Research on Engineering Structures & Materials
www.jresm.org
Research Group

Research Article

Digital-twin-supported quality-aware multimodal fusion for damage diagnosis of identically detailed reinforced concrete frames across concrete batches and loading paths

Fen Ye [1,a], Jing Liu [*,2,b]

[1]*School of Architecture, Road and Bridge Engineering, Guangxi Eco-engineering Vocational & Technical College, Liu Zhou, China*
[2]*School of Civil Engineering and Architecture, Guangxi University of Science and Technology, Liu Zhou, China*

Article Info

Article History:

Received 13 Aug 2026
Accepted 08 Oct 2026

Keywords:

Structural health monitoring;
Digital twin;
Multimodal data fusion;
Reinforced concrete frame

Abstract

Reinforced concrete frame structures accumulate crack growth, local spalling and irreversible deformation under cyclic loading, and manual inspection alone stays sensitive to viewing angle and recording completeness. This study proposes a building information modeling (BIM) component-level vision-language-sensor fusion model supported by a digital twin, evaluated on three one-third scale two-story two-bay frames that share geometry and reinforcement but differ in concrete batch and damage path. Images, inspection text, acceleration and strain are aligned to ten beam-column nodes through component identifiers, and a low-parameter quality-aware neuro-fuzzy gate assigns sample-level modal weights from computable quality indicators. In-fold finite element domain-randomized pretraining, component topology propagation and a training-phase physical direction constraint form the digital twin support, with strict outer isolation of every test specimen. Under a leave-one-specimen-out protocol across F1, F2 and F3, in which the held-out frame differs from the training frames only in concrete batch and loading path, the quality-aware gate keeps higher relative performance than fixed-weight fusion under image occlusion, sensor degradation and modality absence. Ablation shows that finite element pretraining, component topology and the direction constraint carry distinct diagnostic roles within a three-specimen feasibility range.

© 2026 MIM Research Group. All rights reserved.

1. Introduction

Reinforced concrete building structures may develop crack propagation, local spalling and irreversible deformation under cyclic loading, material ageing and environmental action, and inspection that relies only on human patrol is easily influenced by observation angle, recording completeness and differences in field experience. In engineering practice such damage is judged against performance levels, namely Immediate Occupancy, Life Safety and Collapse Prevention in ASCE/SEI 41-17 [1] and Damage Limitation, Significant Damage and Near Collapse in Eurocode 8 Part 3 [2], so a data-driven damage grade is useful only when it can be read against these levels. A digital twin has gradually shifted from a static digital model toward a dynamic mapping between the physical entity and its virtual counterpart, and building information modeling can serve as a carrier of component semantics, spatial topology and damage information [3]. Construction-oriented research has outlined how a semantic digital twin should connect design, monitoring and maintenance information across the asset life cycle [4], while broad reviews of digital twin applications in the construction industry report that most implementations still separate the physical model, the real-time data stream and the diagnostic model rather than fusing them [5].

*Corresponding author: jingliu5495@outlook.com
DOI: http://dx.doi.org/10.17515/resm2026-2179vk0813rs

Studies that couple a building information model with a real-scene three dimensional model have improved the geometric fidelity of structural health tracking [6], yet the topology stored in the model is rarely allowed to participate directly in multimodal reasoning.

Intelligent digital twin frameworks that combine wireless sensing, digital signal processing and structural analysis have brought real-time response data closer to the diagnostic loop [7], and model-data integrated frameworks now support full-field virtual sensing by blending a calibrated finite element model with measured signals [8]. Post-earthquake evaluation using unmanned aerial imagery within a model-based twin illustrates how visual evidence and a structural model can be organized around the same asset [9], and deep-learning-based monitoring has become a mature route for extracting damage-sensitive patterns from raw measurements [10]. General surveys of enabling technologies emphasize that data quality, model updating and open validation remain unresolved obstacles for engineering twins [11], and a paradigmatic move from digital shadows toward genuine digital twins has been argued for advanced rail transit monitoring [12]. Across these efforts the finite element model and the component topology are frequently used for visualization of results, and comparatively little work lets the physical model and the topological graph act as active constraints inside a component-level diagnostic model that must generalize to unseen physical specimens.

Visual recognition can extract crack orientation, distribution and spalling as surface features, and a review of computer-vision monitoring at local and global scales has mapped its strengths and its blind spots under occlusion and reflection [13]. Combining deep learning with finite element model updating has improved the physical plausibility of detected damage [14], and unsupervised anomaly detection has been used to localize damage without dense labels [15]. Acceleration and strain monitoring perceive dynamic response and local force redistribution, and a physics-augmented spatial-temporal graph network has localized damage from guided-wave records [16], while optimized stacked autoencoders with multi-sensor fusion have diagnosed cracking in concrete jack arch beams [17]. Multichannel fusion with convolutional networks has been compared systematically for frame structures [18], decision-level fusion of several vibration signals has raised robustness [19], and a combined one dimensional convolution and long short-term memory model has exploited multi-sensor time series [20]. A recurring limit is that different damage combinations can produce similar dynamic responses, so response data alone carries a non-uniqueness that surface images cannot always resolve.

Inspection text records component position and the change trend relative to the previous visit, and a large language model has been applied to post-earthquake structural damage assessment [21]. Multimodal chain-of-thought reasoning has been coupled with large models for damage identification [22], natural language processing has turned bridge inspection reports into machine-readable condition indicators [23], and named entity recognition with domain pretraining has extracted defect entities from Chinese bridge records [24]. Component-level defect data can then be managed inside a model-based maintenance system [25], although text stays qualitative, sparse and uncertain, which makes it more suitable as weak evidence than as a direct label. Four problems remain insufficiently solved. Vision, language and sensor data lack a unified component-level semantics. Fixed-weight fusion tends to suffer negative transfer when a modality degrades. Random splits inside one specimen can rewrite re-identification of seven states as generalization. Finite element pretraining that touches test-specimen material or healthy-state data can leak information. This study addresses these gaps on three frames with identical configuration but different concrete batches and damage paths, and it limits every conclusion to a cross-batch and cross-path feasibility range within one two-story two-bay configuration. In this paper the term cross-specimen therefore denotes transfer to a physically separate frame with the same geometry, sections and reinforcement that differs only in concrete batch and loading path, and it does not cover changes in bay number, story height or reinforcement ratio. Within this scope the work concentrates on three propositions. The first proposition examines whether the in-fold finite element pretraining, the building information modeling component topology and the training-phase direction constraint separately affect cross-batch adaptation, component localization and physical consistency. The second proposition examines whether a low-parameter quality-aware fuzzy gate keeps a higher relative performance than fixed-weight fusion once images are occluded,

sensors degrade or a modality is absent on unseen physical specimens. The third proposition examines whether the sensor and the language modality supply effective evidence when surface information is insufficient and when vision and sensors degrade at the same time. Establishing a component-level semantic anchor for the three modalities and keeping every test specimen strictly outside the pretraining and the hyper-parameter selection together form the methodological core of the study.

2. Materials and Methods

2.1 Scaled Specimens, Loading Regime and Damage Truth

The research runs from September 2024 to May 2026. The early stage completed the design of three specimens, material property testing, building information modeling and sensor calibration. The middle stage completed graded loading and multimodal acquisition, and the late stage completed label review, model training and cross-specimen validation. The three specimens are plane reinforced concrete frames at about one-third scale, each a two-story two-bay frame with a total width near 3.0 m and a total height near 2.2 m, a beam section near 120 mm by 180 mm and a column section near 150 mm by 150 mm. Geometry, reinforcement and measurement positions stay identical, while the concrete batch differs. F1 uses a relatively symmetric horizontal cyclic loading, F2 applies an added vertical load on the left bay before loading, and F3 applies an added load on the right bay and changes the cyclic order, so the three specimens form a balanced, a left-concentrated and a right-concentrated damage path. The measured cube compressive strength of the three batches ranged from about 32 MPa to 38 MPa and the measured elastic modulus varied by close to 12 percent, and this batch difference is the main source of the cross-specimen response shift that the model must tolerate. The longitudinal reinforcement ratio of the columns stayed near 1.8 percent and that of the beams near 1.2 percent, and the stirrup spacing was kept identical so that the difference between specimens came from the concrete batch and the loading path rather than from the detailing.

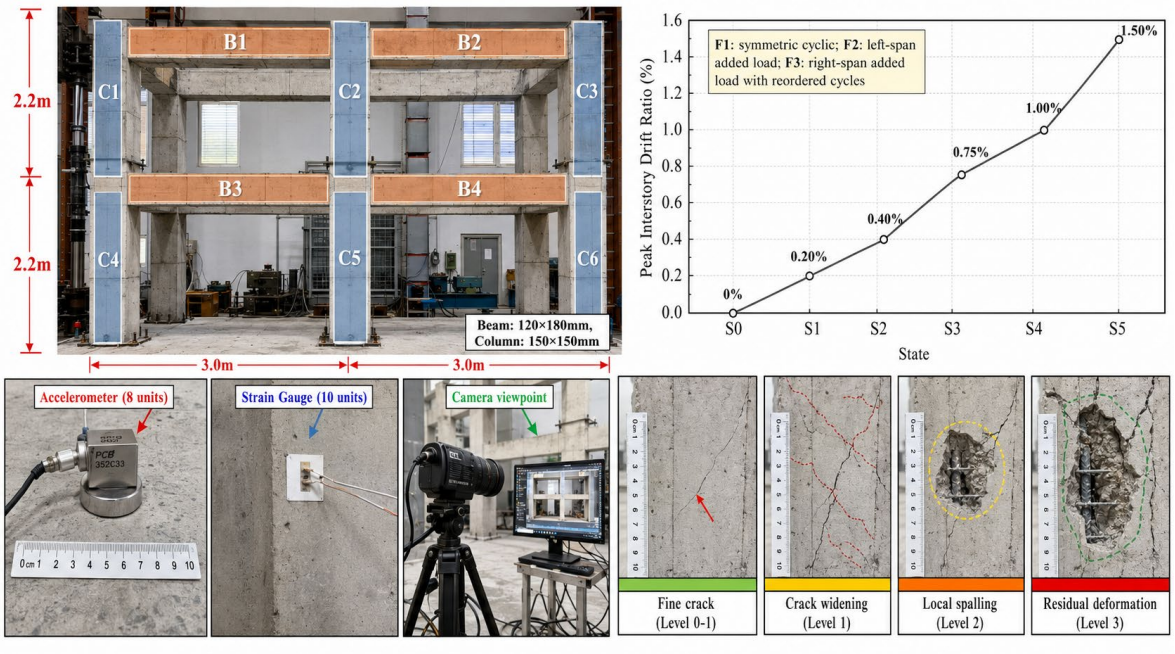


Fig. 1. Test system, sensor layout and damage evolution

The frames use prototype materials, so the length scale is 1/3 and the modulus and stress scales are 1, which satisfies stress similitude for the quasi-static cyclic loading. No artificial mass was added for Froude-type dynamic similitude, which would have required a density scale of 3, or about 14 kN of supplementary mass. The resulting scales are therefore 1/27 for mass, 1/3 for stiffness and time, and 3 for frequency and acceleration. The added vertical loads on F2 and F3 (about 6.0 kN of steel weights) shape the damage path and enter the finite element model only as added mass.

