DAMEQ-BENCH: A PHYSICS-AWARE BENCHMARK FOR GRAVITY-DAM SEISMIC VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating physically plausible seismic-response videos of gravity dams remains challenging for modern video generation models. Unlike active motion in natural videos, gravity-dam vibration is a passive forced response governed by the prescribed seismic excitation, reservoir condition, and structural state; consequently, visual realism does not necessarily imply physically correct dynamics. To address this gap, we introduce **DamEQ-Bench**, a comprehensive benchmark for state-conditioned gravity-dam seismic video generation. DamEQ-Bench is constructed from controlled large-scale shaking-table experiments and comprises **198 seismic conditions** spanning empty-reservoir and impounded-reservoir states, together with approximately **1,300 multi-view UAV recordings** and synchronized sensor measurements as experimental references. We formulate gravity-dam seismic video generation as visually conditioned I2V/V2V tasks driven by prescribed seismic conditions and an initial image or video. To jointly assess visual quality and physical-response fidelity, we establish a comprehensive evaluation framework combining **six general visual-quality metrics** with **eight response-oriented metrics** that characterize **structural dynamic response and modal characteristics**. Under a unified protocol, we further evaluate **sixteen representative open- and closed-source video generation models**, together with **two non-generative static control baselines**, and investigate whether lightweight LoRA-based domain adaptation can improve seismic-response fidelity. DamEQ-Bench provides a unified, interpretable, and reproducible testbed for distinguishing visually plausible motion from structural responses that are consistent with prescribed seismic conditions.

1 INTRODUCTION

Gravity dams are critical hydraulic infrastructures for flood control, water supply, and hydropower generation, and their seismic safety is closely related to downstream protection and infrastructure resilience (Tinawi et al., 2000; Huang, 2015). Accordingly, extensive experimental and numerical studies have investigated their dynamic response under earthquake excitation (Donlon & Hall, 1991; Mridha & Maity, 2014; Wang et al., 2018).

Despite rapid advances in video generation, gravity-dam seismic video generation poses a fundamentally different problem from conventional motion synthesis. Dam vibration is a *passively forced structural response* driven by prescribed seismic excitation and governed by dam–reservoir coupling and boundary conditions (Chandrashaker & Humar, 1993; Akköse & Şimşek, 2010; Miquel & Bouaanani, 2010). Consequently, visually similar videos may correspond to substantially different physical responses in terms of amplitude, phase, modal behavior, peak response, and height-wise acceleration amplification. The key question is therefore not whether a generated dam merely *appears to shake realistically*, but whether it reproduces the *measurable structural response* associated with the prescribed seismic and reservoir conditions.

Recent video generation research has mainly focused on Image-to-Video (I2V) and Video-to-Video (V2V) paradigms (Ho et al., 2022b; Xing et al., 2024b). I2V models generate temporal sequences from an initial image, whereas V2V models exploit an observed video prefix to predict or transform subsequent motion (Blattmann et al., 2023; Xing et al., 2024a; Voleti et al., 2022; Chen et al., 2024). However, these models are primarily designed for general visual scenarios and optimized for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

perceptual realism and temporal coherence rather than physically constrained engineering responses (Singer et al., 2023; Ho et al., 2022a; Yang et al., 2025).

Earthquake-Bench (Bao et al., 2026) established a general benchmark for earthquake video generation, examining whether generated motion responds consistently with prescribed seismic loading and observation conditions. DamEQ-Bench addresses a more specific engineering question: whether video generation models can reproduce the condition-dependent dynamics of a coupled gravity-dam–reservoir system. In particular, reservoir impoundment can alter response amplitude, modal characteristics, and spatial amplification even under comparable seismic loading. Thus, DamEQ-Bench moves beyond *seismically plausible motion* toward quantitatively consistent *engineering response generation*.

Constructing such a benchmark involves three major challenges:

- **Scarce Systematic Seismic Observations.** Real gravity-dam earthquake responses are difficult to acquire under repeatable loading and observation conditions, motivating controlled large-scale experiments.
- **Condition-Dependent Dam–Reservoir Dynamics.** Structural response depends jointly on seismic loading and reservoir condition, requiring models to reproduce meaningful response variations rather than generic shaking patterns.
- **Insufficient Engineering-Level Evaluation.** Conventional visual and motion metrics cannot determine whether generated videos preserve measurable dynamic characteristics such as modal behavior, extreme response, and acceleration amplification.

To address these challenges, we introduce **DamEQ-Bench**, a physics-aware benchmark constructed from large-scale gravity-dam shaking-table experiments. It contains **198 controlled seismic conditions** under **empty-reservoir** and **impounded-reservoir** conditions, with diverse seismic loading conditions and multi-view UAV observations. Synchronized sensor measurements provide physically traceable references for evaluation rather than generation inputs.

