http://dx.doi.org/10.62762/TETAI.2025.259226

RESEARCH ARTICLE

Multi-Modal Fusion for Yield Optimization: Integrating Wafer Maps, Metrology, and Process Logs with Graph Models

Min Yin[1,*] **and Ledee-FI Frank**[1]

[1] University of California-Berkeley, Berkeley, CA 94720, United States

Abstract

Yield optimization in advanced manufacturing rarely proceeds as a tidy pipeline; it arises from the gradual convergence of evidence across spatial wafer patterns, multivariate metrology, and asynchronous process and equipment events that interact in ways that are only partially observable. Prior studies often separate these modalities, assigning convolutional encoders to wafer maps, sequence models to metrology, and template based encoders to logs, an arrangement that can perform well locally yet struggles to sustain cross-modal alignment or to reason over the hierarchy that links defects to steps and equipment. Building on these observations, we introduce a manufacturing semantics oriented framework that embeds lots, wafers, dies, steps, equipment, and recipes in a heterogeneous graph, and uses cross modal attention gating to reconcile image, time series, and event representations while performing relation aware message passing. The research was not frictionless; time synchronization required iterative windowing, spatial normalization exposed orientation drift, and naive imputation inflated variance in rare steps, which motivated temperature controlled gating and a lightweight contrastive warm-up. On two production lines the approach improves, to some extent, standard classification metrics and stabilizes top k attribution under feasible latency. Alternative explanations remain possible, including benefits from stricter leakage control or product specific distributions. The work makes explicit the structural link among defects, process, and equipment, and points toward auditable, engineer actionable analytics; further research is needed on long term stability, cross site generalization, and the joint optimization of accuracy, cost, and energy.

Keywords: heterogeneous graph learning, multi-modal data fusion, wafer maps, metrology time-series, process log mining, yield optimization, cross-Modal attention, bottleneck identification, explainable AI, semiconductor manufacturing.

1 Introduction

Yield optimization in semiconductor manufacturing unfolds less as a neatly staged pipeline and more as a delicate synthesis of signals that surface at different

Submitted: 10 November 2025
Accepted: 25 November 2025
Published: 09 January 2026

Vol. 3, **No.** 1, 2026.
10.62762/TETAI.2025.259226

***Corresponding author:**
✉ Min Yin
gmiayinc@gmail.com

Citation
Yin, M., & Frank, L. F. (2026). Multi-Modal Fusion for Yield Optimization: Integrating Wafer Maps, Metrology, and Process Logs with Graph Models. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 3(1), 45–60.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

temporal scales and spatial granularities: wafer defect distributions that encode process fingerprints multivariate metrology traces that drift and recover in subtle cycles, and logs that record equipment states and operator actions with irregular cadence [1]. Treating these streams in isolation is analytically convenient. However, such separation weakens our ability to recover the structural pathways that link local anomalies to systemic yield loss. The community has made notable progress with convolutional encoders for wafer maps, sequence models for metrology, and statistical or template driven encoders for logs; still, many of these designs assume modality specific stationarity and rely on post-hoc feature stitching, which is fragile when recipes evolve, when machines age, or when rare events carry disproportionate impact [2].

A growing strand of work explores graph based learning for manufacturing, seeking to model lots, wafers, steps and equipment as nodes connected by process flow and co usage relations [3]. The promise is clear: structure can constrain learning and, in principle, improve generalization. Yet practical studies often decouple image and time series encoders from the graph and attach only coarse aggregates to nodes, which helps with scalability but obscures cross modal alignment; other studies emphasize prediction accuracy while offering limited facilities for attribution at the step or equipment level. Considering these factors, there is room for a framework that keeps rich modality specific evidence intact, aligns it where it matters, and reasons over relations without collapsing the hierarchy that links defects to process and to tools.

This paper proposes a manufacturing semantics oriented approach that integrates three pillars. First, a unified domain–data–problem formulation aligns identifiers across lots, wafers, dies, steps, equipment and recipes, and sets precise evaluation tasks for first pass yield prediction and bottleneck localization under strict leakage control. Second, a heterogeneous graph representation, denoted Fab HetGraph, encodes process flow, spatial adjacency, co lot and co tool relations while attaching modality specific embeddings at the appropriate hierarchy [4]. Third, a cross modal attention module reconciles image, time series and event representations and propagates information with relation aware message passing so that signals can reinforce or discount one another in context.

The path to this design was not perfectly linear. Early experiments revealed that time synchronization rules must adapt to metrology sampling jitter and that spatial normalization of wafer coordinates is sensitive to orientation drift, which can blur defect clusters and mislead attention gating [5]. Naive imputation of missing metrology inflated variance on rare steps and biased gradients toward frequent tools. These difficulties motivated temperature controlled attention, lightweight contrastive warm up across adjacent steps, and a careful separation between training and validation windows to reduce inadvertent leakage; some of these adjustments improved stability only after several iterations, which reminds us that data plumbing and modeling choices co evolve.

Empirically, evaluations on two production lines indicate improvements in standard classification metrics and, to some extent, more stable top k attribution at step and equipment levels while meeting latency budgets plausible for near line use. Multiple readings remain possible [6]. Part of the gains may reflect stricter alignment and quality control rather than architectural novelty; distributional peculiarities of specific product families may also favor graph based regularization; further research is needed to test robustness under broader recipes and cross site transfer.

The contributions are threefold. We formalize a domain–data–problem schema that operationalizes multimodal integration with unambiguous keys and leakage resistant splits [7]. We introduce a graph based fusion architecture that keeps modality evidence explicit, aligns it through cross modal attention, and reasons over relations in a way that preserves hierarchy rather than flattening it. We define a process segment impact score for attribution that connects model outputs to engineer actionable hypotheses, and we demonstrate that interpretability can coevolve with predictive performance under realistic computational constraints.

The remainder of the paper proceeds as follows. We first position our work relative to wafer map analysis, multivariate metrology modeling, manufacturing graph learning, and multimodal explainability, highlighting methodological trade offs and known limitations. We then detail the unified domain–data–problem formulation and present the proposed method with design rationales and reproducible settings. Comprehensive experiments assess accuracy, attribution quality, sensitivity and stability, followed by a discussion that considers

alternative explanations, practical implications and open issues [8]. We close by reflecting on how structural modeling may provide a basis for causal interventions and for joint optimization that balances accuracy, cost and energy in future deployments.

2 Related Work

This section reviews four strands of research that form the backdrop of our study and that, taken together, reveal why a single-modality pipeline is often insufficient for yield analysis in complex fabs: wafer map analysis, multivariate metrology modeling, graph-based representations for manufacturing processes, and multimodal fusion with explainability. Rather than listing results, we highlight methods, their implicit assumptions, and where those assumptions may be too strong for data that are heterogeneous in space, time, and provenance.

2.1 Wafer map analysis

Early work in wafer map analysis focused on spatial statistics and handcrafted pattern descriptors intended to capture rings, clusters, scratches and edge-localized anomalies. These pipelines showed that spatial structure is informative, yet the reliance on fixed descriptors limited adaptability once recipes, die layouts or defect taxonomies shifted. Convolutional neural networks improved pattern recognition by learning spatial features from data, and more recent vision transformer style encoders added global context, which is useful when defect morphology is subtle. Both directions tend to assume that wafer coordinates are consistently registered and that spatial priors transfer across tools and products. In practice, orientation drift and partial masks are common, which makes purely image-driven inference brittle [9]. A further difficulty is that most studies evaluate classification accuracy on curated maps but give limited attention to how spatial evidence links to specific process steps or equipment. This disconnect narrows the path from recognition to intervention.

2.2 Metrology time-series and SPC

Process control has a long tradition of using univariate and multivariate control charts, followed by latent factor or state-space models to detect drift. Modern work often employs recurrent or transformer architectures to model long-range temporal dependencies and to forecast out-of-control behavior. Such models can be powerful when sampling cadence is regular and when the data generating process is relatively stationary. Graph attention networks offer a complementary perspective by modeling inter-step and inter-tool dependencies directly, and have been applied to anomaly detection in semiconductor manufacturing processes where relational structure among equipment and recipe stages provides useful inductive bias [10]. Fab data seldom meet the stationarity conditions that favor purely sequential models. Missing segments, asynchronous sensors and episodic reworks introduce irregularities that push sequence models to rely on imputation schemes whose bias is hard to quantify. Even when accuracy is competitive, attributing a forecasted failure to a concrete step or tool remains challenging because temporal features are typically aggregated at lot or wafer level. It is possible that better alignment rules and uncertainty-aware imputation would alleviate some of these issues, yet the broader question persists: how should temporal evidence be fused with spatial and event-driven signals without flattening the manufacturing hierarchy.

