Earthquake-Bench: Video Generation Benchmark for Earthquake Simulation

Supplementary Material

Contents

A. Seismic Wave Processing

In this study, seismic acceleration records are first obtained from strong-motion observation networks, such as the PEER NGA-West2 database and the USGS Strong Motion Archive. The raw records are typically represented as discrete acceleration time histories $\{a_i\}_{i=1}^{N}$ with a corresponding sampling interval Δt_0. As field measurements may contain instrument drift, low-frequency noise, or inconsistent sampling rates, systematic preprocessing is required.

Specifically, baseline correction is performed using trend removal or polynomial fitting. The baseline trend is approximated as

$$p(t) = \sum_{k=0}^{m} c_k t^k, \tag{1}$$

where $p(t)$ denotes the fitted polynomial trend. This trend is then subtracted from the original signal to obtain the corrected acceleration:

$$a_{\text{corr}}(t) = a(t) - p(t), \tag{2}$$

thereby effectively eliminating cumulative drift.

Subsequently, to ensure a uniform temporal resolution, the original acceleration series is resampled to a target time step Δt, typically using linear interpolation:

$$a(t) = a(t_i) + \frac{a(t_{i+1}) - a(t_i)}{t_{i+1} - t_i}(t - t_i), \quad t_i \le t \le t_{i+1}, \tag{3}$$

resulting in a consistent and evenly sampled acceleration sequence.

To match a target seismic intensity, peak normalization or amplitude scaling is applied. Given a target peak ground acceleration PGA_{tar}, the scaling factor is defined as

$$\alpha = \frac{\text{PGA}_{\text{tar}}}{\max |a_{\text{corr}}(t)|}, \tag{4}$$

and the scaled ground motion is expressed as

$$a_{\text{scaled}}(t) = \alpha\, a_{\text{corr}}(t). \tag{5}$$

The processed acceleration is stored as a single-column time series with a uniform time step Δt, which serves as the external excitation input in numerical simulation platforms such as OpenSees. The structural response is governed by the dynamic equilibrium equation:

$$\mathbf{M}\ddot{\mathbf{u}}(t) + \mathbf{C}\dot{\mathbf{u}}(t) + \mathbf{K}\mathbf{u}(t) = -\mathbf{M}\mathbf{r}\, a_g(t), \tag{6}$$

where $\mathbf{M}$, $\mathbf{C}$, and $\mathbf{K}$ denote the mass, damping, and stiffness matrices, respectively, and $\mathbf{r}$ is the influence vector.

Under the excitation of the input ground motion, the structure produces displacement, velocity, and acceleration responses. The final outputs include time histories of key node displacements, inter-story drifts, acceleration responses, internal forces, and hysteretic behavior. These results enable verification of the effectiveness of the input seismic motion and provide a reliable basis for subsequent dynamic response analysis and structural performance evaluation.

B. Detailed Evaluation Metrics

B.1. Structure-Level Displacement Difference (D-Structure)

This metric extracts the overall displacement trajectory of structures from videos using optical flow methods and com-

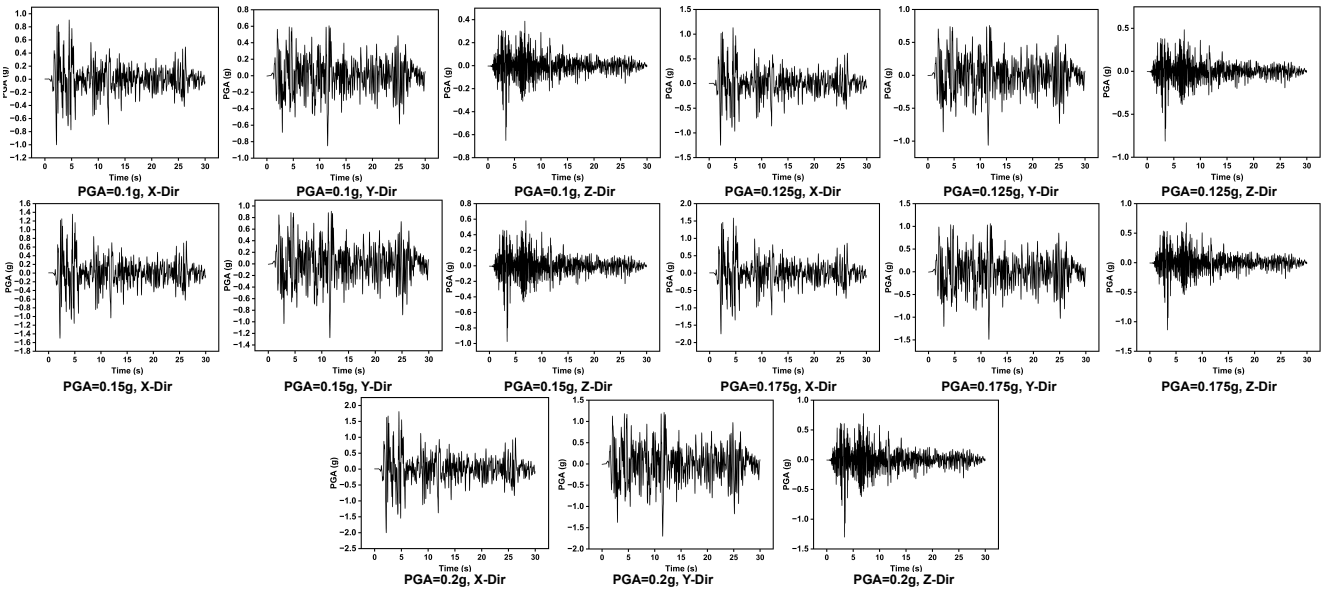

Figure 1. Directional ground-motion acceleration time histories under varying PGA levels.

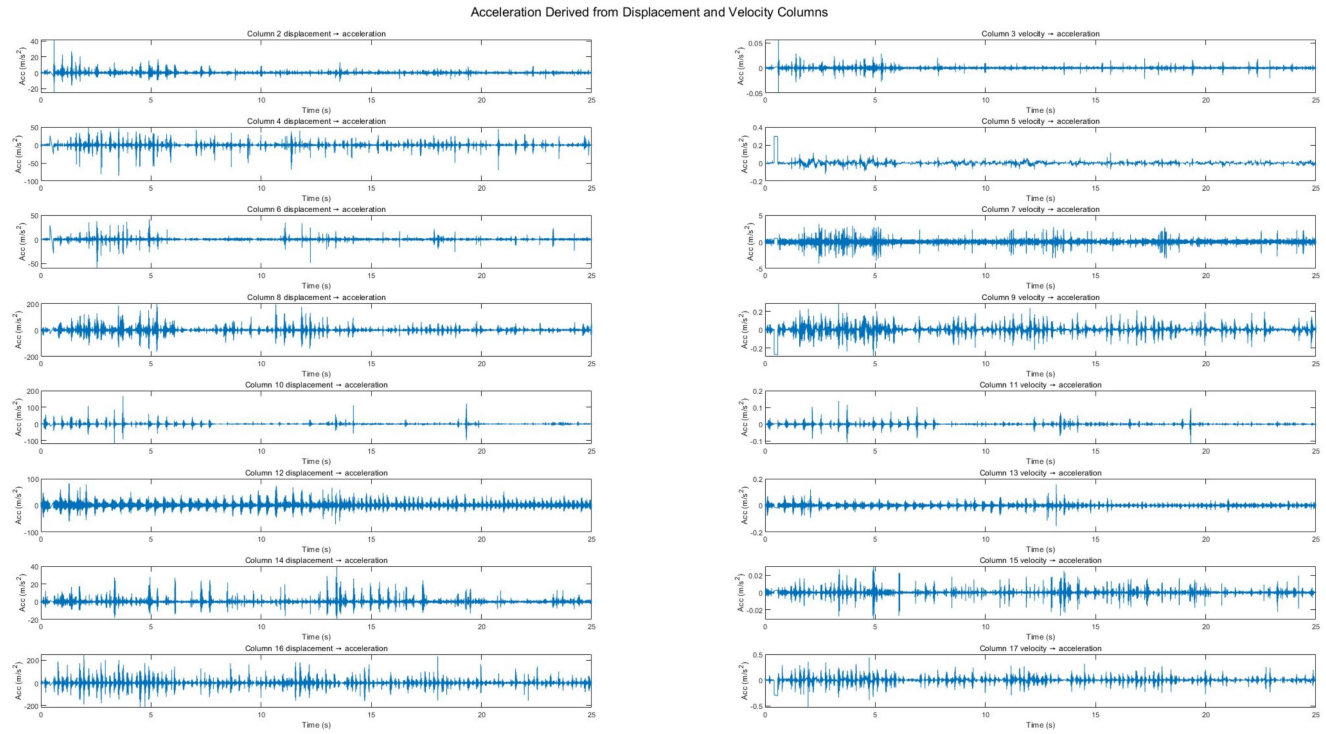

Figure 2. Acceleration Derived from Displacement and Velocity Columns

pares it with sensor measurements to comprehensively evaluate the physical consistency of generated videos at the structural scale.

B.1.1. Optical Flow Extraction and Displacement Calculation

Given a generated video sequence $V_g = \{I_t^g\}_{t=1}^T$ and ground truth video $V_{gt} = \{I_t^{gt}\}_{t=1}^T$, we compute dense optical flow fields using the Farnebäck algorithm. For con-

secutive frames I_t and I_{t+1}, the optical flow field $\mathbf{F}_t \in \mathbb{R}^{H\times W\times 2}$ represents pixel-wise displacement:

$$\mathbf{F}_t(x,y) = [u_t(x,y), v_t(x,y)]^\top, \tag{7}$$

where u_t and v_t denote the horizontal and vertical displacement components at pixel location (x, y), respectively. The flow magnitude is:

$$m_t(x,y) = \sqrt{u_t(x,y)^2 + v_t(x,y)^2}, \tag{8}$$

We extract structural displacement trajectories by spatially averaging the optical flow within the segmented structure region $\mathcal{R}_t \subset \mathbb{R}^{H\times W}$:

$$\bar{\mathbf{F}}_t = \frac{1}{|\mathcal{R}_t|} \sum_{(x,y)\in\mathcal{R}_t} \mathbf{F}_t(x,y), \tag{9}$$

where $|\mathcal{R}_t|$ denotes the region area in pixels. The cumulative displacement trajectory is:

$$\mathbf{D}_t = \sum_{i=1}^{t} \bar{\mathbf{F}}_i, \quad t = 1,\ldots,T, \tag{10}$$

B.1.2. Sensor Data Alignment

Shake table accelerometers record three-axis acceleration $\mathbf{a}(t) = [a_x(t), a_y(t), a_z(t)]^\top$ at sampling frequency $f_s = 1000$ Hz. To align with the video frame rate $f_v = 30$ fps, we perform temporal downsampling via double integration:

$$\mathbf{d}_k = \iint_{t_k}^{t_{k+1}} \mathbf{a}(t)\, dt^2, \quad k = 1,\ldots,T, \tag{11}$$

where $t_k = k/f_v$ denotes the timestamp of the k-th frame. The camera projection matrix $\mathbf{P} \in \mathbb{R}^{3\times 3}$ maps 3D displacement to 2D pixel coordinates:

$$\mathbf{d}_{2D}^k = \mathbf{P}\mathbf{d}_k, \tag{12}$$

B.1.3. Normalized Error and Weighted Scoring

The structure-level displacement difference between generated and ground truth videos is computed as:

$$\text{D-Structure}_{\text{video}} = 1 - \frac{1}{T}\sum_{t=1}^{T} \frac{\|\mathbf{D}_t^g - \mathbf{D}_t^{gt}\|_2}{\max(\|\mathbf{D}_t^{gt}\|_2, \epsilon)}, \tag{13}$$

where $\epsilon = 1e^{-6}$ prevents division by zero for static frames. We additionally compute the correlation with sensor measurements:

$$\rho_{\text{sensor}} = \text{Corr}(\mathbf{D}^g, \mathbf{d}_{2D}), \tag{14}$$

where $\text{Corr}(\cdot)$ denotes the Pearson correlation coefficient.