DamEQ-Bench organizes these experimental observations into standardized I2V and V2V tasks. I2V uses an experiment-specific initial frame together with a standardized seismic-condition prompt, while V2V additionally uses an observed response prefix. Ground-truth videos, seismic inputs, reservoir labels, viewpoint metadata, and sensor responses are aligned under a unified protocol, allowing generated results to be evaluated from both computer-vision and structural-dynamics perspectives.

Our evaluation explicitly distinguishes **visual plausibility** from **physical-response fidelity**. Beyond general visual and motion consistency, we evaluate displacement response, the first three modal characteristics, peak and crest response, and height-wise acceleration amplification. These measurements distinguish videos that merely look plausible from those that reproduce experimentally observed structural dynamics.

Our main contributions are summarized as follows:

- **Gravity-Dam Seismic Video Generation Benchmark.** We introduce **DamEQ-Bench**, a standardized I2V/V2V benchmark built from 198 controlled large-scale shaking-table conditions, aligning multi-view UAV videos, seismic inputs, reservoir conditions, and synchronized sensor references.
- **Engineering-Oriented Physics-Aware Evaluation.** We extend seismic-video evaluation from generic motion consistency to measurable gravity-dam dynamics, including displacement, multi-order modal behavior, extreme response, and height-wise acceleration amplification.
- **Systematic Model Benchmarking and Adaptation.** We evaluate representative open- and closed-source I2V/V2V models under a unified protocol and further investigate lightweight LoRA adaptation, examining whether improved visual quality is accompanied by improved physical-response fidelity.

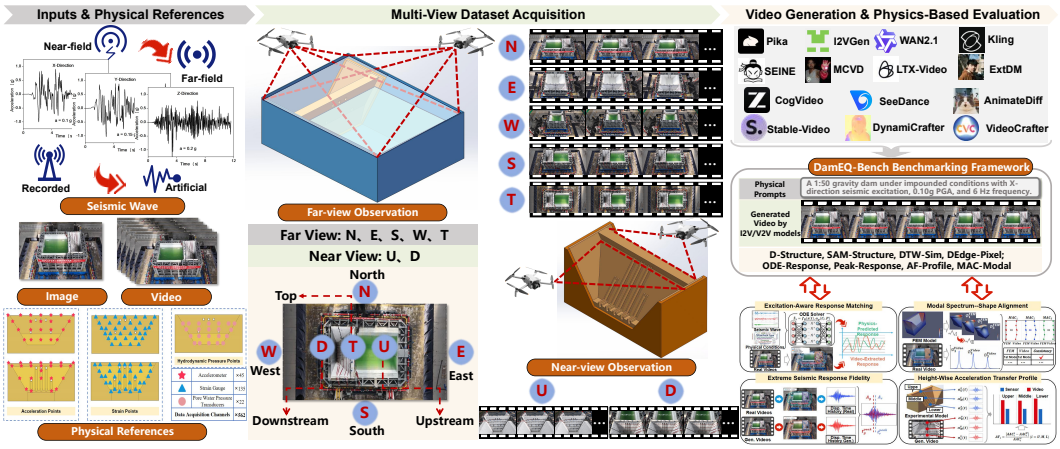


Figure 1: **Overview of DamEQ-Bench.** We propose DamEQ-Bench, a comprehensive physics-based video generation benchmark for gravity-dam seismic response simulation. The benchmark incorporates multi-view UAV observations and diverse seismic conditions, including ground motion, PGA, excitation direction, reservoir and structural state, component ratio, viewpoint, and input frequency when available. We propose four engineering response metrics—ODE-Response, Peak-Response, AF-Profile, and MAC-Modal. We evaluate representative image-to-video and video-to-video generation models on DamEQ-Bench to quantify their visual quality, physical consistency, and engineering-response fidelity.

2 Related Work

2.1 Seismic Response Modeling of Gravity Dams

Earthquake ground motion exhibits pronounced temporal, spectral, and directional variability (Shahi & Baker, 2014), while gravity-dam response is jointly governed by dam inertia, reservoir hydrodynamics, foundation flexibility, and their coupled interactions (Chopra & Chakrabarti, 1981; Fenves & Chopra, 1985b;a). Numerical approaches, including finite-element, nonlinear, and fluid–structure interaction models, provide systematic tools for seismic response analysis (Fenves & Chopra, 1984; 1987; Ghaemian & Ghobarah, 1999), but their predictions remain sensitive to modeling assumptions, boundary conditions, and dam–reservoir–foundation interaction (Ghaemian & Ghobarah, 1999; Arici & Soysal, 2024). Shake-table experiments provide complementary physical evidence by directly observing structural dynamics under controlled seismic excitation (Donlon & Hall, 1991; Mridha & Maity, 2014), although their high cost and operational complexity generally limit the scale and diversity of available observations.