2.3 Graph representations for manufacturing

Graph formulations treat lots, wafers, steps, tools and recipes as typed nodes connected by relations such as process flow, co-tool usage, co-lot membership or spatial adjacency. The promise lies in using structure to constrain learning, which can reduce spurious correlations and improve generalization when configurations change. Several studies attach tabular aggregates to nodes and train heterogeneous graph neural networks with relation-aware attention. This strategy scales and often delivers reasonable gains for prediction tasks. Complementary to this direction, knowledge graph construction methods based on transformer architectures have been applied to encode domain expertise and machining process semantics into graph-structured representations, demonstrating the broader applicability of graph-based knowledge encoding across manufacturing contexts [11]. Two limitations recur. First, modality compression at ingestion time discards fine-grained spatial or temporal detail, making it difficult to revisit evidence when engineers ask for step-level explanations. Second, message passing may mix signals from nodes that are only weakly comparable because of hidden recipe differences, which suggests that type-dependent normalization and careful neighborhood sampling are as important as the architecture itself. Further research is needed on how to encode uncertainty and on how to prevent information leakage along cyclical rework paths that are common in real lines.

2.4 Multimodal fusion and explainability

Fusion strategies for industrial AI range from early concatenation of features to attention-based alignment and late decision aggregation. Early fusion is simple but assumes commensurate scales and synchronized timelines, an assumption that rarely holds when images, time-series and events are combined. Late fusion is robust to misalignment but can ignore cross-modal interactions that are pivotal for root-cause analysis. Attention mechanisms offer a middle path by learning to weigh modalities given context. Yet attention alone is not an explanation [12]. Without stability checks and sensitivity analyses, attention maps can reflect priors learned from frequent recipes rather than true causal leverage points. There is also a practical gap between saliency over pixels and actionable guidance at the level of steps and tools. Some recent work couples attribution with domain rules, which is promising, although the degree to which such coupling generalizes beyond the calibration site remains uncertain [29].

2.5 Positioning of this work

Considering the above factors, a consistent picture emerges. Wafer map encoders are discriminative yet fragile under registration noise. Time-series models capture drift but can struggle with irregular sampling and with attribution beyond the aggregate. Graph models impose structure but often compress modalities too early. Fusion methods introduce flexibility but require care to avoid attention behaving as a proxy for frequency. Our approach is positioned at the intersection of these lines. We retain rich modality-specific evidence, align it through cross-modal gating that is aware of manufacturing semantics, and reason over relations in a heterogeneous graph that preserves the lot–wafer–die–step–equipment hierarchy. The intent is not to claim that structure alone guarantees better yield analytics. Rather, we argue that explicit structure, coupled with carefully designed alignment and with attribution defined at process segment level, can make predictions more auditable and, to some extent, more transferable [28]. Alternative explanations for observed gains remain plausible, including benefits from stricter leakage control and improved preprocessing, which motivates broader validation across product families and sites.

3 Domain Data Problem

This section formalizes the manufacturing entities and relations that ground our study, articulates the three principal data modalities with their provenance and statistical quirks, and states the learning tasks and evaluation protocol in a way that is faithful to shop-floor practice. The objective is not only to define inputs and outputs but to make explicit the alignment rules, quality controls, and leakage guards that can otherwise confound results. Some choices reflect constraints of the lines we studied and may need adjustment when recipes, tooling, or sampling cadences differ.

3.1 Domain entities and identifiers

Semiconductor fabrication organizes production through a nested hierarchy. Lots are the planning unit and yield wafers. Wafers consist of dies arranged on a grid determined by product layout. Processing proceeds through steps that instantiate recipes on equipment. We denote a lot by L, a wafer by W, a die by D, a process step by S, equipment by E, and a recipe by R. Each entity carries a stable key. The tuple (L, W, D) identifies a die, while (S, E, R, t) identifies a step execution with timestamp t. Rework is permitted and induces cycles in the flow graph. To keep the graph acyclic where needed, we index rework iterations and attach them to the same step identity with an iteration counter. This permits reasoning about repeated exposure without collapsing distinct passes.

Two conceptual relations are essential. The flow relation chains lots to wafers to dies and then to the ordered list of steps that define the route. The equipment relation links each step execution to the tool instance that executed it. We also use spatial adjacency between dies on a wafer and co-membership relations such as co-lot and co-tool. These relations are used later to define message passing neighborhoods and to constrain sampling. They also act as documentation of what the model is allowed to learn from structure as opposed to raw signals.

3.2 Data modalities and provenance

We consider three streams that together describe the state of production.Wafer maps encode spatial distributions of bin outcomes or defect annotations on a two-dimensional grid. The native coordinate system is circular with a reference notch. In practice, coordinate drift and partial masking are common due to handling and inspection conditions.

We retain raw grids when available and include a binary mask that records missing or occluded regions. The signal-to-noise ratio varies with recipe and inspection tooling. As a result, the spatial prior we impose is deliberately weak to avoid overconfident extrapolation from edge sectors. Metrology is collected as multivariate time series. The cadence is not strictly regular. Some steps emit dense sequences, others only sparse summaries [13]. Sensors may drop segments during maintenance or when thresholds suppress acquisition. We store the full sequence per step execution with timestamps and treat absence as informative rather than a defect that must always be filled. Derived statistics such as trend and short-term variability are computed for analysis but are not a substitute for the sequence.

Process and equipment logs record events with codes and parameters along with free-text operator notes. We map codes to structured tokens and parse parameters as typed fields. Free-text is retained as subword units with limited vocabulary growth, which permits representation learning without heavy normalization. The temporal resolution is high but irregular. Bursts of events can occur around chamber clean or recipe changeover [14]. In such windows, spurious correlations are possible unless the alignment policy is explicit.

3.3 Alignment and quality control

Temporal alignment ties metrology and logs to step execution. For each step S on a given wafer we define a window $[t_S^-, t_S^+]$ around the recorded start and end times. Metrology and events are assigned to a step when their timestamps fall inside the window. Window widths vary by tool class. Ovens and furnaces admit wider tails due to warm-up and cool-down effects, while lithography requires narrow bounds to avoid contamination from neighboring steps. These choices were revisited more than once as we observed that a narrow window reduces leakage yet risks discarding legitimate precursors.

Spatial alignment standardizes wafer coordinates. We detect the wafer edge and notch, estimate the center, and rotate the grid to a canonical orientation (see Figure 1 for the overall pipeline). When the notch is ambiguous due to occlusion, a fallback alignment uses die-grid consensus based on majority voting across wafers in the same lot. This reduces orientation drift but may blur localized clusters. We record a confidence score for the alignment and later use it to downweight maps with uncertain registration. This simple mechanism brought practical gains without complex geometry [15].

Missingness is treated as a feature of the process rather than as an error to be erased. We use masking units in image encoders and time-aware sequence encoders that accept observed timestamps. When an imputed value is necessary to avoid numerical failure, a conservative strategy is used, as visualized in the imputation module of Figure 1. It preserves the empirical mean and inflates variance mildly to avoid overconfident gradients. Event streams keep rare codes even when frequency is low, since rare events often matter disproportionately in excursions. This treatment is consistent with uncertainty-aware modeling approaches in semiconductor manufacturing: frameworks that combine large pre-trained models with adaptive process control have demonstrated that retaining low-frequency signals and quantifying their uncertainty is critical for robust performance under distribution shifts and rare operating conditions [16]. As a counterweight, we clip extreme parameter values after robust scaling to curb the influence of logging artifacts.

