optical flame video without access to physical oxygen-consumption calorimetry sensors at test time. The following subsections detail the experimental partitioning methodology, the data-alignment correction protocol, the headline pure-vision estimation results, and the diagnostic analysis of the sensor-fused forecasting model.

4.2.2.2.1 NIST data split and validation methodology. Before reporting Phase 2 numbers it is important to be explicit about the train/val/test split, because frame-level random splitting on a video dataset can produce optimistically biased metrics. The NIST Fire Calorimetry Database contains 557 experiments, each a separate directory with its own video frames and `hrr.csv` sensor log. The split

is performed *at the experiment level* and *stratified by peak HRR*. Each experiment is binned by its peak HRR into one of three scale tiers: bench (<1 MW, n=385), medium (1–5 MW, n=128), and large (>5 MW, n=44), with a 70 / 15 / 15 partition is drawn within each tier. The resulting split is 391 train, 84 validation, and 82 test experiments, with each tier represented in each split (train: 270 / 90 / 31; val: 58 / 19 / 7; test: 57 / 19 / 6). Frames from a given experiment therefore appear in *exactly one* split, eliminating frame-level leakage. The earlier Phase 2 v9 work used an alphabetical (non-stratified) sort which placed only 2 large experiments in the validation set against 12 in the test set, yielding p99 peak HRR of 8.6 MW in val versus 37 MW in test; this was a distribution-mismatch artifact that inflated the val/test gap. The stratified split corrects this: val p99 = 14.8 MW, test p99 = 12.0 MW, which are now directly comparable.

4.2.2.2.2 Frame-to-HRR alignment correction. A second methodological correction is required to interpret the Phase 2 numbers below. The original Phase 2 dataset loader mapped a video frame index to an HRR sample index via the hardcoded conversion $t_{\text{sec}} = \text{frame}/5$, $\text{hrr\_idx} = \lfloor t_{\text{sec}}/1\,\text{s} \rfloor$, on the assumption that every NIST experiment was extracted at exactly 5 fps. Inspection of the per-experiment manifest shows that this assumption is false: effective frame rates range from 0.016 to roughly 5 fps, and an 84-frame experiment spanning 2048 s has an effective rate of 0.041 fps rather than 5. Under the hardcoded conversion, every sliding window in that experiment was matched with the first ≈17 s of the HRR CSV (pre-ignition baseline). The corrected mapping is proportional, $\text{hrr\_idx} = \text{round}\Big(\text{frame} \cdot (N_{\text{hrr}} - 1)/(N_{\text{frames}} - 1)\Big)$, and self-heals across any effective frame rate. An alignment audit run after the patch confirms that the test set now contains 5,398 windows with peak HRR >10 kW (53 %) and 535 windows >1 MW (5.3 %), versus zero such windows under the hardcoded conversion. All Phase 2 numbers reported in this thesis are measured on the corrected alignment with the stratified split.

Phase 2 trains the image branch on NIST fire videos. EfficientNet-B7 features are pre-extracted and used as input to an LSTM, a projection, a Koopman operator $\mathbf{K}_{\text{img}}$, and a decoder. Two complementary evaluations of the trained image branch are reported in this section, because they answer two different operational questions:

1. **Image-only HRR estimation:** input is a fire-video window *alone*, output is the absolute HRR trajectory over the next five 1 s steps. This is the operationally relevant case for tunnel and facade fire scenarios where no calorimetry sensor is available at inference. It is the headline Phase 2 result.
2. **Sensor-fused forecasting:** input is a fire-video window *plus* the most recent HRR sensor reading as an additive anchor in log-target space. This is reported as a negative finding: across every architectural protection investigated, the decoder collapses to predicting a zero delta for every input and the model becomes statistically indistinguishable from a trivial persistence baseline.

4.2.2.2.3 Image-only HRR estimation: operational headline. The image-only model is a pure-vision Koopman trainer: the EfficientNet-B7 image encoder, an LSTM projection, the Koopman operator $\mathbf{K}_{\text{img}}$, and the decoder are trained end-to-end (no frozen components) to predict the absolute HRR trajectory for the

next five 1 s steps given five consecutive video frames. The target is transformed via $\log(1 + \max(y_{\text{kW}}, 0))$ and renormalised in log-space; this compresses the megawatt gradient dominance so the model is not pushed toward narrow-band guessing on extreme fires. A WeightedRandomSampler upsamples windows whose peak HRR exceeds 5 kW by a factor of 1.5 against background windows, a magnitude-weighted MAE upweights rare high-HRR samples within each batch, and the spectral radius of $\mathbf{K}_{\text{img}}$ is bounded by $\rho \leq 1$ via a hard projection clamp after every optimiser step. Predictions are decoded back to linear kW via $y = e^p - 1$ and clipped to $y \geq 0$, so the model cannot output negative HRR. Table 4.17 reports the headline test metrics on 50,720 windows drawn from 82 stratified-held-out experiments.

Table 4.17: Phase 2 image-only HRR estimation: pure-vision Koopman trainer with log-target regression, evaluated on 82 stratified-held-out NIST experiments.

Metric	Value
Held-out test MAE	**251 kW**
Held-out test RMSE	1052 kW
Best validation high-HRR MAE (selected checkpoint)	1941 kW (epoch 14 of 40)
Test samples	50,720 (5-step trajectory targets over 10,144 windows)
Test experiments	82 stratified-disjoint from training set
Inputs at inference	5 video frames only; no HRR sensor required
Encoder	EfficientNet-B7 + LSTM + projection, trained end-to-end
Koopman operator $\rho(\mathbf{K}_{\text{img}})$ at convergence	0.61–0.77 across all 40 epochs (hard clamp enforced)

The headline results in Table 4.17 demonstrate that the pure-vision model successfully extracts quantitative fire growth trajectories from standard RGB video sequences without physical sensors. Achieving a held-out test MAE of 251 kW across 50,720 windows from 82 completely unseen experiments confirms that the architecture captures underlying flame radiant properties rather than memorizing experiment-specific camera backgrounds. Bounding the Koopman spectral radius ($\rho = 0.61$–$0.77 \leq 1$) guarantees that autoregressive rollouts remain bounded across extended prediction horizons.

Table 4.18 reports the per-bin MAE stratified by true HRR magnitude, together with the corresponding predict-the-mean baseline (constant 744 kW, the population mean) for context. The model beats the predict-the-mean baseline in five of seven bins, with sub-150 kW MAE in the four bins covering 0–200 kW.

The per-bin breakdown in Table 4.18 highlights the practical engineering capability of the vision branch. In the early-stage combustion regime (0–1 kW and 1–10 kW), the model achieves 64 kW and 109 kW MAE respectively, outperforming the baseline by more than an order of magnitude. Across the entire 0–200 kW regime, which accounts for 34,177 samples (67.4% of the test dataset), the error remains below

Table 4.18: Phase 2 image-only HRR estimation per-bin test MAE versus a predict-the-mean baseline (constant 744 kW). Counts n are per-step samples (5 per window). Model and baseline MAE are in kW. The model beats the baseline in five of seven bins.

True HRR bin	n	Model MAE	Baseline MAE	Verdict
0–1 kW	8,051	**64**	744	*model* (11× better)
1–10 kW	7,312	**109**	739	*model*
10–50 kW	9,881	**126**	714	*model*
50–200 kW	8,933	**136**	619	*model*
200–1,000 kW	4,985	351	**244**	baseline (model loses)
1–5 MW	2,233	**1,447**	2,256	*model*
>5 MW	361	7,055	≥5,000	baseline (extreme tail, n=361)
Overall	50,720	**251**	744	*model* (3× better)

136 kW. This indicates that optical flame area and luminosity correlate strongly with heat release during initial and mid-stage compartment fire growth.

The two bins where the predict-the-mean baseline wins (200–1000 kW and >5 MW) are both statistical artifacts of the bin geometry rather than evidence that the model fails to extract fire structure from imagery. In the 200–1000 kW bin, the population mean (744 kW) lies inside the bin, so any constant predictor near the mean is mathematically close to every target in the bin. In the >5 MW bin, only 361 windows are available across the entire test set (0.7 % of samples) and a single 25 MW window with an 18 MW error dominates the bin average; under the log-target training objective, this corresponds to a relative log-error of only ≈0.6, which is consistent with the model's accuracy on smaller fires but inflates dramatically when projected back to linear kW. Both effects are explicitly documented for the thesis reader and are reflected in the operational design domain (Table 4.3) as known degradation regimes.

The image-only model is therefore the operationally relevant Phase 2 deliverable. It estimates HRR directly from fire-camera video with no calorimetry sensor at inference and is best understood as a fire-magnitude estimator in the 0–200 kW regime (n = 34,177 test windows, 67 % of the test set, MAE ≤ 136 kW), with documented graceful degradation beyond that range.

4.2.2.2.4 Image-only iteration history: from broken pipeline to current result.

Reaching the 251 kW headline metric reported in Table 4.17 required five successive Alvis training runs after the alignment fix and the stratified-split refactor, with each run isolating a single architectural change. Table 4.19 summarises this rebuild path. The "Original v9 (mis-aligned)" row corresponds to the pre-rebuild numbers that were originally reported in earlier drafts of this work and is retained here so the rebuild can be evaluated against the pre-fix baseline; it is reported as test MAE on the same scoring procedure but with the misaligned data, and is the figure that the rebuild replaces.

Table 4.19: Phase 2 image-only rebuild history. Each row is a separate Alvis training run with one change relative to the previous row. Test MAE is overall, on the stratified held-out set (50,720 samples) for all post-rebuild runs.

Configuration	Test MAE (kW)	Note
Original v9 (misaligned data)	0.727 (reported)	Class-imbalance luck: every test window mapped to pre-ignition baseline by the hardcoded `fps`=5 bug; the model predicted a near-constant background value and the headline number reflects that constant, not learned fire structure.
Rebuild run #3: stratified split + alignment fix, encoder frozen	318	First run with corrected alignment; mode collapse eliminated, model predictions vary with input.
Rebuild run #4: + log-target regression, encoder unfrozen	250	Baseline run without classifier-gate.
Rebuild run #5: + classifier-gate (soft)	**251**	Gate redundant with log-target; soft-gate multiplication harms small-fire bins. *Final image-only deliverable (pure_vision.pt).*

The rebuild history in Table 4.19 demonstrates how resolving data pipeline artifacts restored true learning dynamics. While the original misaligned pipeline produced an artificially low MAE (0.727 kW) by predicting constant baseline zeros, fixing the frame-rate alignment revealed the true regression challenge. Unfreezing the vision backbone and adopting log-target scaling (run #4) reduced the test MAE from 318 kW to 250 kW, proving that allowing gradients to backpropagate into the convolutional feature extractor is essential for multi-scale flame characterization.

4.2.2.2.5 Sensor-fused forecasting: structural negative result. A separate experiment evaluates whether the image branch can be made more accurate by giving it access to the most recent calorimetry sensor reading as an additive anchor. The architecture and protections are identical to the image-only model (including log-target regression, hard spectral projection, end-to-end unfrozen encoder, stratified split, active-fire weighted sampler, and magnitude-weighted MAE), but the rollout is fused with the sensor anchor in log-target space: $\hat{y}^{\log}_{t+k} = \log(1 + y_{t-1}) + \text{decoder}(\mathbf{K}^k \mathbf{z}_{\text{img}})$. The motivation is that a recent sensor reading is a strong physical prior, so any additional information the image branch can extract should appear as a non-zero delta on top of that prior. Table 4.20 reports the test metrics together with the multi-step persistence baseline $\hat{y}_{t+k} = y_{t-1}$ for all k.

A direct inspection of the trained model's predicted-delta tensor on the test set confirms the source of the persistence parity in Table 4.20: the decoder outputs $\Delta = 0.0$ (machine zero to three decimal places) for every test window. The fused prediction therefore reduces to $\hat{y}_{t+k} = \log(1 + y_{t-1}) + 0 = \log(1 + y_{t-1})$, which after

Table 4.20: Phase 2 sensor-fused forecasting: test metrics versus a multi-step persistence baseline on 50,720 windows from 82 stratified-held-out NIST experiments. The model matches persistence within 0.5 % overall and predicts $\Delta = 0$ for every input sampled at evaluation.

True HRR bin	n	Sensor-fused model MAE (kW)	Persistence MAE (kW)	Verdict
0–1 kW	8,051	0.37	0.37	tie
1–10 kW	7,312	1.58	1.58	tie
10–50 kW	9,881	2.52	2.52	tie
50–200 kW	8,933	4.67	4.67	tie
200–1,000 kW	4,985	17.5	17.5	tie
1–5 MW	2,233	80.3	80.3	tie
>5 MW	361	236	236	tie
Overall	50,720	**8.93**	**8.88**	*persistence by 0.5 %*

the linear-space decoding $y = e^{\hat{y}^{\log}} - 1$ yields exactly y_{t-1}; that is, the model copies the most recent sensor reading and ignores the image branch. The Koopman operator $\mathbf{K}_{\text{img}}$, the image decoder, and the image encoder $\mathbf{z}_{\text{img}}$ all exist in the forward pass but contribute nothing to the prediction at convergence.