2.2 Video Generation Models

Existing video generation methods can be broadly categorized into Text-to-Video (T2V), Image-to-Video (I2V), and Video-to-Video (V2V) paradigms (Ma et al., 2025). Representative T2V models, such as Phenaki, Imagen Video, and VideoPoet, generate dynamic content from textual conditions (Villegas et al., 2023; Ho et al., 2022a; Kondratyuk et al., 2023), whereas I2V and V2V methods introduce visual conditions to better preserve appearance, structure, and temporal coherence (Blattmann et al., 2023; Zhang et al., 2023; Esser et al., 2023). Despite substantial advances in visual quality, recent physics-oriented benchmarks show that visually plausible generation does not necessarily preserve physically valid dynamics (Meng et al., 2025; Guo et al., 2025).

Gravity-dam seismic video generation differs fundamentally from conventional video synthesis because the observable structural motion is jointly governed by seismic excitation, reservoir interaction, foundation properties, and structural and boundary conditions (Silveira et al., 2021; Gorai & Maity, 2022; Khoshnafas Anari et al., 2024). Consequently, visually smooth vibration may still deviate from the physically admissible seismic response, while variations in reservoir, foundation, or structural state can produce markedly different dynamic responses under similar seismic excitations.

2.3 Evaluation Metrics

Video-generation models are commonly evaluated through human judgments and automated metrics. Human evaluation can directly assess perceptual realism, temporal coherence, subject preservation, and condition alignment, but is costly to scale and difficult to standardize across large-scale model comparisons (Liu et al., 2024; He et al., 2024). Consequently, recent benchmarks increasingly employ automated and multi-dimensional evaluation protocols, such as EvalCrafter and VBench++, to characterize visual quality, temporal consistency, motion quality, and condition alignment (Liu et al., 2024; Huang et al., 2026).

Full-reference metrics further quantify differences between generated and reference observations from complementary perspectives. PSNR measures pixel-level reconstruction fidelity (Huynh-Thu & Ghanbari, 2008), SSIM evaluates structural similarity (Wang et al., 2004), DISTS combines structure and texture similarity (Ding et al., 2022), gradient-based metrics characterize local structural variations (Liu et al., 2012), and LPIPS measures perceptual similarity in deep feature space (Zhang et al., 2018). Together, these metrics provide effective measures of visual and generic temporal quality, but they do not determine whether the generated motion is the physically correct structural response to a prescribed seismic input.

3 DamEQ-Bench Formulation

In this section, we present the design of DamEQ-Bench by first defining the gravity-dam seismic video generation problem. We then introduce the data and equipment infrastructure. Finally, we describe the construction of the benchmark.

3.1 Problem Definition

Physical Foundation. We consider a gravity dam–reservoir system subjected to multi-component seismic base excitation, where each experimental condition is characterized by the earthquake record, target peak ground acceleration (PGA), excitation direction, and structural–reservoir state. Its forced response can be compactly expressed as:

$$\underbrace{\mathcal{M}_s\ddot{\mathbf{z}}_c(t)}_{\text{inertia}}+\underbrace{\mathcal{C}_s\dot{\mathbf{z}}_c(t)}_{\text{dissipation}}+\underbrace{\mathcal{K}_s\mathbf{z}_c(t)}_{\text{restoring response}}+\underbrace{\mathcal{F}_s\left(\mathbf{z}_c,\dot{\mathbf{z}}_c\right)}_{\text{state-dependent effects}}=-\underbrace{\mathcal{B}_s\mathbf{a}_c(t)}_{\text{seismic forcing}}, \tag{1}$$

Here, $\mathcal{M}_s$, $\mathcal{C}_s$, and $\mathcal{K}_s$ denote the state-dependent mass, damping, and stiffness operators, respectively, while $\mathcal{F}_s$ captures dam–reservoir coupling and other state-dependent effects.

Video Generation Formulation. We formulate gravity-dam seismic video generation as a conditional synthesis problem that preserves both the experimental scene and the seismic-response characteristics associated with condition c. For an observation viewpoint v, the target video is represented as

$$V_{c,v}=\left\{I_{c,v}^1,\ldots,I_{c,v}^N\right\}, \tag{2}$$

Since pure T2V generation cannot uniquely specify the scene-specific configurations, we focus on visually conditioned I2V and V2V generation tasks.

I2V Generation. Given an experiment-specific initial frame $I_{c,v}^0$ and the standardized condition prompt $T(c,v)$, the I2V generation process is formulated as

$$\underbrace{\left(I_{c,v}^0,\,T(c,v)\right)}_{\substack{\text{initial structural frame}\\+\text{ seismic condition prompt}}}\xrightarrow{f_{\text{I2V}}(\,\cdot\,;\theta_{\text{I2V}})}\underbrace{\hat{V}_{c,v}=\left\{\hat{I}_{c,v}^1,\ldots,\hat{I}_{c,v}^N\right\}}_{\text{generated seismic-response video}}, \tag{3}$$

V2V Generation. For V2V generation, the observed visual condition is the first 12 frames of the same experimental sequence,

$$V_{c,v}^{1:12}=\left\{I_{c,v}^1,\ldots,I_{c,v}^{12}\right\}, \tag{4}$$

Table 1: Technical specifications of the large-scale seismic simulation shaking table.

Parameter	Specification
Platform dimensions	$20\ \mathrm{m} \times 16\ \mathrm{m}$
Maximum payload	$1350\ \mathrm{t}$
Horizontal peak acceleration	$\pm 1.5g$
Vertical peak acceleration	$\pm 2.0g$
Effective frequency band	0.1–25 Hz

Conditioned on this observed response prefix and the same standardized seismic prompt, the V2V model predicts the subsequent dynamic response:

$$\underbrace{\left(V_{c,v}^{1:12},\, T(c,v)\right)}_{\substack{\text{12-frame response prefix} \\ +\ \text{seismic condition prompt}}} \xrightarrow{f_{\mathrm{V2V}}(\,\cdot\,;\theta_{\mathrm{V2V}})} \underbrace{\hat{V}_{c,v}^{13:N} = \left\{\hat{I}_{c,v}^{13}, \ldots, \hat{I}_{c,v}^{N}\right\}}_{\text{predicted future seismic response}}, \tag{5}$$

3.2 Data and Equipment Infrastructure

Seismic Waveforms. Each ground-motion component is normalized to the prescribed PGA and projected onto the commanded shaking-table direction. multi-axisexcitation is then constructed by superposing the scaled directional components:

$$\mathbf{a}_{\mathrm{cmd}}(t) = \sum_{i\in\{x,y,z\}} \underbrace{\alpha_i}_{\text{directional ratio}} \underbrace{\mathbf{P}_i}_{\text{direction mapping}} \underbrace{\frac{p_i g}{\max\limits_{\tau} \left\|\mathbf{a}^{(i)}(\tau)\right\|_{\infty}}}_{\text{PGA scaling}} \underbrace{\mathbf{a}^{(i)}(t)}_{\text{ground motion}}, \tag{6}$$

Large-scale Shake Table System. The prescribed seismic motions are reproduced by a large-scale three-directional, six-DOF shaking table:

$$\underbrace{\mathbf{a}_{\mathrm{cmd}}(t)}_{\text{target acceleration}} \xrightarrow{\int\int} \underbrace{\mathbf{x}_{\mathrm{des}}(t)}_{\text{platform trajectory}} \xrightarrow{\mathcal{K}^{\dagger}} \underbrace{\mathbf{u}(t)}_{\text{actuator commands}}, \tag{7}$$

where $\mathcal{K}^{\dagger}$ denotes the inverse kinematic mapping implemented by the control system.

The key technical specifications of the large-scale shaking-table system are provided in Table A1.

Scaled Gravity-Dam Model. Mounted on the shaking-table system is a geometrically scaled gravity-dam model comprising the dam body, foundation, and reservoir domain. The geometric similarity ratio between the model and prototype is

$$\lambda_L = \frac{L_m}{L_p} = \frac{1}{50}, \tag{8}$$

where subscripts m and p denote model- and prototype-scale quantities, respectively.

Let $\lambda_\rho = \rho_m/\rho_p$ and $\lambda_E = E_m/E_p$ denote the density and elastic-modulus similarity ratios. Under elastic similitude, the principal dynamic scaling relationships are

$$\lambda_t = \lambda_L\sqrt{\frac{\lambda_\rho}{\lambda_E}}, \quad \lambda_f = \frac{1}{\lambda_L}\sqrt{\frac{\lambda_E}{\lambda_\rho}}, \quad \lambda_a = \frac{\lambda_E}{\lambda_\rho\lambda_L}, \quad \lambda_v = \sqrt{\frac{\lambda_E}{\lambda_\rho}}, \quad \lambda_d = \lambda_L. \tag{9}$$

where λ_t, λ_f, λ_a, λ_v, and λ_d denote the time, frequency, acceleration, velocity, and displacement similarity ratios, respectively.

The dam–foundation–reservoir system is housed in a rigid model container and securely connected to the shaking table to minimize relative slip and ensure stable transmission of seismic excitation.

UAV-based Video Acquisition. To mitigate camera vibration and motion blur induced by strong shaking-table excitation, DamEQ-Bench employs UAV-based stabilized imaging to provide clearer and mechanically decoupled structural observations.

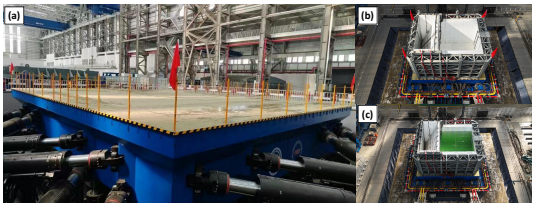


Figure 2: Experimental setup: (a) large-scale three-directional six-degree-of-freedom shaking table; (b) impounded gravity-dam model; and (c) dry gravity-dam model.

3.3 Dataset Construction Methodology

State-Conditioned Seismic Configurations. DamEQ-Bench is constructed exclusively from physically executed shaking-table experiments.

The benchmark contains two reservoir conditions spanning three experimental stages:

- **Empty reservoir**: dam response without reservoir-water interaction;
- **Impounded reservoir**: response with dam–reservoir interaction.

Camera Viewpoint Variations. DamEQ-Bench organizes UAV recordings using both far- and near-view configurations.

For far-view observations, four lateral viewpoints are represented by nominal azimuths

$$\phi \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}, \tag{10}$$

together with an additional top view. Here, ϕ denotes the nominal viewing azimuth relative to a fixed benchmark reference direction.

Near-view observations focus on the two physically important dam faces,

$$v_{\text{near}} \in \{\text{Upstream}, \text{Downstream}\}, \tag{11}$$

Summary. DamEQ-Bench comprises approximately **1,300 multi-view UAV recordings** across **198 seismic conditions**, including **70 empty-reservoir** and **128 impounded-reservoir** conditions.

4 Evaluation Metrics

DamEQ-Bench emphasizes the evaluation of **physical-response consistency** by integrating computer-vision-based motion analysis, sensor-referenced response analysis, and time-series comparison. To jointly assess the **visual quality and physical-response fidelity** of generated videos, we establish a comprehensive evaluation framework that systematically characterizes gravity-dam seismic video generation from multiple complementary dimensions, as illustrated in Figure 3.

- **Modal Spectrum–Shape Alignment.** We extract multi-point displacement responses along the dam height and estimate the first three modal frequencies and mode shapes. Frequency similarity and MAC jointly measure agreement with the reference vibration patterns.
- **Excitation-Aware Response Matching.** Video-derived responses are compared with continuous-time trajectories predicted by a Neural ODE conditioned on the recorded seismic excitation and structural–reservoir state. The Neural ODE is trained only on training data, allowing excitation-aware physical evaluation.
- **Extreme Seismic Response Fidelity.** We compare peak displacement amplitude, peak timing, and crest factor between generated and reference responses to evaluate the fidelity of critical strong-motion extremes.
- **Height-Wise Acceleration Transfer Consistency.** At lower, middle, and upper dam elevations, we compare acceleration amplification, height-wise gradient, spatial profile, and peak timing to evaluate whether the generated response reproduces the characteristic amplification along the dam height.

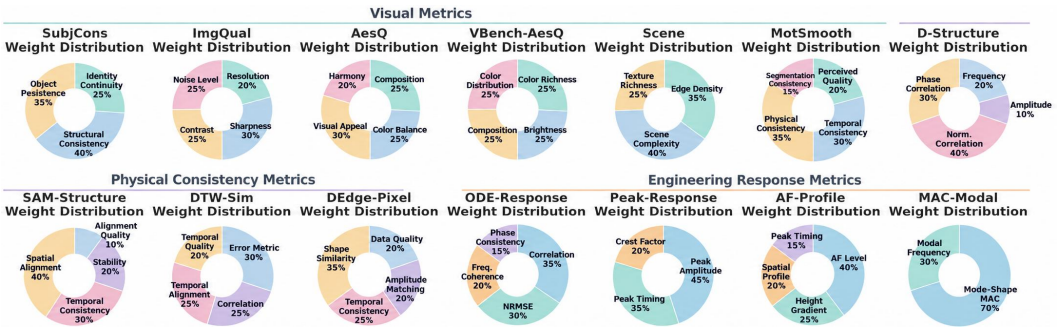


Figure 3: Component Diagram of Evaluation Metrics.

Due to space limitations, detailed procedures for these metrics are provided in the Appendix.

5 EXPERIMENTS

We benchmark a diverse set of open- and closed-source I2V/V2V models on **DamEQ-Bench**, assessing both visual quality and physical-response fidelity, with additional analyses of LoRA adaptation and performance across structural–reservoir states.

5.1 BASELINES

I2V Models. I2V models generate gravity-dam seismic response videos from a single experiment-specific initial frame conditioned on standardized seismic prompts, representing a controlled "static-to-dynamic" generation setting. We evaluate a diverse set of open-source and commercial I2V models, including **AnimateDiff-SparseCtrl**, **CogVideoX-I2V**, **DynamiCrafter**, **I2VGen-XL**, **LTX-Video-2B-distilled**, **Stable-Video**, **VideoCrafter1-I2V**, **WAN2.1**, **Kling**, **Seedance**, and **Veo**. **StaticRepeat-I2V** is included as a non-generative static control baseline.

