CRAFT: Cross-modal Representation with Adaptive Fusion Transformer for Operational Defect Detection

Aman Gulati[1], Virinchi Srinivas[1], Gokul Swamy[1], and Vikalp Gajbhiye[1]

Amazon.com Inc.

Abstract. When fusing heterogeneous modalities for classification, a central challenge is cardinality heterogeneity: modalities often produce token sequences of vastly different lengths, yet standard symmetric fusion wastes attention capacity under this asymmetry. We present CRAFT, a modality-agnostic fusion framework that selects a high-density attention backbone using token cardinality and standalone task relevance, then conditions this backbone through asymmetric cross-attention and gated residual fusion. This design avoids the attention dilution that arises when cardinality-mismatched sequences interact symmetrically, while enabling inference over arbitrary modality subsets. CRAFT further introduces a class-frequency-adaptive Supervised Cross-Modal Alignment objective (SupCMA) for rare-class alignment and Counterfactual Cross-Modal Interaction Scores (CMIS), a model-internal counterfactual score for cross-modal feature interactions. On a large-scale proprietary deployment and a public benchmark dataset, CRAFT outperforms 12 baselines, improving PR-AUC by up to 7.6% over the strongest baseline ($p < 0.01$) while reducing fusion latency by 1.9×. In a four-week controlled A/B test, CRAFT reduced operational costs by 6.4%, corresponding to $6.51M in annualized savings.

Keywords: Multimodal fusion · Defect detection · Interpretability

1 Introduction

E-commerce logistics at scale produces operational defects (package losses, delivery failures, item condition issues, and fulfillment routing errors) stemming from network inefficiencies, supply-chain gaps, carrier breakdowns, or customer-side factors. Detecting and classifying these defects early enables targeted interventions: rerouting at-risk shipments, flagging supply-chain quality issues, or proactively communicating delays to customers. Accurate detection requires integrating evidence across fundamentally different data types: structured tabular features capturing order attributes and fulfillment history, unstructured text sequences encoding shipment tracking events and order patterns, and graph embeddings representing relational structure across shared logistics signals (e.g., shared carriers, delivery routes, seller networks). This integration must operate

under strict latency constraints (sub-1-second SLA for real-time order evaluation). Existing production systems treat each modality independently, missing cross-modal signals. For instance, an unusual delivery-route deviation (weak tabular signal) combined with anomalous scan-event timing in shipment tracking (weak text signal) at a hub with elevated loss rates (weak graph signal) may each individually appear benign, but their conjunction indicates a systemic fulfillment defect.

Standard multimodal fusion methods [3,13,1,11] assume comparable representational granularity across modalities. However, many applied settings exhibit *cardinality heterogeneity*: one modality may produce hundreds of tokens while others yield only a handful or a single pooled vector. This pattern arises in financial risk (100+ application features + credit report text), healthcare (structured EHR + clinical notes + patient similarity networks), and recommendation (user features + review text + interaction graphs). We formalize this as the *asymmetric-cardinality fusion* problem and show that symmetric approaches waste attention capacity on the lower-cardinality direction, with the inefficiency growing as the cardinality ratio increases. We propose CRAFT (Cross-modal Representation with Asymmetric Fusion Transformer), a modality-agnostic framework that explicitly accounts for this asymmetry. Our contributions are:

- We formalize *asymmetric-cardinality fusion*, a common production setting where one modality produces many fine-grained tokens while others provide compact embeddings.
- We propose density-guided asymmetric fusion, which selects a backbone using token cardinality and standalone task relevance (Proposition 1), avoiding redundant symmetric attention.
- We introduce class-frequency-adaptive SupCMA, a supervised cross-modal contrastive objective that strengthens alignment for rare classes without adding inference cost, and Counterfactual CMIS, a model-internal attribution method for cross-modal feature interactions.
- We validate CRAFT on a large-scale production deployment and a public benchmark, demonstrating higher PR-AUC and lower latency than 12 baselines, robustness to missing modalities, and a 6.4% online cost reduction ($6.51M annually, $p<0.01$).

2 Related Work

Multimodal fusion architectures. Multimodal fusion is commonly performed through early feature concatenation [3], late prediction aggregation [9], or intermediate representation fusion using cross-attention [6], dual-stream pathways [7], or mixture-of-experts routing [13,1]. Representative models such as MulT [11], MMBT [20], Perceiver IO [21], UniS-MMC [22], and AutoMM [23] improve fusion through symmetric attention, pseudo-token injection, latent bottlenecks, contrastive pre-alignment, or automated encoder selection. Related tabular-text methods either serialize structured rows into text [5,2,12] or assume

schema-text alignment [14], while FT-Transformer [15] preserves per-feature tokens but remains largely unimodal. These approaches optimize *how* modalities interact, but not *which* modality should provide the key-value memory under token-cardinality imbalance. Perceiver IO's latent bottleneck is modality-agnostic rather than cardinality- or task-guided, treating all inputs uniformly regardless of their representational richness. CRAFT addresses this gap through density-guided asymmetric backbone selection.

Contrastive cross-modal alignment. Contrastive learning has been widely used for representation alignment, including instance-level vision-language alignment in CLIP [16] and label-based supervised contrastive learning [17]. Cross-modal supervised contrastive objectives naturally extend this idea, but typically apply uniform alignment pressure across classes. This is limiting in imbalanced settings where minority classes provide far fewer positive pairs. CRAFT introduces class-frequency-adaptive alignment to strengthen rare-class supervision.