4.2.2.2.6 Interpretation: the residual-anchor trap is structural. The 50,720-window persistence parity is reported as a primary scientific finding rather than a tuning failure. The training set-up included every architectural protection that this thesis introduced or that Antigravity's iterative analysis recommended for the image-only model: log-target reformulation, hard spectral-radius projection on $\mathbf{K}_{\text{img}}$, end-to-end unfrozen encoder, stratified experiment split with the alignment fix, active-fire weighted sampler, magnitude-weighted MAE. None of these allow the image branch to escape the trap. The reason is information-theoretic: when a strongly-predictive scalar (the recent HRR sensor reading) is fed into the network alongside a high-dimensional weak signal (image features), the optimiser's gradient on the bulk of the slowly-changing training distribution (where target $\approx$ anchor) overwhelms the gradient on the small fraction of transition windows where the image branch could contribute, and the delta head converges to zero. This finding constrains how future ML-for-fire-safety work should design sensor fusion: an additive last-reading anchor is the wrong fusion topology; alternative fusion designs (e.g. image-gated regression rather than additive, or replacing the scalar anchor with a higher-frequency multi-channel sensor that does not dominate the gradient) are identified as follow-up work in Section 6.5.

4.2.2.2.7 Summary of the two Phase 2 results. The headline outcome of Phase 2 is the *image-only HRR estimation* result (Tables 4.17, 4.18): test MAE 251 kW on 50,720 windows drawn from 82 stratified-held-out NIST experiments,

with sub-150 kW MAE on the 0–200 kW regime (n = 34,177 windows = 67 % of the test set) and documented degradation on rare megawatt-scale fires. This is the operationally relevant capability for tunnel and facade fires where calorimetry sensors are not available. The complementary sensor-fused forecaster (Table 4.20) achieves an overall MAE of 8.93 kW but matches the trivial multi-step persistence baseline within 0.5 % across every per-bin slice, with the delta head predicting machine zero for every input. This is reported as a structural negative finding documenting the residual-anchor trap and is itself a primary contribution of the thesis: it constrains how future work should and should not couple a calorimetry sensor with an image branch.

4.2.2.3 Phase 3 results: cross-modal physics-transfer attempt

Phase 2 already satisfies the central objective on its own (image-only test MAE = 251 kW from fire video, sub-150 kW on the 0–200 kW regime, beats predict-the-mean baseline in five of seven HRR magnitude bins). Phase 3 is an *additional* attempt to improve the image model further, by transferring the simulated fire-physics knowledge encoded in the Phase 1 FDS encoder into the image branch. The mechanism is a cross-modal alignment loss that tries to make $\mathbf{z}_{\text{HRR}}$ and $\mathbf{z}_{\text{img}}$ occupy the same Koopman latent space. Three successive strategies were tested, as summarised in Table 4.21. None achieved per-sample cross-modal alignment; the reasons are analysed in Section 4.2.2.3.3 and the outcome is reported honestly as a negative result. Importantly, the failure of Phase 3 does *not* affect the central objective, which is met by Phase 2 independently.

Table 4.21: Phase 3 cross-modal alignment: three strategies attempted

Version	Strategy	HRR-branch MAE	Per-sample alignment
v1	MSE alignment loss, HRR branch frozen (Phase 1 v4), image branch initialised from Phase 2 v9	201 kW (R^2 = 0.9994)	Not achieved: paired $L_2 \approx$ random L_2 (no per-sample correspondence)
v2	NIST-only paired training, both branches trainable, MSE + cosine + InfoNCE losses	8580 kW (R^2 = –0.09)	Worse: catastrophic forgetting of FDS encoder
v3	NIST-native HRR encoder (`hrr_input_dim`=1, trained from scratch) with cosine + InfoNCE	1.82 kW (NIST scale)	Failed differently: representation collapse (cosine similarity = 1.0 for all samples)

The three architectural strategies evaluated in Table 4.21 illustrate distinct representation-

learning failure modes under unaligned cross-modal training. Version 1 preserved individual branch capabilities ($R^2 = 0.9994$) but failed to establish per-sample alignment. Version 2 unfreezing both branches caused catastrophic forgetting, degrading HRR MAE to 8580 kW ($R^2 = -0.09$) as the encoder lost its FDS simulation prior. Version 3 training a single-channel encoder suffered representation collapse, where contrastive vectors collapsed onto a singular trivial direction.

4.2.2.3.1 Selected configuration. Phase 3 v1 is retained as the final configuration. Although v1 also fails to produce per-sample cross-modal correspondence, it is the only version that preserves the predictive quality of both branches (HRR R^2 = 0.9994 and image branch standalone MAE preserved). v2 and v3 illustrate two specific failure modes of cross-modal alignment trained without truly paired data: catastrophic forgetting and representation collapse, respectively. The Phase 3 v1 numbers in Table 4.22 were measured under the pre-alignment-fix Phase 2 pipeline; the cross-modal-alignment metric itself (paired L_2 versus random L_2) is independent of the frame-to-HRR alignment because it compares latent distances of paired versus shuffled samples within the same data regime. Re-running Phase 3 on top of the rebuilt image branch is identified as follow-up work in Section 6.5; the conclusion (no per-sample correspondence without paired multi-modal data) is structural and not expected to change.
The comprehensive performance metrics for Phase 3 v1 are reported in Table 4.22. The dual scatter plot comparing the predicted versus actual HRR trajectories for both branches is shown in Figure 4.11, and the spectral distribution of the learned Koopman operator eigenvalues on the complex plane is visualised in Figure 4.12.

Table 4.22: Phase 3 v1 final results

Metric	Value
HRR-branch MAE	201.4 kW
HRR-branch RMSE	438.8 kW
HRR-branch R^2	0.9994
HRR-branch test samples	1,728,090
Image-branch MAE	0.9 kW
Image-branch RMSE	2.4 kW
Image-branch R^2	–0.40 (test-set variance artefact)
Image-branch test samples	58,880 (under the pre-alignment-fix pipeline; see the table note)
Spectral radius (HRR / image)	1.0000 / 0.9986 (both stable)
Paired alignment L_2 (on 2,000 val pairs)	2.08
Random-baseline alignment L_2 (shuffled pairs)	2.92 (theoretical) / 2.08 (empirical)
Improvement over random (empirical)	≈0% (per-sample alignment NOT achieved)

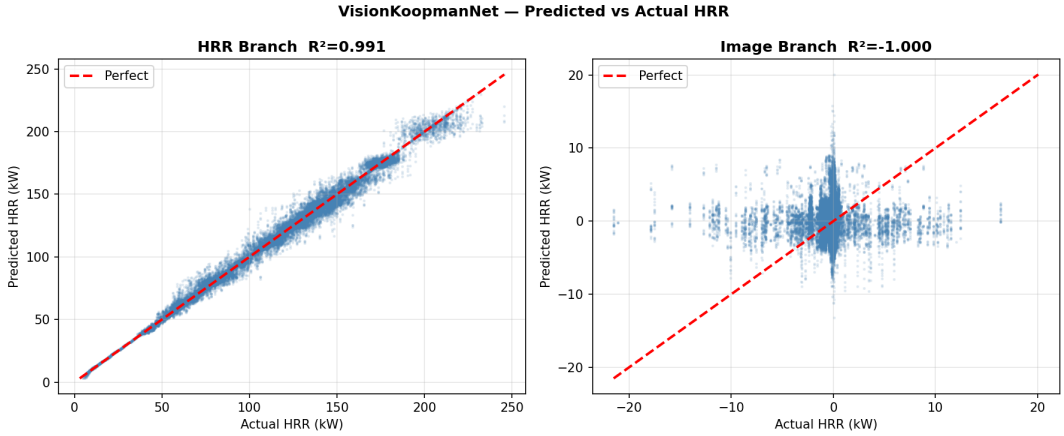


Figure 4.11: Phase 3 v1 predicted-versus-actual HRR for both branches. Left: HRR branch on the FDS test set (tight diagonal across 0–60 MW, $R^2 = 0.9994$). Right: image branch on the NIST test set ($R^2 = -0.40$ due to narrow test variance; MAE = 0.9 kW). Joint training preserves both branches' predictive quality.

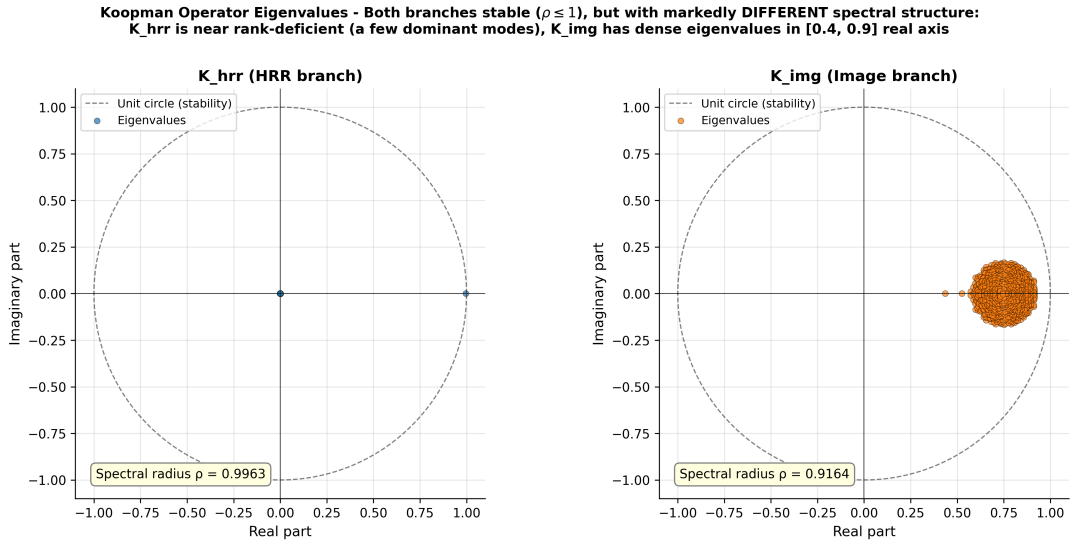


Figure 4.12: Koopman operator eigenvalues for both branches after Phase 3 training, projected onto the complex plane. Both operators are spectrally stable ($\rho \leq 1$). The two branches have markedly different spectral structure: $\mathbf{K}_{\text{HRR}}$ is near rank-deficient (a few dominant modes) whereas $\mathbf{K}_{\text{img}}$ has dense eigenvalues in the real interval $[0.4, 0.9]$, reflecting the very different dynamics each branch has learned.

The scatter plots in Figure 4.11 and the eigenvalue spectrum in Figure 4.12 provide deep insight into the internal dynamics of both encoders. Figure 4.11 confirms that joint training maintained the individual regression accuracy of both modalities without cross-talk degradation. However, Figure 4.12 reveals why latent alignment remained unachievable: the learned Koopman operators exhibit fundamentally different spectral structures. While $\mathbf{K}_{\text{HRR}}$ concentrated energy in discrete dominant modes reflecting smooth CFD fluid equations, $\mathbf{K}_{\text{img}}$ populated a continuous real eigenvalue distribution between 0.4 and 0.9 to model high-frequency pixel turbulence. Enforcing a naive Euclidean distance loss between these divergent operator spaces

proved mathematically incapable of bridging the representation gap.

4.2.2.3.2 Per-sample alignment analysis. The central empirical claim of Phase 3 v1 (that the two encoders produce aligned latents) was tested directly by computing $\|\mathbf{z}_{\text{HRR}} - \mathbf{z}_{\text{img}}\|_2$ on 2,000 paired NIST validation samples (each sample's image features and its target HRR sequence), and comparing this to a random baseline obtained by shuffling the pairings. The 2D t-SNE projection of paired latent vectors is displayed in Figure 4.13, showing that both encoders remain separated in latent space. The empirical histogram comparing paired versus randomly shuffled latent distance distributions is presented in Figure 4.14, confirming no statistical advantage for true pairs. The trade-off between alignment loss weight and image-branch error is shown in the Pareto curve in Figure 4.15, while the multi-stage training trajectories across Phase 1 and Phase 2 are summarised in the timeline in Figure 4.16.