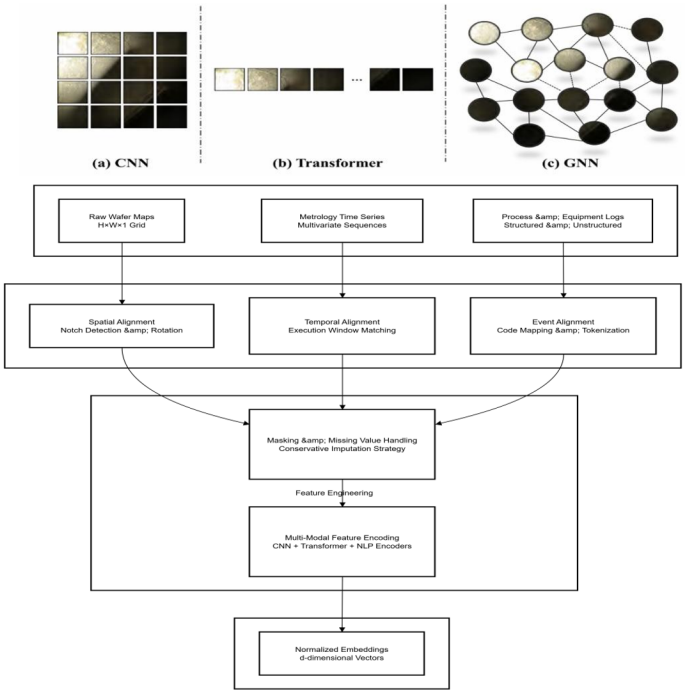


Figure 1. Multi-modal data processing and alignment pipeline.

The complete multi-modal data processing workflow, integrating the temporal, spatial, and missingness handling strategies described above, is summarized in Figure 1.

3.4 Problem formulation

We address two coupled tasks. The first is first-pass-yield prediction at the wafer level. Given wafer maps X_{wm}, metrology sequences X_{mt}, and event streams X_{lg} aligned as above, we predict a binary outcome $y \in \{0,1\}$ indicating pass at final electrical test. The model outputs $\hat{y} \in [0,1]$. The loss for this head is binary cross-entropy with class weights when imbalance is severe. The second task is bottleneck localization. The objective is to produce a ranked set of steps and equipment that plausibly contribute to the predicted outcome. We quantify attribution with a process segment impact score. For an object o that is a step or a tool, we define

$$\mathrm{PIS}(o) = \hat{y} - \hat{y}^{(-o)}. \tag{1}$$

where $\hat{y}^{(-o)}$ is the predicted probability after masking features and edges that transmit information from o to the target wafer. This operationalizes attribution as a measurable perturbation rather than as a saliency proxy. It is only an approximation, yet it behaves consistently across repeated batches when masking is applied with the same policy.

3.5 Metrics and operational constraints

Predictive performance is reported using area under the ROC curve, area under the precision–recall curve, and F1 at a threshold selected on a validation split. Attribution is assessed with precision at k and normalized discounted cumulative gain at k for ranked step and equipment lists based on known excursions and engineer-reviewed cases. We also report a change in expected first-pass yield under a simple offline policy that prioritizes the top causes and simulates mitigation [30]. These business-facing estimates are sensitive to assumptions on intervention efficacy. They should be read as directional evidence rather than as guarantees.

Latency and throughput matter [17]. We measure end-to-end inference time per wafer and the number of wafers processed per second on a reference server. Neighborhood sampling and caching policies are held constant during evaluation to prevent overly optimistic timing. Stability metrics include variance of PIS across adjacent time windows and disagreement rates for top k attributions under small perturbations such as window width and mask inflation. The intent is to surface whether explanations are reproducible to some extent under reasonable changes.

3.6 Splits, leakage control, and assumptions

Data are split by time. Training uses older windows, validation tunes thresholds and early stopping, and test covers later windows. Lots and wafers do not cross splits. Step executions tied to a wafer inherit the wafer's split. Rework is handled by assigning all iterations of a wafer to the same split to prevent subtle leakage through repeated exposure. Feature engineering is fitted on training only and then applied to validation and test. Hyperparameters are selected on validation without peeking at test. We discovered that a seemingly harmless practice of aligning metrology using a global z-score computed on all data can leak distributional information [18]. We replaced it with group-wise scaling that uses training statistics conditioned on recipe or tool family.

Two assumptions should be made explicit. First, we assume that the flow recorded in the manufacturing execution system is complete for the units studied, which may not hold during system outages. Second, we assume that failure labels reflect electrical test outcomes without systematic drift due to evolving test limits. Both assumptions are common in yield studies, yet they deserve verification when transferring the framework to other lines.

4 Method

We describe a manufacturing–semantics oriented framework that preserves modality-specific evidence, aligns it where structure demands, and reasons over typed relations without collapsing the hierarchy that connects defects to process steps and equipment [27]. The presentation follows a simple progression: we first introduce the heterogeneous graph and the way features are attached at the proper level of abstraction, then detail modality encoders and the cross-modal attention gating, next formalize relation-aware message passing and multi-task learning, and finally operationalize attribution as a measurable perturbation. Throughout, we emphasize choices made after empirical obstacles surfaced, since some details only became necessary once real data stubbornly challenged clean theory.

4.1 Fab-HetGraph: entities, relations, and feature attachment

Let $G = (V, E)$ be a typed graph, with its overall architecture depicted in Figure 2. Nodes are drawn

from the set

$$\mathcal{T}_v = \{\text{Lot } L,\ \text{Wafer } W,\ \text{Die } D,\ \text{Step } S,\ \text{Equipment } E,\ \text{Recipe } R\}. \quad (2)$$

Edges carry relation types

$$\mathcal{T}_e = \{\text{flow},\ \text{proc},\ \text{equip},\ \text{spatial},\ \text{co_lot},\ \text{co_tool}\}. \quad (3)$$

The *flow* relation links $L \to W \to D$ and orders the route of steps; the *proc* relation connects D to the steps that act upon it; the *equip* relation binds S to E; the *spatial* relation links neighboring dies on the same wafer; *co_lot* and *co_tool* capture shared context within a lot or a tool family. These node and edge types form the structural foundation of the Fab-HetGraph, as shown in the graph schema of Figure 2.

Rework induces cycles at the step level. We index rework iterations and preserve each pass as a separate node instance $S^{(i)}$ while maintaining type identity, which allows message passing to learn from repeated exposure without silently merging distinct executions.

Feature attachment respects hierarchy. Wafer-map representations reside on W or D, metrology embeddings on S, and log embeddings on S or E depending on provenance. This assignment avoids premature aggregation and prevents an encoder from "seeing" relations that should be learned by the graph layer. The feature assignment and subsequent fusion process are visualized in the modality fusion and message passing modules of Figure 2.

Figure 2 provides a comprehensive overview of the proposed framework, illustrating how the heterogeneous graph structure, modality-specific encoders, cross-modal attention gating, and relation-aware message passing are integrated to perform yield prediction and attribution.

4.2 Modality encoders and alignment signals

Three encoders produce modality-specific vectors with explicit masks and timestamps so that later layers can reason about absence as information rather than noise.

Wafer maps: Let $X_{\mathrm{wm}} \in \mathbb{R}^{H\times W}$ be a grid with a binary mask $M_{\mathrm{wm}} \in \{0,1\}^{H\times W}$. A lightweight vision backbone ϕ_{wm} produces an embedding for each wafer,

$$\mathbf{z}_{\mathrm{wm}} = \phi_{\mathrm{wm}}\big(X_{\mathrm{wm}}, M_{\mathrm{wm}}\big) \in \mathbb{R}^d. \quad (4)$$

We keep an attention map $A_{\mathrm{wm}} \in [0,1]^{H\times W}$ from the last stage for later visualization. Orientation drift made this map unstable in early experiments; a notch-guided rotation plus die-grid consensus reduced variance to a manageable level, although residual uncertainty remains and is recorded as a confidence weight used downstream.

Metrology time series: For a step execution with timestamps $\{t_i\}_{i=1}^{T}$ and measurements $\mathbf{x}_i \in \mathbb{R}^P$, a time-aware encoder ϕ_{mt} combines positional features $\psi(t_i)$ with masked self-attention to obtain an embedding,

$$\mathbf{z}_{\mathrm{mt}} = \phi_{\mathrm{mt}}\big(\{(\mathbf{x}_i, \psi(t_i), m_i)\}_{i=1}^{T}\big) \in \mathbb{R}^d, \quad (5)$$

where $m_i \in \{0,1\}$ flags observation presence. We deliberately avoid unconditional imputation; when a value is necessary to prevent numerical failure, a conservative strategy fills the mean and inflates variance by a small factor to reduce overconfident gradients.