The measured frequencies are model-scale quantities used only as changes relative to each specimen's own healthy state and are not extrapolated to full scale.

Table 1. Quantitative thresholds of the three labeling indicators

Level	Max crack width (mm)	Spalling area (cm^2)	Residual chord rotation (rad)	Indicative performance level
0	< 0.05	0	< 0.001	Undamaged
1	0.05–0.20	< 5	0.001–0.003	IO, before DL
2	0.20–0.80	5–50	0.003–0.006	Between IO and LS, beyond DL
3	> 0.80	> 50 or exposed bar	> 0.006	Approaching LS and SD

Each frame contains six column components and four beam components, which give ten building information modeling diagnostic units. The six acquisition points S0 to S5 correspond to the undamaged state and to peak interstory drift ratios of about 0.20%, 0.40%, 0.75%, 1.00% and 1.50%, as shown in Figure 1. The state index is not a model input and does not directly equal a damage level. Component damage truth is fixed jointly by three measurements that stay outside the diagnostic input. Near-field measurement with a scale bar gives the maximum crack width, three-dimensional surface scanning gives the spalling area, and the unloaded relative displacement of two marked points on a fixed gauge length gives the residual chord rotation. Each indicator is divided into levels 0 to 3, and the final level uses a joint decision that follows the multi-damage identification practice of prior frame studies [26] and the time-frequency severity classification of acceleration data [27] as a methodological basis rather than a transplant of their thresholds. Each indicator is assigned a sub-level by the thresholds in Table 1. Level 0 requires all three sub-levels to be 0, level 3 requires at least two sub-levels of 3, and otherwise the final level is the median of the three sub-levels. A sample whose sub-levels differ by two or more grades is marked as conflicting. It receives the rounded mean of its sub-levels as an intermediate level and keeps an uncertainty flag, which prevents any single visual indicator from directly deciding the final label. For engineering interpretation, the four grades are read against these performance levels. Level 0 corresponds to an undamaged, fully operational state. Level 1, with hairline cracking and no spalling, falls within Immediate Occupancy and before Damage Limitation. Level 2, with widened cracks and local cover spalling but small residual rotation, lies between Immediate Occupancy and Life Safety, beyond Damage Limitation but short of Significant Damage. Level 3, with extensive spalling, exposed reinforcement or clear residual rotation, approaches Life Safety and Significant Damage. Because the peak drift ratio stayed at 1.50% and no frame lost its vertical load-carrying capacity, Collapse Prevention and Near Collapse were not reached and no grade represents them. This correspondence is indicative at the component level, since the grades rest on residual indicators of one-third scale members rather than on the acceptance criteria of a full-scale building.

Table 2. Test states and component-level label distribution

State	Drift ratio	Level 0	Level 1	Level 2	Level 3	Total
S0	0	30	0	0	0	30
S1	0.20%	24	6	0	0	30
S2	0.40%	17	11	2	0	30
S3	0.75%	10	14	5	1	30
S4	1.00%	5	12	9	4	30
S5	1.50%	2	8	12	8	30
Total	-	88	51	28	13	180

Among the 180 component-states the joint decision produced 166 consistent labels and 14 conflicting samples. 10 of the conflicting samples occurred at S3 and S4, and they received 9 level-1 and 5 level-2 labels. The two annotators agreed independently on 12 of them, and the remaining 2 were settled by a third senior reviewer. The same drift ratio produced different grades on adjacent columns, which confirms that the state index and the damage level stay distinct quantities. The state system, sensor layout and representative damage evolution appear in Figure 1, and the resulting component-state label distribution is listed in Table 2.

2.2 Multimodal Acquisition, Text Design and Sample Organization

Each frame carries eight uniaxial acceleration points and ten component-end strain points. The acceleration points cover the two stories and the left and right regions, and the strain points correspond to the ten-building information modeling component nodes. The scale-bar images, the three-dimensional scan and the residual displacement used for labeling do not enter the model. Acceleration and strain are sampled synchronously at 200 Hz, and every observed state runs six independent low-amplitude impact or ambient excitations that each last 45 s. The three specimens therefore form 108 excitation events, each record holds 9000 sampling instants over 18 sensing channels, and the total reaches about 17.496 million raw scalars. The signal passes through zero-point correction, spike detection, 0.5 to 50 Hz band-pass filtering and in-fold standardization, and it is then split by a 4 s window with 50% overlap into 21 windows, which gives 2268 sensor windows in total. The acceleration channels supply the floor-level and region-level dynamic response, the strain channels supply the local deformation at the ten component ends, and the two sensor types are kept aligned in time so that a window carries a consistent global and local view of the same excitation. The quality indicators of every window, namely the signal-to-noise ratio, the outlier ratio and the continuous packet-loss ratio, are computed on the same window so that the fuzzy gate can later read a physically meaningful quality value for each modality. The design of these indicators extends multimodal sensor anomaly diagnosis for structural monitoring toward a joint vision, language and sensor setting [28] rather than staying at sensor channel fusion alone. A synchronization checks across the eight acceleration channels kept the inter-channel time offset below one sample, which matters because a misaligned window would otherwise inflate the apparent response change and mislead the sensor branch.

Table 3. Multimodal data scale, input form and quality indicators

Modality	Data scale	Model input	Automatic quality indicators	Anti-leakage handling
Vision	1080 component-events, 3 views each, about 3240 images	512x512 input, cropped to 448x448 for training	Sharpness, exposure bias, occlusion ratio, target completeness	No manual crack width or level
Language	1080 component-event texts, 180 unique, 25 to 60 words each	Component position, qualitative form, trend and inspection confidence	Effective length, reference completeness, trend coverage	Removes width, spall area, rotation and level words
Sensor	108 excitation events, 8 acceleration and 10 strain channels	9000 points per channel at 200 Hz, split into 21 windows	Signal-to-noise ratio, packet-loss ratio, saturation, drift, sync bias	Standardization and filtering fixed by training fold only
BIM	3 specimens, 30 physical components	Component type, position, floor, bay and adjacency matrix	Mapping completeness, ID conflict, adjacency consistency	Writes no damage label or test-specimen statistics

Each excitation event collects three ordinary RGB views for every component, namely a global, a close and a lateral view, which gives about 3240 diagnostic images, and the model images keep no measurement scale bar. Every event also forms ten short component inspection texts, about 1080 in total. The text carries no crack width value, no spalling area, no residual rotation, no loading state and no damage level, but it may record trend phrases such as a newly appeared diagonal crack, some growth relative to the previous visit, no obvious change, or a locally occluded region. The vision branch receives only the current-event images and no historical images, so the trend text carries temporal information that neither the current images nor the current sensor windows provide explicitly. Modalities are linked by specimen number, acquisition time and component identifier, and the 21 sensor windows of one event share that event image and text. Evaluation first

aggregates the 21 window predictions to the event level and then reports the ten component outputs. The data scale is stated as 1080 component-events, 180 component-states and 30 unique physical components, so repeated excitations and overlapping windows are not treated as independent samples. The 180 unique component-states arise because ten components observed at six states across three specimens give the deduplicated count, and the 1080 component-events arise because each component-state repeats over six excitations. This separation keeps the reporting honest, since a large window count reflects data density rather than a large number of independent structures. The inspection texts were written by two trained annotators who followed a fixed phrase template, and a small share of deliberately vague or partly wrong phrases was kept so that the language quality indicator would face realistic noise rather than an idealized clean text. The scale, input form and quality indicators of every modality are listed in Table 3.

2.3 Digital-Twin-Supported Low-Parameter Multimodal Model

The digital twin comprises the building information modeling component model, a ten-node beam-column connection graph, a training-only finite element model and a database of past diagnostic states, which follows the structural digital twin formulation that integrates physical entity, monitoring data, finite element model and diagnostic algorithm into one system [29] while adding the four-layer support described below. The vision branch uses MobileNetV3-Large pretrained on ImageNet-1k, and the three views of each component are encoded separately and then merged by view attention. To control overfitting on a small sample, the first 13 of 16 bottleneck blocks stay frozen and only the last stage and a 64-dimensional projection layer are trained. The language branch uses RBT3, a three-layer Chinese RoBERTa-wwm-ext encoder, with a frozen main body, and only the projection layers for component semantics, damage description and trend information are trained. The sensor branch first encodes the eight acceleration channels and the ten strain channels, where acceleration forms floor-level and region-level global response features and strain forms local features aligned with the component nodes, and the two are merged by multi-scale one dimensional convolution and a small temporal aggregation network whose low layers stay frozen after finite element pretraining (Table 4). The three modalities map to the building information modeling component nodes, and a two-layer low-dimensional graph attention network propagates information along the shared beam-column nodes, as shown in Figure 2.