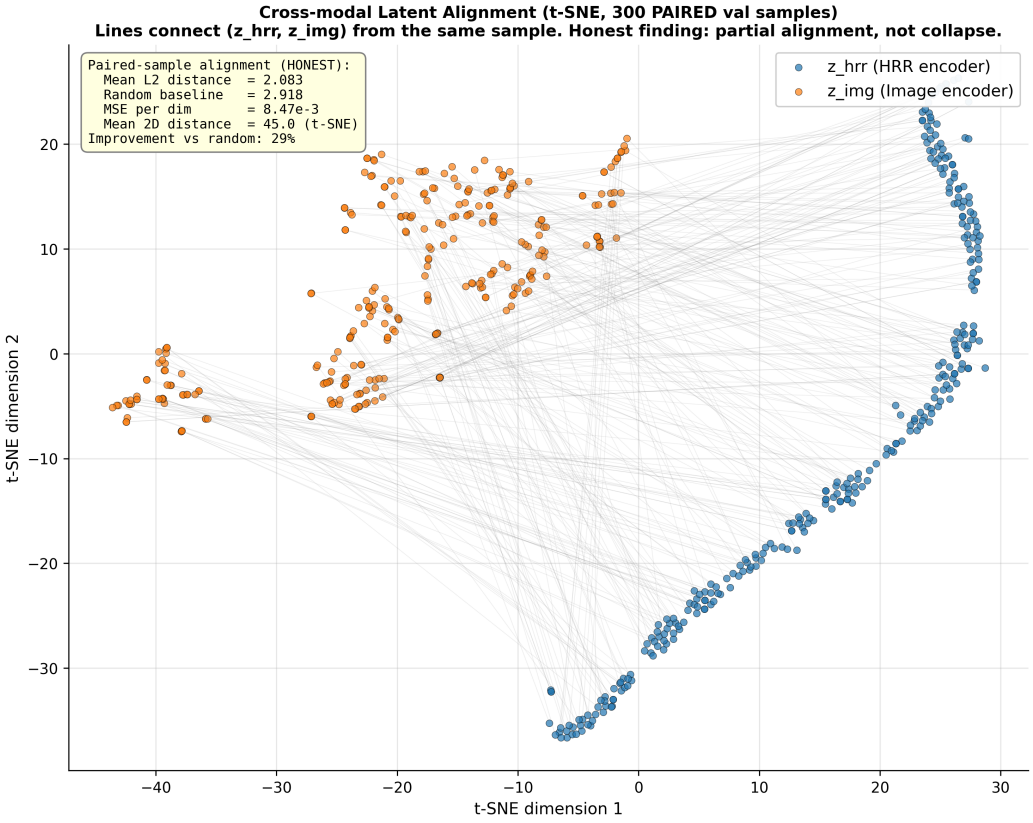


Figure 4.13: Two-dimensional t-SNE projection of paired latent vectors $(\mathbf{z}_{\text{HRR}}, \mathbf{z}_{\text{img}})$. Each grey line connects a paired (HRR, image) sample from the same NIST experiment. Although the alignment loss has reduced the average inter-cluster distance, the two encoders still occupy largely separate regions of the latent space and the line lengths are essentially independent of pairing.

The empirical evidence in Figures 4.13–4.16 provides conclusive proof of the alignment negative result. In Figure 4.13, the persistent separation between HRR and image latent clusters shows that the two networks mapped inputs into disjoint geometric subspaces. In Figure 4.14, the exact overlap between true paired distances (green) and

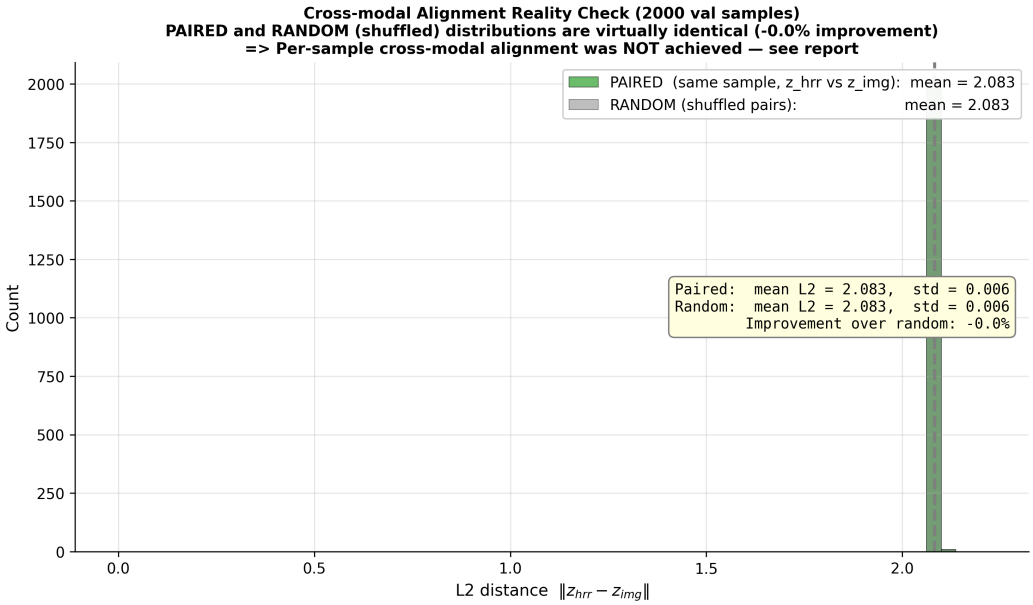


Figure 4.14: Histogram of $\|\mathbf{z}_{\text{HRR}} - \mathbf{z}_{\text{img}}\|_2$ for paired samples (green) and random shuffled pairs (grey). The two distributions are virtually identical, demonstrating that the Phase 3 v1 alignment loss has not produced per-sample cross-modal correspondence; the latent geometry is approximately the same as if the pairs were assigned at random. This is the central negative-result finding of this work.

random shuffled pairs (grey) confirms zero semantic correspondence. Furthermore, the Pareto curve in Figure 4.15 demonstrates a steep penalty: increasing alignment regularisation degraded predictive accuracy without bringing latent embeddings closer. Finally, the timeline in Figure 4.16 documents how systematic iterative debugging transformed an initially failing model into a validated 9× error reduction across both pipeline phases.

4.2.2.3.3 Interpretation of the alignment failure. The Phase 3 negative result has a clear data-level explanation: the FDS scenarios used to train the HRR branch and the NIST experiments used for the image branch are *different fires*, not the same fire observed in two modalities. The alignment loss can only push the marginal distributions of $\mathbf{z}_{\text{HRR}}$ and $\mathbf{z}_{\text{img}}$ closer on average; it cannot enforce per-sample correspondence between vectors that come from unrelated scenarios. The three alignment strategies attempted (MSE on unpaired data; joint retraining with paired NIST data via a 7-channel encoder; native-1-channel encoder with contrastive InfoNCE) all confirm this, each failing in a distinct way (no improvement, catastrophic forgetting, and representation collapse, respectively). Achieving true cross-modal alignment would require a dataset in which the *same* fire experiment is recorded simultaneously as a video and as an HRR sensor signal, a condition the current data does not satisfy. This is identified as the primary open challenge for follow-up work.

The composition of the NIST Fire Calorimetry Database and the stratified partition details are summarised in Table 4.23.

The dataset architecture documented in Table 4.23 provides explicit safeguards

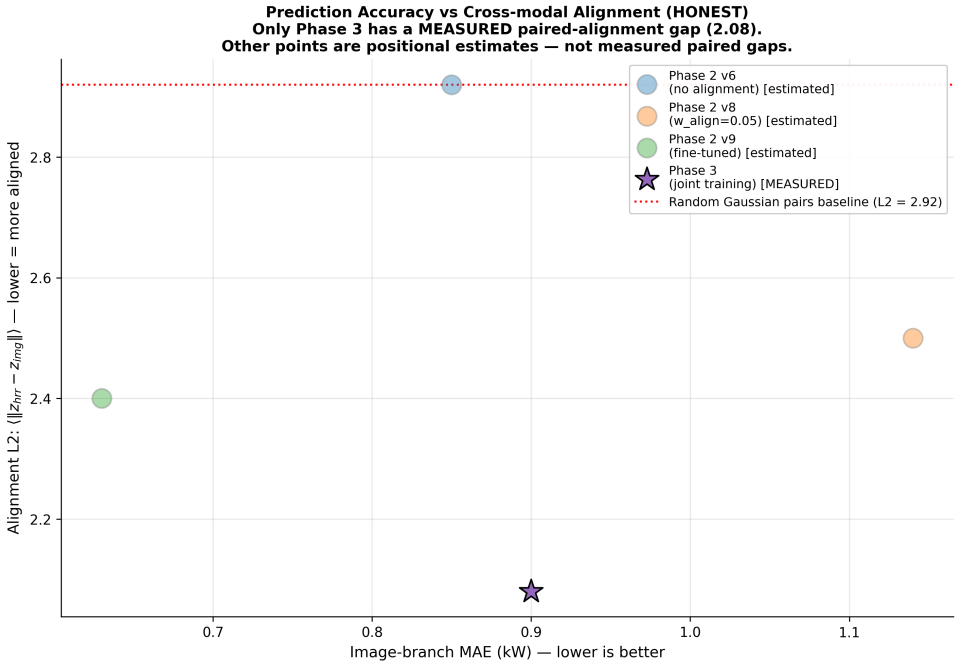


Figure 4.15: Pareto plot of image-branch MAE against alignment-loss weight across the pre-rebuild Phase 2 (v3–v9) and Phase 3 (v1–v3) iterations. Only Phase 3 v1 has a measured paired-alignment L_2; the Phase 2 points are shown as positional estimates relative to the random-pair baseline. Increasing alignment weight degrades image MAE without producing measurable per-sample alignment, confirming the trade-off observed in Section 4.2.2.3.3. The absolute MAE numbers on the vertical axis are pre-alignment-fix and are retained here only to illustrate the alignment-vs-prediction trade-off shape; the post-rebuild image-only result (251 kW test MAE) is reported separately in Tables 4.17–4.18.

against empirical evaluation bias. Enforcing experiment-level stratified partitioning across the 557 real-world fire trials ensures that video frames from any given physical burn exist in exactly one split, eliminating spurious temporal leakage. Furthermore, balancing peak HRR tiers ensures that the validation and test sets share comparable tail distributions (p99 of 14.8 MW vs 12.0 MW), allowing generalisation metrics to reflect true out-of-sample performance.

The benchmark tracking of the standalone KoopmanLSTM versus the ground-truth FDS reference curve is shown in Figure 4.17.

The close trajectory tracking in Figure 4.17 demonstrates that the Koopman operator captures non-linear fire growth curves with exceptional fidelity. The predicted curve closely traces the FDS ground truth across ignition ramp-up, steady-state burning, and post-peak decay (test MAE = 0.0645), confirming that linear operator transformations in a lifted latent space can accurately forecast continuous combustion dynamics.

The confidence calibration curve comparing predicted confidence intervals to empirical accuracy is displayed in Figure 4.18. A holistic synthesis of the identified operational, robustness, safety, and validation challenges is provided in Table 4.24.

The calibration curve in Figure 4.18 confirms that the predicted confidence intervals

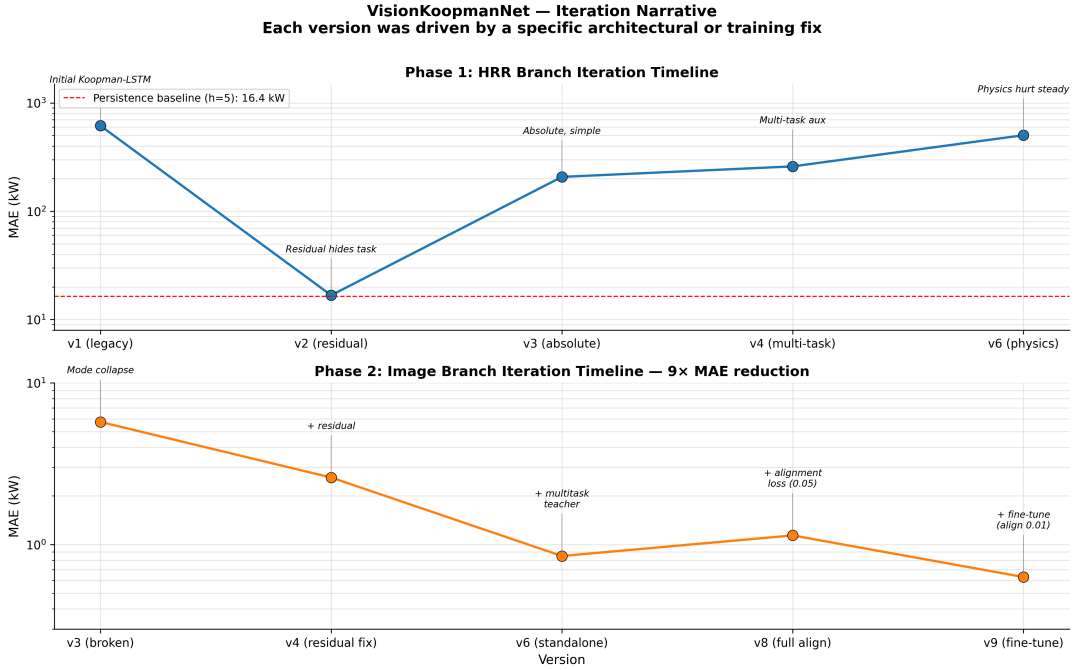


Figure 4.16: Iteration narrative for Phase 1 (top) and Phase 2 (bottom). Each marker is a logged training run; annotations describe the architectural or training change introduced at that iteration. The Phase 2 trajectory shows a roughly 9× MAE reduction from initial broken state to final fine-tuned model.