Multimodal interpretability. Existing interpretability methods, including SHAP [24], GNNExplainer [25], and attention visualization, usually explain each modality independently and therefore miss cross-modal interaction effects. Recent interaction-based methods [6] capture broader multimodal behavior but provide limited feature-level attribution across modalities. CRAFT addresses this gap through counterfactual cross-modal interaction scoring, which measures how perturbing evidence from one modality changes the contribution of features in another, thereby capturing cross-modal synergy and suppression at the feature level.

3 Methodology

3.1 General Framework

CRAFT is a fusion framework for M heterogeneous modalities $\{m_1, \ldots, m_M\}$ where each modality m_i produces n_i tokens of dimension d. The token count n_i depends on both the data and encoder design: a text encoder may produce n_{txt}=1 (mean-pooled), n_{txt}=128 (sequence of sentence embeddings), or n_{txt}=512 (per-token outputs).

Backbone Selection via Information Density. CRAFT selects the backbone using a composite criterion that accounts for both token cardinality and task relevance:

$$m^* = \arg\max_i \ \rho(m_i), \quad \rho(m_i) = \log_2(n_i) \cdot R_i, \tag{1}$$

where n_i is the token count and $R_i \in [0, 1]$ is the validation PR-AUC of a single-modality classifier for modality i, min-max normalized across modalities (for public benchmark, we use validation accuracy; the backbone choice is stable across metric variants). The logarithmic compression of cardinality ensures that when cardinalities are comparable (e.g., n_a=300 vs. n_b=256), the task-relevance

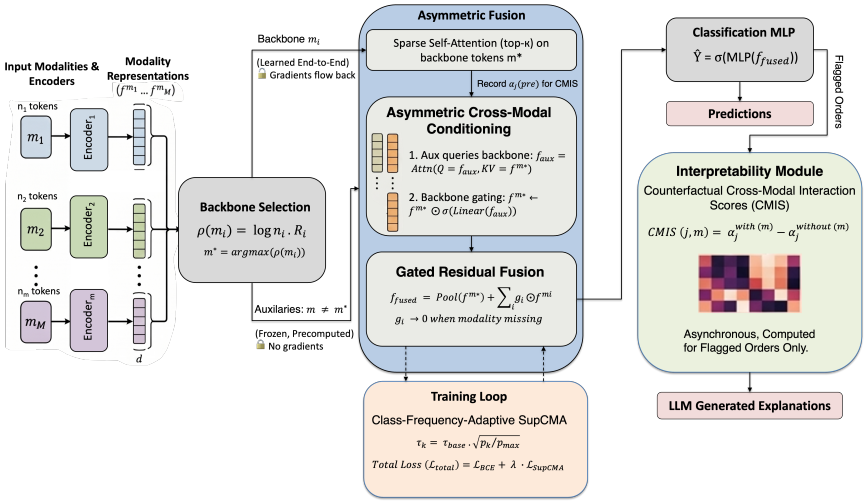

Fig. 1. CRAFT architecture. Frozen text/graph encoders and learned tabular tokenization produce heterogeneous-cardinality representations. Asymmetric cross-attention with gated residual fusion yields predictions. SupCMA aligns during training. CMIS enables interpretability.

term R_i breaks the tie in favor of the more predictive modality, while at extreme ratios (e.g., n_a=300 vs. n_b=1) the cardinality gap naturally dominates. This criterion introduces no additional computational cost since R_i is already available from pre-training.

Given the selected backbone, CRAFT: (1) directs all $M-1$ auxiliary modalities as queries attending over the backbone's key-value tokens; (2) conditions the backbone via per-auxiliary gating; and (3) fuses all streams via gated residual addition. The backbone identity is determined at model initialization. Since fusion layers attend over the backbone's tokens, the backbone tokenization is learned jointly with the fusion parameters. Auxiliary encoders, which serve only as queries, can be frozen without significant loss because their role is to select from the backbone rather than to be selected from. New modalities are added by simply instantiating an encoder and a gated residual branch. We validate CRAFT with M=3 (tabular, text, graph) for operational defect detection and M=2 (tabular, text) on public benchmark, but the framework generalizes to any modality combination (Figure 1).

3.2 Modality-Specific Representations

We describe the three modality encoders used in our production deployment for operational defect detection. The same architectural patterns are applied to the public benchmark in Section 4.7 with dataset-specific encoders.

Tabular. We encode 300+ structured features (order properties, fulfillment attributes, delivery indicators, seller metrics, network health signals) as per-feature tokens via quantile-binned embeddings [15]. Each numeric feature x_j is encoded using a learned bin embedding $E_{\text{bin}}(d_j)$, where d_j is the quantile bin index of

x_j, and a learned residual embedding $E_{\text{res}}(r_j)$ capturing the within-bin position $r_j \in [0, 1]$. The full token representation is:

$$t_j = W_p\big[E_{\text{bin}}(d_j) \parallel E_{\text{res}}(r_j) \parallel g(j)\big] \in \mathbb{R}^d, \tag{2}$$

where $g(j)$ is a learned feature-index embedding identifying which feature j the token represents, and W_p projects the concatenation to d=256. This yields F=300+ tokens per order.