Table 4. Branch configuration and trainable parameters

Branch	Layer	Channels / dim	Kernel / stride	Status
Vision	MobileNetV3-Large blocks 1–13	3→112	—	Frozen
Vision	Last stage (3 bottlenecks + 1×1 conv)	112→960	5×5 dw / 1	Trainable
Vision	Projection + view attention	960→64	—	Trainable
Language	RBT3 encoder	768	—	Frozen
Language	Three projections	768→256→32 each	—	Trainable
Acceleration	Conv1D A1, A2	8→32→64	7/2, 5/2	Frozen
Acceleration	Multi-scale Conv1D A3	64→3×32	3, 5, 9 / 1	Trainable
Acceleration	GRU-A	96→64	—	Trainable
Strain (shared by 10 nodes)	Conv1D S1, S2	1→16→32	7/2, 5/2	Frozen
Strain	Multi-scale Conv1D S3	32→3×16	3, 5, 9 / 1	Trainable
Strain	GRU-S	48→64	—	Trainable

Note: The sensor input is one 4 s window of 800 samples. The trainable parameters are about 2.3 M for vision, 0.62 M for language, 0.10 M for sensor and 0.08 M for alignment, graph, heads and gate, about 3.1 M in total. In the sensor branch, 4 of 8 layers and about 13% of the parameters are frozen.

Each branch outputs a 128-dimensional feature together with a modal evidence vector and an automatic quality score, and the graph propagation lets a reliable neighbor node correct a

component whose own modality has degraded, which is the main reason the topology is treated as an active constraint rather than a display layer. The membership breakpoints of the three quality levels are set from the empirical quantiles of the training fold, so a medium image on one-fold is not forced to match a medium image on another fold, and this keeps the gate calibrated to the actual quality range seen during training. Each raw quality indicator is min-max normalized with bounds from the training fold and oriented so that a larger value means better quality, and the modal quality q_m is their mean. The breakpoints are the 10th, 50th and 90th percentiles (Q10, Q50, Q90) of q_m over the training-fold samples, namely 1512 sensor windows and 720 component-events for vision and language. The low membership equals 1 below Q10 and falls linearly to 0 at Q50, the medium membership is triangular over Q10, Q50 and Q90, and the high membership rises linearly from 0 at Q50 to 1 at Q90. In the inner loop the bounds and quantiles come only from the inner training specimen. After hyper-parameter selection they are recomputed on the two outer training specimens and frozen, and the test specimen is mapped with these frozen values, with out-of-range scores clipped to [0, 1]. Across the three folds the vision breakpoints shifted by at most 0.04. The conflict coefficients are constrained so that lowering a quality score never raises the corresponding weight, which is the monotone condition that gives the gate its predictable behavior under degradation. The entire gate contains only a few dozen trainable parameters, which is negligible compared to the encoders and helps the gate stay stable when the training set is limited to two specimens per fold.

The quality-aware fuzzy gate computes separate quality scores for vision, sensor and language. Vision quality combines sharpness, over-exposure ratio and occlusion ratio, sensor quality combines estimated signal-to-noise ratio, outlier ratio and continuous packet-loss ratio, and language quality combines effective length, reference completeness and trend coverage. Each modality holds low, medium and high membership functions of fixed shape whose breakpoints are set only by the current training fold, and the gate learns no more than a set of cross-modal conflict coefficients that satisfy a monotone condition, so a weight does not increase when quality drops. The gate uses the nine shared rules of Table 5 and re-normalizes the weights when a modality is absent. The forward pass for component j in sensor window w is formulated as follows. Each of the three views v is encoded and the views are merged by attention:

$$\left(h_V^{(v)} = P_V\left(E_V\left(X_V^{(v)}\right)\right), \quad a_v = softmax_v\left(\omega^T h_V^{(v)}\right), \quad h_V = \sum_v a_v\, h_V^{(v)} \in \mathbb{R}^{64}\right) \tag{1}$$

The text is encoded by three projections for component semantics, damage description and trend, and the sensor feature joins the floor- and region-level acceleration feature with the local strain feature of node j:

$$\begin{aligned} h_L &= [P_c(E_L(X_L)); P_d(E_L(X_L)); P_t(E_L(X_L))] \in \mathbb{R}^{96} \\ h_{S,j}^{(w)} &= [G_A(X_A^{(w)})_{f(j),r(j)}; G_\varepsilon(X_{\varepsilon,j}^{(w)})] \in \mathbb{R}^{128} \end{aligned} \tag{2}$$

A modality-specific layer aligns each branch to a shared 128-dimensional space indexed by the component identifier:

$$Z_{m,j} = LN(W_m h_{m,j} + b_m) \in \mathbb{R}^{128}, \quad m \in V, L, S \tag{3}$$

Rule r fires with strength τ_r from the quality of modality m and the best quality among the other available modalities. The gate then forms a zero-order Takagi-Sugeno output, which is masked by the availability indicator $\delta_{m,j}$ and normalized:

$$\tau_r = \mu_{A_r}(q_{m,j}) \cdot \mu_{B_r}(\bar{q} - m, j), \bar{q} - m, \tag{4}$$

$$j = max_{k\neq m,\ \delta_{k,j}=1}\, q_{k,j} w_{m,j} = \beta_m \frac{\sum_{r=1}^{9} \tau_r\, c_r^m}{\sum_{r=1}^{9} \tau_r} W_{m,j} = \frac{\delta_{m,j} w_{m,j}}{\sum_k \delta_{k,j}\, w_{k,j}} z_j^{(0)}$$
$$= \sum_m W_{m,j}\, Z_{m,j}$$

Here c_r^m are the conflict coefficients and β_m is a modality prior, giving 30 trainable parameters. The coefficients are constrained to be non-decreasing in the own-quality level and non-increasing in the other-quality level. $W_{L,j}$ is capped at $\kappa = 0.5$, and any excess is redistributed to vision and sensor in proportion to their weights.

Two graph attention layers propagate the fused features over the neighbors N(j) given by the BIM adjacency matrix:

$$\begin{aligned} \alpha_{ij} &= \underset{i\in\mathcal{N}(j)\cup\{j\}}{softmax}\, LeakyReLU(\boldsymbol{a}^\top[\Theta \boldsymbol{z}_j^{(l)} \| \Theta \boldsymbol{z}_i^{(l)}]) \\ \boldsymbol{z}_j^{(l+1)} &= ELU(\sum_i \alpha_{ij}\, \Theta \boldsymbol{z}_i^{(l)}), \quad l = 0,1 \end{aligned} \tag{5}$$

A detection head and a cumulative ordinal head give the outputs, with $b_0 < b_1 < b_2$:

$$\begin{aligned} p_j^{det} &= \sigma(\boldsymbol{u}_d^\top \boldsymbol{z}_j^{(2)}) \\ P(y_j > k) &= \sigma(\boldsymbol{u}_o^\top \boldsymbol{z}_j^{(2)} - b_k) \\ p_{jk} &= P(y_j > k-1) - P(y_j > k) \end{aligned} \tag{6}$$

The detection and ordinal losses are

$$\begin{aligned} \mathcal{L}_{det} &= -\frac{1}{N}\sum_j [\, y_j^{det} log\, p_j^{det} + (1 - y_j^{det})\, log(\,1 - p_j^{det})] \\ \mathcal{L}_{ord} &= -\frac{1}{N}\sum_j \sum_{k=0}^{2} [\, \mathbb{I}(y_j > k)\, log\, P\,(y_j > k) + \mathbb{I}(y_j \le k)\, log(\,1 - P(y_j > k))] \end{aligned} \tag{7}$$

and the window predictions are averaged to the event level,

$$\bar{p}_j = \frac{1}{21}\sum_w p_j^{(w)} \tag{8}$$

Z_V and Z_L are shared by the 21 windows of an event.

Table 5. Rule base of the quality-aware gate (consequents c_r^m, mean over three folds)

Rule	Own quality q_m	Best other quality $\bar{q}_{-m}$	c^V	c^S	c^L
R1	Low	Low	0.45	0.42	0.30
R2	Low	Medium	0.25	0.22	0.15
R3	Low	High	0.10	0.08	0.05
R4	Medium	Low	0.70	0.66	0.45
R5	Medium	Medium	0.50	0.48	0.30
R6	Medium	High	0.32	0.30	0.18
R7	High	Low	0.92	0.88	0.60
R8	High	Medium	0.80	0.76	0.48
R9	High	High	0.65	0.60	0.35

Every leave-one-specimen outer fold rebuilds the training-only finite element model, which may use only the design parameters, measured material values and training-state responses of the two training specimens for calibration. Domain randomization over material modulus, boundary stiffness, added mass, excitation position and a 5% to 40% component stiffness reduction generate

about 2400 sensor pretraining samples per fold, and the test-specimen material data, healthy-state frequency and any response are excluded.

The finite element model was built in ABAQUS with C3D8R concrete elements and T3D2 truss reinforcement embedded in the concrete. The embedding assumes perfect bond and neglects bond slip, which is acceptable for the low-amplitude responses used here. The mesh size is 25 mm and is refined to 15 mm within 150 mm of each joint, giving about 2.6×10^4 elements. Concrete uses the Concrete Damaged Plasticity model with the measured modulus and strength of the training specimens, a dilation angle of 30°, an eccentricity of 0.1, fb0/fc0 of 1.16, Kc of 0.667 and a viscosity of 0.0005. The damage variables dc and dt are derived from the uniaxial compressive and tensile curves. Reinforcement is bilinear elastic-plastic with von Mises yielding, a measured yield strength of 420 MPa and a hardening modulus of 0.01Es. The column bases are tied to the foundation beam, which is supported by translational and rotational springs whose stiffness is randomized. A quasi-static cyclic analysis of each training frame reproduces its training-state damage and sets the 5% to 40% stiffness reduction of the end zones. The synthetic responses are then obtained by frequency extraction and modal dynamic analysis with 3% Rayleigh damping, sampled at 200 Hz and windowed as the measured data. The calibrated first-mode frequency differs from that of the healthy training frames by less than 4%.