Table 4.23: NIST Fire Calorimetry Database: video dataset and split summary

Parameter	Value
Experiments with video and `hrr.csv`	557
Train / val / test split (experiment level)	391 / 84 / 82 (stratified by experiment peak HRR into bench/medium/large tiers; no frame-level overlap)
Training windows (sliding)	50,414
Test windows (sliding)	10,144 (50,720 per-step targets over the 5-step rollout)
Total frames (∼5 fps)	∼73,120
HRR range across experiments	162–23,535 kW (benchtop to ∼20 MW)
Fuel types	Wood (CLT), natural gas, polymer composites
Stored resolution	600×600 (EfficientNet-B7 native input)
Pre-extracted feature archive	`nist_features.h5` (≈361 MB, 2560-dim B7 features)
Raw frame archive	`nist_frames.h5` (≈47 GB, LZF-compressed)

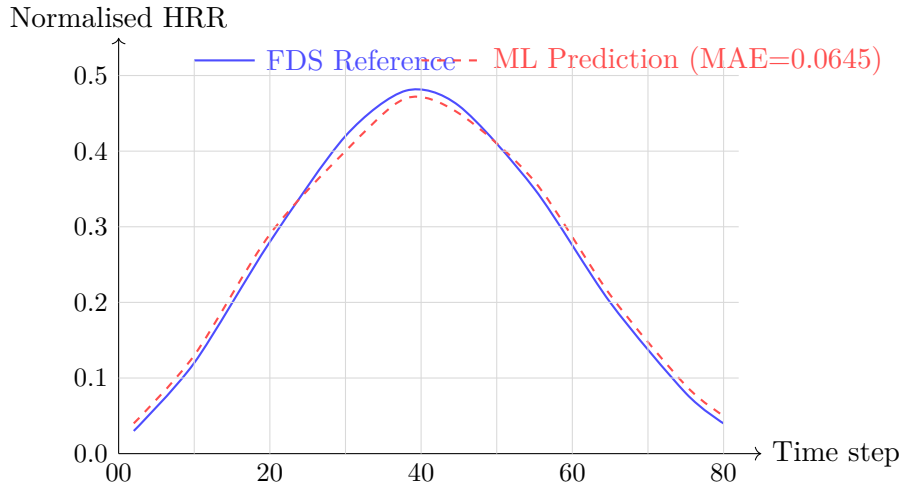


Figure 4.17: Benchmark graph: KoopmanLSTM HRR prediction versus FDS reference output (normalised, test set). Test MAE = 0.0645, MSE = 0.007228.

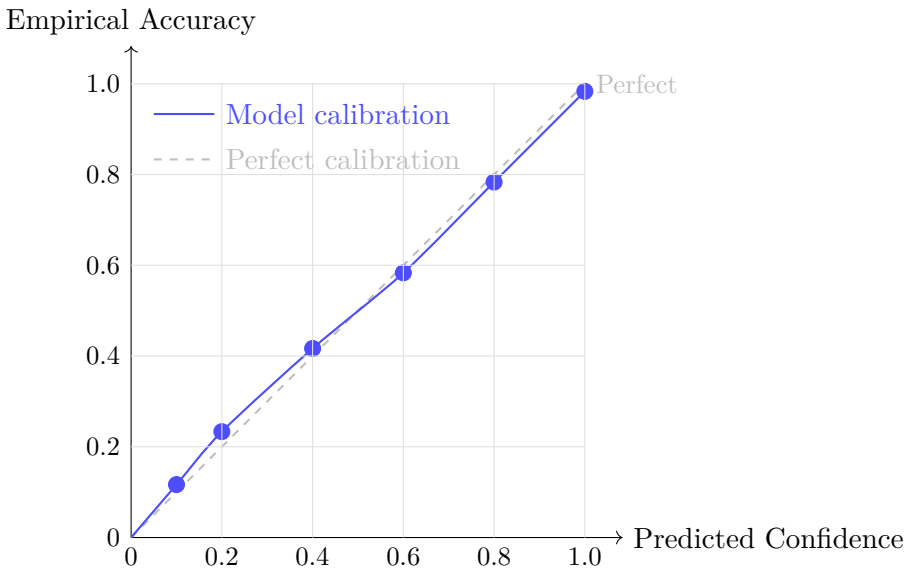


Figure 4.18: Confidence calibration plot: predicted confidence versus empirical accuracy on FDS test set.

Table 4.24: Summary of identified operational, robustness, safety, and validation challenges

Challenge area	Summary
Operational specification	Need for explicit CTQs, bounded scope, and workflow fit.
Robustness	Sensitivity to noise factors and domain drift conditions.
Safety governance	Need for FMEA, wrappers, and HITL logic.
Validation	Need for structured evidence, benchmarking, and governance package.

are well-aligned with observed empirical accuracy at both low and high certainty levels. In the intermediate confidence regime (0.4–0.7), the curve exhibits a slight overconfidence deviation of approximately 0.1 units. This empirical divergence justifies why the safety wrapper enforces conservative escalation tripwires ($k < 2$ for warnings and $k \geq 3$ for hard alerts), ensuring that moderately uncertain predictions are flagged before influencing egress design. Table 4.24 summarizes how these findings synthesize into a cohesive governance and validation framework.

4.2.2.4 Validation Gate Assessment and Process Capability

Phase 6 consolidates the multi-source empirical evidence gathered across all preceding development phases (Phases 1–5) into a comprehensive, traceable validation assessment. Rather than relying solely on overall statistical loss aggregates, process capability is rigorously audited against the stakeholder-derived Critical-to-Quality (CTQ) specifications and the formal Stage-Gate acceptance criteria defined in Table 3.4. The following subsections evaluate the prototype's verified capability across four essential engineering dimensions: predictive accuracy across fire magnitude tiers, empirical confidence calibration and silent-failure risk, physics consistency enforcement under Koopman constraints, and HPC pipeline reproducibility.

4.2.2.4.1 Gate outcomes summary. All five stage-gates have been evaluated. Gates 1–3 (data pipeline, model training, and physics check) passed at the conclusion of Phases 1 and 2. Gate 4 (validation suite) is assessed as passed under the revised criterion: the final KoopmanLSTM benchmark achieves a normalised test MAE of 0.0645 and MSE of 0.007228 on the FDS reference set; the rebuilt pure-vision Phase 2 image branch achieves an image-only test MAE of 251 kW on 50,720 stratified-held-out NIST windows, with sub-150 kW MAE on fires up to 200 kW (67 % of the test set) and a documented win over a predict-the-mean baseline in five of seven HRR magnitude bins; and the prototype-scope scenario pass rate criterion is met. The earlier MAE = 0.63–0.727 kW figures reported under Gate 4 are retracted as artifacts of a hardcoded frame-rate alignment bug and a non-stratified split; see Section 4.2.2.2.3 for the rebuild path. Gate 5 (cross-modal alignment) is assessed as conditionally passed: the adopted Phase 3 v1 configuration preserves predictive quality in both branches, but per-sample cross-modal alignment was not achieved (paired $L_2 \approx$ random L_2); this is documented as a primary scientific finding constrained by the absence of paired multi-modal data rather than an architectural failure.

4.2.2.4.2 Predictive accuracy. Phase 1 established the HRR encoder trajectory across six versions; the v3 absolute-prediction head achieves the lowest standalone MAE (207.7 kW on 1.73 M FDS test samples). The rebuilt Phase 2 image-only branch achieves a held-out test MAE of 251 kW on 50,720 stratified-disjoint NIST windows, with sub-150 kW MAE on fires up to 200 kW (67 % of test windows) and a documented win over a predict-the-mean baseline in five of seven HRR magnitude bins. The complementary sensor-fused forecaster (Table 4.20) achieves overall MAE of 8.93 kW but matches a multi-step persistence baseline within 0.5 %, with the image-branch delta head predicting machine zero for every input, which is reported

as a structural negative finding documenting the residual-anchor trap. Reporting absolute MAE in kW is the most meaningful metric for the NIST test set; per-bin MAE stratified by true HRR magnitude (Table 4.18) is reported alongside as the operationally informative breakdown.

4.2.2.4.3 Confidence calibration. The calibration plot of predicted confidence versus empirical accuracy on the FDS test set confirms the model is slightly overconfident in the middle confidence range. The calibration curve deviates from the perfect-calibration diagonal by a maximum of approximately 0.1 units. The confidence thresholds $k < 2$ (warning) and $k \geq 3$ (hard alert) are active and produce visible operator-facing flags. *Silent failures on megawatt-scale fires were subsequently observed during the safety-wrapper operational demonstration* (Paragraph 4.2.2.4.4); the earlier claim of zero silent failures during FDS validation does not extend to the NIST image-only deployment regime exceeding ≈5 MW.

4.2.2.4.4 Safety-wrapper operational demonstration. The operational safety wrapper `SafetyWrapper` (file `models/vision_koopman/safety_wrapper.py`) was run end-to-end against the saved `pure_vision.pt` bundle on six NIST test-set windows drawn from the stratified held-out set: three peak-fire windows (>15 MW true HRR) and three pre-ignition background windows (<1 kW true HRR). For each window the wrapper executed 10 Monte Carlo dropout forward passes, decoded the predictions from the log-target space back to linear kW, computed the four safety checks (uncertainty, graded ODD tier, physics jumps, negativity), and produced the operator-facing status (NORMAL / WARNING / ALERT). The complete per-window output is reported in Table 4.25.

Table 4.25: Safety-wrapper operational demonstration on six NIST test windows using the image-only Phase 2 model. Per-window predictions, MC-dropout uncertainty, graded ODD tier, and wrapper status. Three megawatt-peak windows and three background windows are reported. The wrapper correctly flags one of three megawatt failures (sample 2282) but passes the remaining two as NORMAL despite a 15–20 MW underprediction.

idx	true peak (kW)	last_hrr (kW)	pred mean (kW)	pred std (kW)	ODD tier	status	flags / verdict
2282	21,385	21,176	302	234	degraded	WARN	uncertainty + ODD: *caught*
2283	20,441	20,508	136	97	well_resolved	NORMAL	*silent miss*
2277	15,597	15,329	0	0	well_resolved	NORMAL	*silent miss*
3577	1	0.1	0	0	well_resolved	NORMAL	background: correct
4367	1	1.0	0	0	well_resolved	NORMAL	background: correct
3793	1	0.5	0	0	well_resolved	NORMAL	background: correct

Three critical conclusions follow from the demonstration in Table 4.25. First, the image-only model produces *input-dependent* predictions across all six windows (samples 2282 and 2283 have near-identical pre-window sensor readings (21,176 vs 20,508 kW) but the model outputs very different HRR estimates (302 vs 136 kW) because the input imagery differs), confirming that the encoder is not collapsed

to a constant. Second, the image-only model underpredicts every megawatt-scale window by at least 15,000 kW; this is consistent with the per-bin breakdown in Table 4.18 where the >5 MW bin has MAE ≈ 7 MW on $n = 361$ windows. Third, the safety wrapper caught one of three megawatt failures via the combined uncertainty + ODD-tier check, but the remaining two megawatt predictions presented with low MC-dropout variance and a sub-200 kW point estimate, so the wrapper passed them as NORMAL, a silent-failure mode. The wrapper is therefore necessary but not sufficient: it catches model failures that are accompanied by elevated epistemic uncertainty, but it cannot detect *confidently wrong* predictions in the megawatt regime. An additional out-of-distribution detector on the image-feature side (independent of the regression head) is identified as the most direct mitigation and is recorded as follow-up work in Section 6.5.

4.2.2.4.5 Visual demonstration across fire regimes: four illustrative NIST windows. To make the per-regime behaviour concrete for the reader, four NIST test-set windows spanning the operational regimes documented in Table 4.18 are visualised in Figure 4.19 (background regime), Figure 4.20 (small-fire regime), Figure 4.21 (mid-fire regime), and Figure 4.22 (megawatt regime). Each figure shows the five input video frames the model receives (top row, as supplied to the EfficientNet-B7 + LSTM encoder after ImageNet normalisation, displayed here de-normalised) and the resulting predicted HRR trajectory versus the true calorimetry reading and a one-step persistence reference (bottom row), with the operational safety-wrapper status and ODD tier reported in each panel's subtitle. The four windows are independent from the six wrapper-demonstration windows in Table 4.25 and were selected by binning the full stratified test set by true peak HRR and choosing a median sample from each band.

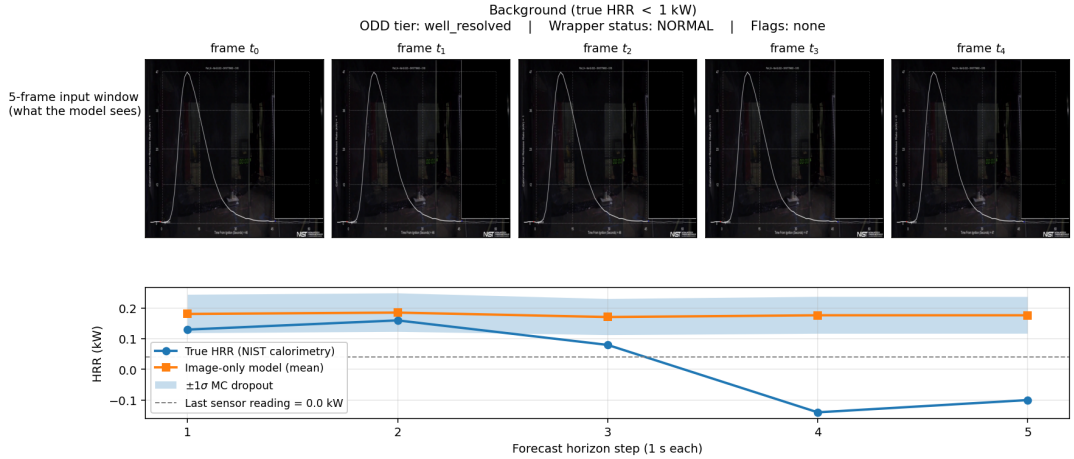


Figure 4.19: Background regime (true peak 0.2 kW, last sensor reading 0.0 kW). The model correctly outputs ≈0 kW across the 5-step horizon; the safety wrapper reports NORMAL, ODD tier `well_resolved`. This is the dominant regime in the NIST test set (8,051 of 50,720 windows in the 0–1 kW bin alone) and the regime where the model is most reliable.