The final metric integrates four sub-dimensions: correlation r (40%), phase alignment PhaseCorr (30%), coherence Coherence (20%), and amplitude similarity Amplitude Similarity (10%), combined as:

$$\mathcal{E}_{\text{phys}} = \sum_{i\in\{r,p,c,a\}} w_i\, \phi_i \tag{15}$$

where w_i denotes the predefined weight associated with each component, satisfying $\sum_i w_i = 1$; ϕ_i represents the corresponding physics-aware similarity measure, including r (Pearson correlation coefficient), p (phase alignment score), c (frequency-domain coherence), and a (amplitude similarity). where r represents the Pearson correlation coefficient between optical flow displacement and sensor displacement; PhaseCorr measures the degree of phase alignment through frequency domain analysis, computed as the normalized peak of the cross power spectrum between the two signals; Coherence quantifies frequency domain structural response similarity, defined as the average of $\gamma^2(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f)P_{yy}(f)}$ over the effective frequency band; Amplitude Similarity evaluates displacement magnitude consistency, computed as $1 - \frac{|\max|\mathbf{D}^g| - \max|\mathbf{D}^{gt}||}{\max|\mathbf{D}^{gt}|}$.

B.2. Pixel-Level Edge Difference ($\mathbf{D}_{\text{Edge-Pixel}}$)

This metric evaluates deformation consistency at pixel-level detail in generated videos through multi-scale feature point tracking and edge detection, while integrating four dimensions: segmentation consistency, physical consistency, temporal consistency, and perceived quality.

B.2.1. Feature Point Selection Strategy

We employ a hybrid strategy combining structural vertices and edge-based interest points. First, corner vertices are identified using the Shi-Tomasi detector:

$$\mathcal{C} = \{(x_i, y_i) : \min(\lambda_1, \lambda_2) > \tau\}, \tag{16}$$

where λ_1, λ_2 are eigenvalues of the structure tensor and $\tau = 0.01$ is the quality threshold. We augment corner points with uniformly sampled edge points along detected structural boundaries, yielding a total of $N_p = 50$ tracked points per frame.

B.2.2. Multi-Scale Edge Detection

We apply multi-scale Canny edge detection with the following procedure: (1) Gaussian blur ($\sigma = 1.4$) for noise reduction; (2) Sobel gradient computation: $G_x = \frac{\partial I}{\partial x}$, $G_y = \frac{\partial I}{\partial y}$; (3) Gradient magnitude: $G = \sqrt{G_x^2 + G_y^2}$; (4) Non-maximum suppression along gradient direction; (5) Double thresholding: $T_{\text{low}} = 50$, $T_{\text{high}} = 150$; (6) Edge tracking by hysteresis. The resulting edge map $E_t \in \{0,1\}^{H\times W}$ is morphologically dilated with a 3×3 kernel to form a tracking mask.

B.2.3. Pixel-Level Displacement Tracking and Pyramid Fusion

For each feature point $\mathbf{p}_i^t = (x_i^t, y_i^t)$, we compute its displacement vector using Lucas-Kanade optical flow:

$$\mathbf{v}_i^t = \mathbf{p}_i^{t+1} - \mathbf{p}_i^t, \tag{17}$$

Trajectories with tracking confidence below $\alpha = 0.8$ are discarded. The displacement sequence for the i-th point is:

$$\mathbf{V}_i = [\mathbf{v}_i^1, \mathbf{v}_i^2, \ldots, \mathbf{v}_i^{T-1}], \tag{18}$$

We construct a three-level Gaussian pyramid with scaling factors $\{1, 0.5, 0.25\}$. At each scale s, we extract displacement trajectories $\mathbf{V}_i^{(s)}$ and compute scale-weighted displacement:

$$\tilde{\mathbf{V}}_i = \sum_{s=1}^{3} w_s \cdot \mathbf{V}_i^{(s)}, \tag{19}$$

where weights $w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$ emphasize finer scales.

B.2.4. Four-Dimensional Comprehensive Evaluation

This metric integrates four core sub-dimensions:

Shape Similarity (35%) Computed based on pixel-level edge differences:

$$\text{Sims} = 1 - \frac{1}{N_p} \sum_{i=1}^{N_p} \frac{\|\tilde{\mathbf{V}}_i^g - \tilde{\mathbf{V}}_i^{gt}\|_F}{\|\tilde{\mathbf{V}}_i^{gt}\|_F + \epsilon}, \tag{20}$$

Temporal Consistency (25%) Evaluated through pixel-level mean absolute error (MAE) sequences between adjacent frames:

$$\text{MAE}_t = \frac{1}{N} \sum_{i=1}^{N} |F_t(i) - F_{t+1}(i)|, \tag{21}$$

The stability score is computed as:

$$S = \frac{255 - \overline{\text{MAE}}}{255}, \tag{22}$$

where $\overline{\text{MAE}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \text{MAE}_t$.

Amplitude Matching (20%) Evaluates displacement peak consistency:

$$\mathcal{A}_{\text{amp}} = 1 - \frac{1}{N_p} \sum_{i=1}^{N_p} \frac{|\max \|\tilde{\mathbf{V}}_i^g\| - \max \|\tilde{\mathbf{V}}_i^{gt}\||}{\max \|\tilde{\mathbf{V}}_i^{gt}\|}, \tag{23}$$

Data Quality (20%) Comprehensively assesses tracking success rate, edge detection robustness, and feature point distribution uniformity:

$$\text{Data Quality} = 0.5 \cdot \frac{N_{\text{tracked}}}{N_p} + 0.3 \cdot \text{EdgeCoverage} + 0.2 \cdot \text{SpatialUniformity}, \tag{24}$$

where N_{tracked} is the number of successfully tracked points, EdgeCoverage is the edge coverage ratio, and SpatialUniformity is quantified by the inverse of feature point Voronoi region variance.

Weighted Combination The final pixel-level edge difference metric is:

$$\begin{aligned} \text{D}_{\text{Edge-Pixel}} = {} & 0.35 \cdot \text{Shape Similarity} \\ & + 0.25 \cdot \text{Temporal Consistency} \\ & + 0.20 \cdot \text{Amplitude Matching} \\ & + 0.20 \cdot \text{Data Quality}, \end{aligned} \tag{25}$$

B.3. Segment-Level Structure Comparison (SAM-Structure)

This metric utilizes SAM2 (Segment Anything Model 2) for frame-by-frame semantic segmentation to evaluate spatial alignment, temporal continuity, contour stability, and segmentation quality by comparing structural masks between generated and ground truth videos.

B.3.1. SAM2 Segmentation Configuration

We employ the following configuration for segmentation: model checkpoint `sam2_hiera_large`; prompt type based on structure centroid points; multi-object mode disabled (single structure per frame); confidence threshold $\tau_{\text{conf}} = 0.85$. For temporal consistency, we propagate the initial mask using the SAM2 video predictor:

$$\mathcal{M}_t = \text{SAM2-Propagate}(\mathcal{M}_{t-1}, I_t, \mathbf{h}_{t-1}), \tag{26}$$

where $\mathbf{h}_{t-1}$ is the hidden state from the previous frame.

B.3.2. IoU Mathematical Definition and Bidirectional Evaluation

For binary masks $\mathcal{M}_t^g$ (generated) and $\mathcal{M}_t^{gt}$ (ground truth) at frame t, the Intersection over Union (IoU) is defined as:

$$\text{IoU}_t = \frac{|\Omega_t^{\text{pred}} \cap \Omega_t^{\text{gt}}|}{|\Omega_t^{\text{pred}} \cup \Omega_t^{\text{gt}}|}, \tag{27}$$

where $|\cdot|$ denotes the number of pixels. The temporal average IoU is:

$$\text{IoU}_{\text{temporal}} = \frac{1}{T} \sum_{t=1}^{T} \text{IoU}_t, \tag{28}$$

To ensure symmetry, we perform bidirectional evaluation:

$$\begin{aligned}\mathrm{IoU}_{\mathrm{fwd}} &= \frac{1}{T}\sum_{t=1}^{T}\frac{|\Omega_t^{\mathrm{pred}} \cap \Omega_t^{\mathrm{gt}}|}{|\Omega_t^{\mathrm{gt}}|},\\ \mathrm{IoU}_{\mathrm{bwd}} &= \frac{1}{T}\sum_{t=1}^{T}\frac{|\Omega_t^{\mathrm{pred}} \cap \Omega_t^{\mathrm{gt}}|}{|\Omega_t^{\mathrm{pred}}|},\\ \mathrm{IoU}_{\mathrm{harm}} &= \frac{2\,\mathrm{IoU}_{\mathrm{fwd}}\,\mathrm{IoU}_{\mathrm{bwd}}}{\mathrm{IoU}_{\mathrm{fwd}}+\mathrm{IoU}_{\mathrm{bwd}}}.\end{aligned} \tag{29}$$

B.3.3. Four-Dimensional Comprehensive Evaluation

Spatial Alignment (40%) Adopts the harmonic mean IoU:

$$\mathcal{S}_{\mathrm{align}} = \mathrm{IoU}_{\mathrm{harm}}, \tag{30}$$

Temporal Consistency (30%) Evaluates the smoothness of IoU across adjacent frames:

$$\mathcal{S}_{\mathrm{temp}} = 1 - \frac{1}{T-1}\sum_{t=2}^{T}\frac{|\mathrm{IoU}_t - \mathrm{IoU}_{t-1}|}{\mathrm{IoU}_{t-1}+\epsilon}, \tag{31}$$

Stability (20%) Measures the inter-frame mask change rate:

$$\mathcal{S}_{\mathrm{stab}}^{t} = 1 - \frac{|\mathcal{P}_t \triangle \mathcal{P}_{t-1}|}{|\mathcal{P}_t \cup \mathcal{P}_{t-1}|}, \tag{32}$$

where $\triangle$ denotes the symmetric difference. The average temporal stability is defined as

$$\mathcal{S}_{\mathrm{stab}} = \frac{1}{T-1}\sum_{t=2}^{T}\mathcal{S}_{\mathrm{stab}}^{t}, \tag{33}$$

Weighted Combination The final segment-level structure comparison metric is:

$$\mathcal{S}_{\mathrm{sam}} = 0.40\,\mathcal{S}_{\mathrm{align}} + 0.30\,\mathcal{S}_{\mathrm{temp}} + 0.20\,\mathcal{S}_{\mathrm{stab}} + 0.10\,\mathcal{S}_{\mathrm{f1}}, \tag{34}$$

B.4. Time Warping Consistency (DTW-Sim)

This metric employs the Dynamic Time Warping (DTW) algorithm to temporally align motion feature sequences from generated and ground truth videos, evaluating temporal dynamic response consistency through optimal warping path computation and multi-scale feature similarity.