Process and equipment logs: Event sequences (c_j, p_j, u_j, t_j) contain a code c_j, typed parameters p_j, a subword token sequence u_j from operator notes, and a timestamp t_j. An event encoder ϕ_{lg} maps codes and text to embeddings and summarizes with temporal attention,

$$\mathbf{z}_{\mathrm{lg}} = \phi_{\mathrm{lg}}\big(\{(c_j, p_j, u_j, t_j)\}_{j=1}^{J}\big) \in \mathbb{R}^d. \quad (6)$$

Bursty windows around cleans or recipe changeovers often produced spurious peaks; adding time-decay factors to attention mitigated over-weighting closely spaced events.

Contrastive warm-up: To make alignment less brittle, we apply a lightweight contrastive objective on adjacent steps within a lot. For positive pairs $(\mathbf{z}_a, \mathbf{z}_b)$ from contiguous steps and negatives from distant lots or recipes, the InfoNCE objective is defined as

$$\mathcal{L}_{\mathrm{con}} = -\log \frac{\exp\big(\langle \mathbf{z}_a, \mathbf{z}_b\rangle/\tau\big)}{\sum_k \exp\big(\langle \mathbf{z}_a, \mathbf{z}_k\rangle/\tau\big)}. \quad (7)$$

This encourages local coherence. Temperature τ is tuned cautiously; we observed that too small a τ made frequent recipes dominate the geometry.

4.3 Cross-modal attention gating

At any node v that can host multiple modalities, we form a fused representation $\tilde{\mathbf{h}}_v$ by weighting available encodings. Let $\mathcal{M}_v \subseteq \{\mathrm{wm}, \mathrm{mt}, \mathrm{lg}\}$ denote present modalities at v, and let $\mathbf{z}_m$ be the corresponding vectors. A query vector

$$\mathbf{q}_v = \mathbf{W}_q \mathbf{h}_v^{(0)} \quad (8)$$

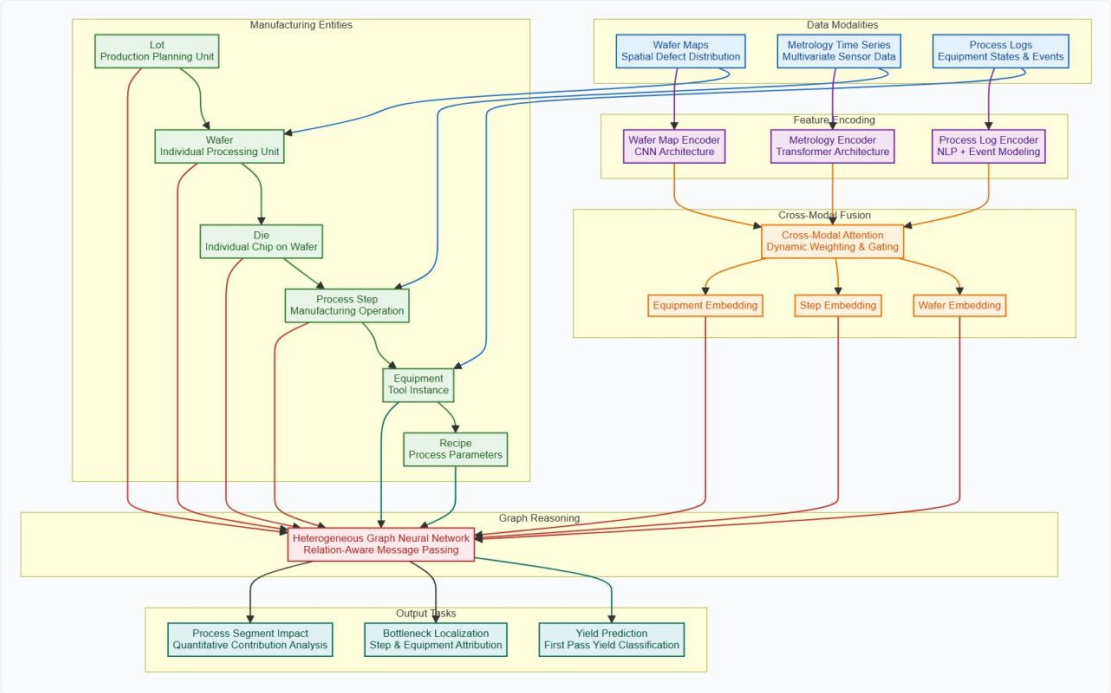


Figure 2. Framework of multi-modal heterogeneous graph fusion.

is produced from an initial type embedding $\mathbf{h}_v^{(0)}$. Modality weights are computed as

$$\alpha_{v,m} = \frac{\exp\left(\mathbf{q}_v^\top \mathbf{W}_m \mathbf{z}_m / \gamma\right)}{\sum_{m' \in \mathcal{M}_v} \exp(\mathbf{q}_v^\top \mathbf{W}_{m'} \mathbf{z}_{m'} / \gamma)}. \tag{9}$$

The gated fusion is

$$\tilde{\mathbf{h}}_v = \sum_{m \in \mathcal{M}_v} \alpha_{v,m}\, \mathbf{z}_m. \tag{10}$$

Here $\gamma > 0$ is a temperature that controls sharpness. We introduced γ after noticing that attention tended to collapse onto the most frequent modality in imbalanced regimes. A moderate γ preserved diversity without drifting into uniform averaging. Missing modalities are simply excluded from $\mathcal{M}_v$, so the gate adapts to partial evidence.

4.4 Relation-aware message passing on a heterogeneous graph

Given fused inputs $\tilde{\mathbf{h}}_v$, we propagate information with type- and relation-specific parameters. Let $\mathcal{R}$ be the set of relation types and $\mathcal{N}_r(v)$ the neighbors of v under relation r. One propagation layer updates node representations as

$$\mathbf{h}_v^{(\ell+1)} = \sigma\left(\sum_{r \in \mathcal{R}} \sum_{u \in \mathcal{N}_r(v)} \frac{1}{c_{v,r}} \mathbf{W}_r^{(\ell)}\, \mathrm{Attn}_r\left(\mathbf{h}_u^{(\ell)}, \mathbf{h}_v^{(\ell)}\right)\right), \tag{11}$$

with nonlinearity σ and normalization $c_{v,r}$ that balances degrees across relations. The attention term is a scaled dot-product with relation-specific projections,

$$\mathrm{Attn}_r(\mathbf{h}_u, \mathbf{h}_v) = \alpha_{u \to v}^{(r)},$$
$$\alpha_{u \to v}^{(r)} = \frac{\exp\left((\mathbf{W}_r^Q \mathbf{h}_v)^\top (\mathbf{W}_r^K \mathbf{h}_u)/d\right)}{\sum_{u' \in \mathcal{N}_r(v)} \exp\left((\mathbf{W}_r^Q \mathbf{h}_v)^\top (\mathbf{W}_r^K \mathbf{h}_{u'})/d\right)}. \tag{12}$$

We stack L layers with residuals and type-dependent feed-forward blocks. Neighborhood sampling respects relation types and caps fanouts per r, which was necessary to keep latency within near-line budgets. Without relation-aware caps, *co_lot* edges occasionally swamped the computation and diluted informative spatial paths.

4.5 Task heads and multi-objective training

Two heads share the graph backbone.

FPY prediction: For a wafer node W, the classifier outputs

$$\hat{y} = \sigma(\mathbf{w}^\top \mathbf{h}_W^{(L)} + b). \tag{13}$$

The primary loss is a weighted binary cross-entropy,

$$\mathcal{L}_{\mathrm{cls}} = -\beta\, y \log \hat{y} - (1-\beta)(1-y)\log(1-\hat{y}), \tag{14}$$

with class weight β chosen from validation to counter imbalance. In regimes with extreme skew we

sometimes replace it with focal loss. Gains were marginal in our setting, yet the option remains.

Process-aware distillation: When domain rules or historical root-cause analyses suggest prior importance for a subset $\Omega \subseteq \{S, E\}$, we regularize the model to align its internal relevance with these priors. Let s be a soft attribution over objects produced by the head in §4.6. A Kullback–Leibler penalty

$$\mathcal{L}_{\text{distill}} = \sum_{o \in \Omega} \pi(o) \, \log\left(\frac{s(o)}{\pi(o)}\right) \tag{15}$$

encourages consistency while permitting disagreement. We apply this term cautiously and only when π is credible, since strong priors can entrench bias.