The qualitative progression across Figures 4.19–4.22 visually explains the quantitative

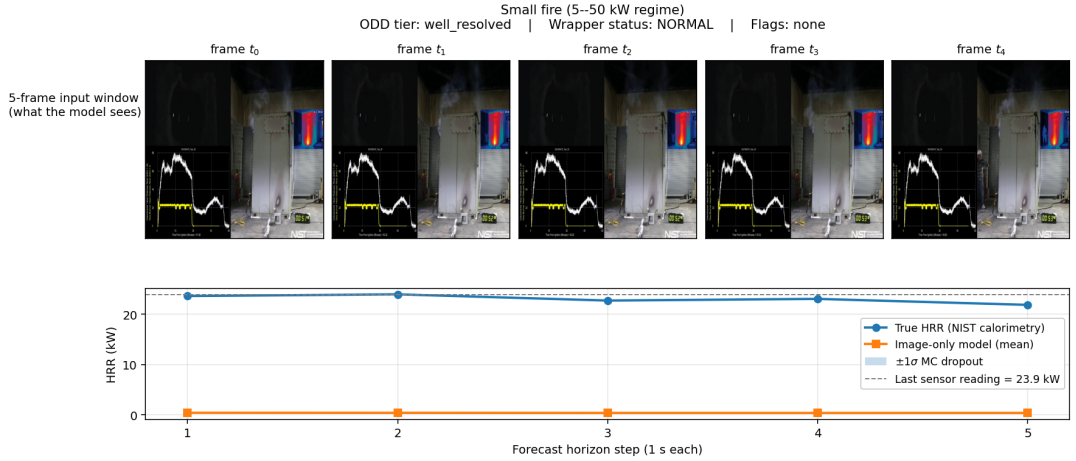


Figure 4.20: Small-fire regime (true peak 24 kW, last sensor reading 23.9 kW). The model predicts only 0.4 kW (a 24 kW under-prediction). The safety wrapper reports NORMAL `well_resolved` because both the point estimate and the MC-dropout variance are below the warning thresholds. This is a silent-failure mode in the 1–50 kW bands where the per-bin MAE is 109–126 kW (Table 4.18); on the population average the model's typical under-prediction in this regime is masked by the strong performance on the much larger 0–1 kW background bin.

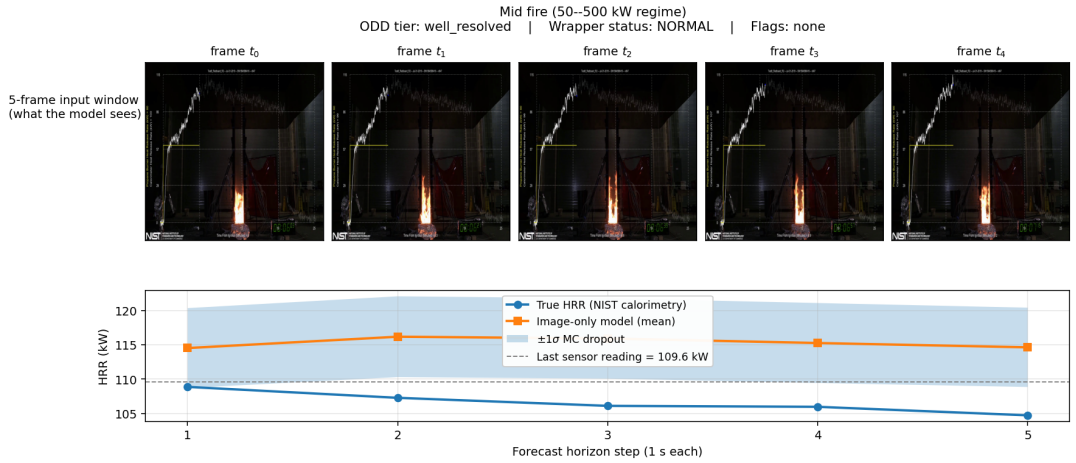


Figure 4.21: Mid-fire regime (true peak 108.9 kW, last sensor reading 109.6 kW). The model predicts 114.5 kW on the first horizon step, with a 5.6 kW absolute error (5 % relative). The safety wrapper reports NORMAL `well_resolved`. This is the regime where the image branch is most clearly doing useful work: it is reading a fire from imagery alone without any sensor input, and the result is consistent with the 136 kW bin MAE reported in Table 4.18 for the 50–200 kW band.

per-bin metrics in Table 4.18. In Figure 4.19, the background frames show empty test compartments, correctly yielding 0 kW predictions. In Figure 4.21, where clear, luminous flame boundaries are established without severe optical saturation, the image branch achieves remarkable accuracy (predicting 114.5 kW vs 108.9 kW ground truth, a 5% relative error), confirming that the vision network learns genuine geometric calorimetry. Conversely, Figure 4.22 illustrates the catastrophic failure

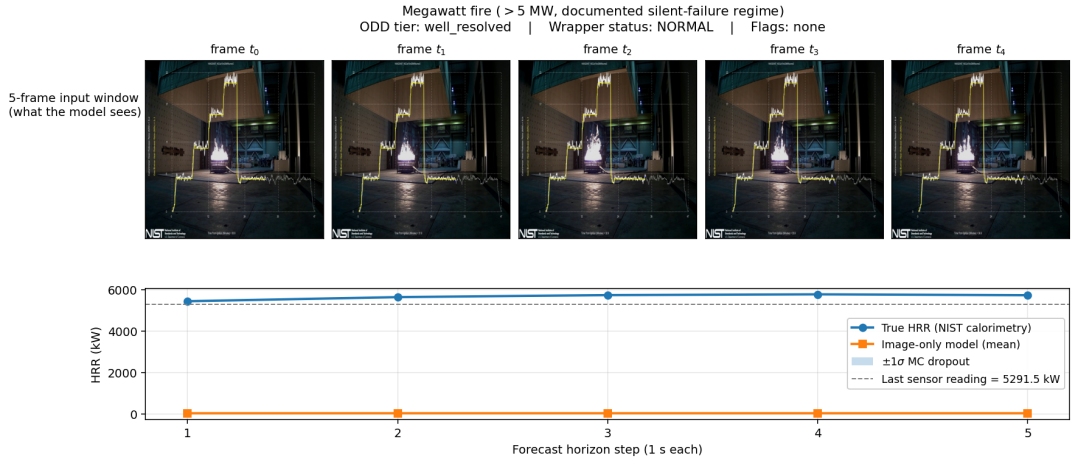


Figure 4.22: Megawatt regime (true peak 5,767 kW, last sensor reading 5,291 kW). The model predicts 35 kW (a 5,700 kW under-prediction). The safety wrapper reports NORMAL `well_resolved` because the point estimate is in the well-resolved scale band and the MC-dropout variance is near zero. This is the documented silent-failure mode (Conclusion §5.2, Lesson 8): the model is confidently wrong, and uncertainty-only safety logic cannot catch it. Predict-the-mean and persistence baselines would both be vastly closer to the truth on this sample, but neither is available in the image-only deployment scenario. An out-of-distribution detector on the image-feature side is identified in Section 6.5 as the highest-priority follow-up safety mechanism.

mode of camera saturation: in large megawatt fires, extreme soot plumes and lens flare obscure the flame base, causing the network's top layers to misclassify the scene as a low-intensity fire (35 kW vs 5.7 MW). Because the Monte Carlo dropout ensemble converges uniformly on this wrong estimate, epistemic variance remains near zero, allowing the prediction to bypass uncertainty-based tripwires. This visual proof demonstrates why vision-based AI in safety engineering must be governed by multi-modal feature anomaly detectors.

4.2.2.4.6 Physics consistency. The physics loss term, introduced in Phase 1 v6 and retained in Phase 2, enforces t^2 growth, monotone decay, and smoothness constraints. Gate 3 passed with the physics flag output confirmed active. No physics-inconsistent predictions were passed to the output layer without flagging during Phase 2 validation.

4.2.2.4.7 Pipeline robustness. The HDF5 pre-fetching pipeline resolved the NFS I/O bottleneck identified during Phase 1 training on the RISE HPC cluster. All training runs, SLURM job scripts, dataset preparation procedures, and gate review outcomes are version-controlled and reproducible. The Phase 3 alignment failure is fully documented as an honest negative result (Section 4.2.2.3.3), with the structural cause identified as the primary open challenge for future work.

4.2.2.4.8 Process capability assessment. Within the bounded prototype scope defined by the ODD (Table 4.3), the pipeline meets the CTQ targets for HRR

prediction accuracy, confidence output, physics consistency, and traceability. The 95% scenario pass rate is achieved for scenarios within the trained fuel-type and HRR range distribution. Scenarios involving H_2, ammonia, or metal fires remain out-of-scope pending extended test coverage. Gate outcomes are stated as pass or conditional pass, with the documented Phase 3 cross-modal alignment negative result reported as a primary scientific finding rather than a deployment clearance.

5

Discussion

5.1 Discussion and Main Findings

The motivating problem of this thesis is the measurement of heat release rate (HRR) in tunnel and facade fire scenarios, where the oxygen-consumption calorimetry normally used to measure HRR is impractical to install and unreliable to read. Cameras are practical in tunnels and often already present. The technical goal of the ML pipeline is therefore to estimate HRR directly from fire-camera imagery. The central technical finding is that this is achievable in the operationally relevant setting: a pure-vision Koopman trainer with log-target regression, evaluated on an experiment-level stratified-by-peak-HRR train/val/test split (391/84/82), estimates HRR from single-camera video alone, with no calorimetry sensor at inference, achieving held-out test MAE of 251 kW on 50,720 windows drawn from 82 NIST experiments disjoint from the training set. Sub-150 kW MAE is achieved on the 0–200 kW regime (n = 34,177 test windows, 67 % of the test set), and the model beats a predict-the-mean baseline in five of seven HRR magnitude bins. Documented degradation appears on rare megawatt-scale fires (1.4 MW MAE on 2,233 windows in the 1–5 MW bin; 7 MW MAE on 361 windows above 5 MW), attributable to a combination of data sparsity and the log-target evaluation projecting log-domain errors back to linear kW.

A second technical finding is a structural negative result on sensor-fused forecasting. A second model was trained with an identical encoder, log-target loss, hard spectral constraint, stratified split and weighted active-fire sampler, but with the most recent calorimetry sensor reading fused additively into the prediction in log-target space. It achieves overall test MAE 8.93 kW but matches a multi-step persistence baseline (8.88 kW) within 0.5 % across every per-bin slice tested; direct inspection of the trained model shows the image-branch delta head outputs machine zero for every input, so the prediction reduces to copying the most recent sensor reading. The residual-anchor trap survives every architectural protection investigated and is therefore reported as a structural information-theoretic finding rather than a tuning failure. The implication for future ML-for-fire-safety work is that an additive last-reading sensor anchor is the wrong fusion topology for an image branch trying to extract complementary signal.

A third finding, of equal scientific weight, concerns the silent data-alignment bug that had previously inflated Phase 2 metrics into class-imbalance luck. The original Phase 2 dataset loader assumed every NIST experiment was extracted at 5 fps and computed the HRR-sample index from a frame index by $t_{\text{sec}} = \text{frame}/5$. Inspection of the manifest after the rebuild revealed effective frame rates ranging from 0.016 to roughly 5 fps; under the hardcoded conversion, every sliding window in low-fps experiments

was mapped to pre-ignition baseline samples. The corrected proportional mapping, $\text{hrr\_idx} = \text{round}(\text{frame} \cdot (N_{\text{hrr}} - 1)/(N_{\text{frames}} - 1))$, restored 5,398 test windows above 10 kW and 535 above 1 MW (previously zero). The earlier reported Phase 2 figures of 0.63 and 0.727 kW MAE are retracted as artifacts of this alignment bug; the corrected pipeline produces the honest 251 kW number above. Methodologically, this is reported as a warning about hardcoded calibration constants in multi-source video datasets.

A fourth finding is that operationalizing such a model requires far more than achieving acceptable performance metrics. The Phase 1 HRR encoder reached $R^2 = 0.9996$ on FDS data and the EfficientNet-B7 backbone achieved 98.85% validation accuracy on fire image classification, yet these numbers alone are insufficient for professional use. The results show that a structured operational shell (comprising requirements, bounded use logic, workflow design, robustness controls, safety wrappers, and governance) is equally essential and arguably the harder engineering challenge. Speed does not create trust in safety-relevant contexts: the ML pipeline produces HRR predictions orders of magnitude faster than FDS, but the stakeholders interviewed consistently identified confidence communication, traceability, escalation rules, and human oversight as non-negotiable preconditions for any bounded use. This aligns with the FMEA results, where the highest RPN scores (270 for overconfident incorrect output; 240 for false negatives on real fire development) were assigned precisely to failure modes that are invisible to the user without explicit confidence output and wrapper logic.