The model is optimized jointly by the detection and ordinal severity losses of Eq. (7) and a training-phase direction consistency loss. The direction constraint only compares whether the frequency and floor-level mode-shape change directions implied by the prediction clearly contradict the training data, and it does not act as a local damage truth. The predicted grade probabilities pjk of component j are first converted into an expected stiffness reduction, which is mapped to frequency and mode-shape changes through sensitivity matrices Sf and Sφ. These matrices are obtained by finite differences of the in-fold finite element model:

$$\begin{aligned} \hat{r}_j &= \sum_{k=0}^{3} p_{jk}\,\rho_k \\ \Delta\hat{f} &= S_f\hat{r}, \quad \delta\hat{\varphi} = S_\varphi\hat{r} \end{aligned} \tag{9}$$

where ρ = [0, 0.05, 0.15, 0.30] is the reduction assigned to levels 0–3. The measured changes Δf_m and $\delta\varphi_m$ of mode m are taken relative to the S0 state of the same training specimen. Modes are paired between states by MAC > 0.90, and a change is counted only when it exceeds the thresholds τ_f = 0.5% and τ_φ = 0.02:

$$\begin{aligned} \mathcal{L}_{dir} = \frac{1}{M}\sum_{m}[\,&\mathbb{I}(|\Delta f_m| > \tau_f)\,ReLU(-sgn(\Delta f_m)\Delta\hat{f}_m) + \beta\mathbb{I}(\|\delta\varphi_m\| \\ &> \tau_\varphi)\,ReLU(-\cos(\delta\hat{\varphi}_m, \delta\varphi_m))] \end{aligned} \tag{10}$$

Here M = 2 floor-level modes and β = 0.5. The penalty is therefore non-zero only when the predicted change has the opposite sign of a clear measured change. It enters the total loss as

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{det} + \lambda_{ord}\mathcal{L}_{ord} + \lambda_{dir}(e)\mathcal{L}_{dir} \\ \lambda_{dir}(e) &= \lambda_{dir}\cdot min(1, e/E_w) \end{aligned} \tag{11}$$

and its gradient reaches the network through the chain rule:

$$\frac{\partial\mathcal{L}_{dir}}{\partial p_{jk}} = \rho_k\sum_m\left(\frac{\partial\mathcal{L}_{dir}}{\partial\Delta\hat{f}_m}S_{f,mj} + \frac{\partial\mathcal{L}_{dir}}{\partial\delta\hat{\varphi}_m}S_{\varphi,mj}\right) \tag{12}$$

The term is used only in training. A prediction whose L_{dir} is positive is counted as physically contradictory. The detection weight is fixed at 1 as the reference. λ_{ord} and λ_{dir} are chosen by grouped inner validation from [0.5, 1, 2] and [0.05, 0.1, 0.2, 0.5], which gives λ_{ord} = 1.0 and λ_{dir} = 0.2 in all three folds. The direction weight is ramped over the first E_w = 10 epochs so that the heads learn the grades before the physical penalty acts. Across λ_{dir} from 0.1 to 0.5, the inner-validation macro-F1 changed by less than 0.02.

Training has two stages. First, the sensor branch is pretrained on the in-fold finite element samples for 30 epochs, and its low layers are then frozen. Second, the whole model is trained with AdamW, using a learning rate of 1×10^{-3} for the heads, gate and graph layers and 1×10^{-4} for the unfrozen encoder blocks, a weight decay of 1×10^{-4}, a batch of 16 component-events and a cosine schedule for at most 80 epochs. Early stopping on the inner-validation macro-F1 uses a patience of 10 epochs. After each step the gate conflict coefficients are projected back onto the monotone feasible set, which keeps the gate monotone throughout training. At test time the model first outputs damage probability, level, modal weight and uncertainty, and it writes back to the model only after evaluation. The ordinal severity loss respects the natural order of the four grades so that a level-3 unit predicted as level 2 is penalized less than one predicted as level 0, which matches the engineering cost of the two errors. The uncertainty output combines the spread of the modal evidence with the gate confidence, so a component that relies on a single degraded modality receives a wider uncertainty band, and this band later drives the transparency of the building information modeling write-back rather than a purely decorative color.

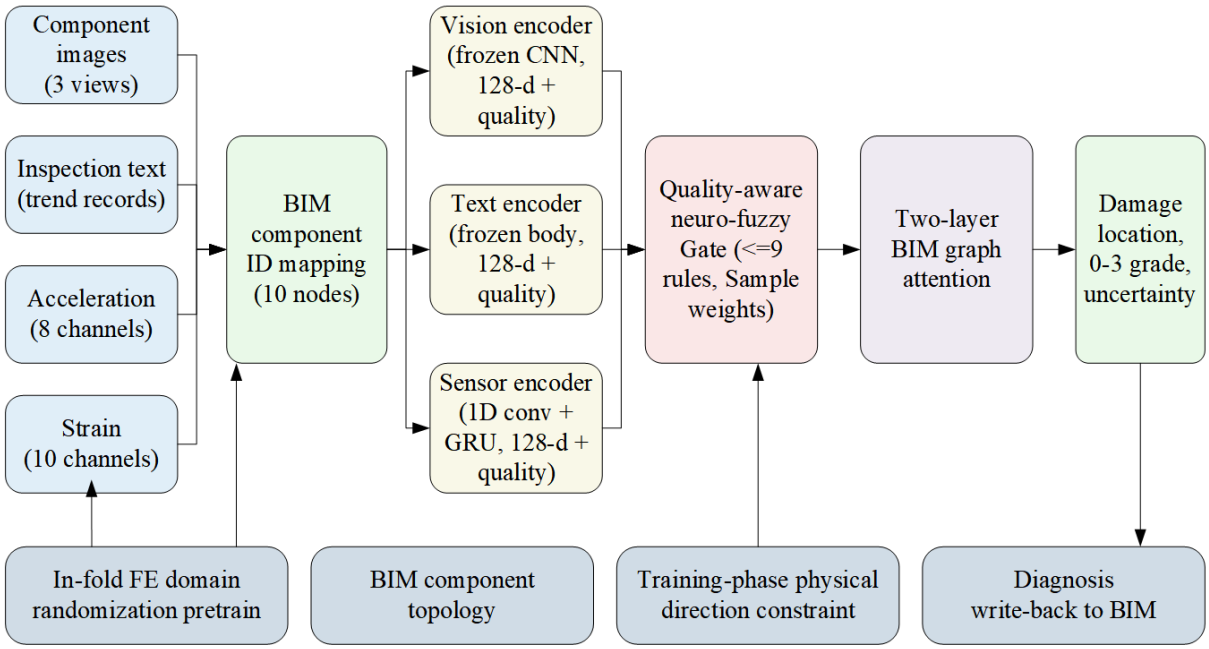


Fig. 2. BIM component-level alignment and the digital-twin-supported multimodal diagnostic model

2.4 Cross-Specimen Protocol, Compact Controls and Degradation Setting

The main evaluation uses a leave-one-specimen protocol, so each round trains on two specimens and tests one whole specimen, which forms the three outer folds F1, F2 and F3, as shown in Figure 3. The outer training set holds 72 excitation events, 720 component-events and 120 component-states, and the test set holds 36 events, 360 component-events and 60 component-states. Hyper-parameters are chosen by grouped inner validation between the two training specimens, and the preprocessing parameters and the finite element pretraining data are rebuilt within the inner training range, which follows the domain adaptation view that any test-specimen information must stay outside calibration and model selection [30]. A same-specimen event split of the full model and of the fixed-weight model is also run as an in-distribution reference of the full model itself and is not called a generalization upper bound. Beyond the main task the protocol also defines a within-specimen reference, an S1 to S4 leave-state grading with S5 anomaly localization, and a semi-supervised recovery of unlabeled component positions, and S5 reports no AUC because it holds only two healthy component-states.

The modality combinations keep only the vision, sensor and language single modalities, the vision-sensor pair and the vision-language-sensor triple. The fusion comparison covers fixed-weight fusion, one published multi-view uncertainty fusion method, a gated multimodal unit (GMU) [31], a two-layer cross-attention Transformer fusion [32] and the present fuzzy gate. The GMU and

Transformer share the encoders, alignment layers, graph attention and heads of the full model, so only the fusion layer differs, and one published sensor spatial-temporal graph method is chosen to compare with the present sensor branch. The digital twin ablation includes only the removal of in-fold finite element pretraining, the removal of the building information modeling component graph and the removal of the direction constraint. The degradation study is limited to 25% and 50% image occlusion, 20 dB sensor noise, 20% continuous packet loss, the absence of the language modality, and a dual degradation that combines 25% occlusion with 20% loss. The evaluation indicators include the macro-averaged F1 value, the component localization F1 value, the balanced accuracy, the per-grade recall, the grade mean absolute error, the confidence calibration error and the performance retention ratio relative to clean input. Every method computes its retention ratio against its own clean baseline to avoid an unfair comparison caused by different clean starting points. The statistical unit stays the specimen rather than the component-degradation cell, so the robustness claims rests on a consistent trend across the three folds rather than on a significance value computed from correlated units. Paired tests are nevertheless reported at two levels. The first is a two-sided paired t-test over the three held-out specimens. The second is a Wilcoxon signed-rank test over the 30 physical components, using each component's macro-F1 averaged over its states and events, with Holm correction across the comparisons against the gate. An exact specimen-level Wilcoxon test is not used, because with three pairs its smallest two-sided p-value is 0.25. Cluster bootstrap intervals are attached to the reported values so that a small overlap between methods under mild degradation is visible rather than hidden by point estimates.