Graph. A production heterogeneous graph (∼500M nodes, ∼2B temporal edges across 9 logistics-signal types including shared carriers, delivery routes, and seller networks) encodes relational patterns. A GAT-based model [4] produces 128D embeddings per order, projected to $\mathbb{R}^{1\times 256}$.

Text. Each order generates two textual streams capturing temporal patterns invisible to point-in-time tabular features. Given a recent history of n orders $S = \{o_1, \ldots, o_n\}$ for a customer, the order-history construction function is:

$$f_p(S) = \text{concat}(\{(o_i, m_i, a_i, q_i, p_i, s_i, l_i)\}),\ i \le n \tag{3}$$

where m_i = item description, a_i = brand/seller, q_i = quantity, p_i = payment method, s_i = shipping address, and l_i = outcome status. The shipment-journey prompt captures the package trajectory:

$$f_t(S) = \text{concat}(\{(sc_e, st_e, sr_e, ss_e, ty_e, d_e, v_e)\}),\ e \in \mathbb{Z}^+ \tag{4}$$

where e indexes each scan event (sc_e = scan type, st_e = relative time, sr_e = reason code, ss_e = status, ty_e = station type, d_e = handler count, v_e = item value).

Text encoder. A fine-tuned ModernBERT-Base [10] (149M params, selected over 11 alternatives for accuracy-latency trade-off, Table 2) produces a pooled embedding $\mathbf{e} \in \mathbb{R}^{768}$, projected to $\mathbb{R}^{1\times 256}$.

In our deployment, tabular yields F=300+ tokens while text and graph each yield 1 token (mean-pooled). Since tabular also has the highest standalone performance ($R_{\text{tab}} > R_{\text{txt}}, R_{\text{graph}}$), the density criterion (Eq. 1) selects tabular as backbone. In general, if a text encoder retains per-token outputs (n_{txt}=512) and achieves strong standalone performance, it may itself become the backbone.

3.3 Cardinality-Aware Asymmetric Fusion

Proposition 1 (Attention-selectivity advantage under cardinality imbalance). Let modality a produce n_a tokens and modality b produce n_b tokens with $n_a \gg n_b$. Cross-attention using a as the key-value backbone provides each auxiliary query access to n_a distinct key-value states. In the reverse direction, each high-cardinality query from a can retrieve from only n_b states, and when n_b=1 the attention distribution is degenerate (softmax over a single key yields 1 identically). Thus, under fixed embedding dimensionality, the high-cardinality-as-backbone direction supports strictly richer attention selectivity than the reverse.

Instantiation. In our deployment (n_a=300, n_b=1), the reverse direction collapses entirely: every tabular query receives the same broadcast value. By contrast, the productive direction ($b \to a$) performs meaningful selective retrieval over the 300-token backbone. Even at moderate asymmetry (n_a=300, n_b=128), the reverse direction forces 300 queries to compete for 128 keys, concentrating attention and reducing effective selectivity.
The ρ criterion (Eq. 1) complements this selectivity argument by incorporating task relevance, so the selected backbone offers not only more attention targets but more informative ones. When $n_a \gg n_b$, the log-cardinality term favors the higher-cardinality modality. When cardinalities are comparable, R_i becomes the deciding factor.

Design. The fusion operates in two asymmetric stages:

(1) Auxiliary queries backbone: Each auxiliary embedding serves as a query attending over all backbone key-value pairs: $\hat{f}^{m_i} = \text{Attn}(Q{=}f^{m_i}, KV{=}f^{m^*})$, producing a context-enriched auxiliary representation per modality.

(2) Backbone conditioning via gating: Each backbone token receives per-auxiliary gating: $f_j^{m^*} \leftarrow f_j^{m^*} \odot \sigma(\text{Linear}(\hat{f}^{m_i}))$, applied sequentially for each auxiliary m_i. This injects global context from auxiliary modalities into the fine-grained backbone stream without the without the attention dilution of the reverse direction.

Sparse backbone self-attention. Before cross-modal conditioning, backbone tokens interact via sparse top-κ self-attention: each query attends to only κ most relevant tokens via a learned routing projection, reducing self-attention from $O(F^2)$ to $O(F \cdot \kappa)$. For our deployment use-case, κ is selected via validation sweep over $\{4, 8, 16, 32, 64\}$. Doubling to κ=32 yields <0.1% gain at 1.5× latency.

Gated residual fusion. Final representations are combined via learned per-modality gates. For M modalities with backbone m^*:

$$f_{\text{fused}} = \text{Pool}(f^{m^*}) + \sum_{i \neq m^*} g_i \odot f^{m_i}, \tag{5}$$

where $g_i = \sigma(W_g^i[\text{Pool}(f^{m^*}) \| f^{m_i}] + b_g^i)$. This formulation is modality-count-agnostic: adding a new modality m_{M+1} requires only a new encoder and gate parameters (W_g^{M+1}, b_g^{M+1}). Cross-modal interaction between auxiliaries is mediated through the shared backbone via sequential gating, avoiding $O(M^2)$ pairwise connections. When a modality is missing at inference, its gate learns to approach zero, allowing the model to fall back to available modalities without architectural changes.

3.4 Class-Frequency-Adaptive SupCMA

Standard fusion relies solely on BCE loss, which provides $O(N_k)$ gradient signal per class k per batch, insufficient for rare defect types. Instance-level contrastive

methods (CLIP [16]) define positives by co-occurrence, not by semantic label. We introduce SupCMA with class-frequency-adaptive temperature.