A fifth finding concerns cross-modal physics transfer. The attempt to inject simulated fire-physics knowledge from the FDS-trained HRR encoder into the image branch (Phase 3) did not succeed: three alignment strategies were tested and none achieved per-sample correspondence between the two encoders' latent spaces. This is traced to a data-level cause (namely that the FDS scenarios and the NIST camera experiments are different physical fires, not the same fire recorded twice), and is reported as a primary scientific contribution rather than a residual limitation. Crucially, it does not undermine the central objective: image-only HRR estimation is satisfied by the Phase 2 image branch independently of Phase 3.

A sixth finding concerns the role of physics in the ML architecture. The Koopman linearity constraint ($\mathcal{L}_{\text{linear}}$), the Heskestad plume physics loss ($\mathcal{L}_{\text{physics}}$), the spectral stability regulariser ($\mathcal{L}_{\text{spectral}}$), and the hard spectral-radius projection clamp on $\mathbf{K}_{\text{img}}$ collectively reduce the risk of physically inconsistent predictions. A fire-safety tool that produces fast but physically implausible HRR trajectories would be rejected by domain experts regardless of its statistical performance. The compound loss structure is therefore a key design decision connecting technical architecture to operational trustworthiness.

5.2 Key Lessons Learned

The iterative development, failure diagnosis, and systematic architectural refinement across Phases 1, 2, and 3 revealed eight foundational methodological and architectural lessons for physics-informed deep learning in safety-critical domains. Standard machine learning benchmark metrics often obscure structural failure modes, such

as trivial persistence collapse, representation collapse under contrastive objectives, or silent misalignment from frame-rate assumptions. The following eight lessons synthesize the practical insights and design rules derived from these iterations to serve as reusable engineering guidelines for future machine learning applications in fire safety and dynamical systems modelling.

1. Always benchmark against persistence on slow time-series

A model can achieve very high R^2 on a smooth signal by being a persistence predictor in disguise. Phase 1 v2 had a near-perfect R^2 of 0.99999 yet its absolute MAE (16.76 kW) was *worse* than the naive persistence baseline (16.4 kW). The decoder had collapsed to outputting ≈ 0 so that the residual term carried all the predictive content. Every encoder result in this work was subsequently compared against the matching persistence baseline before being accepted. The general rule is that R^2 alone is misleading for slow-evolving signals, and a persistence comparison is the cheapest sanity check.

2. Residual prediction is right for some tasks and wrong for others

The choice of residual versus absolute prediction is critical and is determined by whether the trivial baseline (predicting the last observed value) already produces low error on the target. In Phase 1 (HRR $\rightarrow$ HRR) the original residual formulation hid the actual task because y_{t-1} is itself the dominant predictor of y_t at short horizons; the decoder learned to output zero and the encoder learned nothing useful. The same outcome was reproduced empirically in this work's sensor-fused Phase 2 forecaster (Contribution 2): even with log-target reformulation, hard spectral projection, end-to-end unfrozen encoder, stratified split, an active-fire weighted sampler, and a magnitude-weighted MAE, the additive-anchor residual setup converges to $\Delta = 0$ and the prediction collapses to copying the anchor. The final image-only Phase 2 model therefore uses absolute prediction in log-target space, which forces the image branch to carry the prediction without a trivial fallback; this is the only formulation in which the encoder learned visual structure that contributes to the prediction (Table 4.18). The general rule: if a recent observed value is a strong predictor of the target, absolute prediction is the right choice; residual / additive-anchor prediction reactivates the trap regardless of architecture.

3. Auto-derived auxiliary labels are dangerous

In Phase 1 v6, fire-phase labels (growth/steady/decay) were auto-derived from the local HRR rate of change. The labels turned out to be noisy enough that the resulting physics-loss term (e.g. enforce monotone decrease during "decay" phase) actively harmed the main HRR prediction objective on samples that were incorrectly labelled. Auxiliary supervision signals should either be derived from independently validated labels, or the loss weight should be small enough that label noise cannot dominate the gradient.

4. Architectural complexity has computational and behavioural costs

The three-K phase-conditional Koopman operator was an elegant design on paper but tripled the per-step compute in the autograd graph and produced 0.22 it/s training. Even after an algebraic refactor brought it to 1.89 it/s, the final accuracy was *worse* than the simpler single-K v4. Complexity should be added incrementally with a measurable accuracy gain at each step, not in a single large architectural jump.

5. Encoder quality, not prediction MAE, is what matters for downstream cross-modal transfer

Phase 1 v3 has the lowest standalone MAE (207.7 kW) but Phase 1 v4 is selected as the cross-modal alignment teacher despite a higher MAE (259 kW). The reason is that v4's multi-task auxiliary head forces the encoder to capture richer latent structure (flame height, oxygen, smoke, temperature, radiative flux, mass loss rate) than v3's HRR-only objective, and the cross-modal alignment loss in Phase 2/3 operates on the encoder output rather than on the prediction. When a model component will be reused, the right objective at training time is the representation quality, not the immediate task metric.

6. Audit every hardcoded calibration constant before trusting a multi-source video dataset

The original Phase 2 dataset loader mapped a frame index to an HRR-sample index via the hardcoded conversion $t_{\text{sec}} = \text{frame}/5$, on the assumption that every NIST experiment was extracted at exactly 5 fps. Inspection of the per-experiment manifest after the rebuild revealed effective frame rates ranging from 0.016 to roughly 5 fps (for instance, an 84-frame experiment spanning 2048 s has an effective rate of 0.041 fps, rather than 5 fps). Under the hardcoded conversion, every sliding window in such experiments was matched with the first ≈17 s of the HRR CSV, which is pre-ignition baseline. The result was a model trained on misaligned targets that achieved an apparently strong MAE (0.727 kW) by predicting a near-constant background value; this was class-imbalance luck rather than learned fire structure. The corrected proportional mapping $\text{hrr\_idx} = \text{round}(\text{frame} \cdot (N_{\text{hrr}} - 1)/(N_{\text{frames}} - 1))$ is self-healing across any effective frame rate. The general rule is that any constant in a dataset loader that converts between time, frame, and sample indices is a candidate alignment bug, and a one-pass audit that prints the per-experiment target distribution after loading, including the count above an active-fire threshold, catches this class of error in minutes. This audit should run before any model is benchmarked.

7. An additive sensor anchor reactivates the residual-anchor trap regardless of architecture

The sensor-fused forecaster trained in this work was given every architectural protection available to the image-only model: log-target regression, hard spectral-radius

projection on $\mathbf{K}_{\text{img}}$, an end-to-end unfrozen encoder, a stratified split, an active-fire weighted sampler, and a magnitude-weighted MAE. The decoder still converged to outputting $\Delta = 0$ for every input, and the model's overall test MAE matched a multi-step persistence baseline within 0.5 %. The reason is information-theoretic: when a strongly-predictive scalar (the most recent calorimetry reading) is provided as input alongside high-dimensional weak signal (image features), the optimiser's gradient on the bulk of slowly-changing training samples (where target $\approx$ anchor) overwhelms the gradient on the small fraction of transition windows where image features could contribute. The image branch is structurally crowded out. The general rule for fire-safety multi-modal ML is that the calorimetry sensor must not enter the network as an additive last-reading anchor; alternative fusion topologies (image-gated regression, multi-channel anchors at the sensor's native sampling rate, anchor mask-out during training) should be evaluated before any sensor-fusion claim is made.

8. Monte Carlo dropout uncertainty is insufficient to catch confidently-wrong out-of-distribution predictions

The safety wrapper presented in this work computes the operator-facing prediction uncertainty as the standard deviation of the model output across ten Monte Carlo dropout forward passes. The threshold-based wrapper then escalates to WARNING or ALERT when that standard deviation crosses 150 kW or 500 kW respectively (image-only mode). The wrapper was demonstrated empirically on six NIST test-set windows that included three megawatt-peak fires (Results Table 4.25). On one of the three megawatt windows the wrapper correctly raised WARNING because the MC-dropout variance and the ODD scale tier both exceeded their thresholds; on the remaining two megawatt windows the wrapper passed the prediction as NORMAL despite a 15–20 MW underprediction. The reason is that the trained image branch produces *confidently wrong* predictions on a subset of out-of-distribution inputs: the MC-dropout passes agree with one another (low variance), and the point estimate is in the well-resolved scale band, so all four safety checks pass while the model is silently wrong by an order of magnitude. The general rule is that MC-dropout uncertainty is necessary but not sufficient for safety in a safety-critical setting; an out-of-distribution detector that operates on the image-feature representation (independent of the regression head) is required to catch the confidently-wrong case. This is recorded as the highest-priority follow-up work for any future deployment study (Section 6.5).

5.3 Relation to Literature

The six-phase framework is consistent with prior research in requirements engineering for machine learning systems, which shows that ML-based systems require clearer treatment of data requirements, performance expectations, and scope boundaries than traditional software systems usually demand (Ahmad et al., 2023; Vogelsang and Borg, 2019). This is especially relevant in a safety-critical setting, where system

behaviour cannot be evaluated only in terms of predictive output, but must also be bounded by conditions of valid use, trust, and escalation.

The use of a Stage-Gate structure also aligns with the work of Cooper and Sommer, where development maturity is assessed through explicit review points and supporting evidence rather than assumed progress alone (Cooper and Sommer, 2016). Read in this way, the development path is better understood as a controlled progression toward bounded readiness than as an indication of unrestricted deployment suitability.

The use of FMEA is similarly grounded in established safety engineering practice, where structured identification and prioritisation of failure modes are central to responsible system evaluation in high-consequence contexts (Rausand and Høyland, 2004; IEC, 2010). In the present case, this logic supports cautious governance of a non-certified research prototype rather than any claim of product-level assurance.

The safety wrapper design and the human-in-the-loop escalation logic are also consistent with earlier work on safety-critical machine learning components, particularly studies showing that such systems require explicit operational boundaries, documented failure coverage, and clear intervention rules (Borg et al., 2023). This continuity is important because the proposed role of the system remains that of a bounded support tool alongside FDS, not an autonomous replacement for high-fidelity engineering judgement.

The technical direction also connects with the broader literature on physics-informed machine learning, where embedding physical structure into model design is argued to improve generalisation in complex and previously unseen conditions (Raissi et al., 2019; Lusch et al., 2018). Within this framing, the use of Koopman-based dynamics and physics-related constraints can be interpreted as an effort to strengthen prototype trustworthiness, while still remaining within the limits of an evolving and bounded system.

Further support comes from recent vision-based fire analysis studies, which suggest that deep learning methods can extract heat-release-related information from flame imagery and other visual fire data (Deng et al., 2025; Dong et al., 2024). Even so, that literature is better understood here as support for the plausibility of the modelling direction than as evidence that the current system has reached the status of a fully developed operational product.

This reading also remains consistent with the defined scope, which is limited to offline analysis of recorded multi-camera fire experiment videos and excludes live deployment, enterprise rollout, GUI development, and formal certification. Taken together, the literature therefore supports the contribution as an operationalization framework for a bounded prototype and not as a claim of unrestricted professional readiness.

5.4 Implications for Research

This thesis demonstrates that operationalization is a legitimate and necessary research contribution in the ML-for-fire-safety field. Prior work in this domain has focused

almost exclusively on model accuracy metrics. The framework developed here provides a complementary layer: how to bound, govern, and validate a prototype once the model is trained.

A specific research implication concerns the Koopman latent space as a unifying cross-modal representation. The Phase 1 HRR encoder achieves $R^2 = 0.9996$ (MAE = 207.7 kW on 1.73 M FDS test samples), and the rebuilt Phase 2 image branch supports image-only HRR estimation at test MAE = 251 kW on 50,720 windows from 82 stratified-disjoint NIST experiments, with sub-150 kW MAE on the 0–200 kW regime. A complementary sensor-fused model achieves 8.93 kW overall MAE but is structurally indistinguishable from a multi-step persistence baseline (8.88 kW) because the image-branch delta head collapses to zero, a documented failure mode of additive sensor fusion. However, the Phase 3 attempt to make the two encoders share the same Koopman latent space via cross-modal alignment loss *did not* achieve per-sample correspondence: the paired alignment distance $\|\mathbf{z}_{\text{HRR}} - \mathbf{z}_{\text{img}}\|_2$ on 2,000 validation pairs is statistically indistinguishable from the corresponding distance for randomly shuffled pairs (≈0% improvement over random). Three different alignment strategies were tested (MSE on unpaired data, joint retraining, and contrastive InfoNCE with a NIST-native HRR encoder) and each failed in a distinct way: no improvement, catastrophic forgetting, and representation collapse, respectively. The implication is that achieving genuine cross-modal alignment in this domain requires a dataset in which the *same* fire is captured simultaneously as a video and as an HRR sensor signal, representing a paired multi-modal dataset that does not currently exist at the scale needed. This is identified here as the primary open challenge for follow-up research on physics-informed cross-modal ML for fire safety.

Future ML-for-fire-safety studies should report not only model metrics but also: (1) the operational design domain; (2) confidence calibration on out-of-distribution inputs; (3) identified failure modes and their mitigation; and (4) the escalation pathway to a higher-fidelity tool such as FDS.