V2V Models. V2V models take an observed prefix of the real shaking-table video together with the corresponding seismic-condition prompt and predict the subsequent structural response. We evaluate **ExtDM**, **MCVD**, **Pika**, **SEINE**, and **WAN2.1-VACE-1.3B**, with **StaticRepeat-V2V** included as a non-generative static control reference.

5.2 EXPERIMENT SETUP

Preprocessing for ODE-Response Evaluation. Displacement responses $u_i(t)$ are extracted from the same predefined dam locations using an identical motion-tracking pipeline, converted into physical units by fixed view-specific calibration, synchronized with the seismic input $a_g(t)$, and resampled onto a common time grid. Only the DC component is removed, without amplitude normalization. The generated responses $u_i^g(t)$ are compared with Neural ODE predictions $u_i^{\mathrm{ODE}}(t)$ in the time and frequency domains, with multi-location results macro-averaged.

Preprocessing for Peak-Response Evaluation. Displacement responses $u(t)$ at the dam crest or representative locations are uniformly calibrated, synchronized, and resampled. Within the common vibration interval $[t_0, t_1]$, we extract the peak amplitude P, peak occurrence time t_p, and crest factor CF. No min–max or peak-wise normalization is applied, preserving physically meaningful differences in extreme responses.

Preprocessing for AF-Profile Evaluation. Responses are extracted at the lower, middle, and upper elevations z_1, z_2, and z_3. Experimental accelerations $a_i^s(t)$ are obtained from synchronized sensors, while video-derived accelerations $a_i^v(t)$ are recovered from calibrated displacement $u_i(t)$ using identical Savitzky–Golay smoothing and second-order differentiation. Using the base acceleration $a_0(t)$ as reference, we obtain $\mathbf{AF} = [AF_1, AF_2, AF_3]$ for height-wise comparison.

Preprocessing for MAC-Modal Evaluation. Multi-point displacement responses $\{u_i(t)\}_{i=1}^{M}$ along the dam height are uniformly resampled and analyzed using the same CPSD–FDD pipeline. Modal frequencies f_k and mode shapes ϕ_k are identified, with experimental modes used as references and generated modes matched by frequency proximity and MAC_k. Only reliably excited modes $k \in \mathcal{K}$ are retained for final scoring.

5.3 Overall Benchmark Results

Table 2 presents the overall benchmark results for 16 video generation models and two static control baselines.

Table 2: Overall evaluation results of representative I2V and V2V baselines.

	Visual Metrics						Physical Consistency Metrics				Engineering Response Metrics			
Model	**Subject Consistency**	**Aesthetic Quality**	**VBench Aesthetic**	**Temporal Flickering**	**Imaging Quality**	**Scene Coherence**	**D-Structure**	**SAM-Structure**	**DTW-Sim**	**DEdge-Pixel**	**ODE-Response**	**Peak-Response†**	**AF-Profile**	**MAC-Modal**
I2V Models														
AnimateDiff-SparseCtrl	0.815	0.641	0.676	0.915	0.788	0.382	0.123	0.788	0.689	0.621	0.558	0.555	0.370	0.473
CogVideoX-I2V	0.644	0.625	0.671	0.826	0.713	**0.440**	0.128	0.470	0.616	0.468	0.327	0.347	**0.557**	0.571
DynamiCrafter	0.578	0.616	0.602	0.833	0.813	0.363	0.121	0.562	0.643	0.620	0.380	0.496	0.403	0.538
I2VGen-XL	0.521	0.602	0.492	0.825	0.763	0.346	0.129	0.491	0.555	0.551	0.444	0.398	0.387	0.456
LTX-Video-2B-distilled	0.767	0.634	0.649	0.913	0.811	0.381	0.125	0.792	0.714	0.639	0.472	0.596	0.358	0.585
Kling	0.887	0.657	0.733	0.954	0.761	0.410	0.134	0.509	0.668	**0.656**	0.247	0.430	0.335	**0.762**
Seedance	0.935	0.651	0.715	0.975	0.762	0.404	**0.180**	0.509	**0.744**	0.516	0.502	0.535	0.385	0.622
Veo	0.760	0.659	**0.735**	0.918	0.763	0.410	0.115	0.497	0.616	0.515	0.449	0.433	0.305	0.565
StaticRepeat-I2V	**0.997**	0.649	0.677	**0.999**	0.775	0.396	0.132	**0.967**	0.695	0.655	0.504	0.530	0.244	0.527
VideoCrafter1-I2V	0.755	**0.684**	0.728	0.873	**0.829**	0.401	0.126	0.582	0.669	0.601	**0.602**	**0.604**	0.369	0.486
WAN2.1	0.856	0.650	0.675	0.941	0.786	0.397	0.118	0.911	0.700	0.642	0.469	0.565	0.327	0.572
Stable-Video	0.666	0.624	0.630	0.861	0.811	0.370	0.120	0.611	0.645	0.589	0.441	0.528	0.353	0.513
V2V Models														
ExtDM	0.818	0.589	0.558	0.929	0.775	0.360	0.125	0.928	0.682	0.634	0.411	0.487	0.336	0.451
MCVD	0.715	0.644	0.601	0.792	**0.833**	0.368	0.124	0.694	**0.770**	0.484	0.387	0.452	0.488	0.424
SEINE	0.565	0.673	0.685	0.755	0.809	0.405	0.129	0.506	0.721	0.385	0.430	0.489	0.458	0.497
StaticRepeat-V2V	**0.990**	0.650	0.675	**0.996**	0.776	0.396	**0.134**	**0.969**	0.693	0.722	0.522	0.537	0.225	0.596
Pika	0.975	**0.688**	**0.726**	0.990	0.821	**0.414**	0.098	0.534	0.747	**0.822**	**0.589**	**0.690**	**0.627**	**0.700**
WAN2.1-VACE-1.3B	0.872	0.674	0.689	0.940	0.765	0.399	0.113	0.911	0.677	0.638	0.532	0.566	0.223	0.581