Objective. For a mini-batch of N samples, each modality's post-fusion representation is projected and L2-normalized into a shared 128D space: $z_i^a = \pi_a(f_i^a)/\|\pi_a(f_i^a)\|_2$. The SupCMA loss for class k:

$$\mathcal{L}_{\text{SupCMA}}^{(k)} = \frac{-1}{|\mathcal{P}|N} \sum_{(a,b)\in\mathcal{P}} \sum_{i=1}^{N} \frac{1}{|P_i^{(k)}|} \sum_{j\in P_i^{(k)}} \log \frac{e^{\text{sim}(z_i^a, z_j^b)/\tau_k}}{\sum_\ell e^{\text{sim}(z_i^a, z_\ell^b)/\tau_k}} \tag{6}$$

where $\mathcal{P}$ is the set of modality pairs, $P_i^{(k)} = \{j \neq i : y_j^{(k)} = y_i^{(k)} = 1\}$ is the positive set for class k, and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity.

Adaptive temperature. The key novelty over standard Supervised Contrastive Learning (SupCon [17]): we set $\tau_k = \tau_{\text{base}} \cdot \sqrt{p_k/p_{\max}}$, where p_k is the prevalence of class k and $p_{\max} = \max_k p_k$. Lower temperature sharpens the contrastive distribution, amplifying gradients for rare classes without overweighting their contribution in the loss average. The effective gradient magnification for class k is $\sqrt{p_{\max}/p_k}$, providing up to 1.84× stronger alignment for the rarest vs. most prevalent defect type for our production use-case.

Combined objective. $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda\,\mathcal{L}_{\text{SupCMA}}$, with λ=0.3 and τ_{base}=0.07 selected via grid sweep over $\lambda \in \{0.1, 0.2, 0.3, 0.5, 1.0\}$ and $\tau \in \{0.05, 0.07, 0.1, 0.2\}$. Larger λ ($\geq$0.5) degrades performance as the contrastive objective dominates the classification loss. Projection heads (π_m: 2-layer MLP to 128D, L2-normalized) are discarded at inference, adding zero deployed parameters.

Batch construction. We use class-balanced batch sampling with oversampling to ensure sufficient positive pairs for rare classes in each batch of 512 samples. This same sampling strategy is applied across all ablation configurations, including non-SupCMA rows, so that reported SupCMA gains are attributable to the contrastive objective itself rather than the sampling strategy.

3.5 Counterfactual Cross-Modal Interaction Scores

CMIS answers a model-internal counterfactual question: *"How would feature j's importance change within the fusion computation if auxiliary modality m's gate were disabled?"* This distinguishes CMIS from standard attention visualization, which reflects only the final joint distribution.

Definition 1 (Counterfactual CMIS). *For tabular feature j and auxiliary modality m:*

$$\textit{CMIS}(j,m) = \alpha_j^{with(m)} - \alpha_j^{without(m)}, \tag{7}$$

where $\alpha_j^{without(m)}$ is the self-attention importance of feature j when modality m's gating is identity (counterfactual: modality not observed), and $\alpha_j^{with(m)}$ is the importance after modality m's gating is applied. $\textit{CMIS}(j,m) > 0$ indicates synergy: modality m amplifies feature j. $\textit{CMIS}(j,m) < 0$ indicates suppression.

CMIS is interventional with respect to the learned fusion computation, not causal with respect to the external data-generating process. It measures how the model's feature-importance distribution changes when an auxiliary modality's gate is set to identity (disabled) while all inputs remain fixed. CMIS requires $M+1$ self-attention passes per order (one baseline + one per auxiliary modality). Since this is infeasible during real-time inference, CMIS is computed asynchronously only for flagged orders within the operational explanation pipeline (Section 4.6), adding zero inference latency.

CMIS reveals three feature categories. Across the test population, CMIS identifies: (i) *Self-sufficient features* (low |CMIS| for all m): consistently important regardless of context, such as historical defect rates. (ii) *Synergistic features* (high positive CMIS): become important only with cross-modal context, e.g., a delivery-route feature becomes predictive when shipment text reveals anomalous scan timing at a specific hub. (iii) *Suppressed features* (negative CMIS): important in isolation but suppressed when cross-modal evidence provides an alternative explanation, e.g., high delivery delay suppressed when tracking confirms a carrier-wide weather disruption.

3.6 Training Pipeline

Algorithm 1 summarizes the complete training procedure. Text and graph encoders are frozen and their outputs precomputed. The tabular tokenization layer (E_{bin}, E_{res}, $g(j)$, W_p), all attention layers, gating, and the classification head are trained jointly as fusion parameters Θ.

4 Experiments

4.1 Setup

Datasets. *(1) Proprietary:* Orders from a major e-commerce marketplace with multi-label annotations for 5 operational defect types: D1 (delivery failure), D2 (item condition defects), D3 (fulfillment routing errors), D4 (return processing failures), D5 (supply-chain compliance violations) (prevalences in Table 5). Training: 6-month window (12.2% positive rate after undersampling). Out-of-time test: unsampled orders from a subsequent month. *(2) Public:* Airbnb Melbourne [18,19] (23K listings, 61 native tabular features + 28 host-written text fields). Details in Section 4.7.