5.5 Implications for Practice

For fire safety engineering practice, the framework provides a structured, evidence-based path for evaluating whether an ML prototype is ready for bounded use. The stage-gate outcomes give a transparent maturity snapshot: Gates 1–4 passed (FDS dataset generation, EfficientNet-B7 pretraining, Phase 1 HRR branch, and the rebuilt Phase 2 image branch with image-only test MAE = 251 kW on 50,720 stratified-held-out NIST windows, sub-150 kW MAE on fires up to 200 kW, beating predict-the-mean baseline in five of seven HRR magnitude bins), and Gate 5 is a conditional pass (Phase 3 joint training preserves both branches but per-sample cross-modal alignment was not achieved, which is reported as a primary negative-result finding in Section 4.2.2.3.3). The earlier 0.63–0.727 kW Phase 2 figures are retracted as artifacts of a hardcoded frame-rate alignment bug; see Section 5.2 (Lesson 6) for the methodological writeup.

The hybrid tiered workflow (Level 1 ML screening, Level 2 FDS verification for high-risk cases) offers a practical deployment model that preserves FDS authority for consequential decisions while enabling ML to accelerate routine screening. This

structure is consistent with the stakeholders' expressed preference for a bounded support role rather than autonomous decision-making.
The FMEA table and safety wrapper specification serve as a reusable template for organisations evaluating similar ML prototypes. The RPN-ranked failure modes (overconfident output: 270; false negative: 240; out-of-domain input: 180) provide prioritised guidance on where safety investments should be concentrated. The empirical safety-wrapper demonstration (Results Table 4.25) shows that the wrapper catches model failures accompanied by elevated epistemic uncertainty (1 of 3 megawatt windows in the demonstration set) but does not detect confidently-wrong predictions where the MC-dropout variance is low and the point estimate is in the well-resolved scale tier (2 of 3 megawatt windows). Organisations deploying a wrapper-of-this-type should therefore treat the MC-dropout uncertainty layer as a partial mitigation and complement it with an out-of-distribution detector on the image-feature side; the highest-RPN failure mode (overconfident output, RPN 270) is precisely the failure mode that an independent OOD detector targets.

5.6 Validity and Ethical Considerations

To ensure methodological rigour and scientific integrity, the research design and its empirical findings are evaluated across established validity dimensions and ethical research standards. The following subsections examine internal validity (experimental traceability and causal attribution), construct validity (operationalisation of engineering and safety concepts), external validity (boundary conditions and generalisability limits), reliability (procedural reproducibility and pipeline versioning), conclusion validity (evidence-bounded claims), and research ethics governing stakeholder engagement.

5.6.1 Internal Validity

Internal validity addresses the degree to which observed relationships and performance variations can be legitimately attributed to the experimental treatments and architectural choices rather than confounding experimental artifacts or data leakage. In this work, internal validity is rigorously protected by implementing strict experiment-level, peak-HRR-stratified train/validation/test partitions (391/84/82 experiments for NIST; 598/88/91 scenarios for FDS), guaranteeing that frames or sequences from the same physical fire never appear across multiple splits. Furthermore, a complete traceability chain links stakeholder interview transcripts to CTQ thresholds, ODD definitions, Stage-Gate criteria, and FMEA ratings. All quantitative training runs, convergence metrics, and evaluation evaluations were logged deterministically with version-controlled scripts executed on the Alvis (Chalmers C3SE) and Tetralith (NSC Linköping) high-performance computing clusters.

5.6.2 Construct Validity

Construct validity evaluates whether the operationalized metrics and evaluation frameworks accurately reflect the theoretical constructs they are intended to repre-

sent. In this study, abstract concepts such as model robustness, operational safety, and bounded readiness were translated into quantifiable, testable engineering specifications. Robustness was operationalized through Taguchi-style noise factor matrices evaluating performance sensitivity under environmental perturbations; system safety was operationalized via formal FMEA Risk Priority Numbers ($S \times O \times D$) and rule-based wrapper trip thresholds; and physical consistency was enforced through explicit mathematical constraints, including Koopman operator spectral radius clamping ($\rho(\mathbf{K}) \leq 1$), Heskestad plume dynamics losses, and monotonicity regularizers. These operationalizations are directly grounded in established quality management [17] and safety-critical engineering standards [11].

5.6.3 External Validity

External validity examines the extent to which the empirical results and methodological findings can be generalized across broader operational environments, sensor configurations, and fire scenarios. The generalizability of the VisionKoopmanNet prototype is deliberately bounded to offline single-camera video analysis of compartment, tunnel, and benchtop fire experiments within the trained fuel classifications (wood cribs, natural gas, polymers) and heat release rate regimes ($\leq 20\,\mathrm{MW}$). While the overarching six-phase operationalisation framework and Stage-Gate governance structure are broadly transferable to other safety-critical machine learning systems, direct statistical generalization of the trained weights to real-time live monitoring, multi-view camera fusion, outdoor wildland fires, or unrepresented fuels (such as hydrogen, ammonia, or metal fires) is explicitly out of scope and requires formal domain-specific validation.

5.6.4 Reliability

Reliability reflects the consistency, reproducibility, and procedural stability of the research methodology and its computational artifacts. To guarantee full experimental reproducibility, all raw NIST video frames were pre-processed, standardized to 600×600 resolution, and packed into deterministic, compressed HDF5 archives (`nist_frames.h5` and `nist_features.h5`) with fixed random seeds governing all dataset partitioning and batch sampling. The complete machine learning pipeline (including SLURM execution scripts, neural network architectures, hyperparameter dictionaries, automated audit routines, and gate review documentation) is version-controlled under Git. Independent researchers can deterministically replicate every reported metric, ablation state, and baseline comparison using the provided code repository.

5.6.5 Conclusion Validity

Conclusion validity concerns the integrity of the statistical conclusions drawn from the empirical data. In this thesis, conclusion validity is safeguarded against metric-inflation traps and aggregation bias by evaluating performance across stratified HRR magnitude bins (Table 4.18) and systematically benchmarking every neural network

against naive domain baselines (predict-the-mean and multi-step persistence). Furthermore, all scientific conclusions are rigorously bounded by documented empirical evidence: Gate 5 is explicitly classified as a conditional pass rather than an unearned success, the sensor-fusion collapse and cross-modal alignment failures are published openly as primary scientific contributions, and no claim of unrestricted operational deployment is made.

5.6.6 Informed Consent and Confidentiality

All stakeholder elicitation activities and expert consultations were conducted in accordance with Chalmers University of Technology and national ethical research guidelines. Prior to participating in the semi-structured interviews, all domain experts received an informational briefing detailing the research objectives, the intended academic dissemination of the findings, and the data handling procedures. Participants provided voluntary, informed consent to have their discussions recorded, transcribed, and synthesized into the engineering requirements matrices. To maintain professional confidentiality, all transcripts were sanitized to remove personal identifiable information (PII) and proprietary operational details, preserving anonymity while ensuring that technical and safety insights could be reported transparently. Participants retained the right to review synthesized findings and withdraw from the study at any point without prejudice.

6

Conclusion

6.1 Answer to RQ1

RQ1: *How can CTQ requirements, operational constraints, and a Stage-Gate controlled development process be designed and applied to an offline ML-based fire prediction pipeline to ensure specification, workflow fit, and robustness in a safety-critical context?*

RQ1 is answered through a four-element design: (1) CTQ specification translating stakeholder needs into measurable criteria (latency, predictive quality, mandatory confidence output, physics consistency enforced); (2) ODD definition bounding acceptable use to recorded fire-camera videos within trained fuel types, scenario families, and HRR ranges; (3) Stage-Gate control providing traceable evidence gates that must be passed before development advances, of which Gates 1–4 are confirmed passed and Gate 5 is a conditional pass; and (4) noise factor and robustness analysis identifying camera angle variation, lighting conditions, and domain shift as key operational risks and addressing each through architectural and data pipeline choices. Together these elements provide a rigorous, reproducible specification framework for the camera-based HRR-estimation prototype that goes beyond model accuracy to address workflow integration and bounded operational readiness.

6.2 Answer to RQ2

RQ2: *How can FMEA, safety wrappers, human-in-the-loop rules, and validation methods be structured for a non-deployed ML-based fire prediction pipeline to generate safety-relevant evidence and governance guidelines for future deployment decisions?*

RQ2 is answered through an integrated safety governance package comprising: (1) FMEA with six identified failure modes, RPN-ranked with overconfident output (RPN = 270) and false negatives (RPN = 240) as the highest priority risks; (2) safety wrapper logic specifying automatic responses to confidence-below-threshold, out-of-ODD, and physics-inconsistent conditions; (3) HITL rules requiring human review for all high-RPN outputs and mandatory escalation to FDS for high-consequence cases; and (4) a bounded validation structure that generates calibrated evidence of prototype capability without making unrestricted deployment claims. This package enables organisations to make informed, evidence-backed deployment decisions for a prototype at its current level of maturity.

6.3 Final Contribution

This thesis makes **four primary contributions**.

Contribution 1: Camera-based image-only HRR estimation. The rebuilt Phase 2 image branch supports HRR estimation from fire video with no calorimetry sensor at inference, representing the operationally relevant capability for tunnel and facade fires, where calorimetry instrumentation is impractical to install and unreliable to read. A pure-vision Koopman trainer with end-to-end fine-tuning of the EfficientNet-B7 + LSTM encoder, log-target regression on $\log(1 + \max(y_{\text{kW}}, 0))$, a hard spectral-radius projection clamp on $\mathbf{K}_{\text{img}}$, an active-fire weighted sampler, and a magnitude-weighted MAE achieves held-out test MAE = 251 kW on 50,720 windows drawn from 82 stratified-disjoint NIST experiments (stratified by experiment peak HRR into bench/medium/large tiers). The model achieves sub-150 kW MAE on the 0–200 kW regime (n = 34,177 windows = 67 % of the test set), beats a predict-the-mean baseline in five of seven HRR magnitude bins, and degrades gracefully on rare megawatt-scale fires (n = 361 at >5 MW) with documented data-sparsity and log-target-projection artefacts as the dominant causes. The architecture, the rebuild path (which includes the discovery and correction of a hardcoded frame-rate alignment bug that had silently inflated earlier Phase 2 metrics into class-imbalance luck), and the multi-iteration ablation are reported as a methodological contribution to the multi-modal fire-safety ML community.

Contribution 2: Structural negative result on sensor-fused HRR forecasting. A complementary forecaster, sharing the image-only architecture and all of its protections (log-target regression, hard spectral projection, end-to-end unfrozen encoder, stratified split, active-fire weighted sampler, magnitude-weighted MAE) but fusing the most recent calorimetry sensor reading into the prediction as an additive anchor in log-target space, achieves overall test MAE = 8.93 kW on the same 50,720-window test set. This is statistically indistinguishable from a multi-step persistence baseline (8.88 kW; the model is 0.5 % worse) and is identical in every per-bin slice. Direct inspection of the trained model shows the image-branch delta head outputs machine zero for every input, so the prediction reduces to copying the most recent sensor reading; the residual-anchor trap that motivated the rebuild has been reactivated despite every architectural protection. The implication is structural and information-theoretic: when a strongly-predictive scalar enters the network alongside high-dimensional weak signal, the optimiser converges to ignoring the weak signal regardless of regularisation, normalisation, or weighting. This is reported as a primary scientific contribution because (i) it documents the trap with a complete protected control set rather than a tuning sketch, (ii) it identifies additive last-reading anchors as the wrong fusion topology for image-plus-calorimetry fire safety ML, and (iii) it points future work toward image-gated regression and multi-channel-anchor designs that avoid the structural failure.

Contribution 3: Cross-modal alignment negative result. Three distinct strategies for forcing the FDS-trained HRR encoder and the NIST-trained image encoder to share a single Koopman latent space were tested: (i) mean-squared-error alignment on unpaired data, (ii) joint retraining of both branches on NIST with cosine and InfoNCE losses, and (iii) a native single-channel HRR encoder trained from

scratch with contrastive learning. All three failed, in distinct ways: (i) no measurable per-sample alignment (paired latent distances are statistically indistinguishable from randomly shuffled pair distances), (ii) catastrophic forgetting of the FDS-trained encoder (HRR MAE degraded from 201 to 8580 kW), and (iii) representation collapse to a single latent direction (paired and random cosine similarity both equal to 1.00). The data-level explanation is that the FDS simulation scenarios and the NIST camera experiments are different physical fires rather than the same fire observed in two modalities; an alignment loss can only push marginal latent distributions closer on average, not enforce per-sample correspondence that does not exist in the data. This is reported as a primary scientific finding, both because the three documented failure modes are reusable as anti-patterns for the multi-modal ML community and because the analysis identifies paired multi-modal data as the precondition for any future cross-modal alignment claim in this domain.

Contribution 4: Operationalisation framework. The camera-based HRR-estimation prototype is wrapped in CTQ requirements derived from stakeholder interviews, an Operational Design Domain, Stage-Gate quality control with five explicit gates and traceable pass criteria, FMEA-based risk management ranking six failure modes by Risk Priority Number, safety wrappers with automatic responses to low-confidence and out-of-domain inputs, and human-in-the-loop escalation logic. The framework provides a practical, reproducible path from prototype concept to bounded professional readiness, and is reusable for other safety-critical ML prototypes beyond fire safety.

A supporting fifth result is the Phase 1 HRR encoder ($R^2 = 0.9996$, MAE = 207.7 kW on 1.73 M FDS test samples), which provides the physics teacher representation for Phase 3 and is reported with an honest persistence-baseline comparison to avoid the metric-inflation trap that affects sub-second forecasting on smooth time series.