B.4.1. DTW Algorithm Formulation

Given motion feature sequences $X = [x_1, \ldots, x_n]$ and $Y = [y_1, \ldots, y_m]$ extracted from generated and ground truth videos, we construct the cost matrix $C \in \mathbb{R}^{n\times m}$:

$$C_{i,j} = 1 - \frac{x_i^\top y_j}{\|x_i\|_2 \|y_j\|_2}, \tag{35}$$

using cosine distance for magnitude invariance.

The cumulative cost matrix $D \in \mathbb{R}^{n\times m}$ is computed recursively:

$$D_{i,j} = C_{i,j} + \min\begin{cases} D_{i-1,j-1} & \text{(diagonal)}\\ D_{i-1,j} & \text{(vertical)}\\ D_{i,j-1} & \text{(horizontal)}\end{cases}, \tag{36}$$

with boundary conditions $D_{1,1} = C_{1,1}$, $D_{i,1} = \sum_{k=1}^{i} C_{k,1}$, $D_{1,j} = \sum_{k=1}^{j} C_{1,k}$.

The optimal warping path $\mathcal{W} = \{(i_1, j_1), \ldots, (i_K, j_K)\}$ is obtained via backtracking from (n, m) to $(1, 1)$. We apply the Sakoe-Chiba band constraint (window size $w = 0.1 \times \max(n, m)$) to improve computational efficiency:

$$|i - j| \le w, \quad \forall (i, j) \in \mathcal{W}, \tag{37}$$

The normalized DTW distance is:

$$\text{DTW-Dist}(X, Y) = \frac{D_{n,m}}{K}, \tag{38}$$

where $K = |\mathcal{W}|$ is the path length.

B.4.2. Multi-Scale Feature Extraction

We extract motion features at three pyramid scales $\{1\times, 2\times, 4\times\}$ (corresponding to resolutions $\{256 \times 256, 128 \times 128, 64 \times 64\}$). At each scale s, we compute: (1) Optical flow magnitude $m_t^{(s)} = \|\mathbf{F}_t^{(s)}\|_2$; (2) Flow direction $\theta_t^{(s)} = \arctan 2(v_t^{(s)}, u_t^{(s)})$; (3) Directional encoding $\mathbf{d}_t^{(s)} = [\cos\theta_t^{(s)}, \sin\theta_t^{(s)}]^\top$; (4) Frame difference $\Delta I_t^{(s)} = \|I_{t+1}^{(s)} - I_t^{(s)}\|_1$.

The multi-scale feature vector is:

$$\mathbf{f}_t = \bigoplus_{s=1}^{3}[m_t^{(s)}, \mathbf{d}_t^{(s)}, \Delta I_t^{(s)}], \tag{39}$$

where $\bigoplus$ denotes concatenation. Normalization to zero mean and unit variance:

$$\hat{\mathbf{f}}_t = \frac{\mathbf{f}_t - \mu_{\mathbf{f}}}{\sigma_{\mathbf{f}}}, \tag{40}$$

The final feature sequence $X = [\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_T]$ is used for DTW computation.

B.4.3. Four-Dimensional Comprehensive Evaluation

Error Metric (30%) Based on DTW distance:

$$\mathcal{F}_{\text{align}} = 1 - \mathcal{D}_{\mathrm{DTW}}(X_g, X_{gt}), \tag{41}$$

where $\mathcal{D}_{\mathrm{DTW}}(\cdot, \cdot)$ denotes the Dynamic Time Warping distance between the generated sequence X_g and the ground-truth sequence X_{gt}.

Correlation (25%) Computes Kendall's tau coefficient τ for aligned sequences:

$$\tau = \frac{C - D}{C + D}, \tag{42}$$

where C and D are the numbers of concordant and discordant pairs, respectively.

Temporal Alignment (25%) Evaluates warping path smoothness:

$$\mathcal{C}_{\text{align}} = 1 - \frac{1}{K-1} \sum_{k=2}^{K} \left| \frac{\Delta i_k}{\Delta j_k} - 1 \right|, \tag{43}$$

where $\Delta i_k = i_k - i_{k-1}$ and $\Delta j_k = j_k - j_{k-1}$.

Temporal Quality (20%) Comprehensively assesses physical plausibility and evolution stability:

$$\begin{aligned} \text{Temporal Quality} = 0.6 \cdot \text{PhysicalPlausibility} \\ +0.4 \cdot \text{EvolutionStability}, \end{aligned} \tag{44}$$

where PhysicalPlausibility is verified through acceleration boundedness and displacement monotonicity (computed as the fraction of frames satisfying constraints), and EvolutionStability is evaluated through low variance of feature vector temporal gradients.

Symmetric Scoring and Weighted Combination We perform bidirectional DTW computation to ensure symmetry:

$$\begin{aligned} \text{DTW-Sim}_{\text{fwd}} &= \text{DTW-Sim}(X_g, X_{gt}) \\ \text{DTW-Sim}_{\text{bwd}} &= \text{DTW-Sim}(X_{gt}, X_g), \end{aligned} \tag{45}$$

The final time warping consistency metric is the arithmetic mean of both directions:

$$\begin{aligned} \mathcal{S}_{\text{dtw}} = \frac{1}{2} \Big[& 0.30\, \mathcal{E}_{\text{err}} + 0.25\, \tau \\ & + 0.25\, \mathcal{S}_{\text{align}} + 0.20\, \mathcal{S}_{\text{temp}} \Big]_{\text{fwd+bwd}}, \end{aligned} \tag{46}$$

C. Detailed Evaluation Indicator Calculation Method

To ensure comparability and robustness between visual observations and physical responses within the dataset, we adopt a multi-angle camera configuration that explicitly accounts for projection polarity induced by opposite viewing directions. The experimental setup includes 17 distinct camera poses, covering typical orientations such as north, south, east, west, southeast, southwest, northwest, northeast, and an overhead view. A uniform backdrop is employed to stabilize the background, and two fill lights are positioned to maintain consistent illumination while reducing artifacts caused by shadows or reflections that could affect keypoint detection or segmentation. When excitation is applied along the positive X direction, views from the positive and negative X sides produce opposite displacement projections on the image plane, resulting in a clear polarity reversal in motion appearance. To address this phenomenon, we simultaneously record from both positive and negative viewing directions to improve the model's robustness to polarity variations. Through a data acquisition strategy that integrates multi-view coverage, uniform lighting, and consistent background settings, combined with careful calibration and temporal synchronization, we obtain high-quality video sequences suitable for stable visual processing. This setup further provides physically grounded samples for training and evaluating video generation models under fixed or adaptive camera viewpoints.

C.0.1. Video Quality-Based Metrics

Subject Consistency The Subject Consistency Metric is established to evaluate whether a generated seismic video faithfully preserves the intrinsic structural characteristics of the physical model throughout the entire shaking sequence. As structural appearance and identity stability directly affect downstream tasks—including displacement interpretation, crack detection, and motion tracking—this metric is decomposed into rigorously defined sub-dimensions, each assigned weights based on its engineering relevance and impact on structural observability.

In this study, subject consistency is assessed using three weighted sub-indicators: **Object Persistence (35%)**, **Identity Continuity (25%)**, and **Structural Consistency Score (40%)**. Together, these indicators quantify whether the structural model remains physically coherent, visually stable, and temporally consistent across all frames.

(1) Object Persistence (35%) evaluates the continuous visibility and stability of the structure across the entire video duration. This dimension ensures that the structural model remains present from the beginning to the end of the seismic sequence without exhibiting non-physical artifacts such as sudden disappearance, occlusions unrelated to shaking behavior, or abrupt cropping caused by generation instability. Object persistence is quantified by computing a frame-wise visibility ratio,

$$R_{\text{vis}} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(\text{struct visible at } t), \tag{47}$$

where $\mathbb{I}(\cdot)$ is an indicator function determining whether the structure occupies a minimum required proportion of the frame. Higher values indicate stronger temporal stability and reliable structural observability.

(2) Identity Continuity (25%) measures the preservation of key structural identity attributes such as material

texture, surface color, geometric outlines, and structural accessory features (e.g., joints, beams, or boundary fixtures). This dimension penalizes abrupt or unrealistic variations that violate physical expectations, such as sudden texture distortions or changes in material appearance. To quantify visual identity continuity, we compute a temporal texture deviation metric,

$$D_{\text{tex}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\phi(I_t) - \phi(I_{t+1})\|_2 \,, \tag{48}$$

where $\phi(\cdot)$ denotes a feature extractor (e.g., color-histogram or low-level texture encoder). Lower values indicate smoother temporal evolution and stronger identity preservation.

(3) Structural Consistency Score (40%) represents an integrated measure capturing the overall physical and visual fidelity of the structure. This indicator enforces global stability by combining geometry alignment, boundary coherence, and structural-profile similarity across frames. It is quantified using a structural similarity index averaged over the video,

$$S_{\text{struct}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \text{SSIM}(I_t, I_{t+1}), \tag{49}$$

where $\text{SSIM}(\cdot)$ assesses geometric continuity, luminance stability, and structural detail preservation. A higher score indicates that the model behaves consistently under seismic excitation and maintains physically plausible visual responses.