Metrics. For proprietary experiments, we report aggregate PR-AUC and ROC-AUC as ratios (×) relative to the best single-modality baseline. For context, absolute values fall in the 0.30–0.45 range for PR-AUC and 0.80–0.90 for ROC-AUC, reflecting the low natural prevalence of defects. Public benchmarks report absolute values. All results are averaged over 10 independent runs for the proprietary dataset and 5 runs for public benchmarks, with statistical significance assessed via paired t-test.

Algorithm 1 CRAFT: Training Pipeline

Require: M modalities: frozen auxiliary encoders + learned backbone tokenization
Require: Backbone $m^* = \arg\max_i \rho(m_i)$; hyperparams $\lambda, \tau_{\text{base}}, \kappa$
Ensure: Trained fusion parameters Θ

Precompute: frozen text/graph encoder outputs (tabular tokenized on-the-fly)
for each class-balanced mini-batch $\mathcal{B}$ **do**
  // Step 1: Tokenize backbone
  $T^{m^*} \leftarrow \{W_p[E_{\text{bin}}(d_j) \| E_{\text{res}}(r_j) \| g(j)]\}_{j=1}^{n_{m^*}}$
  // Step 2: Sparse self-attention on backbone
  for each query q_j in T^{m^*} **do**
    $f_j^{m^*}$ += Attn(q_j, top-κ keys from T^{m^*})
  end for
  Record $\alpha_j^{\text{pre}} \leftarrow$ backbone attention importance
  // Step 3: Asymmetric cross-modal conditioning
  for each auxiliary modality $m_i \neq m^*$ **do**
    $\hat{f}^{m_i} \leftarrow \text{Attn}(Q{=}f^{m_i},\ KV{=}f^{m^*})$ ▷ aux queries backbone
    $f^{m^*} \leftarrow f^{m^*} \odot \sigma(\text{Linear}(\hat{f}^{m_i}))$ ▷ gated conditioning
  end for
  // Step 4: Gated residual fusion
  $f_{\text{fused}} \leftarrow \text{Pool}(f^{m^*}) + \sum_{i \neq m^*} g_i \odot f^{m_i}$
  // Step 5: Loss computation
  $\hat{y} \leftarrow \sigma(\text{MLP}(f_{\text{fused}}))$
  $\mathcal{L}_{\text{BCE}} \leftarrow -\frac{1}{N}\sum_{i,k} w_k \text{BCE}(y_{ik}, \hat{y}_{ik})$
  $z_i^m \leftarrow \pi_m(f_i^m)/\|\pi_m(f_i^m)\|$ for each modality m
  $\mathcal{L}_{\text{SupCMA}} \leftarrow \frac{1}{K}\sum_k \mathcal{L}_{\text{SupCMA}}^{(k)}$ with $\tau_k = \tau_{\text{base}} \cdot \sqrt{p_k/p_{\max}}$
  Update Θ via $\nabla(\mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{SupCMA}})$
end for

// CMIS (async, flagged orders only)
for each auxiliary m_i **do**
  $\text{CMIS}(j, m_i) \leftarrow \alpha_j^{\text{with}(m_i)} - \alpha_j^{\text{without}(m_i)}$
end for

Training. All experiments conducted on 8× NVIDIA A10G GPUs. Fusion model: d_{model}=256, H=8 heads, dropout 0.2, AdamW with cosine schedule (lr swept over [1e-4, 5e-3]), batch size 512. Text/graph encoders are frozen (precomputed). Tabular tokenization is learned end-to-end. Each method receives 50 tuning trials (Optuna, TPE sampler) with early stopping on validation PR-AUC (patience 3 epochs). All baselines use identical frozen encoders for fair comparison. Reported latency in Table 1 is average per-order fusion latency measured after encoder outputs are available. End-to-end production latency including feature retrieval and encoder inference is reported separately in Section 4.6.

4.2 Main Results

Table 1 compares CRAFT with 12 baselines spanning five paradigms: tree-based (XGBoost+embeddings), concatenation-based, attention-based, MoE-based, and

Table 1. Fusion comparison (× baseline). All methods use identical frozen encoders. Mean, 10 runs.

Architecture	PR ↑	ROC ↑	Lat.
Best single-modality (tab)	1.000×	1.000×	47ms
XGBoost (tab+text+graph emb.)	1.021×	1.008×	51ms
Early Fusion [3]	1.038×	1.014×	50ms
Late Fusion [9]	1.053×	1.015×	64ms
MMBT [20]	1.062×	1.019×	71ms
Low-Rank [8]	1.068×	1.021×	77ms
MulT [11]	1.080×	1.027×	92ms
Dual-Stream [7]	1.098×	1.032×	112ms
InterpretCC [6]	1.104×	1.038×	121ms
Perceiver IO [21]	1.108×	1.040×	145ms
UniS-MMC [22]	1.112×	1.042×	132ms
I^2MoE [13]	1.122×	1.045×	208ms
AutoMM [23]	1.131×	1.049×	185ms
CRAFT (ours)	**1.217×**	**1.071×**	**95ms**

modality-agnostic. All neural methods use identical frozen encoders and quantile-binned tabular tokenization. All baselines received comparable hyperparameter tuning.

CRAFT achieves 1.217× PR-AUC, outperforming AutoMM (the strongest baseline) by 1.076× and I^2MoE by 1.085× ($p<0.01$, paired t-test). Notably, CRAFT is also 1.9× faster than AutoMM (95ms vs. 185ms) and 2.2× faster than I^2MoE due to the efficient asymmetric design. MMBT [20], which injects tabular tokens into the text transformer sequence, underperforms attention-based methods because it loses the fine-grained per-feature structure when mixing modality tokens. Perceiver IO [21], a modality-agnostic architecture using a fixed latent array, reaches 1.108× but cannot exploit the cardinality asymmetry that CRAFT leverages. UniS-MMC [22], which pre-aligns modality spaces via contrastive learning before fusion, is complementary to but less effective than SupCMA's during fusion.

4.3 Text Encoder Selection

We compare 12 encoder-only and decoder-only architectures for the text modality. Table 2 shows the top 4. ModernBERT-Base achieves the strongest encoder performance at 62ms/order, benefiting from RoPE, alternating local–global attention, and Flash Attention. Larger decoder models (Qwen-2.5-1.5B) yield higher standalone performance but at 12× the latency, making them impractical for real-time deployment.

4.4 Ablation Study

Table 3 isolates each component's contribution by incrementally adding them to a symmetric fusion baseline, all using the same modality encoders. The key findings are: *(i)* Asymmetric fusion yields 1.016× over symmetric at 68%

Table 2. Text encoder selection (top 4 of 12 evaluated). Ratios (×) relative to best encoder.

Model	Params	PR (×)	Lat.
BERT-Large	340M	0.949×	66ms
ModernBERT-Base	**149M**	**1.000×**	**62ms**
GPT-2 Large	762M	0.987×	283ms
Qwen-2.5-1.5B	1.5B	1.086×	751ms

Table 3. Component ablation (× baseline).

Configuration	PR ↑	ROC ↑	Lat.
Best single-modality (tab)	1.000×	1.000×	47ms
Symmetric cross-attn.	1.136×	1.046×	295ms
+ Asymmetric (tab backbone)	1.154×	1.055×	95ms
+ Gated residual	1.184×	1.063×	95ms
+ SupCMA (uniform τ)	1.203×	1.068×	95ms
+ Adaptive τ_k (Full CRAFT)	**1.217×**	**1.071×**	95ms

lower latency (95ms vs. 295ms), eliminating the diluted attention direction. *(ii)* Gated residual adds 1.026× by selectively incorporating auxiliary signals only when complementary. *(iii)* SupCMA with uniform temperature yields 1.016×. Adaptive temperature adds a further 1.012×, concentrating gains on rare defect classes.

Topology validation. Table 4 isolates the backbone choice under matched budget with all other components held constant. The tabular backbone outperforms symmetric by 1.04× and text-backbone by 1.06× PR-AUC, confirming Proposition 1: the higher-ρ modality provides richer attention selectivity as the key-value provider.

Adaptive vs. uniform temperature. Table 5 shows the per-class impact. Adaptive τ_k amplifies gains precisely for low-prevalence defect classes while preserving high-prevalence performance.

Gains are inversely correlated with class prevalence: D5 (rarest, 1.5%) sees 5× larger relative improvement from adaptive temperature than D2 (most prevalent, 5.1%), confirming that frequency-adaptive sharpening concentrates gradient signal where it is scarce.

Isolating cross-modal benefit of SupCMA To confirm SupCMA's gains arise from *cross-modal* alignment (not merely additional gradient signal), we compare against a within-modality contrastive baseline (SupCon applied to tabular representations only). Within-modality SupCon provides +0.8% (gradient signal alone), while cross-modal SupCMA doubles this to +1.6% (uniform) and triples it to +2.8% (adaptive), confirming that cross-modal alignment and frequency-adaptive sharpening each contribute value beyond the contrastive loss itself.

4.5 Arbitrary Modality Subsets at Inference

A key advantage of CRAFT's gated residual design is that a *single trained model* handles any subset of modalities at inference time without separate models or

Table 4. Backbone topology ablation (matched parameters). All rows include gated residual + adaptive SupCMA.

Topology	PR (×)	ROC (×)
Symmetric	1.170×	1.057×
Asymmetric, text backbone	1.146×	1.051×
Asymmetric, tabular backbone	**1.217**×	**1.071**×

Table 5. Per-class improvement: uniform vs. adaptive τ_k.

Defect Type	No SupCMA	Uniform τ	Adaptive τ_k
D2 (5.1%, most prevalent)	1.184×	1.194×	1.196×
D4 (2.5%)	1.165×	1.180×	1.189×
D1 (2.1%)	1.190×	1.205×	1.214×
D3 (1.8%)	1.081×	1.107×	1.126×
D5 (1.5%, rarest)	1.140×	1.175×	1.199×

retraining. This is critical in production: new entities lack graph context (cold-start), some records lack text history, and different deployment contexts may have different modality availability. Table 7 evaluates all subsets containing the backbone. Degradation is gradual and proportional to modality informativeness: removing graph costs 6.2% relative, removing text costs 8.1%, and removing both costs 15.8%. The two-modality configurations (1.142×, 1.118×) remain competitive with several full-modality baselines from Table 1, confirming that the gated residual design enables a single model to adapt to partial modality availability without per-subset retraining.

4.6 Deployment and Online Results

System design. The production pipeline is designed to satisfy a sub 1-second service-level agreement (SLA). Feature retrieval from distributed stores accounts for the largest component of latency (∼600ms), after which text encoding (∼62ms) and graph inference (∼128ms) execute in parallel. The fusion layer is then applied sequentially, adding ∼95ms and yielding a critical-path latency of $\sim 600 + \max(62, 128) + 95 \approx 823$ms. This margin depends on the asymmetric fusion design: replacing it with symmetric cross-attention would increase fusion latency to 295ms (Table 3), pushing the end-to-end critical path to ∼1,023ms and exceeding the SLA.