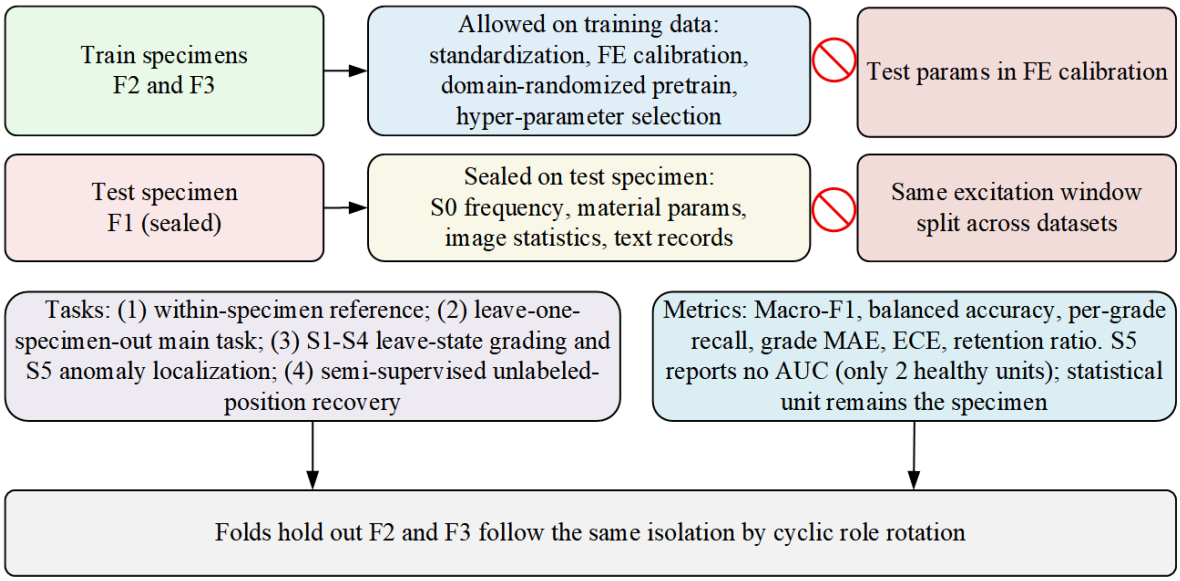


Fig. 3. Cross-specimen evaluation tasks and the in-fold finite element isolation protocol

3. Results

3.1 Data Distribution, Label Reliability and In-Distribution Reference

The component-state counts of the three specimens across S0 to S5 follow the distribution listed in Table 2. At S0 all 30 component-states remain healthy, and the share of level 1, level 2 and level 3 units grows steadily as the drift ratio rises, so that S5 keeps only 2 healthy units while it holds 8 level-3 units. Over all states the 180 component-states split into 88 healthy, 51 light, 28 moderate and 13 severe units, which correspond to shares of 48.9%, 28.3%, 15.6% and 7.2%. The severe class stays clearly smaller than the others, which matches real frame behavior and implies that its confidence interval should be wider than the intervals of the other grades. The joint decision among crack width, spalling area and residual chord rotation produced a set of consistent labels and a smaller set of conflicting samples, and the conflicting samples kept an intermediate level with an uncertainty flag rather than a forced assignment. Fourteen of the 180 component-states fell into this uncertain group, mostly at S3 and S4 where a moderate crack width coincided with a still limited residual rotation. The class imbalance means that a plain accuracy would be dominated by

the healthy and light classes, so the macro-averaged F1 and the per-grade recall are used as the primary indicators throughout the results.

The three loading paths produced distinguishable damage configurations. F1 spread moderate damage across the two bays, F2 concentrated higher grades in the left-bay beams and the adjacent column ends, and F3 concentrated damage on the right side, so the peak damage position, the baseline frequency and the sensor response differed among specimens. The full model and the fixed-weight model were then evaluated under the same-specimen event split as an in-distribution reference. Under that split the full model reached a macro-averaged F1 near 0.86 with a grade mean absolute error near 0.19, which stayed well above the leave-one-specimen results reported below. This gap between same-specimen state re-identification and unseen-specimen diagnosis is treated as a measure of task difficulty, and the higher in-distribution value is not read as generalization capacity, since the repeated excitations of three specimens do not form independent structures. The baseline first-mode frequency of the three healthy specimens differed by close to 6 percent, which reflects the concrete-batch difference and gives the finite element pretraining a realistic domain to randomize over. Under the same-specimen split the fixed-weight model reached a macro-averaged F1 near 0.83, so even the in-distribution reference showed a small gap between adaptive and fixed fusion, and that gap widened once the harder cross-specimen setting was applied. The comparison between the two settings therefore frames the difficulty of the main task rather than serving as a target to be maximized.

3.2 Cross-Specimen Overall Results and Digital Twin Ablation

With F1, F2 and F3 as the test specimen in turn, the cross-specimen results of the single modalities, the fusion methods and the published comparison appear in Table 6. Under clean input the quality-aware vision-language-sensor model reached a macro-averaged F1 of 0.74, 0.69 and 0.72 on the three folds, above the fixed-weight fusion at 0.70, 0.65 and 0.69 and close to the learnable evidence fusion at 0.72, 0.70 and 0.70. The evidence fusion slightly exceeded the present method on F2, which is consistent with the expectation that a single method need not dominate every specimen. The GMU reached 0.71, 0.67 and 0.70, and the cross-attention Transformer reached 0.72, 0.70 and 0.69 with about 0.4 million extra parameters. It matched the gate on F2 but showed the widest inner-to-outer gap among the fusion methods, which is consistent with its larger capacity on two training specimens. Under clean input the gate differed from fixed-weight fusion with $p = 0.008$ (t-test) and $p = 0.004$ (Wilcoxon), and from the GMU with $p = 0.020$ and 0.06. Its differences from the evidence fusion ($p = 0.42$ and 0.31) and the cross-attention Transformer ($p = 0.38$ and 0.27) were not significant. The vision single modality stayed near 0.63 to 0.67 and the sensor single modality near 0.58 to 0.61, while the language single modality remained the weakest near 0.46 to 0.49, which agrees with its role as weak evidence. Among the folds F2 carried the largest material and response shift relative to its training specimens, and that shift caused most of the light-damage misses and the adjacent beam-column confusions. The published sensor spatial-temporal graph method reached 0.59 to 0.63 and stayed below the present sensor branch, which indicates that finite element pretraining and node mapping improved the pure sensor diagnosis. The per-grade recall in Table 6 shows that the main confusion sat between level 1 and level 2, while the level-3 recall of the quality-aware model stayed in a range of 0.42 to 0.54, above the fixed-weight range of 0.38 to 0.48. The severe class stayed the hardest across every method because its sample count is small, and the confidence interval of the level-3 recall was correspondingly wide, which the reported cluster bootstrap intervals make explicit. Of the 41 missed level-3 component-events across the three folds, 34 were predicted as level 2, 7 as level 1 and none as level 0. Thus 91% of the level-3 events were still assigned level 2 or higher, and the missed events carried a mean uncertainty of 0.37, against 0.18 for correctly graded ones.

Table 6. Cross-specimen results of modality configurations and comparison methods

Method	Hold-out F1	Hold-out F2	Hold-out F3	Level-3 recall
Vision single modality	0.67 (0.63, 0.71)	0.63 (0.59, 0.67)	0.66 (0.62, 0.70)	0.34-0.45
Sensor single modality	0.61 (0.57, 0.65)	0.58 (0.54, 0.62)	0.60 (0.56, 0.64)	0.29-0.38
Language single modality	0.49 (0.45, 0.53)	0.46 (0.42, 0.50)	0.48 (0.44, 0.52)	0.17-0.24
Vision + sensor	0.71 (0.67, 0.75)	0.67 (0.63, 0.71)	0.69 (0.65, 0.73)	0.41-0.50
Published sensor method	0.63 (0.59, 0.67)	0.59 (0.55, 0.63)	0.61 (0.57, 0.65)	0.31-0.40
GMU fusion	0.71 (0.67, 0.75)	0.67 (0.63, 0.71)	0.70 (0.66, 0.74)	0.40–0.50
Cross-attention Transformer	0.72 (0.68, 0.76)	0.70 (0.66, 0.74)	0.69 (0.65, 0.73)	0.41–0.52
Learnable evidence fusion	0.72 (0.68, 0.76)	0.70 (0.66, 0.74)	0.70 (0.66, 0.74)	0.44-0.53
Fixed-weight VLS	0.70 (0.66, 0.74)	0.65 (0.61, 0.69)	0.69 (0.65, 0.73)	0.38-0.48
Quality-aware VLS	0.74 (0.70, 0.78)	0.69 (0.65, 0.73)	0.72 (0.68, 0.76)	0.42-0.54

The three digital twin ablations appear in Table 7. Removing the in-fold finite element pretraining lowered the three-fold macro-averaged F1 from 0.74, 0.69 and 0.72 to 0.69, 0.64 and 0.68, which was the largest drop and confirmed that pretraining narrows the cross-specimen sensor gap. Removing the building information modeling graph attention reduced the values to 0.71, 0.67 and 0.69 and mainly increased localization errors at adjacent beam-column joints. Removing the physical direction loss produced the smallest change, to 0.73, 0.68 and 0.71, while it raised the share of predictions that conflict with the frequency and coarse mode-shape change direction. The contribution order from the real ablation is therefore finite element pretraining first, component topology second and direction constraint third, and this order is reported from the data rather than assumed in advance. The effect of the finite element pretraining was strongest on F2, the fold with the largest response shift, which is consistent with the view that pretraining mainly helps when the test specimen sits far from the training specimens in the material and dynamic space. The graph attention removal changed the localization F1 more than the detection F1, since a wrong node mainly moves the predicted damage to a neighbor rather than removing the detection itself. These patterns support treating the three supports as complementary functions instead of interchangeable regularizers. The direction constraint lowered the share of physically contradictory predictions from about nine percent to about four percent without changing the detection F1 much, which is why it ranks third on accuracy yet still earns a place as a physical filter.