These three indicators are fused through weighted aggregation to produce the final Subject Consistency score:

$$S_{\text{sub}} = 0.35\, R_{\text{vis}} + 0.25\,(1 - D_{\text{tex}}) + 0.40\, S_{\text{struct}}, \tag{50}$$

This formulation provides a principled, interpretable, and engineering-aligned measurement of structural fidelity in seismic video generation. The resulting metric ensures that generated shake-table videos preserve structural presence, maintain identity realism, and exhibit coherent physical behavior suitable for downstream structural analysis and experimental validation.

VBench Aesthetic Quality The VBench Aesthetic Quality Metric is introduced to assess the visual presentation quality of videos generated from seismic shaker experiments. Although not directly governed by physical laws, visual fidelity critically influences the effectiveness of observing structural dynamics and damage progression. The metric integrates sub-dimensions tailored to the visual characteristics of seismic testing scenarios, with carefully designed weighting to balance their contributions. In this study, the aesthetic presentation quality of seismic shake-table test videos is evaluated using four sub-indicators, namely **Color Richness (25%)**, **Color Distribution (25%)**, **Brightness Distribution (25%)**, and **Composition Quality (25%)**, which jointly capture the visual fidelity and observational clarity required for seismic experiment analysis. Color Richness evaluates the diversity and saturation characteristics of the video, reflecting whether material colors and structural surfaces remain visually consistent with real-world seismic testing environments. This indicator is quantified by computing the frame-wise color variance,

$$\sigma^2_{\text{color}} = \frac{1}{N} \sum_{i=1}^{N} \left\| C(i) - \overline{C} \right\|^2 , \tag{51}$$

where $C(i)$ denotes the pixel color vector, and contributes 25% to the final score. Color Distribution (25%) assesses the spatial organization of dominant structural and environmental colors, ensuring that the structure occupies the primary visual region while background tones remain stable; this metric is measured through the spatial entropy of color channels,

$$H = -\sum_{k} p_k \log p_k, \tag{52}$$

which captures distribution balance and suppresses configurations where environmental colors obscure structural details.

Brightness Distribution, also weighted at 25%, focuses on the global and local illumination consistency across frames. To quantify this characteristic, we calculate the luminance deviation

$$D_{\text{lum}} = \frac{1}{N} \sum_{i=1}^{N} \left| Y(i) - \overline{Y} \right| , \tag{53}$$

where $Y(i)$ denotes pixel luminance. This indicator penalizes overexposure that removes structural detail or underexposure that conceals vibration states. Finally, Composition Quality (25%) measures the framing logic of the video by evaluating whether the structural model remains properly centered and whether motion trajectories are fully captured. This metric is computed using a structural occupancy ratio,

$$R_{\text{occ}} = \frac{A_{\text{struct}}}{A_{\text{frame}}}, \tag{54}$$

which reflects the visible proportion of the structure within the frame and ensures alignment with conventional seismic observation viewpoints.

These four sub-indicators are aggregated through weighted fusion to produce the final aesthetic quality score. The resulting metric provides a systematic, interpretable, and experimentally aligned evaluation of visual presentation quality, ensuring that generated seismic shake-table videos maintain clarity, stability, and structural observability consistent with real physical testing.

Temporal Flickering In this study, the overall quality of seismic shake-table test videos is evaluated through four core sub-indicators, namely **Segmentation Consistency (15%)**, **Physical Consistency (35%)**, **Temporal Consistency (30%)**, and **Perceived Quality (20%)**, which collectively characterize the realism and reliability of the generated video sequences. Segmentation Consistency assesses the frame-to-frame coherence of structural contours and component boundaries, aiming to detect region misalignment or edge jitter caused by video generation or compression; this indicator is computed by measuring pixel-level discrepancies in structural masks or key spatial regions and contributes 15% to the final score. Physical Consistency, the most critical component (35%), evaluates whether the inter-frame evolution of displacement, velocity, and acceleration adheres to fundamental seismic dynamic principles. It is quantified using the pixel-wise mean absolute error (MAE) between adjacent frames,

$$\mathrm{MAE}(F_t, F_{t+1}) = \frac{1}{N}\sum_{i=1}^{N} \left|F_t(i) - F_{t+1}(i)\right|, \tag{55}$$

and further incorporates the temporal smoothness and physical plausibility of these variations. Temporal Consistency, accounting for 30% of the score, measures the presence of temporal flickering, frame skipping, or discontinuous motion patterns. This metric is implemented by constructing an MAE sequence $\{\mathrm{MAE}_t\}$ and computing its normalized score,

$$S = \frac{255 - \overline{\mathrm{MAE}}}{255}, \tag{56}$$

which reflects the degree of temporal stability. Finally, Perceived Quality (20%) captures human-perceived visual fidelity, including clarity, texture preservation, and color stability, thereby reflecting the subjective credibility of the generated sequences. These four indicators are fused through weighted aggregation to yield the final video quality score, providing a comprehensive and reproducible framework for evaluating seismic video realism in terms of physical validity, dynamic continuity, structural stability, and perceptual quality.

Imaging Quality The Imaging Quality metric is introduced to evaluate the fundamental visual attributes of videos generated from seismic shaker experiments, as these properties directly affect the accuracy of structural detail recognition and the extraction of physically meaningful parameters. The metric integrates four sub-dimensions designed to capture clarity, contrast, noise suppression, and spatial detail, with weighting carefully assigned to balance their contributions.

In this study, imaging quality is assessed using four sub-indicators: **Sharpness (30%)**, **Contrast (25%)**, **Noise Level (25%)**, and **Resolution (20%)**. Sharpness, the primary dimension (30%), evaluates the clarity of structural edges and fine-scale features, which are essential for capturing vibration patterns. This indicator is quantified using the gradient-based sharpness measure

$$S = \frac{1}{N}\sum_{i=1}^{N} \left\|\nabla I(i)\right\|, \tag{57}$$

where $\nabla I(i)$ denotes the spatial gradient at pixel i, reflecting the prominence of structural contours.

Contrast (25%) measures the global and local intensity separation between structural subjects and background regions, ensuring reliable visibility of dynamic responses. It is computed using the normalized intensity variance

$$\sigma^2_{\mathrm{ctr}} = \frac{1}{N}\sum_{i=1}^{N} \left(Y(i) - \overline{Y}\right)^2, \tag{58}$$

where $Y(i)$ represents luminance. Noise Level (25%) quantifies stochastic visual artifacts that obscure structural detail, evaluated using the high-frequency noise energy

$$E_{\mathrm{noise}} = \frac{1}{N}\sum_{i=1}^{N} \left|I(i) - G_\sigma(I(i))\right|, \tag{59}$$

where G_σ denotes Gaussian smoothing, thereby penalizing noise patterns inconsistent with real seismic observations. Resolution (20%) reflects the spatial detail available for capturing subtle structural deformation, measured using the effective spatial frequency bandwidth

$$B = \sum_{f\in\Omega} \left|F(f)\right|, \tag{60}$$

where $F(f)$ is the Fourier component at frequency f.

The final imaging quality score is obtained by weighted fusion of these four sub-metrics, offering a systematic and interpretable evaluation of visual clarity in generated seismic videos.

Scene Scene metrics are introduced to assess the fidelity with which seismic shaker test videos reproduce the essential components of the experimental environment. These metrics evaluate whether the generated content conveys structurally meaningful context, environmental richness, and experimental realism.

In this study, scene fidelity is evaluated using three sub-indicators: **Scene Complexity (40%)**, **Edge Density (35%)**, and **Texture Richness (25%)**. Scene Complexity, the dominant dimension (40%), measures the ability of generated videos to represent the detailed and multi-component conditions of seismic testing. This indicator is computed using the spatial entropy of structural and environmental elements,

$$H_{\mathrm{scene}} = -\sum_{k} p_k \log p_k, \tag{61}$$

capturing the distributional richness required for realistic scene construction.

Edge Density (35%) reflects the abundance of structural edges, sensor outlines, and table boundaries, which jointly define the visual complexity of the experimental setup. It is measured using the normalized edge activation ratio

$$R_{\text{edge}} = \frac{1}{A_{\text{frame}}} \sum_{i=1}^{N} \mathbf{1}\left(\|\nabla I(i)\| > \tau\right), \tag{62}$$

indicating the prevalence of structurally meaningful contours. Texture Richness (25%) evaluates the diversity of surface and material textures across both structures and background elements. This indicator is quantified using the local binary pattern (LBP) variance

$$\sigma_{\text{tex}}^2 = \frac{1}{M} \sum_{j=1}^{M} \left(T(j) - \overline{T}\right)^2, \tag{63}$$

where $T(j)$ denotes the texture descriptor for region j.

The final scene score is computed through weighted aggregation of these three sub-indicators, providing a systematic and interpretable measure of scene realism and environmental completeness in generated seismic videos.

D. Dataset Details

D.1. Structural Model Specifications

In this study, a total of thirteen rigid structural models were selected and constructed to facilitate a controlled investigation of seismic response behavior using shaking table experiments. To ensure that the resulting video data and measured responses primarily reflect rigid-body motion—rather than unwanted flexible deformation—the models were fabricated from materials with exceptionally high stiffness, thereby minimizing structural bending, local deformation, and mode-shape effects during excitation. Each model features well-defined geometric dimensions, mass distributions, and boundary conditions, enabling consistent and repeatable dynamic behavior across all test runs. The model set spans variations in height-to-width ratio, mass arrangement, and plan configuration, allowing the experimental program to capture a diverse range of rigid-body inertial characteristics under multi-directional seismic loading. By intentionally suppressing flexibility through material selection and structural detailing, the experimental outputs focus on translational and rotational rigid-body dynamics, providing clean and physically interpretable data suitable for validating seismic motion prediction methods and supporting video-generation models designed to learn rigid-structure dynamic patterns.

D.2. Seismic waveform

We selected 35 representative ground-motion records from the PEER ground motion database (including the canonical El Centro record) and organized them into a reproducible test matrix for shake-table loading, structural response measurement, and video-generation experiments. The test matrix is formed by combining seven degrees-of-freedom (DOF) configurations with five PGA levels (yielding $7 \times 5 = 35$ conditions). The DOF configurations correspond to directional and planar excitations used in the experiments:

$$\begin{aligned}\mathcal{D}_{\text{dof}} = \Big\{&\text{X},\ \text{Y},\ \text{Z},\\ &\text{XY},\ \text{XZ},\ \text{YZ},\ \text{XYZ}\Big\},\end{aligned} \tag{64}$$

The five PGA target levels are defined as

$$\mathcal{G}_{\text{PGA}} = \{0.10g,\ 0.125g,\ 0.15g,\ 0.175g,\ 0.20g\}. \tag{65}$$

Each experiment is identified in the catalogue by the tuple $\{\text{record name},\ \text{DOF configuration},\ \text{target PGA}\}$. The dataset records include the seismic record source (PEER event and station ID), original sampling rate, original PGA, filtered and scaled PGA, dominant frequency band, and effective duration $T_{5\%-95\%}$.