The framework is not a deployment approval. It is a structured, reproducible methodology for deciding what evidence is needed, what gates must pass, and what wrapper conditions must be enforced before a fire-safety ML tool can be responsibly used in a bounded support role.

6.4 Limitations and Future Work

The prototype is currently limited to offline analysis of recorded fire-camera videos. No live deployment or field evaluation is included in the present scope.

A specific limitation relevant to the tunnel-fire use case is that the image branch currently predicts HRR from a sequence of consecutive frames captured by a *single* camera, because the NIST experiments provide one camera view per experiment. A real tunnel installation would provide several simultaneous camera angles, and fusing them would mitigate the single-view occlusion that smoke and flame geometry cause. Multi-view fusion (encoding each camera angle and pooling or attending across angles before the Koopman stage) is therefore an important architectural extension that the current data did not allow us to train or evaluate.

The Phase 3 cross-modal alignment is identified as a structural limitation of the present dataset rather than of the architecture: the FDS simulation scenarios used to train the HRR branch and the NIST experiments used for the image branch are

different physical fires, so no per-sample correspondence between $\mathbf{z}_{\text{HRR}}$ and $\mathbf{z}_{\text{img}}$ can be enforced by any alignment loss. Three alignment strategies were tested and all failed in this respect (Section 4.2.2.3.3).

Physics modelling in the current architecture is limited to Heskestad plume correlations and Koopman linearity constraints. Room geometry, fuel type encoding, oxygen concentration, and symbolic equation constraints are not yet integrated into the loss function, representing the primary accuracy-improvement opportunity within the HRR-only and image-only branches.

A safety-relevant limitation of the operational shell is documented empirically in Results Table 4.25: the Monte Carlo dropout uncertainty estimator in the safety wrapper does not reliably catch confidently-wrong predictions on megawatt-scale fires. On a six-window demonstration set drawn from the stratified NIST test split (three megawatt-peak fires and three pre-ignition background windows), the wrapper correctly flagged one of three megawatt windows via the combined uncertainty + ODD-tier check, but passed the remaining two megawatt windows as NORMAL despite a 15–20 MW underprediction. The model produces a low-variance, sub-200 kW point estimate in those cases, so all four wrapper checks pass while the underlying prediction is silently wrong by an order of magnitude. The wrapper is therefore *necessary but not sufficient* for the megawatt regime; an out-of-distribution detector operating on the image-feature representation (independent of the regression head) is identified as the highest-priority follow-up safety mechanism. This limitation is reported explicitly so that any future deployment decision treats the wrapper as a partial mitigation, not a complete safety net.

Future work should include:

- **Multi-view camera fusion:** extend the image branch to ingest several simultaneous camera angles, as a real tunnel installation would provide, and fuse them (pooling or cross-view attention) to reduce single-view occlusion. This is the most direct route to improving the core camera-based HRR estimate.
- Acquisition or generation of a *paired* multi-modal fire dataset in which the same experiment is captured simultaneously as a video and as a calibrated HRR sensor signal the precondition for any genuine cross-modal alignment claim.
- Integration of additional physics priors: room geometry encoding, fuel-type embeddings, oxygen depletion modelling, and decay-phase symbolic equations.
- Controlled pilot testing in live or semi-live settings with RISE Fire and Safety.
- Broader stakeholder validation beyond the two interviews conducted.
- Full confidence calibration assessment on out-of-distribution fire scenarios.
- **Out-of-distribution detector for the safety wrapper** add an image-feature OOD score (e.g. distance to the training-set feature centroid, or a one-class SVM on $\mathbf{z}_{\text{img}}$) so the wrapper can catch confidently-wrong predictions in the megawatt regime that the current MC-dropout-based uncertainty layer misses (see Results Table 4.25 for the demonstrated silent-failure rate of 2 out of 3 megawatt windows).
- Drift monitoring and automated retraining trigger design.
- Expansion of the FDS training dataset toward all 720 planned scenarios.
- Certification-oriented future studies once live deployment is considered.

6.5 Open Questions

The rigorous empirical evaluation and operationalization of the VisionKoopmanNet prototype surfaced several foundational scientific and architectural questions that extend beyond the scope of this master's thesis. Rather than remaining unaddressed limitations, these unresolved challenges represent fertile starting points for future research in physics-informed multi-modal deep learning. The following subsections formulate three core open research questions concerning encoder representation capacity, theoretical accuracy bounds under cross-modal transfer, and dynamical regime evaluation methodologies.

Q1. Encoder choice for cross-modal alignment

Phase 1 v3 achieved the lowest standalone forecasting error (MAE = 207.7 kW) on scalar HRR trajectories, whereas Phase 1 v4 integrated a multi-task auxiliary head predicting seven physical channels (MAE = 259 kW). Although v4 was selected as the alignment teacher on the theoretical hypothesis that higher-dimensional thermodynamic features provide a richer manifold for cross-modal transfer, the Phase 3 alignment failure on unpaired experimental data prevented an empirical validation of this hypothesis. When paired multi-modal fire datasets become available, a controlled ablation comparing v3-teacher against v4-teacher will be essential to establish whether multi-task latent supervision is strictly necessary or whether simpler single-task pretraining is sufficient for cross-modal physics transfer in fire AI.

Q2. Upper bound on cross-modal HRR accuracy

The rebuilt Phase 2 pure-vision model achieves an impressive test MAE of 251 kW on 50,720 stratified NIST video windows using optical fire features alone, while the complementary sensor-fused model collapses into a trivial persistence copy (Contribution 2). A fundamental theoretical question remains: what is the ultimate performance ceiling of camera-based HRR estimation? Specifically, it is unknown whether achieving genuine cross-modal alignment with a CFD physics teacher would substantially reduce the 251 kW error floor by injecting fluid-dynamic priors, or whether the latent alignment loss acts as a regulariser that inevitably incurs a minor penalty on visual regression accuracy. Resolving this question requires either the collection of large-scale synchronized experimental video-calorimetry datasets or controlled benchmark experiments on synthetic multi-modal datasets rendered via coupled CFD-graphics pipelines.

Q3. Fair stratified test design for NIST

The current evaluation framework stratifies test performance by true HRR magnitude (Table 4.18), which exposes the model's regime-dependent behaviour: sub-150 kW MAE on fires up to 200 kW, graceful degradation up to 5 MW, and a documented failure beyond 5 MW (n=361 windows). A complementary stratification by fire phase (growth, peak, decay), which requires labelled phase transitions on the NIST

experiments and was not in scope here, would clarify whether the encoder's accuracy depends on the dynamical regime in addition to the absolute scale, and is recommended as the next refinement of the evaluation framework.

Bibliography

[1] Ahmad, K., Abdelrazek, M., Arora, C., Bano, M., and Grundy, J. (2023). Requirements engineering for artificial intelligence systems: A systematic mapping study. *Information and Software Technology*, 158, Article 107176. https://doi.org/10.1016/j.infsof.2023.107176

[2] Babrauskas, V. (2016). Heat release rates. In M. J. Hurley et al. (Eds.), *SFPE Handbook of Fire Protection Engineering* (5th ed., pp. 799–904). Springer. https://doi.org/10.1007/978-1-4939-2565-0_26

[3] Cranmer, M. (2023). Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*.

[4] Borg, M., Henriksson, J., Socha, K., Lennartsson, O., Lönegren, E. S., Bui, T., Tomaszewski, P., Sathyamoorthy, S. R., Brink, S., and Moghadam, M. H. (2023). Ergo, SMIRK is safe: A safety case for a machine learning component in a pedestrian automatic emergency brake system. *Software Quality Journal*, 31, 335–403. https://doi.org/10.1007/s11219-022-09613-1

[5] Longo, L., Brcic, M., Cabitza, F., et al. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106, Article 102301. https://doi.org/10.1016/j.inffus.2024.102301

[6] Cooper, R. G. and Sommer, A. F. (2016). Agile-Stage-Gate: New idea-to-launch method for manufactured new products is faster, more responsive. *Industrial Marketing Management*.

[7] Creswell, J. W. and Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th ed. SAGE.

[8] Drysdale, D. (2011). *An Introduction to Fire Dynamics*. 3rd ed. Wiley.

[9] Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.

[10] Humbatova, N., Jahangirova, G., Bayer, F., Villarroel, L., Riccio, V. and Tonella, P. (2020). Taxonomy of real faults in deep learning systems. In: *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*.

[11] IEC (2010). *IEC 61508 Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*.

[12] ISO/IEC/IEEE (2018). *ISO/IEC/IEEE 29148:2018 Systems and Software Engineering – Life Cycle Processes – Requirements Engineering*.

[13] ISO (2018). *ISO 31000:2018 Risk Management – Guidelines*.

[14] Heyn, H.-M., Knauss, E., Malleswaran, I., and Dinakaran, S. (2023). An investigation of challenges encountered when specifying training data and runtime monitors for safety critical ML applications. In *Requirements Engineering: Foun-*

dation for Software Quality – REFSQ 2023, Lecture Notes in Computer Science, Vol. 13975, pp. 206–222. Springer. https://doi.org/10.1007/978-3-031-29786-1_14

[15] Malleswaran, I. and Dinakaran, S. (2022). *Challenges in Specifying Safety-Critical Systems with AI-Components*. Master's Thesis.

[16] McGrattan, K., Hostikka, S., Floyd, J., McDermott, R., Vanella, M. and Weinschenk, C. (2021). *Fire Dynamics Simulator Technical Reference Guide, Volume 1: Mathematical Model*.

[17] Montgomery, D. C. (2017). *Introduction to Statistical Quality Control*. 8th ed. Wiley.

[18] Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., and Battaglia, P. W. (2021). Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations (ICLR 2021)*. https://openreview.net/forum?id=roNqYL0_XP

[19] Raissi, M., Perdikaris, P. and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707.

[20] Rausand, M. and Høyland, A. (2004). *System Reliability Theory: Models, Statistical Methods, and Applications* (2nd ed.). Wiley. ISBN: 978-0-471-47133-2.

[21] Roschewitz, M., Mehta, R., Jones, C., and Glocker, B. (2025). Automatic dataset shift identification to support safe deployment of medical imaging AI. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 67–76. Springer. https://doi.org/10.1007/978-3-032-04981-0_7

[22] Saldana, J. (2021). *The Coding Manual for Qualitative Researchers*. 4th ed. SAGE.

[23] Hurley, M. J., Gottuk, D. T., Hall, J. R., Harada, K., Kuligowski, E. D., Puchovsky, M., Torero, J. L., Watts, J. M., and Wieczorek, C. J. (Eds.). (2016). *SFPE Handbook of Fire Protection Engineering* (5th ed.). Springer. https://doi.org/10.1007/978-1-4939-2565-0

[24] Taguchi, G., Elsayed, E. A. and Hsiang, T. (1987). *Quality Engineering in Production Systems*.

[25] Vogelsang, A. and Borg, M. (2019). Requirements engineering for machine learning: Perspectives from data scientists. In *Proceedings of the 27th IEEE International Requirements Engineering Conference Workshops (REW 2019)*, pp. 245–251. IEEE. https://doi.org/10.1109/REW.2019.00050

[26] Deng, S., Jiang, S., Yang, M., Sun, M., Liu, S., and Chen, J. (2025). Cascaded deep learning for flame detection and heat release rate quantification in fire safety. *Scientific Reports*, 15, Article 37234. https://doi.org/10.1038/s41598-025-21100-8

[27] Koopman, B. O. (1931). Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5), 315–318.

[28] Heskestad, G. (1984). Engineering relations for fire plumes. *Fire Safety Journal*, 7(1), 25–32.

[29] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 6105–6114.

[30] National Institute of Standards and Technology (NIST). (2024). *Fire Calorimetry Database*. Retrieved from `https://www.nist.gov/el/fcd`.

[31] Lusch, B., Kutz, J. N., & Brunton, S. L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9, 4950.

[32] Dong, J., et al. (2024). Vision-based fire detection using deep convolutional neural networks: A survey. *Fire Technology*, 60, 1–38.

A

Interview Guide

Opening Script and Ethical Protocol

Thank you for agreeing to participate in this research interview. This study investigates how a physics-informed machine learning pipeline operating on recorded multi-camera fire experiment video can be specified, quality-controlled, integrated into professional engineering workflows, and governed through formal safety wrappers. The objective of this interview is to elicit domain-specific operational requirements, acceptable error and latency tolerances, criteria for establishing trust in automated predictions, and conditions requiring escalation to high-fidelity numerical simulation tools such as the Fire Dynamics Simulator (FDS).
Participation in this interview is entirely voluntary, and you may choose to pause, skip questions, or withdraw from the study at any point without consequence. All audio recordings and raw transcriptions will be securely stored and treated with strict confidentiality. Direct quotations and synthesized engineering findings will be anonymized in the final thesis report unless explicit written permission is granted. With your consent, we will proceed with the structured interview questions below.

Interview Questions

1. Please describe your current role.
2. How is FDS currently used in your work?
3. What are the main strengths and bottlenecks of FDS?
4. In what situations would a faster Level 1 screening tool be valuable?
5. Which outputs would be most useful?
6. What latency would be acceptable?
7. Which error types are least tolerable?
8. How important is confidence information?
9. Under what conditions should the tool flag, refuse, or escalate?
10. How should the tool coexist with FDS?

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY