Local-Attention-Based Multilayer Long Short-Term Memory Model for Seismic Response Prediction: A Gantry Crane Case Study

Qihui Peng[1,*], Shihao Hu[1], Hongyu Jia[2], Hui Wang[3]

[1] China Israel Xbot School, Changzhou University, Changzhou, 213000, China

[2] School of Civil Engineering, Southwest Jiaotong University, Chengdu, 610031, China

[3] School of Mechanical and Electrical Engineering, Zhoukou Normal University, Zhoukou, 466001, China

Received 27 April 2026; revised 04 July 2026; accepted 08 July 2026

DOI: https://doi.org/10.46604/ijeti.2026.16408

Abstract

This study aims to develop a local-attention-based multilayer long short-term memory (LSTM) surrogate model for predicting the simulated horizontal displacement response of a gantry crane. The database is generated through incremental dynamic analysis using 100 ground-motion records scaled to 14 peak ground acceleration levels. To isolate the effect of attention scope, no-attention, global-attention, and local-attention configurations use the same LSTM backbone, data partitions, and training protocol. The local-attention mechanism reweights current and recent temporal representations while the recurrent states retain accumulated dynamic information. Across the investigated sample sizes, the local-attention model achieves the lowest minimum RMSE, more concentrated signed-error distributions, and a maximum test-set R^2 of 0.983, approximately 0.04 higher than the baselines in selected smaller-sample settings. Response histories show closer agreement around several peaks and rapidly varying segments. The two attention layers exhibit distinct but complementary allocation patterns, improving feature-processing transparency within the investigated gantry-crane and multi-intensity IDA framework.

Keywords: local attention, LSTM, seismic response prediction, gantry crane, incremental dynamic analysis

1. Introduction

Accurate prediction of structural seismic time-history responses is fundamental to seismic design, structural safety assessment, and post-earthquake performance evaluation. Conventional numerical methods, particularly the finite element method (FEM), have been widely used to simulate the dynamic behavior of complex structures subjected to seismic excitation [1–5]. Through spatial discretization and appropriate constitutive modeling, FEM can represent local mechanical behavior and nonlinear structural response with high fidelity. However, nonlinear time-history analysis generally requires refined numerical models, small integration time steps, and repeated simulations under numerous ground motions and intensity levels [6–12]. The resulting computational cost becomes substantial in incremental dynamic analysis, large-scale parametric studies, and applications requiring rapid response estimation. Consequently, the development of efficient surrogate models for structural seismic-response prediction has become an important research objective.

In recent years, machine-learning and deep-learning methods have received increasing attention in structural seismic engineering because of their ability to perform nonlinear mapping and automated feature extraction. Machine-learning approaches have been investigated for structural seismic-response prediction, damage identification, and rapid post-earthquake assessment [13, 14]. With the development of deep neural networks, convolutional neural networks (CNNs) have been

* Corresponding author. E-mail address: Pengqh@cczu.edu.cn

employed to extract spatial, local temporal, and time–frequency characteristics from seismic signals and structural responses [15–18]. Recurrent neural networks (RNNs) and their variants have also been widely adopted for sequential response modeling. Among these architectures, long short-term memory (LSTM) networks are particularly well suited to seismic time-history prediction because their gated recurrent structure alleviates the vanishing-gradient problem and enables the representation of temporal dependencies and accumulated dynamic states. Representative studies have employed LSTM-based and hybrid recurrent architectures for bridge-response surrogate modeling, transfer-learning-based building response prediction, and seismic-response estimation of railway systems [19–21]. Collectively, these studies demonstrate the ability of recurrent models to effectively establish nonlinear mappings between seismic excitation histories and corresponding time-dependent structural responses.

More recently, attention mechanisms have been incorporated into deep-learning-based seismic-response models to enable adaptive feature weighting and improve the transparency of the prediction process. Liao et al. proposed an attention-based LSTM model, termed AttLSTM, for bridge seismic-response modeling, in which temporal attention was used to identify informative components within sequential representations and improve learning from limited training data [22]. Saida and Nishio developed ExSRNet, an explainable seismic-response prediction framework that integrates SincNet, convolutional processing, frequency-wise feature extraction, and frequency attention to identify influential frequency bands [23]. Attention has also been combined with CNN–LSTM architectures and transfer learning for building seismic-response prediction. These integrated frameworks enable local feature extraction, recurrent temporal modeling, adaptive weighting, and knowledge transfer within a unified architecture. In addition to attention-based approaches, physics-informed or pseudo-labelled learning, pre-training, and transfer-learning strategies have been investigated to improve the effective use of limited structural-response data [24].

Despite these advances, relatively few studies have explored the explicit modeling of short-range temporal relevance within a multilayer recurrent architecture. Existing attention-based models primarily emphasize sequence-level temporal relevance, hybrid convolutional–recurrent representations, or frequency-domain contributions. AttLSTM evaluates informative representations over the encoded sequence, whereas ExSRNet identifies influential information from the perspective of frequency-band contributions. These approaches address different feature domains from the local temporal relationship considered in this study.

At a given time step, structural response reflects both the accumulated dynamic state of the system and the recently evolving excitation and response features. The recurrent hidden and cell states of an LSTM provide an implicit representation of the accumulated history, whereas a localized attention mechanism can explicitly and dynamically evaluate the relative importance of recent temporal representations. Recurrent memory and local temporal attention therefore perform complementary functions: the former preserves historical context, while the latter emphasizes short-range information that is directly relevant to the current prediction. This functional complementarity motivates the use of a restricted temporal attention window embedded at multiple levels of the recurrent architecture.

Based on this motivation, this study develops a multilayer LSTM model with local temporal attention to predict the seismic displacement response of a gantry crane steel structure. The methodological contribution is defined by three coordinated design features. First, the attention domain is explicitly restricted to the current and immediately preceding temporal representations. Second, local feature reweighting is embedded at multiple recurrent feature levels, allowing recent temporal information to be progressively extracted and refined. Third, the influence of attention scope is evaluated through a controlled comparison of no-attention, global-attention, and local-attention configurations, all constructed using the same LSTM backbone, data partitioning, and training protocol. This design enables the effect of attention scope to be examined without simultaneously introducing differences in backbone architecture or data allocation.

A numerical gantry-crane model is adopted as the case-study structure, and the seismic-response database is generated through incremental dynamic analysis using 100 ground-motion records scaled to 14 PGA levels. Model performance is evaluated using RMSE distributions, R^2 scores, signed-error distributions, response-history comparisons, and attention-weight visualization. These complementary analyses are used to assess overall prediction accuracy, local response agreement, error concentration, and hierarchical temporal feature allocation.

The remainder of this manuscript is organized as follows. Section 2 describes the numerical seismic-response database, ground-motion preprocessing, model architectures, and training and evaluation protocol. Section 3 presents the comparative prediction results, including the RMSE and R^2 analyses, signed-error distributions, response-history comparisons, and attention-based interpretation of temporal feature allocation. Section 4 summarizes the principal findings and outlines future extensions of the proposed framework.

2. Materials and Methods

2.1 Data acquisition

The seismic-response database was generated through numerical simulation based on the structural model established in a previous study [14]. Incremental dynamic analysis was performed using 100 ground-motion records selected from the PEER database. Each record was scaled to 14 peak ground acceleration (PGA) levels ranging from 0.1 g to 1.4 g at intervals of 0.1 g, resulting in a total of 1400 seismic input–response samples.

PGA served as the intensity-scaling parameter used to construct the IDA sequence rather than as the sole input descriptor. Each network input consisted of the complete ground-acceleration time history, thereby retaining information associated with waveform shape, duration, frequency content, and pulse characteristics. The corresponding horizontal displacement history at a representative location on the gantry crane, calculated using the OpenSees model, was adopted as the prediction target.

The case study considers a container gantry crane located in Sichuan Province, China. As shown in Fig. 1(a), the structure consists of two trolley girders, two support beams, two sill beams, and four supporting legs, including two rigid legs and two flexible legs. A three-dimensional centerline model was established in OpenSees using a lumped-plasticity formulation.

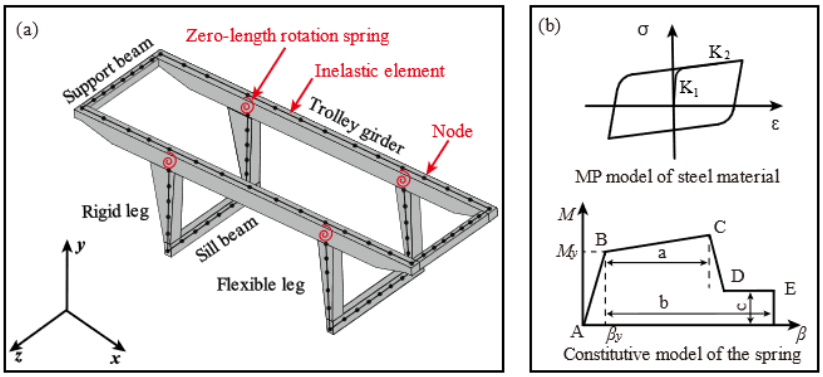


Fig. 1 Numerical modeling of gantry crane components [14]

Structural nonlinearity was represented using the uniaxial Giuffre–Menegotto–Pinto steel material model together with stiffness and strength degradation effects. Second-order P–Δ effects were included, and the connections between the trolley girders and supporting legs were idealized using zero-length rotational springs. As illustrated in Fig. 1(b), the spring constitutive relationship is characterized by the yield rotation βy and yield moment My.

The OpenSees model provides a consistent numerical framework for generating the target displacement histories and comparing the three learning configurations. Accordingly, the prediction task considered in this study is the reproduction of simulated seismic responses under the adopted structural model and IDA conditions.

For the sample-size analysis, six datasets containing 900, 1000, 1100, 1200, 1300, and 1400 input–response samples were constructed from the complete numerical database. Each dataset was randomly divided at the individual-sample level into training, validation, and test subsets using proportions of 80%, 10%, and 10%, respectively. The corresponding numbers of training, validation, and test samples were 720/90/90, 800/100/100, 880/110/110, 960/120/120, 1040/130/130, and 1120/140/140. Because the division was performed at the individual-sample level, samples originating from the same ground-motion record but corresponding to different PGA levels could occur in different subsets.

For each sample-size condition, identical training, validation, and test subsets were used for the no-attention, global-attention, and local-attention models. The training subset was used to optimize the model parameters, the validation subset was used for selecting epoch number and batch size, and the test subset was used only for final performance evaluation. A fixed random seed of 42 was adopted for sample partitioning and model initialization. This common data-allocation procedure ensured that the comparison among the three configurations was conducted under identical sample conditions.

The selected 100 original records from the PEER database exhibited substantial variations in duration and waveform characteristics. Before preprocessing, their sequence lengths ranged from slightly more than 1000 points to more than 10,000 points. Although PGA was used to define the fourteen IDA intensity levels, the network received the complete acceleration histories rather than individual intensity measures. Consequently, temporal duration, frequency content, pulse characteristics, and waveform variation remained available to the model during training and prediction.

The adopted sample-wise evaluation characterizes the ability of the surrogate models to learn the excitation–response relationship within the adopted multi-intensity IDA sample space considered in this study. Generalization to entirely unseen original ground-motion records represents a distinct cross-record evaluation setting.

2.2 Alignment of ground-motion data

The original ground-motion records varied substantially in sequence length, ranging from slightly more than 1000 points to more than 10,000 points. A uniform input length of 2000 points was therefore adopted to support batch-based training and direct comparison among the model configurations. As illustrated in Fig. 2, the alignment procedure was defined using the original sequence length (L) and the index (A) corresponding to the maximum absolute acceleration.

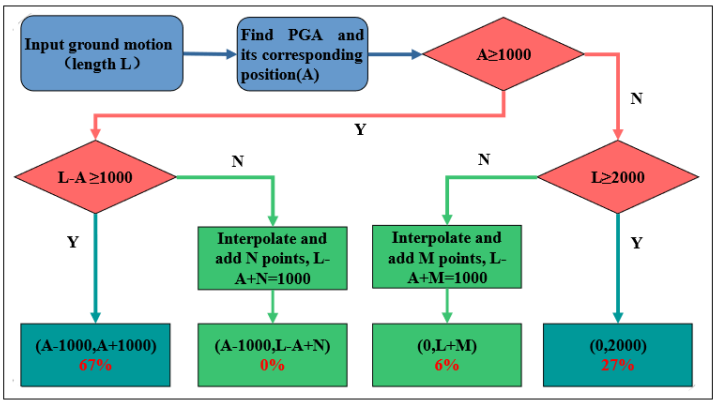


Fig. 2 Flowchart for aligning ground-motion data

The records were classified into four cases according to (L), (A), and the numbers of available points before and after the peak-acceleration index. When sufficient data were available on both sides of the peak, a 2000-point segment centered on the strongest excitation was retained. When the available sequence on either side of the peak was shorter than the required length, additional points were generated through interpolation following the procedure illustrated in Fig. 2. For records with an overall length shorter than 2000 points, interpolation was similarly used to extend the sequence to the prescribed input length. This procedure preserved the peak-acceleration region within the standardized sequence while retaining the available pre-peak and post-peak waveform information.

The same alignment procedure was applied to the ground-motion inputs used by the no-attention, global-attention, and local-attention models. Thus, all three configurations were trained and evaluated using identically preprocessed sequences. Peak-centered alignment was adopted in this study as a consistent sequence-standardization procedure for evaluating the influence of attention scope under otherwise identical input conditions.

2.3 Model architecture

A multilayer long short-term memory framework incorporating local temporal attention was developed to predict the simulated seismic displacement response of the gantry crane. Two reference configurations, namely a global-attention model and a no-attention model, were constructed using the same general recurrent architecture. All three configurations used the same recurrent depth and hidden-state dimension, while differing primarily in the presence and temporal scope of attention-based feature reweighting. This common architectural setting provides a consistent basis for examining the influence of attention scope.

2.3.1 LSTM cell

The LSTM cell is the fundamental recurrent component of the proposed architecture, as illustrated in Fig. 3. At each discrete time step t, a series of gating operations regulates the retention, updating, and transmission of temporal information. Given the current input x_t, the hidden state h_{t-1}, and the cell state C_{t-1} from the preceding time step, the LSTM cell performs the following operations:

(1) Forget gate: The forget gate determines the proportion of information retained from the previous cell state. A sigmoid activation function σ is applied to the concatenated vector $[h_{t-1}, x_t]$ through the corresponding weight matrix W_f and bias term b_f. The forget-gate vector is calculated as $f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right)$, with values ranging from 0 to 1, where values approaching 0 indicate greater information removal and values approaching 1 indicate greater retention.

(2) Input gate: The input gate determines the amount of new information incorporated into the current cell state. Using a similar sigmoid transformation, the input-gate vector is calculated as $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$.

(3) Candidate cell state: The candidate cell state represents the new information available for updating the cell state. It is generated using the hyperbolic tangent activation function as $\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$. Its values lie within the interval $[-1, 1]$, thereby bounding the magnitude of the candidate-state contribution.

(4) Cell-state update: The current cell state is obtained by combining the retained portion of the previous cell state with the gated candidate information, as expressed by $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$.

(5) Output gate: The output gate controls the information transferred from the updated cell state to the current hidden state. The output-gate vector is calculated as $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$.

(6) Hidden-state update: Finally, the current hidden state is obtained by applying the output gate to the transformed cell state, as expressed by $h_t = o_t \odot \tanh(C_t)$. The resulting hidden state serves as the information propagated to the subsequent network layer and the next time step.

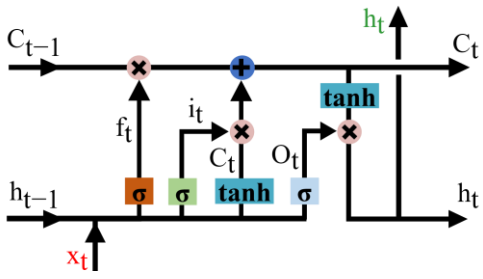


Fig. 3 LSTM cell

2.3.2 Attention mechanisms

Local attention mechanism: The proposed local-attention module operates on the temporal feature representations generated by the recurrent layers. It complements the recurrent memory of the LSTM by providing explicit adaptive weighting of recent hidden representations, rather than compensating for an assumed inability of the hidden state to retain previous information. The LSTM hidden and cell states provide an implicit compressed representation of the accumulated dynamic history, whereas local attention explicitly and dynamically reweights the current and recent hidden representations according to their relevance to the prediction at the present time step.

A local window of size ω is defined such that attention weights are calculated only over the interval [t−ω+1, t]. In the present implementation, ω = 5 was adopted, corresponding to the current hidden representation and the four immediately preceding representations. This value was fixed for all comparative analyses and should be understood as an architectural setting under the adopted temporal discretization rather than as a universal physical constant. Restricting the attention domain enables the model to explicitly emphasize recent temporal information without distributing the attention weights over the complete encoded sequence. For a given time step t, the local feature matrix is defined as:

$$X_t^{local} = \{X_{t-\omega+1}, X_{t-\omega+2}, ..., X_t\} \tag{1}$$

Here, the local feature matrix denotes the sequence of LSTM hidden representations within the selected temporal window. In Eq. (1), each uppercase X denotes the LSTM hidden-feature representation at the corresponding time step, whereas X_t local denotes the feature matrix assembled over the local temporal window. These variables are distinguished from the scalar acceleration input X_t.

$$Q_t = X_t^{local} W_Q \tag{2}$$

$$K_t = X_t^{local} W_K \tag{3}$$

$$V_t = X_t^{local} W_V \tag{4}$$

where WQ, WK, and WV are learnable projection matrices shared across time steps.

$$W_t = \text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) \cdot V_t \tag{5}$$

The scaling factor is determined by the key dimension. The dot-product similarity scores are normalized using Softmax within the local temporal window and are subsequently multiplied by V_t to obtain the local-attention representation. Because the attention module operates on the LSTM hidden-feature sequence rather than on an isolated scalar acceleration value or a PGA descriptor, its output retains the hidden-feature dimension and can be examined across feature channels and time steps.

Global attention mechanism: The global-attention configuration uses the same scaled dot-product formulation and the same general LSTM backbone, but its attention weights are calculated over the entire encoded sequence rather than the restricted interval [t−ω+1, t]. The comparison therefore changes the temporal domain of feature reweighting while retaining the principal recurrent architecture.

No-attention mechanism: The no-attention configuration passes the LSTM hidden representations directly to the subsequent network layers without explicit attention-based reweighting. It therefore provides a reference for evaluating the contribution of temporal feature reallocation under the same data and recurrent-architecture conditions.

2.3.3 Local-attention-based multilayer LSTM architecture

Fig. 4 illustrates the internal prediction architecture at each time step. The model contains three successive fusion levels. The first and second fusion levels share the same general structure, with two LSTM layers followed by a local-attention module. Each LSTM layer contains 32 hidden units. The hidden-state dimension was fixed across the no-attention, global-attention,

and local-attention configurations to maintain a consistent representational setting and prevent comparisons of attention scope from being confounded by differences in network width. The value of 32 is therefore treated as a common architectural setting for the three configurations rather than as an indication of limited recurrent-memory capability.

The first local-attention module reweights the hidden representations generated for the current and four preceding time steps. The local window [t−4, t] is evaluated together with the recurrent states, which continue to propagate accumulated information from earlier time steps. The architecture therefore retains longer-term dynamic context through recurrent state propagation while making the weighting of recent temporal features explicit and inspectable.

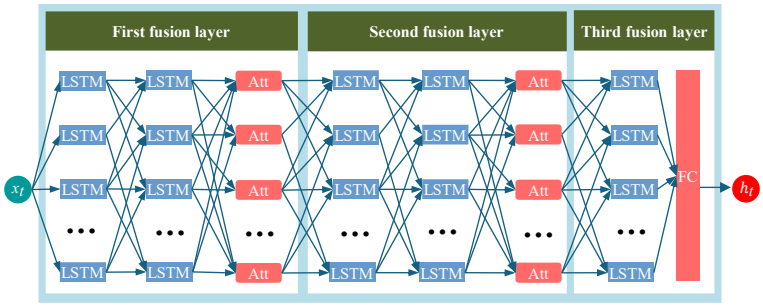


Fig. 4 Local-attention-based multilayer LSTM architecture

The second fusion level receives the representation produced by the first level and processes it through two additional LSTM layers followed by a second local-attention module. This repeated placement is a defining feature of the proposed architecture: local temporal information is reweighted at two recurrent feature levels rather than only once near the network output. The multilayer arrangement enables recent temporal representations to be progressively transformed and refined at different feature depths.

The third fusion level processes the refined representation through a final LSTM layer, after which a fully connected layer generates the predicted displacement at time step t. To reduce overfitting, dropout with a rate of 0.2 is applied after the first and second fusion levels, before their output representations are passed to the subsequent level. The same dropout rate and corresponding dropout positions are used in the no-attention, global-attention, and local-attention configurations. Overall, the proposed architecture is characterized by the integration of local-attention modules at two recurrent feature levels within a common LSTM backbone.

Fig. 5 illustrates the overall temporal-processing architecture. The fusion module shown in Fig. 4 is applied sequentially along the input acceleration history. At each time step, the module generates a displacement prediction and propagates the LSTM hidden state h_t and cell state C_t to the subsequent time step. The complete predicted displacement history is reconstructed from the sequence of stepwise outputs. This explicit propagation of recurrent states provides the accumulated temporal context required for the sequential excitation–response mapping.

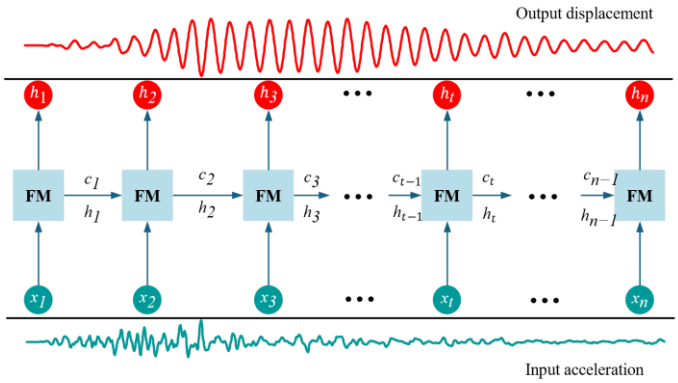


Fig. 5 Temporal processing structure

The LSTM backbone was selected because the seismic excitation–response relationship is inherently sequential and depends on the evolution of recurrent states over time. Convolutional and feature-recalibration modules can effectively extract local patterns, but they do not independently provide the explicit propagation of hidden and cell states used in the present architecture. The proposed local-attention mechanism therefore operates on a temporal neighborhood of recurrent representations rather than on convolutional feature channels. Although alternative recurrent, convolutional, and Transformer-based architectures may provide additional comparison opportunities, the present study retains a common LSTM backbone so that the effect of attention scope can be examined without simultaneously changing the state representation, receptive-field mechanism, and training behavior.

2.4 Training and evaluation protocol

All three model configurations were trained and evaluated using the same data preprocessing procedure, training–validation–test subsets, hidden-state dimension, recurrent depth, dropout setting, candidate training configurations, and evaluation criteria. The principal architectural difference among the models was the presence and temporal scope of attention-based feature reweighting. Keeping the remaining conditions consistent enabled the comparison to focus on the effects of attention scope across the no attention, global attention, and local attention configurations within the same general LSTM framework.

The models were trained by minimizing the mean squared error (MSE) using the Adam optimizer with an initial learning rate of 0.001. As described in Section 2.3.3, dropout with a rate of 0.2 was applied after the first and second fusion levels. The same dropout rate and corresponding dropout positions were used in all three model configurations.

The sensitivity of model performance to training duration and batch size was examined using candidate epoch settings spanning 200–1200 and batch-size settings spanning 1–128. For each sample-size condition and model configuration, the candidate settings were compared using the validation subset. The setting yielding the lowest validation RMSE was selected, after which the corresponding trained model was evaluated on the test subset. The test subset was not used for training or configuration selection. A fixed random seed of 42 was used for sample partitioning and model initialization. The principal training and model-selection settings are summarized in Table 1.

Table 1 Training and model reproducibility settings

Training item	Setting
Data partition	Sample-wise random split: 80% training, 10% validation, and 10% testing
Loss function	Mean squared error (MSE)
Optimizer	Adam
Initial learning rate	0.001
Dropout rate	0.2
Dropout position	After the first and second fusion levels
Investigated epoch settings	200–1200
Investigated batch-size settings	1–128
Model-selection criterion	Minimum validation RMSE
Random seed	42

Model performance was evaluated using the root mean square error (RMSE) and coefficient of determination (R^2), with R^2 reported in decimal form. RMSE and R^2 characterize the overall agreement between the predicted and reference displacement histories. Signed-error distributions and enlarged response-history comparisons were additionally used to examine error concentration, local peak agreement, rapidly varying response segments, and temporal deviations. Accordingly, R^2 was interpreted as a global goodness-of-fit measure and was considered together with the local response-history and signed-error analyses.

3. Results and Discussion

3.1 RMSE analysis across varying sample sizes and attention mechanisms

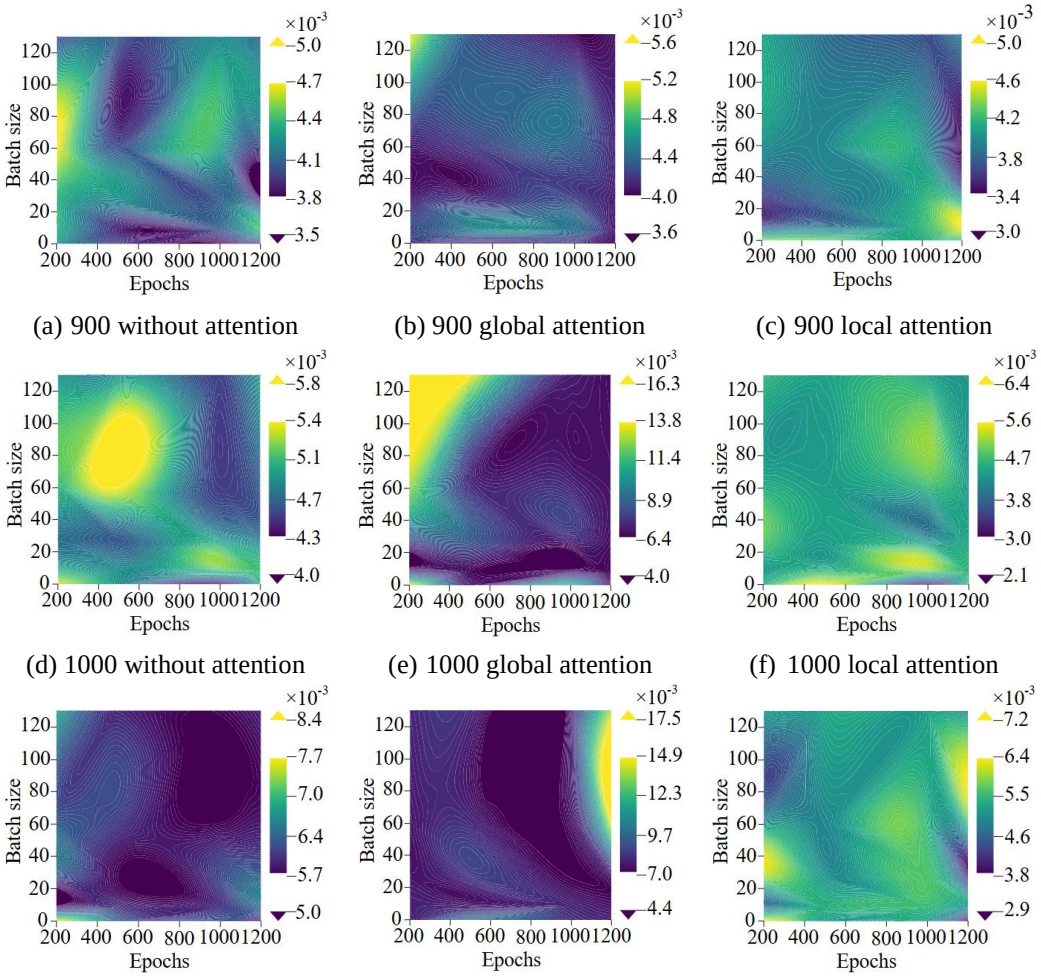
Figs. 6 and 7 present the RMSE surfaces obtained for the no-attention, global-attention, and local-attention models under the investigated epoch and batch-size settings. In addition to the lowest RMSE attained by each model, the surfaces illustrate the sensitivity of prediction performance to alternative training configurations. Darker regions indicate lower RMSE values.

For the 900-sample dataset, the RMSE ranges of the no-attention, global-attention, and local-attention models are 0.00345–0.00518, 0.00358–0.00588, and 0.00286–0.00537, respectively. The local-attention model attains the lowest RMSE of 0.00286 and exhibits multiple low-error regions across different epoch and batch-size combinations.

For the 1000-sample dataset, the minimum RMSE values of the no-attention, global-attention, and local-attention models are 0.00397, 0.00396, and 0.00218, respectively. The global-attention model exhibits the broadest RMSE range, whereas the local-attention model presents a concentrated low-error region at small batch sizes and intermediate-to-high epoch settings.

For the 1100-sample dataset, the corresponding minimum RMSE values are 0.00497, 0.00441, and 0.00333. Although the prediction errors of all three models vary with the training configuration, the local-attention model continues to attain the lowest RMSE.

For the 1200-sample dataset, the RMSE surfaces show broader variation across the investigated training settings. Nevertheless, the local-attention model achieves the lowest minimum RMSE of 0.00281, although the performance gap among the best-performing configurations becomes smaller.



(a) 900 without attention (b) 900 global attention (c) 900 local attention

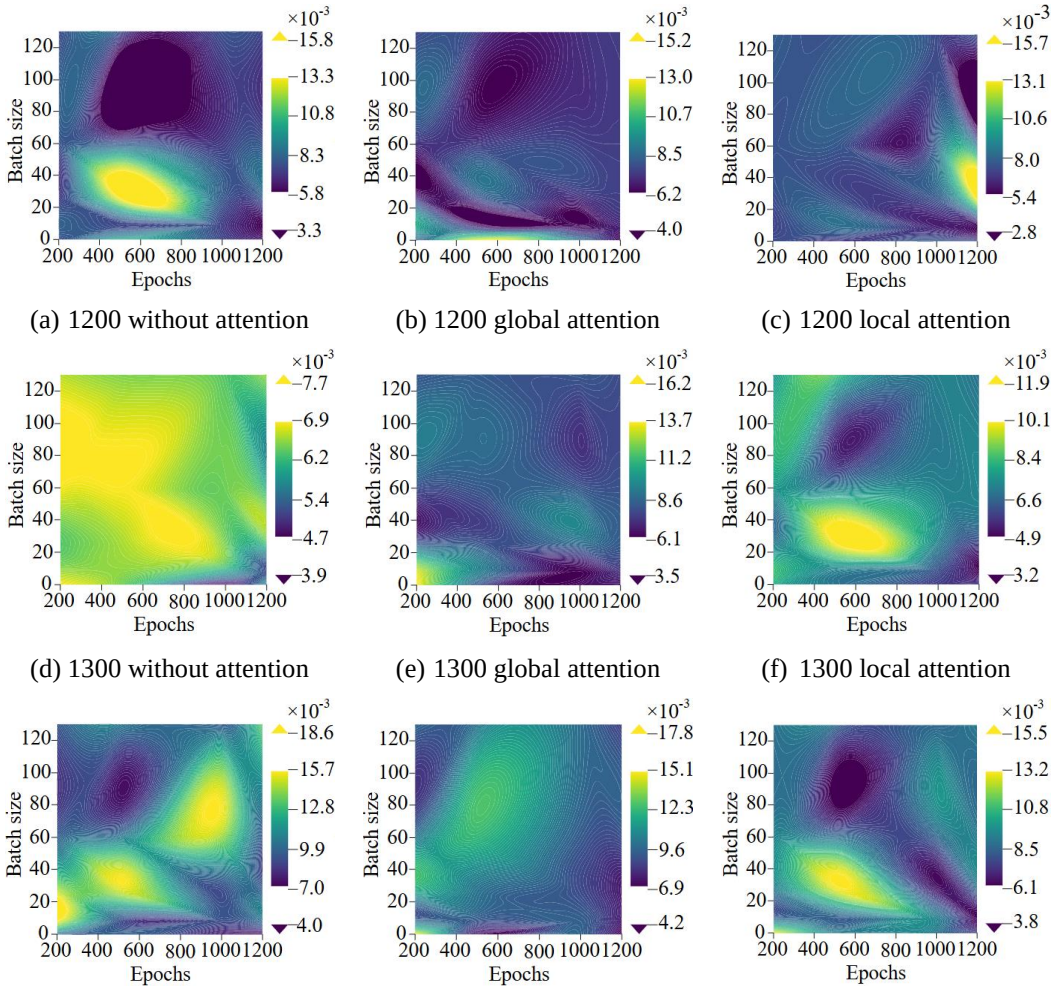
(d) 1000 without attention (e) 1000 global attention (f) 1000 local attention

(g) 1100 without attention (h) 1100 global attention (i) 1100 local attention

Fig. 6 RMSE surfaces of the three model configurations under different epoch and batch-size settings (900, 1000, and 1100 samples)

For the 1300-sample dataset, the RMSE of the local-attention model ranges from 3.23×10^{-3} to 11.9×10^{-3}. Its minimum value remains slightly lower than those obtained by the no-attention and global-attention models, while the performance of the three configurations becomes more comparable.

For the 1400-sample dataset, the minimum RMSE values of the no-attention, global-attention, and local-attention models are 0.00403, 0.00448, and 0.00388, respectively. Although the difference is narrower than that observed for the 900–1100 sample datasets, the local-attention model retains the lowest attained RMSE.



(a) 1200 without attention (b) 1200 global attention (c) 1200 local attention

(d) 1300 without attention (e) 1300 global attention (f) 1300 local attention

(g) 1400 without attention (h) 1400 global attention (i) 1400 local attention

Fig. 7 RMSE surfaces of the three model configurations under different epoch and batch-size settings (1200, 1300, and 1400 samples)

Across all six sample-size conditions, the local-attention model consistently achieves the lowest minimum RMSE among the three configurations. Its relative advantage is more pronounced for the 900–1100 sample datasets and decreases as the available dataset becomes larger. This repeated pattern suggests that explicit local temporal reweighting is particularly effective when the number of available input–response samples is relatively limited. Because all three models use the same data partitions, preprocessing procedure, recurrent depth, and hidden-state dimension, the observed differences provide consistent evidence of the effect of attention scope within the unified LSTM framework. The RMSE surfaces characterize sensitivity to epoch and batch-size selection and are not interpreted as estimates of stochastic uncertainty across repeated random seeds.

3.2 Comparative prediction performance across sample sizes

Table 2 reports the test-set R^2 values of the three model configurations. For the 900–1100 sample datasets, the no-attention and global-attention models achieve R^2 values ranging from 0.920 to 0.948, whereas the local-attention model maintains values between 0.972 and 0.983. The highest test-set R^2, 0.983, is obtained by the local-attention model for the 1000-sample dataset.

For the 1200–1400 sample datasets, the performance of the three configurations becomes more comparable; nevertheless, the local-attention model retains the highest R^2 in each case and remains at or above 0.970 across all six sample-size conditions.

Table 2 R^2 scores of the three model configurations

Sample size	No attention	Global attention	Local attention
900	0.948	0.944	0.972
1000	0.931	0.941	0.983
1100	0.920	0.937	0.973
1200	0.970	0.955	0.977
1300	0.963	0.969	0.970
1400	0.966	0.963	0.970

The relative advantage of local attention is therefore more pronounced for the 900–1100 sample datasets and decreases as the available dataset becomes larger, consistent with the RMSE results. Because R^2 summarizes agreement over the complete response history, it does not independently characterize error concentration or local agreement around response peaks. The R^2 results are therefore considered together with the signed-error distributions and response-history comparisons presented below.

The signed prediction errors are further examined using the equal-width intervals listed in Table 3 and the corresponding distributions shown in Figs. 8–13. For each sample-size condition, the range of observed signed prediction errors was divided into six equal-width intervals (Bins 1–6), ordered from the most negative to the most positive error. The same interval boundaries were used for all three model configurations within each sample-size condition, enabling consistent comparisons of error direction and concentration. Because the intervals were defined separately for each sample-size condition, they are intended only for within-condition comparisons and should not be interpreted as common engineering acceptance thresholds. Errors closer to zero indicate smaller prediction deviations; therefore, a larger proportion of errors in the interval or intervals nearest zero indicates a more concentrated prediction-error distribution. Bins 1–5 are defined as left-closed and right-open, whereas Bin 6 includes both endpoints. All interval boundaries are rounded to two decimal places. In Figs. 8–13, the horizontal-axis values (1–6) correspond to the bin numbers defined in Table 3.

Table 3 Equal-width signed prediction error intervals (mm) for different sample-size conditions

Bin no.	Sample size					
	900	1000	1100	1200	1300	1400
1	[-5.50, -4.08)	[-10.00, -5.33)	[-7.00, -5.00)	[-12.00, -8.50)	[-6.01, -3.84)	[-12.00, -7.00)
2	[-4.08, -2.67)	[-5.33, -0.67)	[-5.00, -3.00)	[-8.50, -5.00)	[-3.84, -1.67)	[-7.00, -2.00)
3	[-2.67, -1.25)	[-0.67, 4.00)	[-3.00, -1.00)	[-5.00, -1.50)	[-1.67, 0.50)	[-2.00, 3.00)
4	[-1.25, 0.17)	[4.00, 8.67)	[-1.00, 1.00)	[-1.50, 2.00)	[0.50, 2.66)	[3.00, 8.00)
5	[0.17, 1.58)	[8.67, 13.33)	[1.00, 3.00)	[2.00, 5.50)	[2.66, 4.83)	[8.00, 13.00)
6	[1.58, 3.00]	[13.33, 18.00]	[3.00, 5.00]	[5.50, 9.00]	[4.83, 7.00]	[13.00, 18.00]

Figs. 8–13 compare the local-attention predictions with the global-attention and no-attention predictions using the same reference displacement history. The enlarged windows highlight representative peaks, rapidly varying segments, low-amplitude regions, and visible temporal deviations. Because all six figures use the same layout, the following discussion focuses on the principal differences associated with each sample-size condition.

For the 900-sample dataset, the local-attention errors are concentrated primarily in the two bins nearest zero, whereas the global-attention errors are distributed across four bins and the no-attention errors are dispersed across all six bins. As shown in Fig. 8, all three configurations reproduce the response trend. However, the local-attention prediction follows several local extrema and rapidly varying segments more closely and shows reduced visible temporal deviation in the enlarged windows.

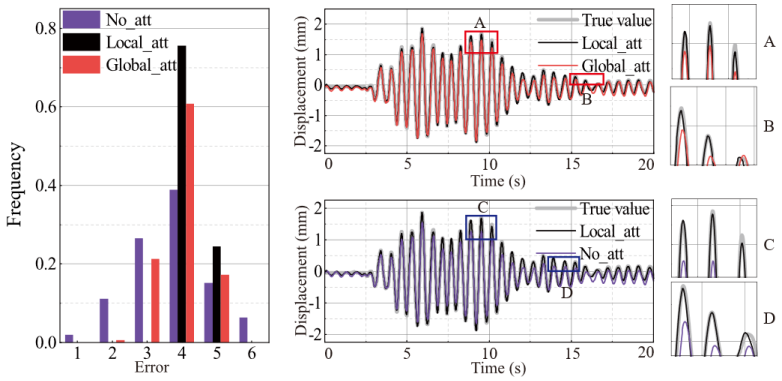


Fig. 8 Comparison of prediction errors and displacement histories for three attention mechanisms (900 samples)

For the 1000-sample dataset, more than 98% of the local-attention errors fall within the two bins nearest zero. Although the global-attention and no-attention models reproduce several large-response peaks, their deviations are more evident in portions of the response with smaller amplitudes. Together with the R^2 value of 0.983 and the minimum RMSE of 0.00218, the concentrated error distribution indicates the strongest combined performance for the local-attention model under this sample-size condition.