D.3. Preprocessing and PGA scaling

All raw acceleration time series $a_{\text{orig}}(t)$ were pre-processed to ensure compatibility with the dynamic range and frequency response of the shake table and to maintain physical fidelity when used as inputs for video-driven modeling. The main preprocessing stages were: baseline correction, anti-aliasing / resampling, bandpass filtering (zero-phase), PGA-based amplitude scaling, and derivation of velocity / displacement traces (if needed). We describe each step in detail below.

Baseline correction and detrending. To remove sensor drift and low-frequency baseline offsets that corrupt numerical integration, every record is detrended and baseline-corrected. Let $a_{\text{raw}}(t)$ be the raw recorded acceleration. We first remove the mean and a low-order polynomial trend:

$$a_{\text{detr}}(t) = a_{\text{raw}}(t) - \overline{a_{\text{raw}}} - P_k(t), \tag{66}$$

where $P_k(t)$ is a polynomial fit of order k (we use $k = 1$ or $k = 2$ depending on the record). After detrending, a high-pass correction (or additional baseline-removal step) is applied to minimize secular drift prior to integration.

Anti-aliasing and resampling. If the original sampling rate f_s^{orig} does not match the processing sampling rate f_s used in the experiment pipeline, we first apply an

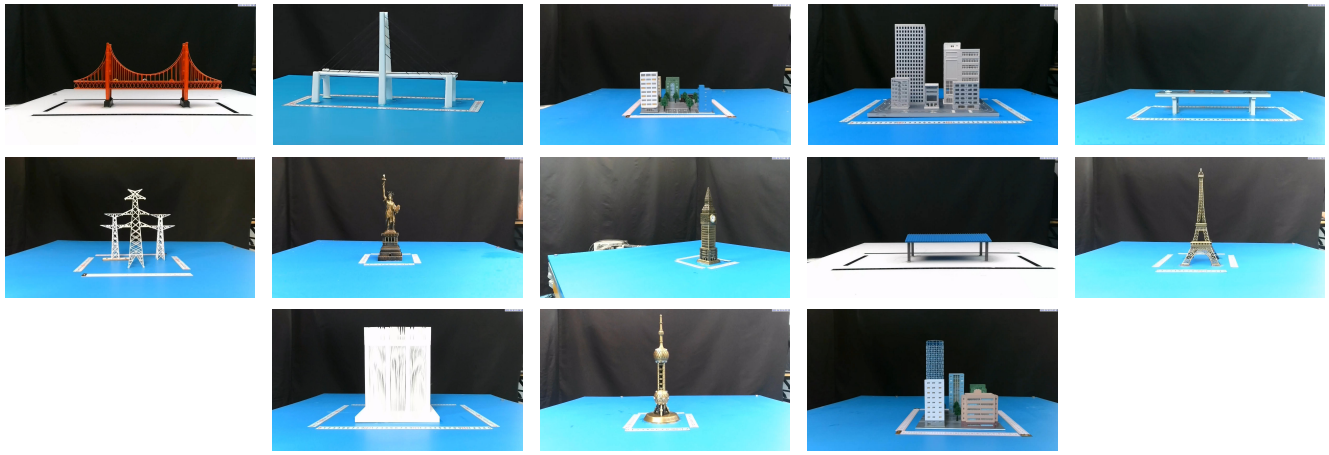

Figure 3. Structural Models

anti-aliasing low-pass filter and then resample. In practice we use a digital Butterworth low-pass with cutoff $f_{\text{nyq}} = 0.45 f_s$ and perform resampling with zero-phase filtering (e.g., `scipy.signal.resample_poly` or `filtfilt`). This sequence prevents aliasing while preserving phase relationships.

Bandpass filtering (device-aware). To ensure the input signals lie within the operational frequency band of the shake table, we apply a zero-phase Butterworth bandpass filter $\mathcal{H}_{\text{BP}}(f)$ with lower and upper cutoffs f_ℓ and f_h. The squared magnitude of the low-pass prototype is

$$|H_{\text{LP}}(j\omega)|^2 = \frac{1}{1 + \left(\dfrac{\omega}{\omega_c}\right)^{2n}}, \tag{67}$$

and a bandpass is formed by cascading a high-pass and a low-pass section (or designing an equivalent bandpass). In discrete-time implementation we design an n-th order Butterworth bandpass using bilinear transformation and apply it with forward–backward (zero-phase) filtering:

$$a_{\text{filtered}}(t) = \text{filtfilt}(\mathcal{H}_{\text{BP}}(f_\ell, f_h, n, f_s), a_{\text{detrended}}(t)), \tag{68}$$

Typical parameter choices in this study were $n = 2$ or $n = 4$ and a default device-aware passband of $f_\ell = 0.1$ Hz to $f_h = 25$ Hz, unless the shaker specification required a narrower band.

PGA scaling (amplitude normalization). To generate multiple loading intensities while preserving waveform shape and frequency content, we apply linear amplitude scaling. Let the original peak acceleration be

$$\text{PGA}_{\text{orig}} = \max_t |a_{\text{filtered}}(t)|, \tag{69}$$

For a desired target PGA $\text{PGA}_{\text{target}} \in \{0.10g,\ 0.125g,\ 0.15g,\ 0.175g,\ 0.20g\}$ we compute the scaling factor

$$\alpha = \frac{\text{PGA}_{\text{target}}}{\text{PGA}_{\text{orig}}}, \tag{70}$$

and produce the scaled acceleration time series

$$a_{\text{scaled}}(t) = \alpha\, a_{\text{filtered}}(t), \tag{71}$$

After scaling we re-check the shaker limits: maximum displacement $u_{\max}$ and maximum velocity $v_{\max}$ obtained by numerical integration (see below) must not exceed equipment specifications. If limits are violated, the record is either rejected or further spectrally modified (e.g., low-pass truncation) to meet safety constraints.

Integration to velocity/displacement. When velocity $v(t)$ and displacement $u(t)$ traces are required (for sensor comparisons or for commanding the actuator), we numerically integrate the scaled acceleration with careful baseline control. Using the detrended acceleration $a_{\text{scaled}}(t)$, discrete-time integration with baseline correction is implemented as:

$$v(t) = \int_0^t a_{\text{scaled}}(\tau)\, \mathrm{d}\tau, \qquad u(t) = \int_0^t v(\tau)\, \mathrm{d}\tau, \tag{72}$$

with Akima or trapezoidal integration after applying a high-pass filter (or baseline-removal) to avoid secular drift. We verify energy-consistency by comparing integration results against frequency-domain-derived displacement estimates where appropriate.

D.4. Time-domain and frequency-domain summary metrics

For every processed record we compute a set of summary metrics used in experiment planning and in later physical-consistency comparisons with video-extracted motion:

- **Peak ground acceleration (PGA)**: $\max_t |a_{\text{scaled}}(t)|$.
- **Dominant frequency band**: estimated from the power spectral density (PSD) via Welch's method; we report the frequency f_{peak} at which PSD attains its maximum and the 50% energy bandwidth.
- **Arias intensity** and effective duration: Arias intensity

$$I_A = \frac{\pi}{2g} \int_0^T a_{\text{scaled}}^2(t)\, \mathrm{d}t, \tag{73}$$

  and the effective duration $T_{5\%-95\%}$ defined as the time interval during which cumulative energy (normalized cumulative integral of a^2) grows from 5% to 95%.
- **Spectral compatibility checks**: compare target and scaled spectra to ensure scaling preserved the relative spectral shape; this is used to detect pathological scaling that alters frequency content significantly.

D.5. Reproducible processing pipeline and safety checks

All preprocessing steps are implemented with reproducibility in mind and recorded in a processing log attached to each catalogue entry. We recommend implementations using standard scientific libraries and saving intermediate artifacts (detrended, filtered, scaled time series, PSD, Arias curve). Before issuing commands to a shake table, the following safety checks are enforced automatically:

1. **Equipment limits**: verify $u_{\max}$ and $v_{\max}$ computed from $a_{\text{scaled}}(t)$ are within shaker specifications.
2. **Energy consistency**: check that spectral energy concentration lies within the intended passband and that no spurious low-frequency energy was introduced during filtering/scaling.
3. **Time alignment metadata**: stamp each processed file with sampling rate and time origin to preserve alignment with video frames and sensor timestamps.

Summary By combining device-aware bandpass filtering, robust baseline correction, conservative resampling/anti-aliasing, and controlled PGA linear scaling (with mandatory shaker-limit checks), the pipeline yields 35 experiment-ready ground-motion records (covering all DOF configurations and PGA levels specified above). These processed records serve as the input library for both the physical shake-table tests and for conditioning the video generation models; the standardized processing ensures reproducibility and valid comparisons between video-extracted motion and instrumented measurements.

D.6. Camera Configuration

D.6.1. Hardware Specifications

This study employs nine MF-500 2K industrial-grade camera modules equipped with 2.8mm distortion-free lenses. Each camera achieves 5-megapixel resolution (2592×1944 pixels) at 30 frames per second, with a wide field of view (FOV) of 85°. The key specifications are summarized in Table 1.

Table 1. Camera Module Specifications (MF-500 2K)

Parameter	Value
Model	MF-500 2K
Sensor	1/3-inch CMOS OV
Resolution	2592×1944 (5MP, 2K)
Pixel Size	1.4μm × 1.4μm
Frame Rate	30 fps
Focal Length	2.8mm (distortion-free)
Field of View	85° (diagonal)
Dynamic Range	72.5dB
SNR	34dB
Power Supply	USB bus power (5V)
Current Draw	150-200mA
Interface	USB 2.0 (UVC protocol)
Output Format	MJPG (YUY2/YUYV)

The 2.8mm lens provides optimal coverage for the shake table experimental setup, with a recommended shooting distance of 0-5m and an effective viewing angle of 85°, ensuring complete structural coverage without geometric distortion.