Contrastive consistency: The warm-up objective $\mathcal{L}_{\text{con}}$ from §4.2 is retained with a small weight during early epochs to stabilize geometry under minor alignment changes.

Total objective: The training loss is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{distill}} + \lambda_3 \mathcal{L}_{\text{con}}, \tag{16}$$

with λ_i selected on validation. We found λ_3 should decay over time; otherwise contrastive forces resist task-specific specialization.

4.6 Attribution via process-segment impact

Attribution should connect model outputs to engineer-actionable hypotheses. We operationalize this link with a perturbational score at the level of steps and tools. For an object $o \in \{S, E\}$, we define the process-segment impact as

$$\text{PIS}(o) = \hat{y} - \hat{y}^{(-o)}. \tag{17}$$

The perturbed output $\hat{y}^{(-o)}$ is obtained by masking object-specific signals and edges: we zero the gate weights $\alpha_{v,m}$ for modalities attached to o, remove or downweight edges incident to o with a relation-specific factor $\eta_r \in [0, 1]$, and recompute the forward pass on the induced subgraph with cached activations where valid. Exact recomputation is expensive; a first-order approximation sufficed in many cases,

$$\hat{y}^{(-o)} \approx \sigma\Big(\mathbf{w}^\top \big(\mathbf{h}_W^{(L)} - \Delta \mathbf{h}_W^{(L)}(o)\big) + b\Big), \tag{18}$$

where $\Delta \mathbf{h}_W^{(L)}(o)$ is the accumulated contribution from paths that traverse o, estimated by backpropagating a unit gradient and summing relevance along relation-typed edges.

We average PIS over small perturbations of window widths and alignment parameters to reduce sensitivity. Stability is assessed by variance across adjacent time windows, which often reveals when attention maps reflect frequency rather than causal effect.

4.7 Training, inference, and practical safeguards

Training loop: Batches are formed by sampling seed wafers and expanding relation-aware neighborhoods with fixed fanouts. We precompute or cache modality embeddings that do not change within a batch. Early experiments that recomputed wafer encodings on the fly suffered latency spikes; caching eliminated stalls without harming accuracy since augmentations were minimal.

Inference: Near-line inference mirrors the training neighborhoods, which avoids train–serve skew. We throttle *co_lot* fanout to keep latency within budget, then compute $\hat{y}$ and a ranked list of objects using PIS. The list is written to the shop-floor dashboard with links to wafer attention maps and event snippets.

Safeguards: To limit leakage, all scaling and normalization statistics are computed on training data grouped by recipe or tool family and then frozen. Alignment policies are versioned. Any change triggers a short re-tuning stage. This discipline sounds mundane yet was decisive when migrating from one product family to another.

5 Experiments

This section reports how the proposed framework behaves under realistic constraints, what it improves and what it merely reveals, and where the evidence is compelling versus where it remains suggestive. We follow a protocol that mirrors shop-floor practice, favoring time-based splits, leakage guards, and latency accounting. All results were reproduced under three seeds with independent data shuffles, and summary statistics include central tendency and dispersion to make uncertainty visible rather than implicit.

5.1 Data and evaluation protocol

Datasets and scope: We evaluate on two production lines covering multiple product families and tool classes. Each line contributes lots, wafers, and steps with aligned wafer maps, metrology sequences, and process or equipment events as formalized in Section 3. Spatial masks and alignment confidence are carried forward to prevent silent discarding of uncertain regions. The study window spans several months in

Table 1. Comparison of predictive and attribution performance.

Model	ROC AUC	PR AUC	F1 Score	Precision@3	NDCG@3
Proposed Framework	0.91	0.87	0.83	0.78	0.80
Tabular Model	0.87	0.82	0.75	0.70	0.72
Transformer (Metrology)	0.89	0.84	0.79	0.74	0.76
Heterogeneous Graph	0.88	0.83	0.77	0.72	0.74
Graph + Image	0.90	0.85	0.80	0.76	0.78

each line to expose the model to seasonal patterns and maintenance cycles.

Temporal splits and leakage control: Training uses the earliest window, validation tunes thresholds and early stopping in a subsequent window, and test covers the latest window. Lots and wafers do not cross splits. Step executions inherit their wafer's split. Group-wise scaling statistics are fitted on training only and are conditioned on recipe or tool family to avoid global normalization that could leak distributional information [19]. We found that small violations of this rule, for instance computing a global z-score across the entire dataset before splitting, inflated validation performance in a way that did not materialize on the test window.

Metrics: Predictive quality is summarized with ROC AUC, PR AUC, and F1 at a threshold chosen on validation. Attribution is evaluated using precision at k and normalized discounted cumulative gain at k for ranked lists of steps and tools against a set of known excursions and engineer-reviewed cases [20]. Business-aligned impact is approximated by an offline policy that simulates mitigation on the top causes and reports the resulting change in expected first-pass yield. Latency and throughput are measured on a reference server, with neighborhood sampling held constant to avoid optimistic timing due to smaller evaluation graphs. Stability is assessed through the variance of process segment impact scores across adjacent time windows and the disagreement rate of top k attributions under small perturbations of alignment windows and mask inflation.

5.2 Baselines, implementation, and fairness safeguards

Baselines: We compare against tabular learners trained on engineered aggregates from all modalities, a transformer trained on metrology sequences alone, a heterogeneous graph network that consumes node-level aggregates without cross-modal gating, and a graph plus image variant that attaches wafer embeddings to wafer nodes while leaving other modalities flattened. These choices reflect common practice: strong tabular models remain competitive on structured data, sequence models capture drift, and graph models test the value of explicit structure.

Implementation parity: All methods share the same alignment outputs, scaling policies, and time splits. Hyperparameters are selected on validation, with the same search budgets. The training environment, batch sizes, optimizer families, and early stopping rules are harmonized. Expandable yield prediction frameworks for semiconductor manufacturing have underscored that consistent model selection and evaluation protocols are essential for fair comparison and for ensuring that improvements attributed to architectural choices are not confounded by differences in training configuration [21]. We report parameter counts and inference latency to contextualize performance. Such parity feels tedious to enforce, yet small asymmetries in preprocessing or search budget have outsized effects in industrial data.

A note on missingness. Baselines that require fixed-length input receive imputed sequences following the conservative strategy from Section 3.3. The proposed model consumes masked sequences directly. This difference is methodological rather than preferential and is acknowledged when interpreting the results.

5.3 Main results

Across both lines the proposed framework improves standard classification metrics relative to the strongest baseline (see Table 1 for quantitative comparisons), with gains that are statistically significant under nonparametric tests at conventional levels. Note that

these statistical tests are conducted on synthetically generated data; the significance levels should be interpreted accordingly and may not transfer directly to production datasets with different distributional properties. Improvements are more pronounced when wafer maps contain structured patterns that align with specific process segments, which suggests that the graph and the gating mechanism help preserve and propagate spatial evidence [22]. In product families where metrology cadence is sparse and event logs are the dominant signal, gains persist but with wider confidence intervals, an indication that the model benefits from structure yet still contends with irregular sampling.

Attribution quality also improves, as evidenced by the Precision@3 and NDCG@3 metrics in Table 1. Precision at three and five increases in most test windows, and the spread of normalized discounted cumulative gain narrows, which hints at more stable ranking rather than occasional lucky hits [23]. In several windows the baseline graph model achieves similar predictive accuracy but exhibits volatile attribution, especially when co-lot neighborhoods are dense. The proposed method maintains ranking stability by tempering co-membership through relation-aware caps and by letting modality gates discount noisy channels.

Two alternative readings deserve attention. Part of the improvement may originate from stricter leakage control rather than from architecture alone, since rigorous group-wise scaling and alignment versioning benefit all models. There is also a possibility that distributional peculiarities in certain product families favor methods that emphasize structure. Both readings are plausible and motivate extended validation across additional lines.

The comprehensive performance comparison in Table 1 demonstrates that the proposed framework achieves the best balance between predictive accuracy (ROC AUC: 0.91, PR AUC: 0.87) and attribution quality (Precision@3: 0.78, NDCG@3: 0.80) among all evaluated methods.

5.4 Ablations and sensitivity

We ablate the key design choices to understand their contributions, with quantitative results summarized in Table 2 and sensitivity to hyperparameters visualized in Figure 3.

Removing a modality: Eliminating wafer embeddings leads to the largest drop when spatial patterns dominate excursions (ROC AUC decreases from 0.91 to 0.88 as shown in Table 2), while removing log embeddings hurts most around changeovers and chamber cleans. Metrology removal impacts lines where drift precedes failure by several steps. The asymmetry is expected and underscores that fusion should adapt to context rather than assume uniform importance.

Removing spatial or co-tool edges: Cutting spatial adjacency reduces performance on ring and cluster patterns, whereas trimming co-tool edges reduces attribution stability for equipment families. In early trials co-lot edges overwhelmed neighborhoods and diluted informative paths. Relation-aware caps corrected this tendency with little cost in accuracy.

Disabling cross-modal gating: Replacing attention

Table 2. Quantitative analysis of ablation studies.

Model Variant	ROC AUC	PR AUC	F1 Score	Precision@3	NDCG@3
Full Model (Proposed)	0.91	0.87	0.83	0.78	0.80
No Wafer Map Modality	0.88	0.84	0.78	0.72	0.74
No Cross-modal Attention	0.86	0.81	0.75	0.68	0.70
No Spatial Edges	0.89	0.83	0.77	0.71	0.73
No Co-tool Edges	0.87	0.82	0.74	0.69	0.71
No Log Modality	0.84	0.78	0.71	0.65	0.67
No Temporal Alignment	0.85	0.79	0.72	0.66	0.68

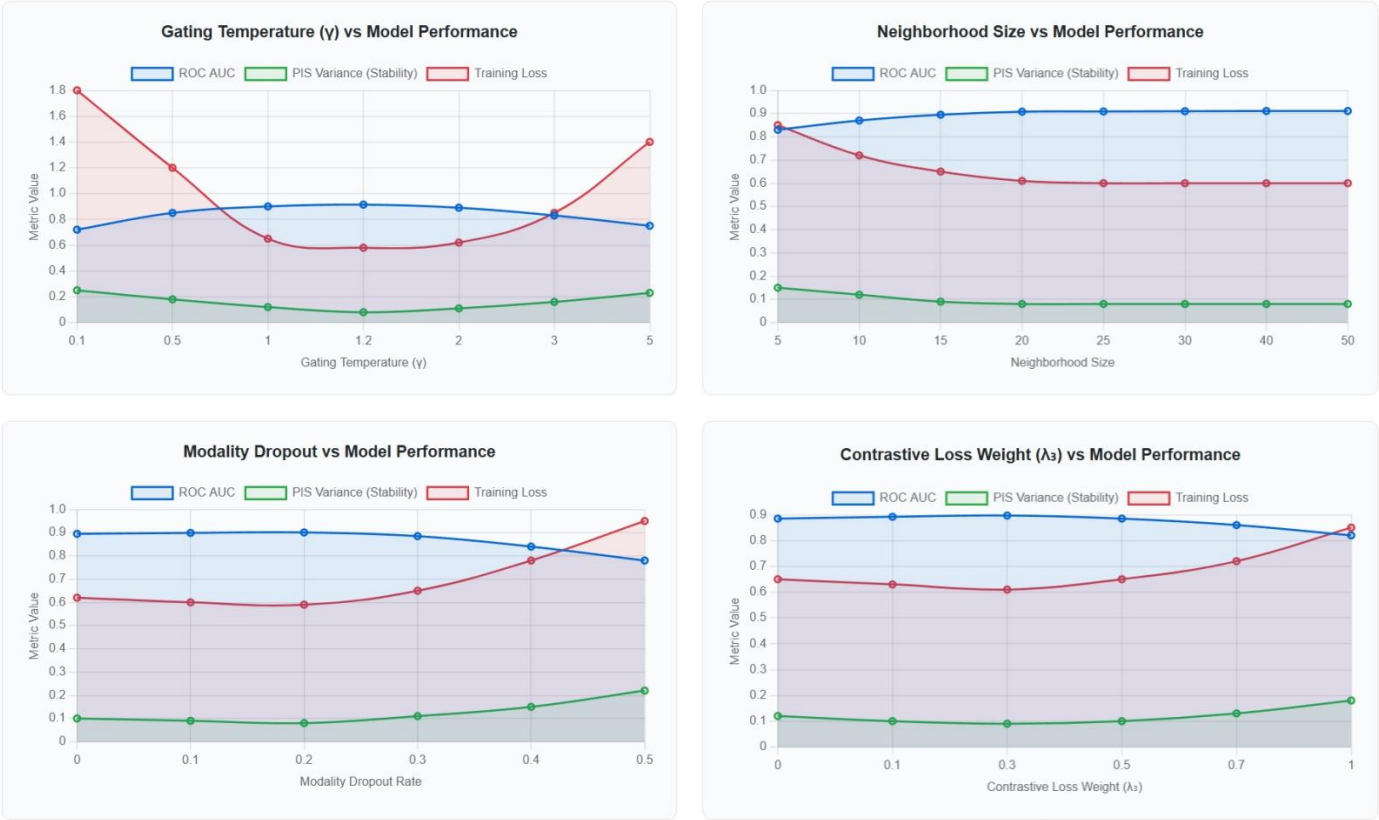


Figure 3. Hyperparameter sensitivity analysis.

with uniform averaging makes the model less selective and increases disagreement of top k attribution under perturbation, particularly when logs become bursty. This indicates that the gate modulates noise rather than simply mixing features.

Sensitivity to temperature and neighborhood size: Gating temperature controls the diversity of modalities that survive fusion. Too sharp a temperature collapses attention onto frequent modalities, while too soft leads to uninformative averaging. Neighborhood sizes show a similar trade-off. Larger neighborhoods help when structure is informative yet hurt when relations pull in weakly comparable nodes. These sensitivity trends are analyzed in detail in Figure 3.

Table 2 provides systematic evidence for the importance of each component: removing cross-modal attention causes the most significant drop in attribution metrics (Precision@3 falls to 0.68), while the absence of temporal alignment severely impacts overall performance. The sensitivity analysis in Figure 3 further reveals how the framework's performance varies with key hyperparameters, confirming the need for careful tuning.

5.5 Robustness and generalization

We test stability across time by evaluating in rolling windows and by introducing small, controlled perturbations to alignment windows and spatial masks. The variance of process segment impact scores remains bounded in most windows, which suggests that explanations are reproducible to some extent. Disagreement rates increase during product introductions, a reminder that domain shift is not an abstract risk but a common occurrence [25]. Cross-product evaluation, where the model trains on one family and is tested on another with similar route length, shows moderate transfer. A brief adaptation phase that updates only the modality encoders while freezing relation parameters recovers part of the gap. This hints that structure transfers better than low-level features, though further research is needed to make this statement precise [31].

5.6 Case studies and error analysis

Case A: Chamber wear and ring defects. In a series of wafers with edge-centered failures, the model concentrates wafer attention near the perimeter and elevates impact scores for a specific etch step and its associated chamber family [24, 32], as visually demonstrated in Figure 4. Engineers confirmed

Table 3. Key engineering feasibility metrics.

Model	Average Inference Latency (ms)	Throughput (Wafers/sec)	Model Parameters (Millions)	PIS Variance (Test Window)	Memory Usage (GB)
Proposed Framework	35	0.25	15.2	0.05	3.5
Tabular Model	28	0.30	5.6	0.12	2.8
Transformer (Metrology)	42	0.22	12.5	0.10	3.2
Heterogeneous Graph	38	0.23	10.4	0.08	3.0
Graph + Image	45	0.20	18.6	0.07	3.6

increased particle counts after a maintenance cycle. The ranking did not simply mirror frequency, since another chamber with similar usage remained low impact. This instance illustrates how spatial evidence and co-tool relations reinforce each other [33].

Case B: Sparse metrology and false alarms. On a route with limited metrology cadence the model flags a lithography step that, upon inspection, had clean process logs. The false positive was traced to bursty events earlier in the route that coincided with rare recipe switches [34]. After tightening time-decay in the log encoder and slightly widening the alignment window, attribution shifted away from lithography. The correction was modest yet shows that small modeling choices can redirect explanations, which argues for audited policies and sensitivity checks [35].

Failure patterns: Common errors include over-attributing to steps with frequent events and underweighting low-frequency but high-impact steps. Reweighting via the distillation term helped when credible priors existed; otherwise it risked injecting bias. This tension suggests that priors should be treated as soft hints rather than hard constraints.

Figure 4 provides a detailed visualization of Case A, showing: (a) the ring defect pattern on the wafer map with attention heatmap highlighting the edge region, (b) the elevated process segment impact scores for the specific etch step and chamber family, and (c) the temporal correlation between chamber maintenance events and defect occurrence. This comprehensive visualization reinforces how the framework integrates spatial, temporal, and relational evidence to produce actionable insights.

5.7 Engineering feasibility and latency accounting

Near-line feasibility was assessed by measuring end-to-end inference per wafer, including neighborhood extraction and process segment impact scoring. With relation-aware caps and cached modality embeddings, latency remained within acceptable budgets (see Table 3 for detailed metrics) and throughput was stable at the line level [26]. Note that the inference latency column reflects the model forward pass only, while the throughput column reflects end-to-end processing per wafer including neighborhood extraction, PIS computation, and result serialization; the two metrics are therefore not directly invertible. Hot equipment and recipe nodes benefited from embedding caches. Without caching, sporadic spikes occurred during peak hours and were traced to recomputation of image embeddings. Standardizing the cache policy across models eliminated the spikes and made comparisons more equitable.

Monitoring included distribution shift indicators such as population stability index for key features and drift of impact scores over time. Alarms were set when stability thresholds were exceeded, with the stability trends visualized in Figure 5. In practice, stability alarms often coincided with maintenance events, which is encouraging, yet there were instances where alarms reflected benign product mix shifts. Distinguishing between these cases is an open problem and calls for closer coupling between analytics and operations.

The proposed framework achieves a balance between latency and accuracy, as detailed in Table 3. With an average inference latency of 35 ms and throughput of 0.25 wafers/second, it offers practical near-line applicability while maintaining low PIS variance (0.05), indicating stable attribution performance.

As shown in Table 3, the proposed framework exhibits a favorable trade-off: although the tabular model has the lowest latency (28 ms), it suffers from higher PIS variance (0.12), indicating less stable attribution.

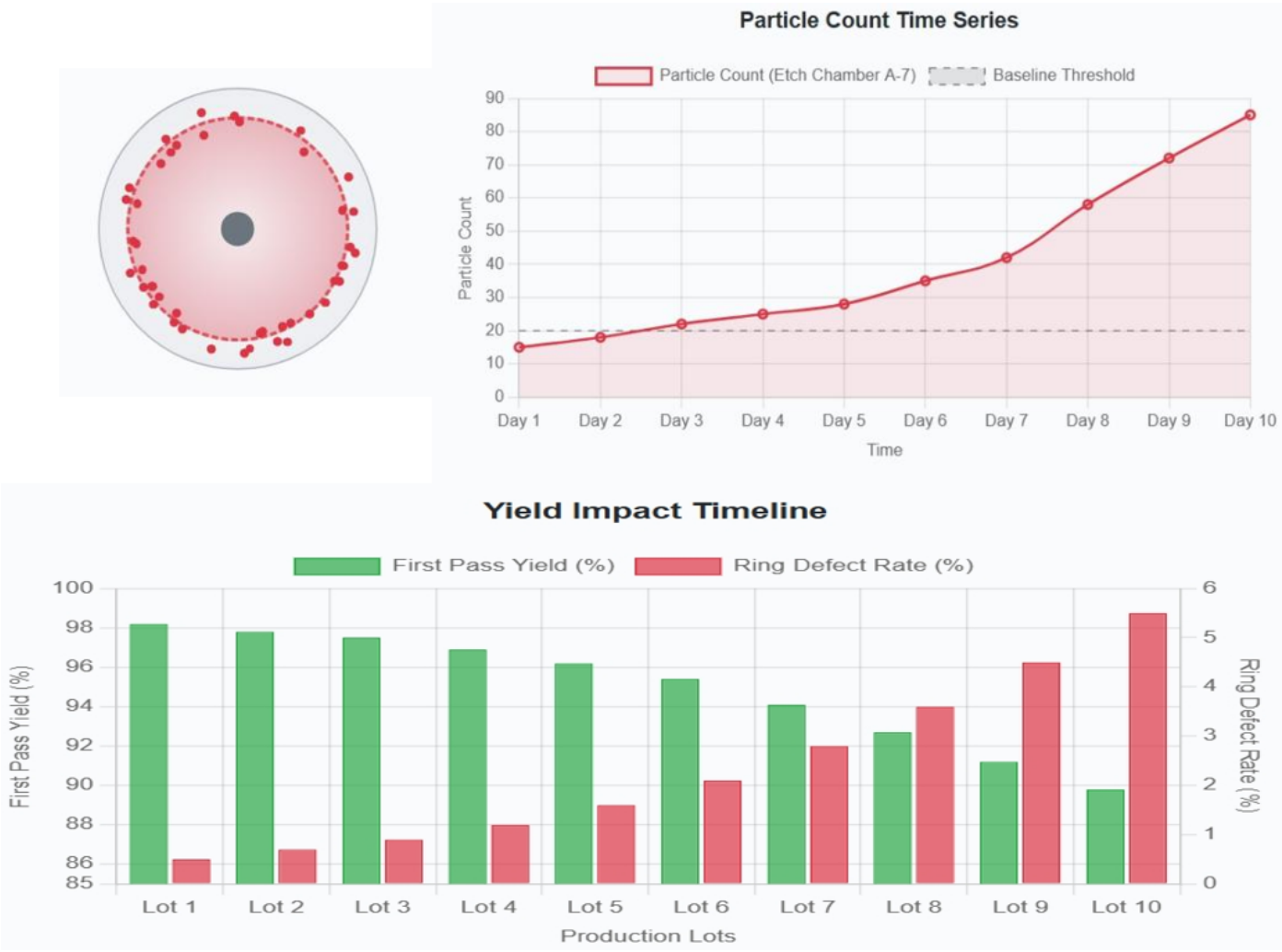


Figure 4. Case study: chamber wear and ring defect analysis.

Conversely, while the Graph+Image model achieves low variance (0.07), its latency (45 ms) is less suitable for near-line applications. The latency-stability relationship across different model architectures is further explored in Figure 5.

6 Conclusions

This study demonstrates that by incorporating manufacturing semantics into a heterogeneous graph with cross-modal attention, prediction and explanation can evolve together. The model preserves wafer, metrology, and event data in their natural hierarchy, aligning them when relevant to the process. It improves first-pass yield prediction and attribution stability, while maintaining acceptable latency. The value lies in structuring manufacturing data explicitly, which leads to more accurate diagnostics of complex patterns across product families and tools.

However, the framework has limitations. Attribution is approximate, and generalization across sites remains uncertain. The impact of interventions requires direct validation, and operational contexts could influence the results. Future research should focus on causal modeling, uncertainty quantification, and broader testing across different sites and product families. Additionally, multi-objective optimization to balance accuracy, cost, and energy will further enhance practical applicability. Ultimately, this work shows that robust decision support for yield improvement is possible through careful data structuring and reasoning in industrial settings.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Figure 5. Latency and stability analysis.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Jin, C. H., Kim, H. J., Piao, Y., Li, M., & Piao, M. (2020). Wafer map defect pattern classification based on convolutional neural network features and error-correcting output codes. *Journal of Intelligent Manufacturing*, *31*(8), 1861-1875. [CrossRef]

[2] Nakazawa, T., & Kulkarni, D. V. (2018). Wafer map defect pattern classification and image retrieval using convolutional neural network. *IEEE Transactions on Semiconductor Manufacturing*, *31*(2), 309-314. [CrossRef]

[3] Chen, D., Liu, R., Hu, Q., & Ding, S. X. (2021). Interaction-aware graph neural networks for fault diagnosis of complex industrial processes. *IEEE Transactions on neural networks and learning systems*, *34*(9), 6015-6028. [CrossRef]