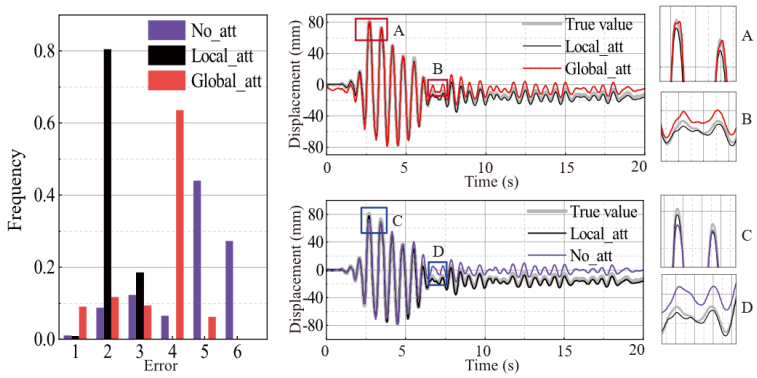


Fig. 9 Comparison of prediction errors and displacement histories for three attention mechanisms (1000 samples)

For the 1100-sample dataset, 90% of the local-attention errors occur in the bin nearest zero, whereas the errors of the other two configurations are more widely distributed. The enlarged windows in Fig. 10 show that the global-attention model overestimates several response segments and that the no-attention model underestimates some local extrema. The local-attention prediction remains closer to the reference history in these selected regions.

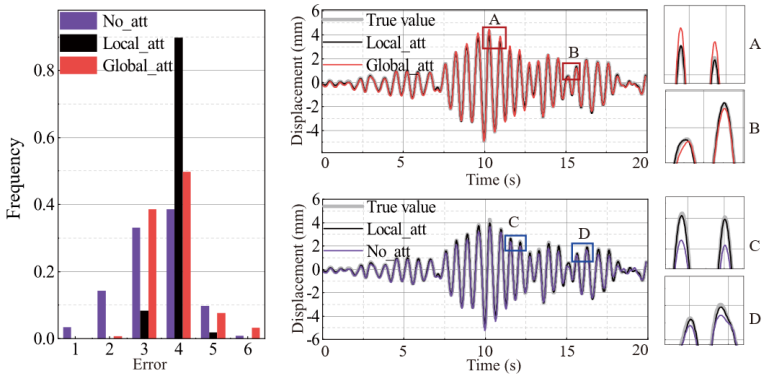


Fig. 10 Comparison of prediction errors and displacement histories for three attention mechanisms (1100 samples)

For the 1200-sample dataset, all three models reproduce the overall response history with relatively high and more comparable accuracy. Nevertheless, 72% of the local-attention errors remain in the bin closest to zero. The enlarged windows in Fig. 11 also show closer local agreement around several peaks and low-amplitude segments. The advantage of local attention therefore remains observable, although it is smaller than that obtained for the 900–1100 sample datasets.

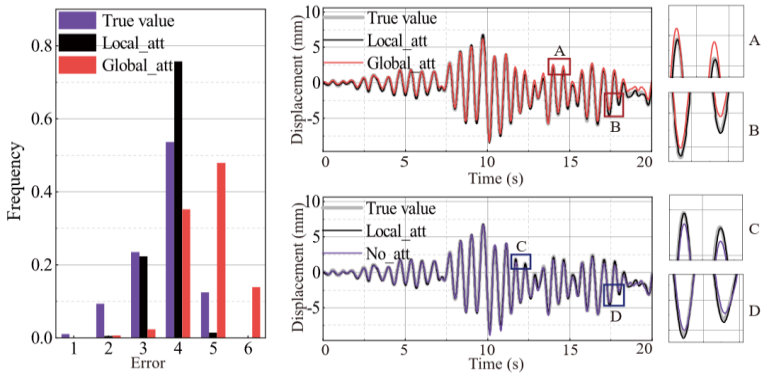


Fig. 11 Comparison of prediction errors and displacement histories for three attention mechanisms (1200 samples)

For the 1300-sample dataset, 83% of the local-attention errors fall within the two bins nearest zero, compared with approximately 72% for the no-attention model and 71% for the global-attention model. The predicted response histories in Fig. 12 are visually similar at the global scale. At this sample size, the benefit of local attention is therefore expressed mainly through a more concentrated error distribution rather than through a substantial difference in the overall response-history shape.

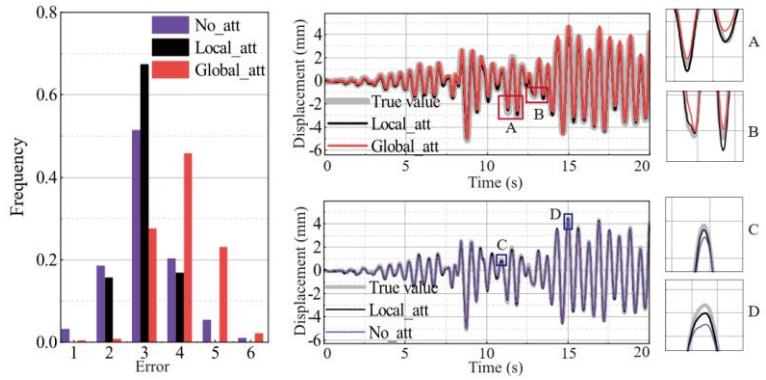


Fig. 12 Comparison of prediction errors and displacement histories for three attention mechanisms (1300 samples)

For the 1400-sample dataset, the three configurations again exhibit high overall agreement with the reference response. The local-attention model places 75% of its errors in the bin nearest zero, compared with approximately 65% for the global-attention model and 64% for the no-attention model. Fig. 13 shows similar global response trends, while the local-attention prediction provides closer agreement around several selected response peaks.

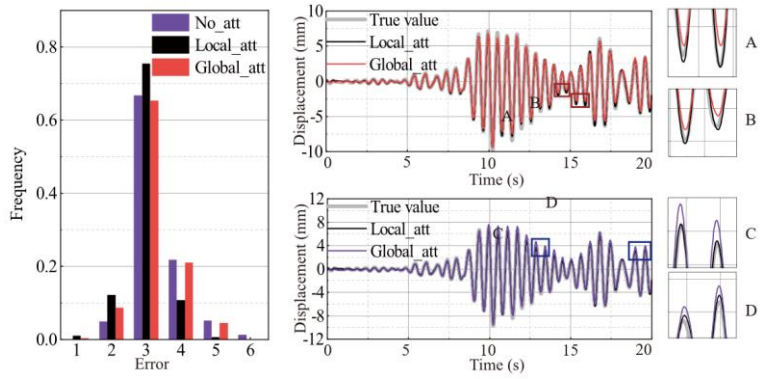


Fig. 13 Comparison of prediction errors and displacement histories for three attention mechanisms (1400 samples)

Taken together, the RMSE, R^2, signed-error, and response-history results provide complementary evidence. Across all six sample-size conditions, the local-attention model achieves the lowest minimum RMSE, maintains test-set R^2 values of at least 0.970, and produces the most concentrated signed-error distributions. Its relative advantage is most evident for the smaller datasets and decreases as the available sample size increases. The enlarged response-history windows further show closer

agreement around several local peaks and rapidly varying segments. These results collectively support the effectiveness of local temporal feature reweighting within the unified LSTM framework and the investigated multi-intensity IDA sample space.