D.6.2. Multi-Camera Spatial Arrangement

To achieve comprehensive 360° structural monitoring with multi-view geometric consistency, we deploy nine cameras in a strategic configuration comprising eight cardinal/intercardinal directions plus one overhead view (Figure **??**). The placement follows a cylindrical coordinate system centered on the shake table:

Horizontal Ring (8 cameras): The eight cameras are positioned at equal angular intervals of 45° around the structure at a fixed distance $R = 3.5$m from the table center, maintaining a horizontal elevation $h_{\text{ring}} = 1.5$m to align with the mid-height of typical test specimens. The camera positions in Cartesian coordinates are:

$$\mathbf{C}_i = \begin{bmatrix} R\cos\theta_i \\ R\sin\theta_i \\ h_{\text{ring}} \end{bmatrix}, \quad \theta_i = \frac{\pi}{4}(i-1), \quad i = 1, \ldots, 8, \tag{74}$$

where:

- $\mathbf{C}_1$ (East, $\theta = 0$): $(3.5, 0, 1.5)$ m

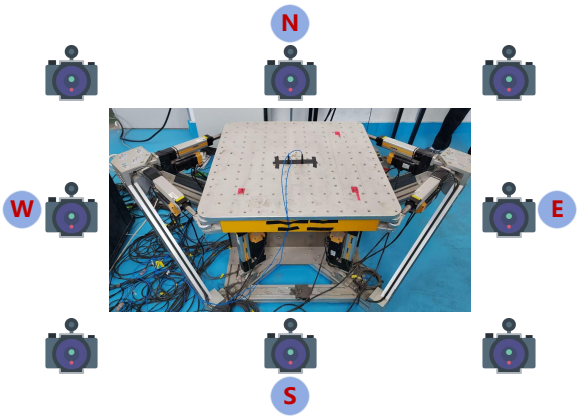

Figure 4. Data Collection and Capture Methods

- $\mathbf{C}_2$ (Northeast, $\theta = 45$): $(2.47, 2.47, 1.5)$ m
- $\mathbf{C}_3$ (North, $\theta = 90$): $(0, 3.5, 1.5)$ m
- $\mathbf{C}_4$ (Northwest, $\theta = 135$): $(-2.47, 2.47, 1.5)$ m.
- $\mathbf{C}_5$ (West, $\theta = 180$): $(-3.5, 0, 1.5)$ m
- $\mathbf{C}_6$ (Southwest, $\theta = 225$): $(-2.47, -2.47, 1.5)$ m.
- $\mathbf{C}_7$ (South, $\theta = 270$): $(0, -3.5, 1.5)$ m
- $\mathbf{C}_8$ (Southeast, $\theta = 315$): $(2.47, -2.47, 1.5)$ m.

Each horizontal camera is tilted downward by $\alpha = 5$ to ensure the structure remains centered in the frame:

$$\mathbf{R}_{\text{tilt}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix}, \tag{75}$$

Overhead Camera ($\mathbf{C}_9$): The ninth camera is mounted directly above the shake table at position $(0, 0, 4.5)$m, oriented vertically downward (nadir view, $\beta = 90$) to capture top-down structural deformation. Its rotation matrix is:

$$\mathbf{R}_{\text{nadir}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \tag{76}$$

This overhead view is critical for measuring horizontal displacement components and detecting out-of-plane rotations invisible to side-mounted cameras.

D.6.3. Camera Synchronization Mechanism

Temporal alignment across all nine cameras is essential for multi-view consistency. We implement a hardware-software hybrid synchronization strategy:

Hardware Trigger: A Raspberry Pi 4B microcontroller generates a master clock signal at 30 Hz via GPIO pins. The trigger pulse $T(t)$ is a square wave:

$$T(t) = \begin{cases} 5V, & t \bmod \frac{1}{30} < \tau_{\text{pulse}} \\ 0V, & \text{otherwise} \end{cases}, \tag{77}$$

where $\tau_{\text{pulse}} = 100\mu$s. This signal is distributed to all cameras via a synchronized USB hub with shared V-sync lines.

Software Timestamp Synchronization: Each frame is timestamped using the Network Time Protocol (NTP)-synchronized system clock with microsecond precision. For camera i capturing frame j at time $t_{i,j}$, we compute the inter-camera synchronization error:

$$\epsilon_{\text{sync}} = \max_{i,k} |t_{i,j} - t_{k,j}|, \quad i, k \in \{1, \ldots, 9\}, \tag{78}$$

In our setup, $\epsilon_{\text{sync}} \leq 1.2$ms, which is negligible compared to the 33.3ms frame interval (30 fps).

Post-Capture Alignment: We apply cross-correlation-based temporal alignment using the shake table acceleration signal $\mathbf{a}(t)$ as reference. For each camera stream V_i, we extract brightness variation $B_i(t) = \frac{1}{HW}\sum_{x,y} I_i^t(x, y)$ and compute lag:

$$\tau_i^* = \arg\max_{\tau} \text{Corr}(B_i(t), \|\mathbf{a}(t + \tau)\|_2), \tag{79}$$

then shift frames accordingly: $V_i^{\text{aligned}}(t) = V_i(t - \tau_i^*)$.

D.6.4. Camera Calibration Parameters

We perform geometric calibration using a 10×7 checkerboard pattern with 30mm square size. Zhang's method yields intrinsic and extrinsic parameters for each camera.

Intrinsic Matrix: For camera i, the intrinsic matrix $\mathbf{K}_i \in \mathbb{R}^{3\times3}$ relates 3D camera coordinates to 2D image coordinates:

$$\mathbf{K}_i = \begin{bmatrix} f_{x,i} & 0 & c_{x,i} \\ 0 & f_{y,i} & c_{y,i} \\ 0 & 0 & 1 \end{bmatrix}, \tag{80}$$

where $f_{x,i}, f_{y,i}$ are focal lengths in pixels, and $(c_{x,i}, c_{y,i})$ is the principal point. For the MF-500 2K with 2.8mm lens, typical values are:

$$f_x \approx f_y \approx 1850 \text{ pixels}, \quad c_x \approx 1296, \quad c_y \approx 972, \tag{81}$$

Distortion Coefficients: Despite the distortion-free lens design, residual radial and tangential distortions are modeled as:

$$\begin{cases} x_{\text{distorted}} = x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ \qquad + 2p_1 xy + p_2(r^2 + 2x^2), \\ y_{\text{distorted}} = y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ \qquad + p_1(r^2 + 2y^2) + 2p_2 xy, \end{cases} \tag{82}$$

where $r^2 = x^2 + y^2$. Calibration reveals $|k_1|, |k_2|, |k_3| < 0.01$ and $|p_1|, |p_2| < 0.001$, confirming minimal distortion.

Extrinsic Parameters: The extrinsic matrix $[\mathbf{R}_i|\mathbf{t}_i]$ transforms world coordinates (shake table frame) to camera coordinates:

$$\mathbf{x}_{\text{camera}} = \mathbf{R}_i\mathbf{X}_{\text{world}} + \mathbf{t}_i, \tag{83}$$

where $\mathbf{R}_i \in \text{SO}(3)$ is the rotation matrix derived from Euler angles $(\phi_i, \theta_i, \psi_i)$, and $\mathbf{t}_i \in \mathbb{R}^3$ is the translation vector. For camera $\mathbf{C}_1$ (East), calibration yields:

$$\begin{cases} \mathbf{R}_1 = \begin{bmatrix} 0.9962 & 0 & -0.0872 \\ 0 & 1 & 0 \\ 0.0872 & 0 & 0.9962 \end{bmatrix}, \\ \mathbf{t}_1 = \begin{bmatrix} -3.48 \\ 0.02 \\ -1.51 \end{bmatrix} \text{ m}, \end{cases} \tag{84}$$

Reprojection Error: Calibration quality is assessed via mean reprojection error:

$$e_{\text{reproj}} = \frac{1}{NM}\sum_{i=1}^{9}\sum_{j=1}^{M} \|\mathbf{p}_{i,j} - \hat{\mathbf{p}}_{i,j}\|_2, \tag{85}$$

where $\mathbf{p}_{i,j}$ are detected checkerboard corners and $\hat{\mathbf{p}}_{i,j}$ are reprojected points. Our system achieves $e_{\text{reproj}} = 0.23$ pixels, well below the 0.5-pixel threshold for high-precision applications.

Multi-View Consistency Validation: To verify geometric consistency, we triangulate 3D positions of checkerboard corners from multiple views using the Direct Linear Transform (DLT):

$$\mathbf{X}_j = \arg\min_{\mathbf{X}} \sum_{i=1}^{9} \|\mathbf{p}_{i,j} - \mathbf{P}_i\mathbf{X}\|_2^2, \tag{86}$$

where $\mathbf{P}_i = \mathbf{K}_i[\mathbf{R}_i|\mathbf{t}_i]$ is the projection matrix. The mean 3D reconstruction error is $\sigma_{3D} = 1.8$mm, demonstrating sub-millimeter spatial accuracy suitable for structural displacement measurement.

D.7. Dataset Overview.

This study constructs a dataset consisting of **7,735 high-resolution seismic shake-table video sequences**, systematically organized to support physics-consistent video generation and structural dynamic analysis. The dataset provides comprehensive coverage across structural configurations, excitation conditions, multi-view observations, and spatial loading directions, ensuring both diversity and representativeness of real experimental scenarios.

It includes all **13 rigid-body and scaled building models** used in laboratory testing, ranging from simple calibration blocks to multi-story frame structures exhibiting inter-story coupling and higher-mode responses, as well as models with plan, torsional, or vertical irregularities that introduce torsion–translation coupling and stress concentration. Structures equipped with base-isolation systems and tuned mass dampers are also included, enabling the dataset to span a broad spectrum of mass–stiffness combinations and nonlinear damping behaviors.

For ground-motion excitation, **35 earthquake loading cases** are generated from the El Centro 1940 NS record through band-pass filtering (0.1–25 Hz) and PGA scaling. These cases cover seven degrees-of-freedom configurations (X, Y, Z, XY, XZ, YZ, XYZ) and five PGA levels (0.10g, 0.125g, 0.15g, 0.175g, 0.20g), thereby encompassing dynamic responses ranging from elastic regimes to moderately nonlinear deformation.

Each seismic condition is captured by **17 fully synchronized cameras**, including 16 azimuthal views arranged along a 3.5 m circular array at 22.5° intervals for dense observation of lateral deformation, inter-story drift, and torsional effects, as well as a vertical overhead view positioned 4.5 m above the structure to record horizontal trajectories and in-plane rotational modes. This high-density geometric configuration provides exceptionally rich multi-perspective constraints for perspective-conditioned video generation and cross-view geometric consistency evaluation.

The dataset further exhibits a wide distribution of excitation directions, consisting of **3,315** unidirectional sequences (42.86%), **3,315** bidirectional sequences (42.86%), and **1,105** triaxial sequences (14.29%), enabling the capture of coupled dynamic behaviors that are rarely included in existing video-generation datasets.