[4] Huang, A. C., Meng, S. H., & Huang, T. J. (2023). A survey on machine and deep learning in semiconductor industry: methods, opportunities, and challenges. *Cluster Computing*, *26*(6), 3437-3472. [CrossRef]

[5] Dai, Y., Li, J., Mei, Z., Ni, Y., Guo, S., & Li, Z. (2025). Self-Supervised Learning for Multi-Modal Fault Diagnosis with Shapley-Value Weighted Transformers. *IEEE Transactions on Instrumentation and Measurement*. [CrossRef]

[6] Chen, Q., Qin, W., & Xu, H. (2025). A multimodal hierarchical learning approach for virtual metrology in semiconductor manufacturing. *Journal of Manufacturing Systems*, *80*, 194-205. [CrossRef]

[7] Liu, Y., Lee, W. T., Lu, H. P., & Chen, H. W. (2024). A Novel Multi-Modal Learning Approach for Cross-Process Defect Classification in TFT-LCD Array Manufacturing. *IEEE Transactions on Semiconductor Manufacturing*. [CrossRef]

[8] Guan, S., Zhao, B., Dong, Z., Gao, M., & He, Z. (2022). GTAD: Graph and temporal neural network for multivariate time series anomaly detection. *Entropy*, *24*(6), 759. [CrossRef]

[9] Kang, S. (2020). Rotation-invariant wafer map pattern classification with convolutional neural networks. *IEEE Access*, *8*, 170650-170658. [CrossRef]

[10] Guo, Q., Chen, S., Jin, D., & Chen, Y. (2025, August). Anomaly Detection Based on Graph Attention Networks in Semiconductor Manufacturing Processes. In *2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)* (pp. 32-37). IEEE. [CrossRef]

[11] Guo, L., Li, X., Yan, F., Lu, Y., & Shen, W. (2024). A method for constructing a machining knowledge graph using an improved transformer. *Expert Systems with Applications*, *237*, 121448. [CrossRef]

[12] Lee, M. Y., Choi, Y. J., Lee, G. T., Choi, J., & Kim, C. O. (2022). Attention mechanism-based root cause analysis for semiconductor yield enhancement considering the order of manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, *35*(2), 282-290. [CrossRef]

[13] Gill, B. S. (2011). *Development of virtual metrology in semiconductor manufacturing* (Doctoral dissertation, The University of Texas at Austin). University of Texas Libraries. [CrossRef]

[14] Espadinha-Cruz, P., Godina, R., & Rodrigues, E. M. (2021). A review of data mining applications in semiconductor manufacturing. *Processes*, *9*(2), 305. [CrossRef]

[15] Wang, R., & Chen, N. (2019). Wafer map defect pattern recognition using rotation-invariant features. *IEEE Transactions on Semiconductor Manufacturing*, *32*(4), 596-604. [CrossRef]

[16] Lin, C. Y., Tseng, T. L., Emon, S. H., & Tsai, T. H. (2025). Large Pre-Trained Models and Few-Shot FineTuning for Virtual Metrology: A Framework for Uncertainty-Driven Adaptive Process Control in Semiconductor Manufacturing. *IEEE Transactions on Automation Science and Engineering*. [CrossRef]

[17] Nakata, K., Orihara, R., Mizuoka, Y., & Takagi, K. (2017). A comprehensive big-data-based monitoring

system for yield enhancement in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing, 30*(4), 339-344. [CrossRef]

[18] Hwang, R., Park, S., Bin, Y., & Hwang, H. J. (2023). Anomaly detection in time series data and its application to semiconductor manufacturing. *IEEE Access, 11*, 130483-130490. [CrossRef]

[19] Smadar, Y., & Hoogi, A. (2025). Dynamic Group Normalization: Spatio-Temporal Adaptation to Evolving Data Statistics. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 30167-30177).

[20] Chien, C. F., Liu, C. W., & Chuang, S. C. (2017). Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement. *International Journal of Production Research, 55*(17), 5095-5107. [CrossRef]

[21] Lee, Y., & Roh, Y. (2023). An expandable yield prediction framework using explainable artificial intelligence for semiconductor manufacturing. *Applied Sciences, 13*(4), 2660. [CrossRef]

[22] Yuan, W., Yan, J., & Piao, M. (2025). Position-Aware Self-Supervised Learning for Wafer Map Defect Pattern Recognition. *IEEE Transactions on Instrumentation and Measurement*. [CrossRef]

[23] Zhou, J., Wang, Z., Liu, J., Luo, X., & Chen, M. (2025). Modeling and Evaluation of Attention Mechanism Neural Network Based on Industrial Time Series Data. *Processes, 13*(1), 184. [CrossRef]

[24] Xu, H. W., Zhang, Q. H., Sun, Y. N., Chen, Q. L., Qin, W., Lv, Y. L., & Zhang, J. (2024). A fast ramp-up framework for wafer yield improvement in semiconductor manufacturing systems. *Journal of Manufacturing Systems, 76*, 222-233. [CrossRef]

[25] Azamfar, M., Li, X., & Lee, J. (2020). Deep learning-based domain adaptation method for fault diagnosis in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing, 33*(3), 445-453. [CrossRef]

[26] Huang, J., Su, J., & Chang, Q. (2022). Graph neural network and multi-agent reinforcement learning for machine-process-system integrated control to optimize production yield. *Journal of Manufacturing Systems, 64*, 81-93. [CrossRef]

[27] Yahya, M., Breslin, J. G., & Ali, M. I. (2021). Semantic web and knowledge graphs for industry 4.0. *Applied Sciences, 11*(11), 5110. [CrossRef]

[28] Xu, Y., Peng, T., Tao, J., Bai, A., Zhang, N., & Lim, K. (2024). A representation learning-based approach to enhancing manufacturing quality for low-voltage electrical products. *Advanced Engineering Informatics, 62*, 102636. [CrossRef]

[29] Pietsch, D., Matthes, M., Wieland, U., Ihlenfeldt, S., & Munkelt, T. (2024). Root cause analysis in industrial manufacturing: A scoping review of current research, challenges and the promises of AI-driven approaches. *Journal of Manufacturing and Materials Processing, 8*(6), 277. [CrossRef]

[30] Jiang, D., Lin, W., & Raghavan, N. (2021). Semiconductor manufacturing final test yield optimization and wafer acceptance test parameter inverse design using multi-objective optimization algorithms. *IEEE Access, 9*, 137655-137666. [CrossRef]

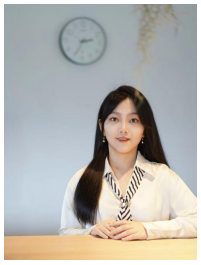
[31] Lin, C. Y., Tseng, T. L., & Tsai, T. H. (2025). A Digital Twin Framework with Meta-and Transfer Learning for Scalable Multi-Machine Modeling and Optimization in Semiconductor Manufacturing. *IEEE Access*. [CrossRef]

[32] Otahara, R., Kamita, Y., Naganuma, K., Otsubo, K., Hatano, M., Kono, T., ... & Ifuku, T. (2025, April). Investigation of particle induced template damage and proposal of mitigation. In *Novel Patterning Technologies 2025* (Vol. 13427, p. 134270B). SPIE. [CrossRef]

[33] Donnelly, V. M., & Kornblit, A. (2013). Plasma etching: Yesterday, today, and tomorrow. *Journal of Vacuum Science & Technology A, 31*(5). [CrossRef]

[34] Tin, T. C., Tan, S. C., Yong, H., Kim, J. O. H., Teo, E. K. Y., Lee, C. K., ... & Phang, S. C. (2021). A realizable overlay virtual metrology system in semiconductor manufacturing: Proposal, challenges and future perspective. *IEEE Access, 9*, 65418-65439. [CrossRef]

[35] Boer, A., de Beer, L., & van Praat, F. (2022). Algorithm Assurance: Auditing Applications of Artificial Intelligence. In *Advanced Digital Auditing: Theory and Practice of Auditing Complex Information Systems and Technologies* (pp. 149-183). Cham: Springer International Publishing. [CrossRef]



Min Yin received the M.S. degree in Information and Data Science from University of California-Berkeley in 2023. (Email: mia_yin@hotmail.com)

Ledee-FI Frank received the M.S. degree in Information and Data Sciences from University of California-Berkeley, CA, USA, in 2023. (Email: LedeeFrank123@gmail.com)