3.3 Attention-based interpretability analysis

In this study, attention-based interpretability refers to the transparency and inspectability of temporal feature allocation within the trained network. The learned attention weights show how the two local-attention layers distribute relative internal emphasis across hidden-feature channels and time steps. This model-level interpretation does not treat an attention weight as direct evidence that a particular input interval physically or causally determines the structural response. Instead, the analysis focuses on the temporal organization of the learned weights, the differences between the two attention layers, and their qualitative correspondence with observable waveform characteristics.

The heatmaps contain 32 hidden-feature channels and 2000 time steps sampled at an interval of 0.01 s. Brighter colors represent relatively larger attention weights, whereas darker colors represent smaller weights. The maps therefore visualize the internal allocation of temporal features produced by the trained network and are interpreted as representations of hierarchical feature processing rather than as direct physical-importance measures.

Fig. 14 presents a record with an initial low-amplitude stage, a pronounced high-amplitude stage at approximately 9–11 s, and a subsequent decay stage. The first local-attention layer assigns relatively high weights to multiple hidden representations around the concentrated high-amplitude region and exhibits alternating weight patterns during the subsequent oscillatory interval. This visual correspondence suggests that the first layer is sensitive to prominent amplitude changes and rapidly varying temporal features.

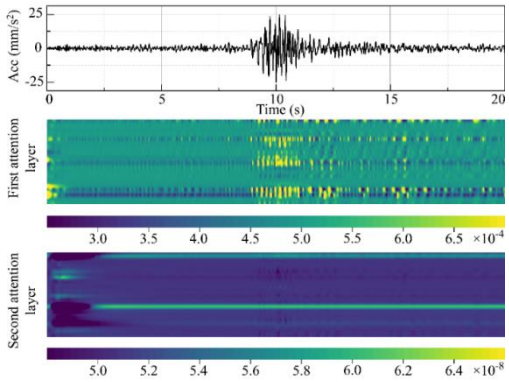


Fig. 14 Seismic waveform and attention weights

The second local-attention layer exhibits a clearly different distribution. The high-amplitude interval is not emphasized in the same manner as in the first layer, while several feature channels retain comparatively persistent weights over other portions of the record. The second layer therefore does not merely reproduce the first-layer allocation; instead, it reorganizes the intermediate representation and assigns emphasis to complementary temporal feature patterns.

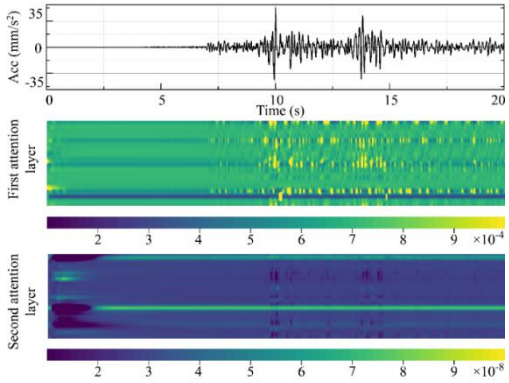


Fig. 15 Seismic waveform and attention weights

Fig. 15 presents a record containing two main stages of increasing and decreasing acceleration amplitude. The first local-attention layer assigns relatively strong weights around both higher-amplitude regions. By contrast, the second layer exhibits a more distributed allocation and retains weight in portions that are not dominated by the largest acceleration values. The difference between the two maps again indicates progressive feature processing: the first layer emphasizes prominent local variations, whereas the second layer refines the resulting representation and preserves additional information outside the dominant peaks.

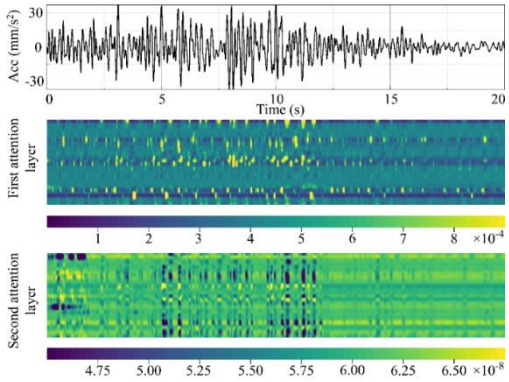


Fig. 16 Seismic waveform and attention weights

Fig. 16 presents a record for which the acceleration amplitude remains large over much of the 20-s duration. The first local-attention layer contains multiple high-weight regions during approximately the first 13 s, followed by fewer high-weight regions as the acceleration amplitude decreases. The second layer again produces a different allocation pattern, with reduced emphasis on several peak-dominated portions and greater continuity across other feature channels and time intervals.

Across the three representative records, the first and second local-attention layers consistently exhibit distinct but complementary weight structures. The first layer is more closely associated with prominent amplitude changes and local temporal variations, whereas the second layer reorganizes and refines the intermediate representation while retaining information beyond the dominant peaks. This differentiation between successive attention layers provides a clearer view of the internal feature-processing procedure relative to a recurrent representation without explicit attention-based reweighting. The results therefore support attention-based interpretability within the model-level definition adopted in this study.

3.4 Controlled comparison, study scope, and limitations

The central research question of this study is how the temporal scope of attention influences seismic-response prediction within a common multilayer LSTM framework. Accordingly, the no-attention, global-attention, and local-attention configurations use the same general recurrent backbone, hidden-state dimension, recurrent depth, data-partitioning procedure, and training protocol. Their primary architectural difference is the presence and temporal range of attention-based feature reweighting. This controlled design enables the observed performance differences to be attributed primarily to differences in attention scope.

Comparisons involving heterogeneous recurrent, convolutional, encoder–decoder, or Transformer-based architectures would address the broader question of backbone selection. Such comparisons would simultaneously introduce differences in state representation, receptive-field construction, parameterization, and training behavior. They are therefore complementary to, rather than substitutes for, the present comparison, which is intentionally designed to examine local, global, and absent attention within a unified recurrent framework.

The IDA database uses PGA as a consistent intensity-scaling coordinate, while the network input consists of the complete ground-acceleration history. Consequently, waveform shape, duration, frequency content, pulse characteristics, and other temporal variations remain available to the model even though PGA defines the intensity sequence. In addition, the same peak-

centered alignment procedure is applied to all three configurations. The input length, retained excitation region, and preprocessing conditions are therefore identical across the models, maintaining the consistency of the attention-scope comparison.

The adopted sample-wise partition evaluates surrogate prediction within the multi-intensity IDA sample space constructed from the selected ground-motion records. It should therefore be interpreted as an assessment of the excitation–response mapping under the investigated data organization rather than as a strict test of generalization to entirely unseen original ground-motion records. Within this evaluation setting, the repeated advantages observed in RMSE, R^2, signed-error concentration, and local response-history agreement provide consistent support for the effectiveness of local temporal attention.

The numerical gantry-crane model provides a controlled reference system for generating nonlinear displacement histories and evaluating the three learning configurations under identical structural conditions. The conclusions should therefore be interpreted within the context of the investigated simulated structure, model architecture, and IDA sample space. Future studies may broaden the validation domain through record-wise evaluation, additional crane and structural configurations, measured response data, alternative preprocessing procedures, and other sequence-model backbones. Such extensions would assess broader generalizability while preserving the controlled comparative framework established in the present study.

4. Conclusions

This study developed a multilayer LSTM model incorporating local temporal attention for predicting the simulated seismic displacement history of a gantry crane. The local-attention model was compared with global-attention and no-attention configurations using the same general recurrent backbone, data partitions, hidden-state dimension, recurrent depth, and training protocol, enabling the influence of attention scope to be examined under consistent evaluation conditions. The principal findings are summarized as follows:

(1) Across all six sample-size conditions, it attained the lowest minimum RMSE, maintained test-set R^2 values of at least 0.970, and produced more concentrated signed-error distributions than the global-attention and no-attention models. The maximum test-set R^2 reached 0.983.

(2) The benefit of local attention was more pronounced for the 900–1100 sample datasets, including an R^2 advantage of approximately 0.04 over the reference configurations in selected cases. Differences became smaller for the 1200–1400 sample datasets, indicating that explicit local temporal reweighting is particularly beneficial when the available database is relatively limited.