Note: *Blue and pink cells indicate the best performance among I2V and V2V models, respectively. Underlined values indicate the second-best result within each generation paradigm. Tied second-best values are underlined simultaneously.*

5.4 Domain Adaptation with LoRA

The zero-shot benchmark examines whether general-purpose video models possess sufficient priors for seismic-response generation. We further adapt three representative I2V models on the training split using LoRA. Table 3 uses eight response-oriented metrics to assess whether lightweight domain adaptation improves gravity-dam seismic-response modeling.

Table 3: Comparison of original and LoRA-adapted I2V models.

		Physical Consistency Metrics				Engineering Response Metrics			
Backbone	**Setting**	**D-Structure**	**SAM-Structure**	**DTW-Sim**	**DEdge-Pixel**	**ODE-Response**	**Peak-Response**	**AF-Profile**	**MAC-Modal**
CogVideoX-I2V	Original	0.128	0.470	0.616	0.468	0.327	0.347	0.557	0.571
	LoRA	0.122	0.416	0.613	0.509	0.331	0.358	0.424	0.520
WAN2.1	Original	0.118	0.911	0.700	0.642	0.469	0.565	0.327	0.572
	LoRA	0.118	0.905	0.685	0.649	0.449	0.575	0.331	0.584
Stable-Video	Original	0.120	0.611	0.645	0.589	0.441	0.528	0.353	0.513
	LoRA	0.123	0.489	0.692	0.404	0.470	0.565	0.480	0.470

5.5 Performance across Seismic Conditions

To assess whether I2V models preserve dynamic responses across reservoir states, Table 4 compares their response consistency under empty- and impounded-reservoir conditions.

6 Insights and Discussion

DamEQ-Bench highlights a key distinction between conventional video generation and seismic-response generation: *visual plausibility does not guarantee physical consistency*. Gravity-dam motion is a forced response governed by seismic excitation and the dam–reservoir state, so visually realistic videos may still exhibit incorrect amplitude, timing, dynamics, or spatial response.

Table 4: Dynamic-Response Consistency under empty and impounded reservoirs.

Model	Empty-Reservoir Condition				Impounded-Reservoir Condition			
	ODE-Response	Peak-Response	MAC-Modal	AF-Profile	ODE-Response	Peak-Response	MAC-Modal	AF-Profile
VideoCrafter1-I2V	**0.600**	**0.564**	0.460	0.410	**0.603**	**0.627**	0.498	0.397
LTX-Video-2B-distilled	0.445	0.556	0.600	0.410	0.488	0.617	**0.573**	0.374
WAN2.1	0.430	0.542	0.538	0.420	0.476	0.584	0.545	0.334
CogVideoX-I2V	0.328	0.359	**0.782**	**0.587**	0.325	0.337	0.496	**0.545**

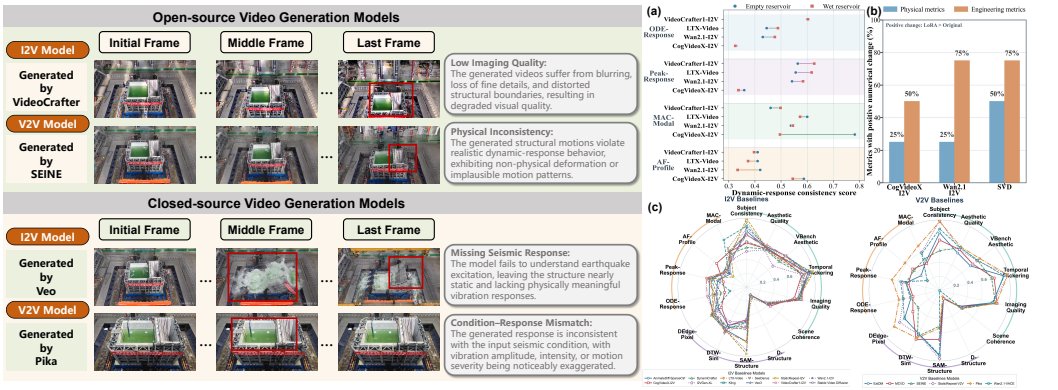


Figure 4: Qualitative and Quantitative Results on DamEQ-Bench. (a) reservoir-state response comparison; (b) LoRA metric improvements; (c) multi-metric performance profiles.