A/B test. We evaluated CRAFT in a 4-week controlled experiment with a 50/50 traffic split (randomized at customer-ID level) against the previous production system, with identical downstream intervention logic. The primary metric was net operational defect cost per order. Guardrails included customer escalation rate and order cancellation rate. CRAFT reduced the primary metric by 6.4% relative to control (95% CI: [5.7, 7.1]%, bootstrap, $p<0.01$), corresponding to $6.51M annualized net savings (after accounting for false-positive costs). No guardrail showed statistically significant degradation; the smallest guardrail p-value was >0.3.

Table 6. Isolating cross-modal contribution: within-modality SupCon vs. cross-modal SupCMA.

Contrastive Objective	PR (×)	Δ vs. None
None (BCE only)	1.184×	—
Within-modality SupCon (tabular)	1.193×	+0.8% rel.
Cross-modal SupCMA (uniform τ)	1.203×	+1.6% rel.
Cross-modal SupCMA (adaptive τ_k)	**1.217**×	+2.8% rel.

Table 7. Arbitrary modality subsets at inference.

Available Modalities	PR (×)	Δ vs. Full
All (backbone + text + graph)	1.217×	—
Backbone + text (no graph)	1.142×	−6.2% rel.
Backbone + graph (no text)	1.118×	−8.1% rel.
Backbone only	1.025×	−15.8% rel.

Operational explanation pipeline. CMIS produces structured evidence (synergistic features, self-sufficient features, perturbation-based importance) consumed by an asynchronous explanation pipeline. An LLM renders these scored signals into natural-language rationales for analyst review; the LLM is used only for language generation, not for scoring or attribution. In a post-deployment survey, analysts (N=100) scored these explanations higher than prior feature-importance lists on a 5-point actionability scale (4.2 vs. 3.1), particularly valuing visibility into which modality triggered specific feature importance shifts.

4.7 Public Benchmark: Airbnb Melbourne

To validate generalizability on publically available multimodal data, we evaluate on the Airbnb Melbourne dataset [18,19] from the multimodal tabular benchmark of Shi et al. [19]. This benchmark tests whether CRAFT's asymmetric design generalizes beyond operational defect detection to a natural tabular-text classification setting with a different label space and domain.

Dataset and setup. The dataset contains ∼23K property listings with two natively co-occurring modalities: *(i) Tabular:* 61 features (37 categorical + 24 numerical) including neighborhood, room type, accommodates, bedrooms, bathrooms, host response rate, review scores, and amenity indicators. Each feature is tokenized via quantile-binned embeddings (61 tokens per listing). *(ii) Text:* 28 host-written text fields (listing name, summary, space description, neighborhood overview, house rules), concatenated into a single input and encoded via BERT-Base into a mean-pooled embedding ($\mathbb{R}^{1\times256}$). The task is multiclass price-bracket classification following the label definition in [19]. We use the standard 80/20 train/test split from [19] with 10% of training data held out for validation. All methods use identical text encoders and hyperparameter search budgets (50 Optuna trials).

Results. Table 8 presents results. CRAFT (M=2) achieves 50.1% accuracy, outperforming AutoMM by 2.9 accuracy points and I^2MoE by 7.4 points (p<0.01). The 61 tabular features form the backbone ($\rho_{\text{tab}} > \rho_{\text{txt}}$ since n_{tab}=61 $\gg$

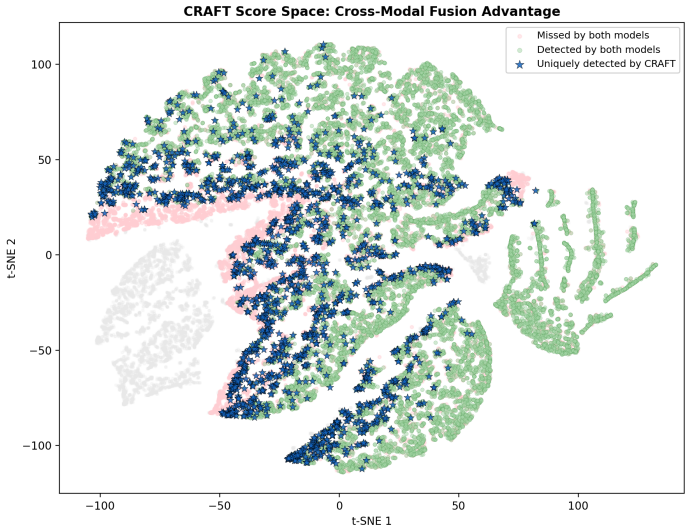

Fig. 2. CRAFT's score-space advantage. Blue stars mark defects uniquely detected by CRAFT but missed by the strongest baseline, concentrated near decision boundaries where symmetric methods are less discriminative.

n_{txt}=1), and the asymmetric design directs the text embedding as a query over tabular key-value tokens. These results suggest that CRAFT's asymmetric fusion transfers to domains and modality configurations beyond our primary deployment setting.