Table 7. Ablation of digital twin components, fusion mechanism and modalities

Configuration	Macro-F1 F1/F2/F3	Surface-response subset	Trend text and occlusion subset
Full model	0.74/0.69/0.72	0.70/0.64/0.68	0.68/0.62/0.66
Remove in-fold FE pretrain	0.69/0.64/0.68	0.65/0.59/0.62	0.64/0.57/0.61
Remove BIM graph attention	0.71/0.67/0.69	0.67/0.62/0.65	0.65/0.60/0.63
Remove direction loss	0.73/0.68/0.71	0.69/0.63/0.67	0.67/0.61/0.65
Change to fixed weight	0.70/0.65/0.69	0.65/0.59/0.63	0.62/0.56/0.60
Remove language branch	0.73/0.69/0.71	0.69/0.63/0.67	0.59/0.54/0.57
Remove sensor branch	0.67/0.63/0.66	0.56/0.51/0.54	0.64/0.58/0.61

3.3 Core Robustness Results under Modality Degradation

This part concerns the central proposition. Each method first fixed its own clean baseline, and the degradation was then applied on the three unseen test specimens. As shown in Figure 4, under 25% and 50% image occlusion the vision quality weight of the present gate fell and the macro-averaged F1 retention stayed near 0.97 and 0.90, above the fixed-weight retention near 0.92 and 0.81. Under 20 dB noise and 20% continuous packet loss the sensor weight fell and the retention of the gate stayed near 0.94 and 0.92, while the fixed-weight retention dropped to about 0.86 and 0.83. When

the language modality was absent the explicit mask kept the retention near 0.97 for both methods, which shows that the mask prevented an invalid feature from entering the fusion. Under the dual degradation that combines 25% occlusion with 20% loss the gate retention fell to a range of 0.79 to 0.83 across the three folds, while the fixed-weight retention fell to a range of 0.67 to 0.71, so the adaptive gate kept clearly more of its clean performance in the hardest condition. The learnable evidence fusion sat between the two methods on most conditions, so its retention stayed near 0.87 under 50% occlusion and near 0.76 under the dual degradation, which shows that an adaptive but heavier method captures part of the benefit without matching the low-parameter gate at the hardest point. The GMU and the cross-attention Transformer behaved similarly, retaining about 0.85 and 0.83 under 50% occlusion and 0.72–0.76 and 0.70–0.75 under the dual degradation. Their gates are learned from the features themselves, so a corrupted but confident feature can still receive a high weight. Under mild degradation the methods stayed close and their bootstrap intervals partly overlapped, which is the expected behavior when the removed information is small, and only the moderate to severe conditions separated the methods clearly. Under the dual degradation the retention of the gate exceeded that of fixed-weight fusion with $p = 0.002$ (t-test) and $p < 0.001$ (Wilcoxon), and that of the evidence fusion with $p = 0.026$ and 0.003. Under 25% occlusion and 20 dB noise, no comparison reached $p < 0.05$. The 20% packet-loss condition affected the sensor localization more than the 20 dB noise, because a continuous gap removes a whole-time segment while additive noise leaves the response envelope partly readable, and the gate tracked this difference by lowering the sensor weight more sharply under loss than under noise.

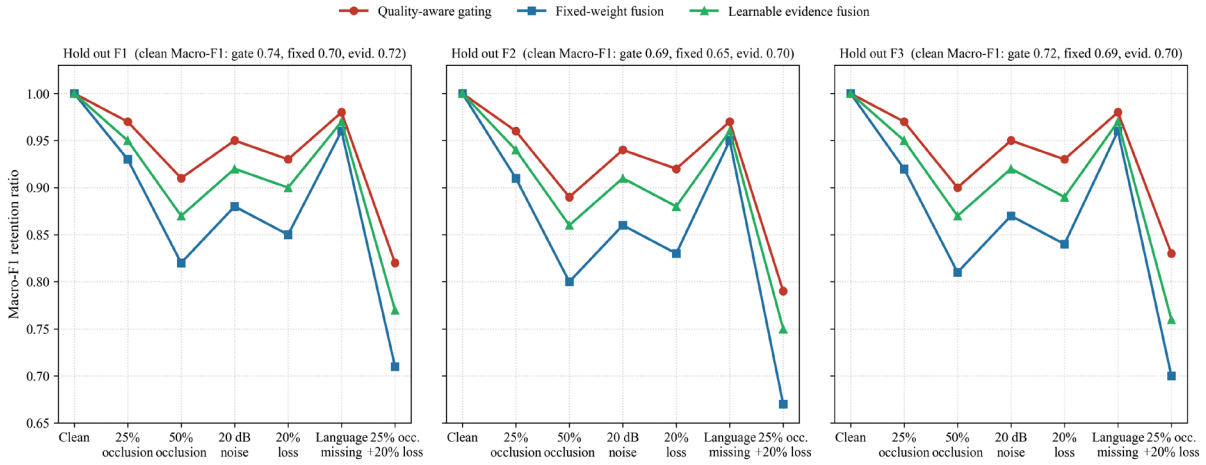


Fig. 4. Degradation performance retention ratio of the three cross-specimen folds

The gate also satisfied the monotone requirement, since each modal weight fell when its quality indicator fell and the higher-quality modality received a relatively higher weight. Because the direction of the improvement held in the F1, F2 and F3 folds, the result is written as a direction-consistent robustness improvement across three unseen physical specimens, and it is reported as consistent-trend evidence within the three specimens rather than as an overall significance computed from component-degradation units. Against the multi-view uncertainty fusion, the practical advantage of the gate lay in a comparable clean accuracy, a slower decay under degradation, a smaller parameter counts and a more interpretable weight change, rather than a large gain under clean input. The clean macro-averaged F1 of each method is annotated in every subplot of Figure 4 so that the retention ratio does not hide the difference in baseline performance among the compared methods. The drop from the moderate 25% occlusion to the severe 50% occlusion reached about 7 percentage points of retention for the gate and about 11 points for fixed-weight fusion, which shows that the advantage of the gate grows as the degradation deepens rather than staying constant. Under the language-missing condition the two fusion methods stayed almost identical, because the explicit mask removes the language path cleanly and the remaining vision and sensor weights are re-normalized in both methods, so the gate offers no special benefit when a modality is simply switched off rather than degraded. This contrast clarifies that the gate helps most when a modality is present but unreliable, which is the common field situation of a blurred image or a noisy channel. The calibration error followed the same trend, since the gate kept its

expected calibration error near 0.06 under moderate degradation while the fixed-weight model rose toward 0.11, which means the adaptive weights not only preserved accuracy but also kept the reported confidence closer to the observed correctness. Across the three folds the ordering of the methods never reversed under any single degradation type, and this repeated ordering is the specific evidence behind the direction-consistent robustness statement, which stays a trend claim within three specimens rather than a population significance result.

3.4 Modality Necessity, Whole-Model Overfitting and Failure Cases

Under clean input the modality comparison in Table 7 separates the roles of the branches. Vision contributed most to surface damage grading, the sensor contributed to response-change recognition, and language contributed to trend description. Using the training fold to set the surface-damage and residual-deformation thresholds, the test specimens revealed components with weak surface features but high residual deformation, and on that subset the vision-sensor model lowered the grade mean absolute error and the localization error relative to the vision-only model, which shows that the sensor supplied response information that vision cannot obtain directly. On the subset with an explicit change-from-previous phrase and simultaneous vision-sensor degradation, the full three-modal model reduced the miss rate and improved the confidence relative to the vision-sensor model, so language is positioned as a conditional backup modality that acts mainly in this specific condition. On the surface-response subset the vision-sensor model lowered the grade mean absolute error from about 0.41 to about 0.29 relative to the vision-only model, and this gain was concentrated on the components with high residual rotation but thin surface cracks, which is exactly where an image underrepresents the damage. On the trend and occlusion subset the removal of the language branch lowered the three-fold macro-averaged F1 from 0.68, 0.62 and 0.66 to 0.59, 0.54 and 0.57, a much larger loss than on the clean set, which confirms that the value of language is conditional rather than uniform. To approximate field reports, a supplementary test injected field-style noise into 30% of the test texts, without retraining. The noise consisted of homophone typos, dropped words, colloquial abbreviations, omitted component references and paraphrasing by a large language model. The mean language quality score fell from 0.71 to 0.52 and the mean language weight from 0.21 to 0.14. The clean macro-F1 retention stayed near 0.98, and on the trend and occlusion subset the three-fold macro-F1 fell to 0.64, 0.58 and 0.62, still above the values without the language branch. The indicator reacted to incomplete or garbled text, but not to fluent text with a wrong trend, which remains the main language risk. The overfitting curves showed that the vision branch held the largest share of trainable parameters and the widest train-to-test gap, while the language branch stayed almost flat between training and testing because most of its body was frozen, which indicates that the small-sample risk concentrates in the vision path and justifies the heavy freezing applied there. The failure cases in Figure 5 also showed that a wrong inspection phrase could pull a healthy component toward a light grade when vision and sensor were both weak, which is why the language weight is capped and the uncertainty output rises in that situation rather than the prediction being trusted outright. The near real-time inference and write-back timing indicate that the whole pipeline could run within a routine inspection cycle rather than as an offline batch.

The overfitting check covered the vision, language, sensor, graph and multi-task heads. The full model held about 3.1 million trainable parameters, and the training macro F1 near 0.93, the inner validation near 0.81 and the outer test near 0.72 showed a moderate gap without a case where the training value kept rising while the cross-specimen value fell. The typical failures, illustrated in Figure 5, include adjacent beam-column co-cracking, a fine crack hidden by occlusion, a failed strain point, a wrong inspection phrase, a shortage of severe samples and a baseline frequency shift of the test specimen. The adjacent beam-column confusion partly reflects an instrumentation limit. The ten strain gauges are uniaxial along the member axes at the component ends, and no sensor lies in the joint cores. Diagonal shear cracking in a core is therefore seen only indirectly, through the neighboring end strains and the close and lateral images. Diagonal core cracks appeared in 3 of the 6 joints of each frame at S4 and S5, with widths of at most 0.30 mm, and they did not govern the damage of any frame. Nevertheless, 14 of the 22 adjacent confusion events involved such a joint. Rosette gauges or digital image correlation on the joint cores would be needed to resolve these cases. A representative underestimation assigned a level-3 column on F3 at S5 to level 2 near the

severity boundary, and a representative false alarm raised a healthy unit on F2 at S1 to level 1 under noise. The inference time of a single excitation event stayed near 42 ms and the building information modeling state update stayed near 15 ms, which supports near real-time write-back.