Overall, the dataset satisfies

$$N_{\text{total}} = 13 \times 35 \times 17 = 7{,}735 \text{ sequences},$$

representing one of the most comprehensive and systematically designed seismic video datasets to date. Its scale, diversity, and controlled experimental dimensionality provide a robust foundation for training and evaluating physics-aware video generation models.

E. Technical Specifications of the Vibration Table System

The propagation of seismic wave energy starts from the seismic source, and classical seismology proposes the concept of energy flux based on the theory of elastic wave propagation. According to the definition by Aki and Richards, the flux rate of energy transmission in a plane wave refers to the total amount of energy transferred per unit area perpendicular to the propagation direction per unit time. Based on

Table 2. Technical specifications of the high-frequency linear motor.

Technical Indicator	Parameter
Rated Output Force	665 N
Rated Current	2.873 A
Rated Speed	1720 mm/s
Encoder Resolution	20 μm
Maximum Displacement	±45 mm
Maximum Speed	1892 mm/s

this definition, when a seismic wave passes through a given surface area A within a specific time period t, its energy flux E_f can be calculated by the following formula:

Seismic Wave Energy Flux (Plane Wave, Used to Estimate the Energy Input by Earthquakes to Local Sites)

$$E_{\text{flux}} = \rho A c \int_0^t \dot{u}(T)^2 dT, \tag{87}$$

Meaning: The seismic wave energy passing through area A per unit time, where ρ is the medium density, c is the wave velocity, and $\dot{u}$ is the vibration velocity.

E.1. Calculation of Energy Input

Energy input is used to verify the calculation results of energy dissipation. According to the first law of thermodynamics (i.e., the law of conservation of energy), the sum of all energy forms in the soil-structure interaction (SSI) model should be equal to the total energy input into the system. When deriving a new energy dissipation equation, this condition is always checked, which will be demonstrated in the next chapter.

It should be noted that the scope of this study is limited to the interior and nearby areas of the local soil-structure interaction (SSI) system. The energy release at the seismic source or the energy flow propagating with seismic waves is not within the scope of this study; for the energy calculation of these issues, refer to the studies by Argyris and Mlejnek and Trifunac et al. The energy input in this study is defined as the mechanical energy that reaches and enters the local soil-structure interaction site. This definition is consistent with most recent studies on energy analysis of soil-structure interaction systems.

Calculation of Energy Input In finite element simulation, all external loads are converted into nodal forces and then applied to the model. After the system solution is completed, the nodal displacements at all positions can be obtained. For dynamic or static working conditions involving multi-step loads, the applied nodal forces and the solved

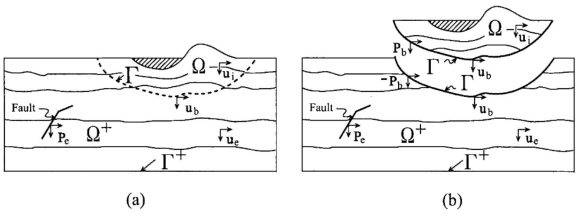

Figure 5. Seismic Wave Application Method

nodal displacements are all time series. Using this information, the energy input E_I can be easily calculated as:

$$E_I(t) = \int_0^t \sum_i F_i^{ex}(\boldsymbol{x}, T)\dot{u}_i(\boldsymbol{x}, T)dT, \tag{88}$$

where F_i^{ex} is the external force, and $\dot{u}_i$ is the displacement calculated at the location where the load is applied at a given time step. The summation term in the integral represents multiple loads.

Plastic Energy Dissipation As mentioned earlier, plastic work and energy dissipation are not the same physical quantity. The confusion of these two concepts often leads to erroneous results and conclusions, especially in seismic energy dissipation analysis. This paper mainly focuses on the calculation of plastic dissipation, which will be elaborated in detail in this section.

Based on the decoupling assumption, the second law of thermodynamics (positive entropy production) directly derives the dissipation inequality, which states that: the dissipated energy caused by the difference between the plastic power and the rate of change of the plastic part of the free energy must be non-negative.

$$\Phi = \sigma_{ij}\dot{\epsilon}_{ij}^{pl} - \dot{\Psi}^{pl} = \sigma_{ij}\dot{\epsilon}_{ij}^{pl} - \rho\dot{\psi}^{pl} \geq 0, \tag{89}$$

where $\dot{\psi}^{pl}$ is the rate of plastic free energy per unit mass, and ρ is the mass density. In addition, ψ^{pl} represents the plastic free energy density, which is usually not constant at different positions of the object. This expression is closer to physical reality and also facilitates subsequent derivation.

Now we proceed to consider how to calculate the plastic free energy, which can then be used to calculate the dissipation. According to the study by Feigenbaum and Dafalias, the plastic free energy density ψ^{pl} is assumed to be additively decomposed into parts corresponding to isotropic, kinematic, and distortional hardening mechanisms, as follows:

$$\begin{cases} \psi^{\text{pl}} = \psi^{\text{pl-iso}} + \psi^{\text{pl-ani}}, \\ \psi^{\text{pl-ani}} = \psi^{\text{pl-kin}} - \psi^{\text{pl-dis}}, \end{cases} \tag{90}$$

where ψ^{pl-iso}, ψ^{pl-ani}, ψ^{pl-kin}, and ψ^{pl-dis} are the isotropic, anisotropic, kinematic, and distortional parts of

the plastic free energy, respectively. The anisotropic part is assumed to be decomposed into kinematic and distortional parts, corresponding to different hardening models. ψ^{pl-kin} is subtracted by ψ^{pl-dis} (rather than added) to obtain the overall anisotropic part ψ^{pl-ani} of the plastic free energy, a new concept proposed by Feigenbaum and Dafalias. This expression fits experimental data better and meets reasonable expectations for anisotropy development limitations. The distortional part ψ^{pl-dis} relates to the directional distortion of the yield surface and exists only when the material model includes distortional strain hardening, which is not considered in this study's formulas and examples. As pointed out by Dafalias et al., each internal variable has a thermodynamic conjugate quantity, and each part of the plastic free energy can be assumed to be a function of these conjugate quantities only. The explicit expressions for the isotropic and kinematic components of the plastic free energy are:

$$\begin{cases} \psi^{pl-iso} = \psi^{pl-iso}(\bar{k}) = \dfrac{\kappa_1}{2\rho}\bar{k}^2; \\ \psi^{pl-kin} = \psi^{pl-kin}(\bar{\alpha}_{ij}) = \dfrac{a_1}{2\rho}\bar{\alpha}_{ij}\bar{\alpha}_{ij}, \end{cases} \tag{91}$$

where $\bar{k}$ and $\bar{\alpha}_{ij}$ are the thermodynamic conjugate quantities of k (the size of the yield surface) and α_{ij} (the deviatoric backstress tensor representing the yield surface center), respectively. The material constants κ_1 and a_1 are non-negative, with values depending on the choice of the inelastic material model.

According to the definition, thermodynamic conjugate quantities are related to the corresponding internal variables as follows:

$$\begin{cases} k = \rho\dfrac{\partial\psi^{\text{pl-iso}}}{\partial\bar{k}} = \kappa_1\bar{k}, \\ \alpha_{ij} = \rho\dfrac{\partial\psi^{\text{pl-kin}}}{\partial\bar{\alpha}_{ij}} = a_1\bar{\alpha}_{ij}, \end{cases} \tag{92}$$

The plastic free energy can be expressed in terms of internal variables:

$$\begin{cases} \psi^{pl-iso} = \dfrac{1}{2\rho\kappa_1}k^2; \\ \psi^{pl-kin} = \dfrac{1}{2\rho a_1}\alpha_{ij}\alpha_{ij}, \end{cases} \tag{93}$$

As long as the internal variables are given, each component of the plastic free energy can be calculated, and the plastic dissipation of a specific inelastic material can be accurately obtained at any position and time. This method allows engineers and designers to accurately identify energy dissipation in space and time, thereby making reasonable inferences about material behavior.

E.1.1. Mises Plasticity Theory

In this subsection, we revisit the plastic free energy and plastic dissipation equations for the pressure-independent von Mises material model. It should be noted that the von Mises plasticity theory always adopts the associated plastic flow rule, meaning that the direction of plastic flow in the stress space is perpendicular to the yield surface. The expression of the von Mises yield function is as follows:

$$f = \sqrt{(s_{ij} - \alpha_{ij})(s_{ij} - \alpha_{ij})} - \sqrt{\frac{2}{3}}k, \tag{94}$$

where $s_{ij} = \sigma_{ij} - \frac{1}{3}\sigma_{kk}$ is the deviatoric part of the stress tensor.

Once the material yields, plastic strain begins to develop. The calculation formula for the plastic strain increment tensor is:

$$d\epsilon_{ij}^{pl} = m_{ij}d\lambda, \tag{95}$$

where $d\lambda$ is a scalar loading index, whose magnitude is equal to the amplitude of the plastic strain increment. Due to the adoption of associated plasticity (theory), the normalized plastic flow direction vector m_{ij} is calculated by taking the gradient of the yield function in the stress space:

$$m_{ij} = \frac{\partial f}{\partial\sigma_{ij}} = \frac{(s_{ij} - \alpha_{ij})}{\sqrt{(s_{mn} - \alpha_{mn})(s_{mn} - \alpha_{mn})}}, \tag{96}$$

Armstrong-Frederick kinematic hardening (Armstrong and Frederick) is a nonlinear strain hardening criterion, often used to simulate cyclic inelastic behavior of various materials, including metals, alloys, soils, and other structural/geotechnical engineering materials. The evolution of the incremental backstress tensor $d\alpha_{ij}$ is defined as:

$$d\alpha_{ij} = \left[\frac{2}{3}h_a m_{ij} - c_r\alpha_{ij}\sqrt{\frac{2}{3}m_{rs}m_{rs}}\right]d\lambda, \tag{97}$$

where h_a and c_r are non-negative hardening constants. When $h_a > 0$ and $c_r = 0$, the nonlinear Armstrong-Frederick hardening becomes a linear hardening criterion. If $h_a = 0$ and $c_r = 0$, the material model becomes an ideal plastic model without internal variable hardening. It should be noted that isotropic hardening can also be defined in a similar form. To avoid repetition, the remaining part of this paper will focus on kinematic hardening. The plastic free energy density function proposed by Feigenbaum and Dafalias and revised by Yang et al. is:

$$d\Psi^{pl-kin} = \frac{3}{2h_a}\alpha_{ij}d\alpha_{ij}, \tag{98}$$

The plastic dissipation density function can be expressed in terms of material parameters and internal variables:

$$\Phi = \sigma_{ij} d\epsilon_{ij}^{pl} - \frac{3}{2h_a}\alpha_{ij} d\alpha_{ij}$$
$$= s_{ij} m_{ij} d\lambda - \alpha_{ij} m_{ij} d\lambda + \frac{3c_r}{2h_a}\sqrt{\frac{2}{3} m_{rs} m_{rs}}\alpha_{ij}\alpha_{ij} d\lambda, \tag{99}$$

where Φ is the plastic dissipation function, σ_{ij} is the Cauchy stress tensor, $d\epsilon_{ij}^{pl}$ is the increment of plastic strain, α_{ij} is the backstress tensor, h_a is the kinematic hardening modulus, s_{ij} is the deviatoric stress tensor, m_{ij} is the flow direction tensor, $d\lambda$ is the plastic multiplier increment, c_r is the isotropic hardening parameter, and $m_{rs}m_{rs}$ denotes the tensor double contraction of m_{ij}.