5 Limitations and Future Work

Frozen auxiliary encoders. Text and graph encoders are frozen before fusion training, so their representations are optimized for standalone classification rather than cross-modal alignment. This is a deliberate trade-off: freezing enables 10× faster training, independent encoder refresh cycles, and modular deployment. The measured accuracy gap from unfreezing is <0.3%, suggesting that pre-trained encoders already provide sufficiently task-relevant representations. Future work could explore lightweight adapters that fine-tune a small parameter subset without sacrificing modularity.

Backbone selection assumes pre-training alignment. The density criterion uses standalone performance R_i from encoder pre-training. If the pre-training objective differs from the downstream task, R_i may not reflect fusion relevance. In our setting, encoders are pre-trained on the same label set from a different time period, so this assumption holds. For transfer settings, a validation-based R_i estimated on a held-out set is a practical substitute at low cost.

CMIS interpretability scope. CMIS measures attention-based importance shifts under counterfactual gating and is validated against perturbation-based ground truth (r_s=0.81). However, attention importance is a proxy for functional contribution, not a formal causal quantity. We therefore use CMIS alongside

Table 8. Airbnb Melbourne (M=2, naturally multimodal). Accuracy (%), mean ± std over 5 runs.

Method	Accuracy (%)
XGBoost (tabular only)	43.5±0.5
BERT-Base (text only)	35.2±0.6
XGBoost (tab + text emb.)	45.1±0.5
Early Fusion [3]	38.4±0.5
MMBT [20]	39.6±0.4
MulT [11]	40.8±0.4
Perceiver IO [21]	41.3±0.4
I^2MoE [13]	42.7±0.4
AutoMM [23]	47.2±0.3
CRAFT (tab + text)	**50.1**±0.3

perturbation-based importance and self-sufficient feature analysis, rather than as a standalone causal claim. Extending CMIS with gradient-based attribution or formal do-calculus interventions remains future work.

6 Conclusion

We presented CRAFT, a fusion framework addressing cardinality heterogeneity through density-based backbone selection, class-frequency-adaptive SupCMA, and counterfactual CMIS. CRAFT outperforms 12 baselines while being 1.9× faster, handles arbitrary modality subsets from a single model, and delivers 6.4% cost reduction ($6.51M annually) in production. Its modular design makes it directly applicable to other heterogeneous-cardinality settings, with new modalities added as gated branches without architectural changes.

References

1. Zihan Yu, Liang Zeng, Yifan Hao, and Jianguo Chen. FuseMoE: Mixture-of-experts transformers for fleximodal fusion. *arXiv preprint arXiv:2402.03226*, 2024.
2. Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. LIFT: Language-interfaced fine-tuning for non-language machine learning tasks. In *NeurIPS*, 2022.
3. Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2019.
4. Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.
5. Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. In *AISTATS*, 2023.
6. Vinitra Swamy, Jibril Frej, and Tanja Käser. InterpretCC: Intrinsic user-centric interpretability through global mixture of experts. *arXiv:2402.02933*, 2024.
7. Thomas Bonnier. Revisiting multimodal transformers for tabular data with text fields. In *Findings of ACL*, pp. 1481–1500, 2024.

8. Zhun Liu, Ying Shen, Varun B. Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, 2018.
9. Gagan Sharma, R Chinmay, and Raksha Sharma. Late fusion of transformers for sentiment analysis of code-switched data. In *Findings of EMNLP*, pp. 6485–6490, 2023.
10. Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv:2412.13663*, 2024.
11. Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pp. 6558–6569, 2019.
12. Jiahui Wang, Hao Wang, and Jian Pei. UniPredict: Large language models are universal tabular classifiers. *arXiv:2310.03266*, 2023.
13. Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna Ballard, Tianlong Chen, and Qi Long. I^2MoE: Interpretable multimodal interaction-aware mixture of experts. In *ICML*, 2025.
14. Pengcheng Yin and Graham Neubig. TaBERT: Pretraining for joint understanding of textual and tabular data. In *ACL*, pp. 8413–8426, 2020.
15. Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *NeurIPS*, vol. 34, pp. 18932–18943, 2021.
16. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.
17. Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, et al. Supervised contrastive learning. In *NeurIPS*, vol. 33, pp. 18661–18673, 2020.
18. Inside Airbnb. Melbourne Listings Dataset. http://insideairbnb.com/, 2023.
19. Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alexander Smola. Multimodal AutoML on structured tables with text fields. In *NeurIPS Datasets and Benchmarks Track*, 2021.
20. Douwe Kiela, Suvrat Bhatt, Remi Cadene, Davide Testuggine, and Andrej Karpathy. Supervised multimodal bitransformers for classifying images and text. *arXiv:1909.02950*, 2019.
21. Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, et al. Perceiver IO: A general architecture for structured inputs & outputs. In *ICML*, 2022.
22. Heqing Yang, Xin Ju, Wentao Zhu, and Junhao Chen. UniS-MMC: Multimodal classification via unimodality-supervised multimodal contrastive learning. In *Findings of ACL*, 2023.
23. Zhiqiang Tang, Haoyang Dong, Zihan Bai, et al. AutoGluon-Multimodal (AutoMM): Supercharging multimodal AutoML with foundation models. In *AutoML Workshop at ICML*, 2023.
24. Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, vol. 30, 2017.
25. Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *NeurIPS*, vol. 32, 2019.

GenAI Usage Disclosure

The authors used generative AI tools for language editing and brainstorming. All technical claims, derivations, experiments, code, and analyses were authored and verified by the authors. No generated text was used without author review.