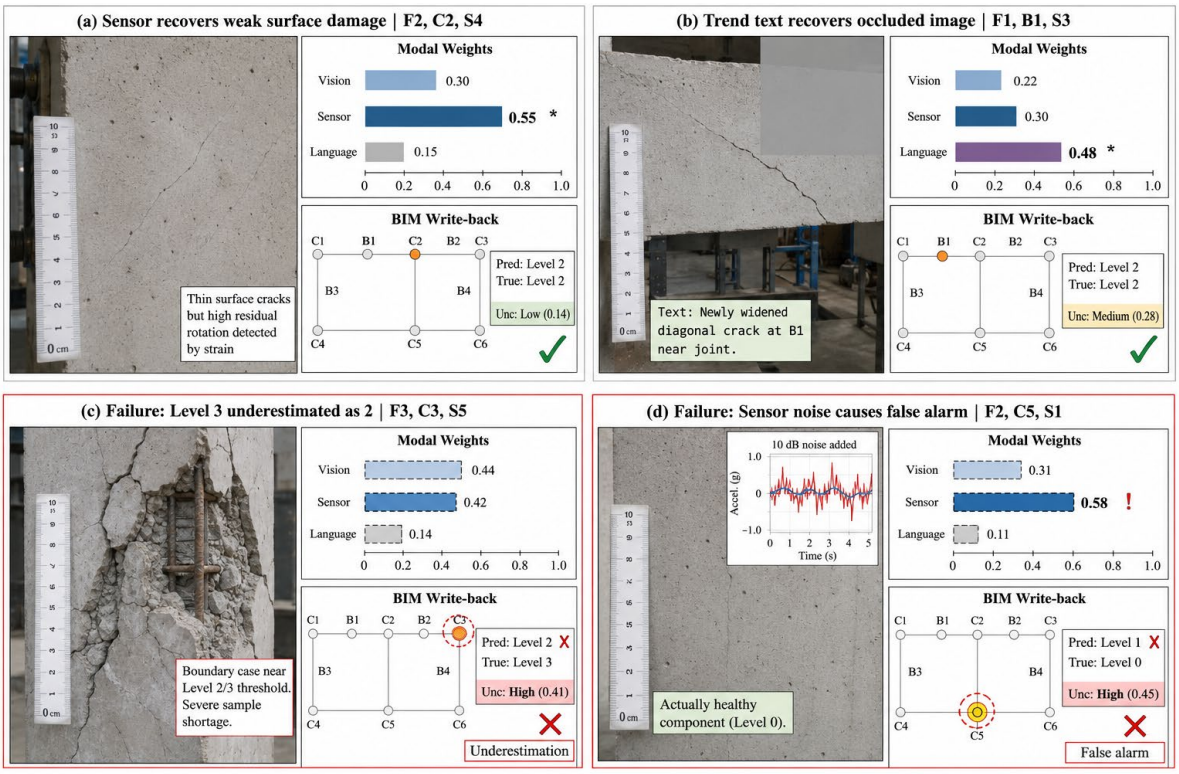


Fig. 5. Representative success and failure cases with modal weights and BIM write-back

4. Discussion

4.1 Robustness Mechanism of Quality-Aware Fusion

The retention ratios of the three outer test specimens explain why fixed-weight fusion loses more performance once a modality degrades. A fixed weight keeps feeding a low-quality visual or sensor feature into the fusion after its image sharpness or its signal-to-noise ratio has clearly dropped, so a negative transfer can appear. The low-parameter fuzzy gate instead adjusts the modal weight from computable indicators such as sharpness, occlusion, signal-to-noise ratio and packet-loss ratio, so the model leans more on the evidence that stays reliable. Trusted multi-view classification with dynamic evidential fusion offers one principled route for such adaptive weighting [33], and multi sensor fault diagnosis through a Markov chain and evidence theory shows a related decision-level treatment of conflicting sources [34]. Because the three folds moved in the same direction, the effect is stated as consistent robustness evidence within the three specimens of one configuration, and if only part of the occlusion or packet-loss conditions had held the claim would be restricted to that degradation type. Work on uncertainty under nonuniform measurement [35] and on sparse learning from uncertain and limited data [36] frames why a low-parameter gate can reach a comparable robustness with fewer parameters and a more interpretable weight change in a small sample. The comparison further connects to robust learning under missing data [37] and to heterogeneous response recovery through multimodal learning [38], which supports the role of the explicit modality mask.

4.2 Distinct Roles of the Digital Twin Components

The real ablation results guide the discussion of the digital twin, without a preset ranking. In-fold finite element pretraining narrows the material and dynamic-response gap between the training specimens and the test specimen, and mechanics-informed learning has been shown to detect and localize unforeseen structural damage in a related setting [39]. The physical basis of this step

follows the broader principle of physics-informed machine learning, in which governing relations regularize a data-driven model [40]. The building information modeling graph confines each modality to propagate only between beam-column nodes that share a real connection, and digital-twin monitoring systems that fuse real-time data with a structural model provide the practical frame for such topology-aware updates [41], including large reservoir and hydro-steel structures where the twin drives maintenance decisions [42]. The direction constraint reduces predictions that clearly contradict the frequency and floor-level mode-shape change, and immersive bridge twin platforms show how a corrected diagnosis can be written back into the asset model for closed-loop management [43]. Since the consistency check derives from a physical summary of the same sensor input, it stays an internal consistency analysis rather than an independent external validation, a distinction that mirrors the open concerns about model error and updating in engineering twins [44]. Because the ablation placed pretraining above topology and topology above the direction constraint, the three supports are read as a layered mechanism in which pretraining aligns the response domain, topology routes the aligned evidence to the correct component, and the direction constraint filters out the residual predictions that contradict the coarse dynamic behavior.

4.3 Multimodal Necessity and Applicability Boundary

Vision holds a natural advantage in the severity task, because crack width and spalling area partly define the grades, and indirect deep-learning identification confirms that response data alone struggles to recover fine surface grades [45]. The necessity of the sensor shows mainly on components where surface phenomena underrepresent the irreversible deformation, and smart-infrastructure twins that integrate dense response data support this complementary reading [46]. Language mainly supplies the change trend that the current image and sensor windows do not contain explicitly, and it serves as a backup when vision and sensor degrade together, which aligns with critical reviews of twin-enabled integrity management that value inspection history as weak but non-redundant evidence [47]. The inspection texts, however, follow a fixed template. The text quality indicator therefore checks form rather than truth, and its behavior on unstructured field reports needs verification with real inspection records. Because the three specimens share geometry, section and reinforcement, the study supports only a cross-batch and cross-path feasibility claim within one frame configuration, and broader civil twin workflows for existing infrastructure caution against extrapolating from a small population [48]. The 1080 component-events contain repeated excitations, and the truly independent physical objects are only 30 components and three frames, so a component-level count cannot claim overall statistical significance, and the reported p-values describe differences within the three specimens rather than across a population, and population-based domain adaptation shows that transfer to different stories, systems or deterioration mechanisms needs a much larger and more varied population [49]. The low share of level-3 samples carries a direct engineering risk. An underestimated severe component may be judged repairable when it already approaches Life Safety, which could delay shoring or occupancy restriction. In the present results this risk is bounded rather than removed: missed level-3 events fell mainly to level 2, which still lies beyond Damage Limitation and calls for inspection, and they carried elevated uncertainty. The output should therefore serve inspection triage rather than release decisions, and any component predicted at level 2 or above with high uncertainty should be escalated to on-site verification. The four evaluation tasks stay distinct as well, since same-specimen re-identification, cross-specimen transfer, leave-state extrapolation and semi-supervised position recovery test different abilities, and a strong result on one task cannot stand in for another. Reading them together rather than separately would inflate the apparent generalization, which is the very error the strict outer isolation was designed to prevent, so the boundaries stated here are treated as part of the result rather than as a disclaimer added afterward.

5. Conclusions

This study organized three one-third scale two-story two-bay reinforced concrete frames into building information modeling component-level multimodal data, supported the diagnosis with strictly outer-isolated digital twin finite element pretraining, and fused the modalities with a low-parameter quality-aware fuzzy gate. On the three unseen physical specimens the full model reached

a macro-averaged F1 of 0.74, 0.69 and 0.72 with matched localization and grading, and the quality-aware gate kept a higher relative performance than fixed-weight fusion under image occlusion, sensor degradation and modality absence, with the largest margin under the dual degradation. The ablation ranked the digital twin roles as finite element pretraining first, component topology second and direction constraint third, each carrying a distinct diagnostic function. The sensor contribution stayed bound to the surface-response inconsistency subset and the language contribution to the dual-degradation trend scenario. All conclusions remain within the feasibility range of one two-story two-bay configuration and three specimens, and a larger and more varied population is needed before any wider generalization. The three propositions raised at the outset are therefore answered in a bounded form. The digital twin supports carry distinct and additive roles, the quality-aware gate decays more slowly than fixed-weight fusion once a present modality becomes unreliable, and the sensor and language modalities each occupy a specific niche rather than a general advantage. Future work will proceed in two steps. First, a shaking table test of a three-story three-dimensional RC frame at 1/2 scale, with artificial mass for dynamic similitude and rosette gauges on the joint cores, will test transfer across story number, spatial configuration and scale. Second, field monitoring of two in-service RC buildings with routine inspection records will test whether the component-level anchor and the low-parameter gate keep their behavior on full-scale structures with real sensing and text noise. Together with frames of different bay numbers, story heights and reinforcement ratios, this will raise the number of independent physical structures well beyond three.

Data and Code Availability

The data, code, trained weights and BIM models are not publicly available because of restrictions of the funding project and the laboratory data agreement. They are available from the corresponding author on reasonable request for academic, non-commercial use. To support reproduction, the manuscript gives the labeling thresholds, the gate rule base, the network configuration, the model equations, the finite element settings and all training hyper-parameters.