E.1.2. Associated Drucker-Prager Plasticity

Drucker-Prager type plasticity is often used to simulate pressure-dependent material behavior. Studies have shown that the plastic free energy function requires an additional pressure-dependent term to ensure that the plastic dissipation calculated for pressure-dependent materials is thermodynamically correct. The Drucker-Prager yield function is:

$$f = \sqrt{(s_{ij} - p\alpha_{ij})(s_{ij} - p\alpha_{ij})} - \sqrt{\frac{2}{3}} k_p, \tag{100}$$

where $p = -\frac{1}{3}\sigma_{kk}$ is the mean stress acting on the material, i.e., hydrostatic pressure. The negative sign ensures that the pressure p is positive when the material is in compression. It should be noted that in Drucker-Prager plasticity, the internal variables k and α_{ij} are dimensionless, while in von Mises plasticity, they have the dimension of stress. The associated plastic flow is:

$$m_{ij} = \frac{(s_{ij} - p\alpha_{ij}) + \frac{1}{3}\delta_{ij}\,\alpha_{pq}(s_{pq} - p\alpha_{pq})}{\sqrt{(s_{rs} - p\alpha_{rs})(s_{rs} - p\alpha_{rs})}} \; + \; \sqrt{\frac{2}{27}}\, k\, \delta_{ij}\,, \tag{101}$$

where δ_{ij} is the Kronecker delta symbol. Due to the presence of the pressure term in the yield function, the plastic flow has both deviatoric and volumetric components:

$$\begin{cases} m_{ij}^{dev} = \dfrac{s_{ij} - p\alpha_{ij}}{\sqrt{(s_{rs} - p\alpha_{rs})(s_{rs} - p\alpha_{rs})}} \\ m_{ij}^{vol} = \dfrac{\delta_{ij}\alpha_{pq}(s_{pq} - p\alpha_{pq})}{3\sqrt{(s_{rs} - p\alpha_{rs})(s_{rs} - p\alpha_{rs})}} + \sqrt{\dfrac{2}{27}} k\delta_{ij}, \end{cases} \tag{102}$$

Viscous Damping The dissipative interaction between the solid inside the pores of a porous solid and the viscous fluid, or between the solid outside the solid/structure and the viscous fluid, also dissipates a large amount of mechanical energy. This type of dissipation is called viscous energy dissipation or viscous damping. Viscous energy dissipation is proportional to velocity. When modeling viscous damping without explicitly considering the interaction between the solid/structure and the fluid, Caughey damping is usually adopted.

Rayleigh damping is a special case of Caughey damping, and its damping matrix is obtained by the linear combination of the mass matrix and the stiffness matrix. Hall pointed out that the damping force obtained by Rayleigh damping may be unrealistically high. This can lead to unconservative analysis results, so the damping parameters need to be carefully calibrated.

Sometimes, a high level of Rayleigh damping is used to simulate inelastic material behavior. That is, linear elastic materials and non-physical high-level Rayleigh damping are used to simulate material nonlinear/inelastic behavior. The energy dissipation generated by plastic dissipation and Rayleigh viscous damping will lead to differences in responses, especially when there is significant material nonlinearity/inelastic behavior.

E.1.3. Energy Dissipation Due to Viscous Damping

To calculate the energy dissipation caused by viscous damping, we start with the general form of the equation of motion:

$$M_{ij}\ddot{u}_j(t) + C_{ij}\dot{u}_j(t) + K_{ij}^{elpl}(t)u_j(t) = f_i(t), \tag{103}$$

where $u_j(t)$ is the generalized displacement vector, M_{ij} is the mass matrix, C_{ij} is the damping matrix, $K_{ij}^{elpl}(t)$ is the inelastic stiffness matrix that usually evolves with time, and $f_i(t)$ is the external load vector. For linear viscous damping of the Rayleigh type, the damping matrix can be expressed as:

$$C_{ij} = a_M M_{ij} + a_K K_{ij}^{el}, \tag{104}$$

where a_M and a_K are damping constants with units s^{-1} and s, respectively. The stiffness matrix used to construct the damping matrix is usually the initial tangent stiffness matrix, which, for inelastic materials, is the elastic stiffness matrix K_{ij}^{el}. During the entire simulation, the damping matrix C_{ij} is constant. Hall gave the expressions for calculating the damping constants to achieve the desired damping ratio ξ within a specific frequency range (from $\hat{\omega}$ to $R\hat{\omega}$):

$$\begin{cases} a_M = 2\xi\hat{\omega}\dfrac{2R}{1 + R + 2\sqrt{R}}, \\ a_K = 2\xi\dfrac{1}{\hat{\omega}}\dfrac{2}{1 + R + 2\sqrt{R}}, \end{cases} \tag{105}$$

Rayleigh damping includes a part proportional to mass and a part proportional to stiffness. Combining these two parts of damping can achieve the expected control of the modal damping ratio. However, as Hall pointed out, the use of classical Rayleigh damping must be combined with appropriate damping coefficients, which should enable all modes

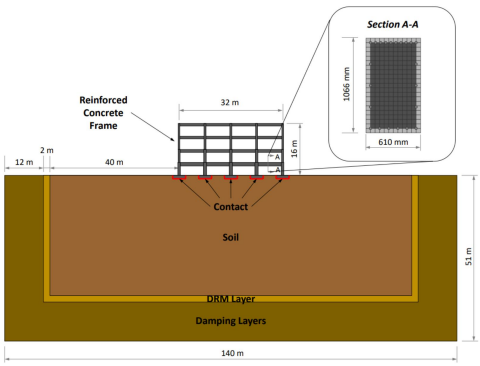

Figure 6. Structure-Interaction Diagram

of interest frequencies to have nearly constant damping values. For modes beyond the specified frequency range, the damping ratio will be unrealistically high. The incremental form of the energy balance for a dynamic system with viscous damping can be expressed as:

$$\Delta E_I = \Delta E_K + \Delta D_V + \Delta W_M, \tag{106}$$

The left side of the equation is the increment of external input work:

$$\Delta E_I = f_i \Delta u_i \tag{107}$$

The three terms on the right side are the kinetic energy increment ΔE_K, the viscous energy dissipation increment ΔD_V, and the material work increment ΔW_M of the system, respectively:

$$\Delta W_M = K_{ij}^{elpl} u_j \Delta u_i, \tag{108}$$

$$\Delta W_M = \Delta E_S + \Delta E_P + \Delta D_P. \tag{109}$$

It should be noted that the material work W_M can be decomposed into elastic strain energy E_S and the plastic work of the system, respectively. Then, as mentioned in the previous section, the plastic work can be further decomposed into plastic free energy E_P and plastic energy dissipation D_P. The energy terms E_S, E_P, and D_P can be calculated by integrating the energy density functions Ψ^{el}, Ψ^{pl}, and Φ over the volume, respectively.

F. Societal Impacts

The community should be aware of the potential negative societal impact that can arise from the misuse of physically consistent video generation models in structural engineering contexts. While our evaluation framework advances the capability to generate realistic structural dynamics videos, such technology could be exploited to fabricate false evidence of structural failures, generate misleading safety assessment reports, or create deceptive visualizations that undermine public trust in building safety evaluations. This could exacerbate issues related to misinformation in disaster response scenarios and compromise the integrity of structural forensic investigations.

Additionally, the biases inherent in the training data may lead to systematic evaluation disparities across different structural types, construction materials, or geographic building codes. Models trained predominantly on data from modern reinforced concrete structures may inadequately evaluate traditional masonry buildings or emerging sustainable construction methods, thereby perpetuating technological disparities and potentially excluding underrepresented construction practices from reliable safety assessments. Such biases could influence policy decisions regarding building codes and retrofit prioritization, affecting vulnerable communities disproportionately.

Therefore, it is imperative that the evaluation of video generation models for structural dynamics not only assesses their technical performance in capturing physical consistency but also considers broader societal implications, ensuring that these technologies contribute positively to public safety and disaster preparedness while mitigating potential risks of misuse. To this end, we plan to incorporate evaluation dimensions focused on cross-domain generalization and fairness across diverse structural typologies in future work, and we advocate for establishing ethical guidelines for the deployment of such models in safety-critical applications.

G. Limitations and Future Work

While our proposed metrics represent the first comprehensive framework for evaluating physically consistent structural dynamics in video generation, several limitations and future directions merit discussion:

- A primary limitation of our work is the reliance on domain-specific sensor data for ground truth validation. Our metrics require synchronized shake table measurements and high-speed camera recordings, which may not be available for all structural engineering scenarios. Future work could explore self-supervised consistency evaluation methods that do not depend on physical sensor measurements, potentially leveraging physics-informed neural networks or inverse dynamics models to estimate ground truth from visual observations alone.
- Our evaluation framework focuses on videos with durations between 2 to 10 seconds at 30 fps, which covers typical seismic response events but may not adequately capture long-duration phenomena such as aftershock sequences or progressive collapse scenarios. For structure-level metrics, longer videos introduce challenges in optical flow accumulation and drift correction. For pixel-level tracking, extended sequences may lead to feature point loss and trajectory fragmentation. We leave the development of robust evaluation protocols for long-duration

structural response videos as important future work.

- The current multi-scale feature extraction operates at three fixed pyramid levels, which may not optimally capture structural behaviors across significantly different spatial scales (e.g., high-rise buildings versus small-scale specimens). Future research could investigate adaptive pyramid construction strategies that automatically determine optimal scale selections based on structure geometry and expected deformation patterns.
- While our metrics effectively evaluate physical consistency for rigid and semi-rigid structures, they may have limited applicability to highly flexible or fluid-structure interaction scenarios where material nonlinearity and large deformations dominate. Extending our framework to such complex scenarios would require incorporating constitutive models and computational mechanics simulations as additional validation references.

To split the supplementary pages from the main paper, you can use Preview (on macOS), Adobe Acrobat (on all OSs), as well as command line tools.