(3) The enlarged response-history comparisons showed closer overall agreement with reference responses around several peaks and rapidly varying segments, while the signed-error distributions exhibited greater concentration near zero. These results complement global accuracy metrics by providing additional insight into local response agreement and error concentration.

(4) The two local-attention layers exhibited distinct but complementary temporal allocation patterns. The first attention layer emphasized prominent amplitude changes and local temporal variations, whereas the second layer reorganized and refined the intermediate representations while retaining information beyond the dominant peaks. These complementary patterns improve the interpretability of temporal feature processing within the trained network, although the attention weights are not interpreted as direct evidence of physical causality.

(5) The adopted IDA database preserved complete ground-motion histories as network inputs while using PGA as a consistent intensity-scaling coordinate, enabling evaluation across multiple seismic-intensity levels under identical preprocessing conditions.

Overall, within the investigated numerical gantry-crane and multi-intensity IDA setting, the proposed local-attention framework improves prediction accuracy, local response representation, and the transparency of temporal feature allocation. Future studies may consider record-wise partitions, other structures, measured data, alternative backbones, and attribution analyses.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Nos. 52178169 and NSF12301467). The authors would like to express their sincere gratitude to the sponsors for their financial support.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

[1] X. Lu, L. Xie, H. Guan, Y. Huang, and X. Lu, "A Shear Wall Element for Nonlinear Seismic Analysis of Super-Tall Buildings Using OpenSees," Finite Elements in Analysis and Design, vol. 98, pp. 14-25, 2015.

[2] S.-Y. Zhai, Y.-F. Lyu, K. Cao, G.-Q. Li, W.-Y. Wang, and C. Chen, "Seismic Behavior of an Innovative Bolted Connection with Dual-Slot Hole for Modular Steel Buildings," Engineering Structures, vol. 279, article no. 115619, 2023.

[3] S. Mangalathu, H. Jang, S.-H. Hwang, and J.-S. Jeon, "Data-Driven Machine-Learning-Based Seismic Failure Mode Identification of Reinforced Concrete Shear Walls," Engineering Structures, vol. 208, article no. 110331, 2020.

[4] H. Jia, L. Ma, B. Jiang, Y. Zhai, Y. Li, and S. Zheng, "Life-Cycle Fragility Analysis of Near-Fault Simply Supported and T-Shaped Girder Bridges Considering Corrosion of Piers," Structures, vol. 86, article no. 111477, 2026.

[5] H. Jia, J. Hou, H. Bai, Z. Xu, K. Jia, and S. Zheng, "Coupled Effects of Corrosion and Fault-Crossing Ground Motions on Continuous Rigid Frame Bridges: Nonlinear Dynamic Response and Failure Mechanisms," Steel and Composite Structures, vol. 59, no. 5, pp. 609-629, 2026.

[6] Q. Peng, W. Cheng, H. Jia, and P. Guo, "Fragility Analysis of Gantry Crane Subjected to Near-Field Ground Motions," Applied Sciences, vol. 10, no. 12, article no. 4219, 2020.

[7] Q. Peng, W. Cheng, H. Jia, D. Guo, G. You, and C. Zhang, "Seismic Analysis of Uplift-Available Gantry Crane Subjected to Extreme Earthquake Loads," Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería, vol. 37, no. 3, article no. 33, 2021.

[8] Q. Peng, W. Cheng, P. Guo, and H. Jia, "Assessing Seismic Performance of Gantry Crane Subjected to Near-Field Ground Motions Using Incremental Dynamic and Endurance Time Analysis Methods," Shock and Vibration, vol. 2022, no. 1, article no. 6624530, 2022.

[9] H. Jia, Z. Liu, L. Xu, H. Bai, K. Bi, C. Zhang, et al., "Dynamic Response Analyses of Long-Span Cable-Stayed Bridges Subjected to Pulse-Type Ground Motions," Soil Dynamics and Earthquake Engineering, vol. 164, article no. 107591, 2023.

[10] H. Jia, W. Wu, L. Xu, Y. Zhou, S. Zheng, and C. Zhao, "Numerical Simulation and Damaged Analysis of a Simply-Supported Beam Bridge Crossing Potential Active Fault," Engineering Structures, vol. 301, article no. 117283, 2024.

[11] H. Jia, S. Chen, D. Guo, S. Zheng, and C. Zhao, "Track-Bridge Deformation Relation and Interaction of Long-Span Railway Suspension Bridges Subject to Strike-Slip Faulting," Engineering Structures, vol. 300, article no. 117216, 2024.

[12] N. G. Wariyatno, A. L. Han, Y. Haryanto, G. H. Sudibyo, Sumiyanto, Nastain, et al., "Formulating Seismic Intensity Scale (JMA-SIS) Using Response Spectrum: A New Approach for Structural Engineering Design," Advances in Technology Innovation, vol. 10, no. 2, pp. 157-173, 2025.

[13] Y. Hu, H. H. Tsang, N. Lam, and E. Lumantarna, "Physics-Informed Neural Networks for Enhancing Structural Seismic Response Prediction with Pseudo-Labelling," Archives of Civil and Mechanical Engineering, vol. 24, no. 1, article no. 7, 2024.

[14] Q. Peng, W. Cheng, H. Jia, P. Guo, and K. Jia, "Rapid Seismic Damage Assessment Using Machine Learning Methods: Application to a Gantry Crane," Structure and Infrastructure Engineering, vol. 19, no. 6, pp. 779-792, 2023.

[15] E. A. Moscoso Alcantara, M. D. Bong, and T. Saito, "Structural Response Prediction for Damage Identification Using Wavelet Spectra in Convolutional Neural Network," Sensors, vol. 21, no. 20, article no. 6795, 2021.

[16] Q. Peng, X. Wen, H. Jia, Y. Pan, X. Gu, C. Yin, et al., "A Novel Sensor-Independent Convolutional Neural Network for Structural Damage Detection: Illustrated by a Case Study on Gantry Crane," Structures, vol. 71, article no. 107971, 2025.

[17] H. S. Park, S. H. Yoo, D. Y. Yun, and B. K. Oh, "Investigation on Employment of Time and Frequency Domain Data for Predicting Nonlinear Seismic Responses of Structures," Structures, vol. 61, article no. 105996, 2024.

[18] F. Guo, Y. Dong, H. Tian, X. Zhang, and Q. Su, "Structural Seismic Response Prediction Based on Convolutional Neural Networks," Vibroengineering Procedia, vol. 51, pp. 56-62, 2023.

[19] H. Zhao, B. Wei, P. Zhang, P. Guo, Z. Shao, S. Xu, et al., "Safety Analysis of High-Speed Trains on Bridges under Earthquakes Using a LSTM-RNN-Based Surrogate Model," Computers & Structures, vol. 294, article no. 107274, 2024.

[20] Y. Shen, G. Ma, H. J. Hwang, D. J. Kim, and Z. Zhang, "Prediction of Seismic Response of Building Structures Using a CNN-LSTM-ATT Network with Transfer Learning," Advances in Structural Engineering, vol. 28, no. 14, pp. 2710-2725, 2025.

[21] X. Zhang, X. Xie, S. Tang, H. Zhao, X. Shi, L. Wang, et al., "High-Speed Railway Seismic Response Prediction Using CNN-LSTM Hybrid Neural Network," Journal of Civil Structural Health Monitoring, vol. 14, no. 5, pp. 1125-1139, 2024.

[22] Y. Liao, R. Lin, R. Zhang, and G. Wu, "Attention-Based LSTM (AttLSTM) Neural Network for Seismic Response Modeling of Bridges," Computers & Structures, vol. 275, article no. 106915, 2023.

[23] T. Saida and M. Nishio, "ExSRNet: Explainable Deep Learning Model for Seismic Response Prediction with Frequency Attention Mechanism," Engineering Structures, vol. 343, part A, article no. 120953, 2025.

[24] W. J. Tang, D. S. Wang, H. B. Huang, J. C. Dai, and F. Shi, "A Pre-Trained Deep Learning Model for Fast Online Prediction of Structural Seismic Responses," International Journal of Structural Stability and Dynamics, vol. 24, no. 14, article no. 2450160, 2024.