References

[1] Committee A. ASCE/SEI 41-17: Seismic evaluation and retrofit of existing buildings. American Society of Civil Engineers. 2017. https://doi.org/10.1061/9780784414859
[2] EN C. Eurocode 8: Design of structures for earthquake resistance—Part 3: Assessment and retrofitting of buildings. Comité Européen de Normalisation: Brussels, Belgique. 2005:89.
[3] Boje C, Guerriero A, Kubicki S, Rezgui Y. Towards a semantic Construction Digital Twin: Directions for future research. Automation in Construction. 2020;114:103179. https://doi.org/10.1016/j.autcon.2020.103179
[4] Opoku DGJ, Nguyen ET, Perera S, Jin R, Zhou L. Digital twin application in the construction industry: A literature review. Journal of Building Engineering. 2021;40:102726. https://doi.org/10.1016/j.jobe.2021.102726
[5] Lu Q, Xie X, Parlikad AK, Schooling JM. Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. Automation in Construction. 2020;118:103277. https://doi.org/10.1016/j.autcon.2020.103277
[6] Xu J, Liu X, Becerik-Gerber B. Developing a digital twin model for monitoring building structural health by combining a building information model and a real-scene 3D model. Measurement. 2023;217:112955. https://doi.org/10.1016/j.measurement.2023.112955
[7] Hu X, Olgun G, Assaad RH. An intelligent BIM-enabled digital twin framework for real-time structural health monitoring using wireless IoT sensing, digital signal processing, and structural analysis. Expert Systems with Applications. 2024;252:124204. https://doi.org/10.1016/j.eswa.2024.124204
[8] Sun L, Shang Z, Xia Y, Bhowmick S, Nagarajaiah S. Hybrid monitoring methodology: A model-data integrated digital twin framework for structural health monitoring and full-field virtual sensing. Advanced Engineering Informatics. 2024;60:102386. https://doi.org/10.1016/j.aei.2024.102386
[9] Levine NM, Spencer BF Jr. Post-earthquake building evaluation using UAVs: A BIM-based digital twin framework. Sensors. 2022;22(3):873. https://doi.org/10.3390/s22030873
[10] Cha YJ, Buyukozturk O. Deep learning-based structural health monitoring. Automation in Construction. 2024;161:105328. https://doi.org/10.1016/j.autcon.2024.105328
[11] Fuller A, Fan Z, Day C, Barlow C. Digital twin: Enabling technologies, challenges and open research. IEEE Access. 2020;8:108952-108971. https://doi.org/10.1109/ACCESS.2020.2998358

[12] Adeagbo MO, Wang SM, Ni YQ. Revamping structural health monitoring of advanced rail transit systems: A paradigmatic shift from digital shadows to digital twins. Advanced Engineering Informatics. 2024;61:102450. https://doi.org/10.1016/j.aei.2024.102450
[13] Dong CZ, Catbas FN. A review of computer vision–based structural health monitoring at local and global levels. Structural Health Monitoring. 2021;20(2):692-743. https://doi.org/10.1177/1475921720935585
[14] Lee Y, Kim J, Kang Y. Structural damage detection using deep learning and FE model updating techniques. Scientific Reports. 2023;13(1):18694. https://doi.org/10.1038/s41598-023-45892-w
[15] Liu J, Zhang X, Zhou Y. Structural damage detection and localization via an unsupervised anomaly detection method. Reliability Engineering & System Safety. 2024;252:110465. https://doi.org/10.1016/j.ress.2024.110465
[16] Sun L, Lin Y, Yuan F. Physics-Augmented Spatial-Temporal graph convolutional network for damage localization using Ultrasonic guided waves. Mechanical Systems and Signal Processing. 2024;221:111738. https://doi.org/10.1016/j.ymssp.2024.111738
[17] Yu Y, Rashidi M, Samali B, Mohammadi M, Nguyen TN, Zhou X. Automated damage diagnosis of concrete jack arch beam using optimized deep stacked autoencoders and multi-sensor fusion. Developments in the Built Environment. 2023;14:100128. https://doi.org/10.1016/j.dibe.2023.100128
[18] Huang T, Liang T, Chen L. A comparative analysis of damage identification methods based on multi-channel data fusion and convolutional neural network for frame structures. Structures. 2024;60:105891. https://doi.org/10.1016/j.istruc.2024.105891
[19] Zhang J, Ma T, Wang L. Structural damage detection based on decision-level fusion with multi-vibration signals. Measurement Science and Technology. 2022;33(10):105112. https://doi.org/10.1088/1361-6501/ac782d
[20] Ahmadzadeh M, Zahrai SM, Bitaraf M. An integrated deep neural network model combining 1D CNN and LSTM for structural health monitoring utilizing multisensor time-series data. Structural Health Monitoring. 2025;24(1):447-465. https://doi.org/10.1177/14759217241243501
[21] Jiang Y, Chen X, Zhang Y. Large language model for post-earthquake structural damage assessment of buildings. Computer-Aided Civil and Infrastructure Engineering. 2025;40(31):6324-6342. https://doi.org/10.1111/mice.13250
[22] Zhang Y, Wang H, Liu Y. SDIGLM: leveraging large language models and multimodal chain of thought for structural damage identification. Structural Health Monitoring. 2025. https://doi.org/10.1177/14759217251348275
[23] Feng DC, Wang WJ, Mangalathu S. Condition Assessment of Highway Bridges Using Textual Data and Natural Language Processing-(NLP-) Based Machine Learning Models. Structural Control and Health Monitoring. 2023;2023(1):9761154. https://doi.org/10.1155/2023/9761154
[24] Liu J, Li P, Zhang H. Chinese named entity recognition for bridge damage and defects based on text mining and natural language pretraining models. Journal of Construction Engineering and Management. 2025;151(6):04025060. https://doi.org/10.1061/JCEMD4.COENG-14250
[25] Li S, Zhao X, Wang Y. Development of a BIM-based bridge maintenance system (BMS) for managing defect data. Scientific Reports. 2023;13(1):846. https://doi.org/10.1038/s41598-023-28012-4
[26] Tran-Ngoc H, Khatir S, De Roeck G, Bui-Tien T, Abdel Wahab M. Damage assessment in structures using artificial neural network working and a hybrid stochastic optimization. Scientific Reports. 2022;12(1):4958. https://doi.org/10.1038/s41598-022-08838-8
[27] Jamshidi M, El-Badry M. Structural damage severity classification from time-frequency acceleration data using convolutional neural networks. Structures. 2023;51:1120-1132. https://doi.org/10.1016/j.istruc.2023.03.085
[28] Nong X, Xu J, Zhang Y. Multimodal deep neural network-based sensor data anomaly diagnosis method for structural health monitoring. Buildings. 2023;13(8):1976. https://doi.org/10.3390/buildings13081976
[29] Chiachío M, Chiachío J, Fuentes R, Beck JL. Structural digital twin framework: Formulation and technology integration. Automation in Construction. 2022;140:104333. https://doi.org/10.1016/j.autcon.2022.104333
[30] Gardner P, Liu X, Worden K. On the application of domain adaptation in structural health monitoring. Mechanical Systems and Signal Processing. 2020;138:106550. https://doi.org/10.1016/j.ymssp.2019.106550
[31] Arevalo J, Solorio T, Montes-y-Gómez M, González FA. Gated multimodal networks. Neural Computing and Applications. 2020;32(14):10209-10228. https://doi.org/10.1007/s00521-019-04559-0
[32] Xu P, Zhu X, Clifton DA. Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;45(10):12113-12132. https://doi.org/10.1109/TPAMI.2023.3275156

[33] Han Z, Zhang C, Fu H, Zhou JT. Trusted multi-view classification with dynamic evidential fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;45(2):2551-2566. https://doi.org/10.1109/TPAMI.2022.3171980
[34] Wang K, Liu Y, Zhang M. Multisensor fault diagnosis via Markov chain and evidence theory. Engineering Applications of Artificial Intelligence. 2023;126:106851. https://doi.org/10.1016/j.engappai.2023.106851
[35] Zhu S, Xiang T. Structural damage identification considering uncertainties in nonuniform measurement conditions based on convolution neural networks. Structural Control and Health Monitoring. 2023;2023(1):8325686. https://doi.org/10.1155/2023/8325686
[36] Ding Z, Li J, Hao H. Structural damage identification by sparse deep belief network using uncertain and limited data. Structural Control and Health Monitoring. 2020;27(5):e2522. https://doi.org/10.1002/stc.2522
[37] Deng F, Zang C, Dong L. A robust deep learning-based damage identification approach for SHM considering missing data. Applied Sciences. 2023;13(9):5421. https://doi.org/10.3390/app13095421
[38] Du B, Zhang Y, Liu X. Heterogeneous structural responses recovery based on multi-modal deep learning. Structural Health Monitoring. 2023;22(2):799-813. https://doi.org/10.1177/14759217221098675
[39] Li X, Liu Z, Zhang H. Mechanics-informed autoencoder enables automated detection and localization of unforeseen structural damage. Nature Communications. 2024;15(1):9229. https://doi.org/10.1038/s41467-024-53550-y
[40] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nature Reviews Physics. 2021;3(6):422-440. https://doi.org/10.1038/s42254-021-00314-5
[41] Li X, Wang Y, Zhang J. Structural health monitoring system based on digital twins and real-time data-driven methods. Structures. 2024;62:106120. https://doi.org/10.1016/j.istruc.2024.106120
[42] Li H, Zhao Y, Liu M. Hydro-steel structure digital twins: application in structural health monitoring and maintenance of large-scale reservoir. Advanced Engineering Informatics. 2024;62:102922. https://doi.org/10.1016/j.aei.2024.102922
[43] Fawad M, O'Byrne M, Ghosh B, Pakrashi V. Development of immersive bridge digital twin platform to facilitate bridge damage assessment and asset model updates. Computers in Industry. 2025;164:104189. https://doi.org/10.1016/j.compind.2024.104189
[44] Wagg D, Worden K, Barthorpe RJ, Gardner P. Digital twins: state-of-the-art and future directions for modeling and simulation in engineering dynamics applications. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering. 2020;6(3):030901. https://doi.org/10.1115/1.4046739
[45] Hajializadeh D. Deep learning-based indirect bridge damage identification system. Structural Health Monitoring. 2023;22(2):897-912. https://doi.org/10.1177/14759217221105301
[46] Broo DG, Bravo-Haro M, Schooling J. Design and implementation of a smart infrastructure digital twin. Automation in Construction. 2022;136:104171. https://doi.org/10.1016/j.autcon.2022.104171
[47] Li S, Brennan F. Digital twin enabled structural integrity management: Critical review and framework development. Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment. 2024;238(4):707-727. https://doi.org/10.1177/14750902231215401
[48] Pregnolato M, Gunner S, Kilsby C. Towards Civil Engineering 4.0: Concept, workflow and application of Digital Twins for existing infrastructure. Automation in Construction. 2022;141:104421. https://doi.org/10.1016/j.autcon.2022.104421
[49] Gardner P, Lord C, Barthorpe RJ. Domain-adapted Gaussian mixture models for population-based structural health monitoring. Journal of Civil Structural Health Monitoring. 2022;12(6):1343-1353. https://doi.org/10.1007/s13349-022-00609-w