Complementarity of visual and physics-based evaluation. DamEQ-Bench uses fourteen metrics to characterize visual quality and physical response. Visual metrics assess appearance and temporal stability, while eight response-oriented metrics evaluate displacement, local motion, geometry, temporal alignment, modal behavior, conditioned dynamics, peak response, and acceleration amplification. Reporting them separately provides interpretable diagnosis than a single aggregate score.

Complementarity of visual and physics-based evaluation. DamEQ-Bench uses fourteen metrics to characterize visual quality and physical response. Visual metrics assess appearance and temporal stability, while eight response-oriented metrics evaluate displacement, local motion, geometry, temporal alignment, modal behavior, conditioned dynamics, peak response, and acceleration amplification. Reporting them separately provides interpretable diagnosis than a single aggregate score.

State dependence is a major source of difficulty. Gravity-dam response varies with reservoir state, loading history, and damage, even under similar visual conditions. Robust models must therefore learn state-dependent response variation rather than memorize recurring motion patterns, making unseen condition generalization particularly challenging.

Can domain adaptation learn seismic dynamics? The LoRA experiments test whether lightweight adaptation can inject seismic-response knowledge into general-purpose video models. Improvements in appearance alone are insufficient; meaningful adaptation should instead improve response-oriented metrics such as temporal alignment, conditioned response, peak behavior, acceleration amplification, and modal consistency.

7 Conclusion

We introduced **DamEQ-Bench**, a state-conditioned benchmark for gravity-dam seismic video generation built from 198 controlled shaking-table experiments. By combining visual-quality and response-oriented evaluation, DamEQ-Bench reveals that perceptual realism alone is insufficient for physically grounded seismic video generation. Our results establish reproducible baselines and highlight condition-aware structural dynamics as a key direction for future video generation models.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing, translation, and improving the clarity and readability of the manuscript. They were not used to generate experimental data, design the experiments, develop the benchmark methodology, implement the evaluation methods, perform quantitative analysis, or produce the reported results. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REFERENCES

M. Akköse and E. Şimşek. Non-linear seismic response of concrete gravity dams to near-fault ground motions including dam-water-sediment-foundation interaction. *Applied Mathematical Modelling*, 34(11):3685–3700, 2010.

Y. Arici and B. F. Soysal. Predicting seismic damage on concrete gravity dams: A review. *Structure and Infrastructure Engineering*, 20(9):1354–1373, 2024.

L. Bao, H. Chen, Y. Chen, et al. Earthquake-bench: Video generation benchmark for earthquake simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4055–4066, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Rune Brincker, Lingmi Zhang, and Palle Andersen. Modal identification of output-only systems using frequency domain decomposition. *Smart Materials and Structures*, 10(3):441–445, 2001. doi: 10.1088/0964-1726/10/3/303.

R. Chandrashaker and J. L. Humar. Fluid–foundation interaction in the seismic response of gravity dams. *Earthquake Engineering & Structural Dynamics*, 22(12):1067–1084, 1993.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. SEINE: Short-to-long video diffusion model for generative transition and prediction. In *International Conference on Learning Representations*, 2024.

Anil K. Chopra and P. Chakrabarti. Earthquake analysis of concrete gravity dams including dam–water–foundation rock interaction. *Earthquake Engineering & Structural Dynamics*, 9(4):363–383, 1981.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

W. P. Donlon and J. F. Hall. Shaking table study of concrete gravity dam monoliths. *Earthquake Engineering & Structural Dynamics*, 20(8):769–786, 1991.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7346–7356, 2023.

Gregory Fenves and Anil K. Chopra. Simplified earthquake analysis of concrete gravity dams: Combined hydrodynamic and foundation interaction effects. *Journal of Engineering Mechanics*, 111(6):736–756, 1985a.

Gregory Fenves and Anil K. Chopra. Simplified earthquake analysis of concrete gravity dams: Separate hydrodynamic and foundation interaction effects. *Journal of Engineering Mechanics*, 111(6):715–735, 1985b.

Gregory Fenves and Anil K. Chopra. Simplified earthquake analysis of concrete gravity dams. *Journal of Structural Engineering*, 113(8):1688–1708, 1987.

Gregory L. Fenves and Anil K. Chopra. Earthquake analysis and response of concrete gravity dams. Technical report, Earthquake Engineering Research Center, University of California, Berkeley, Berkeley, CA, 1984.

Mohsen Ghaemian and A. Ghobarah. Nonlinear seismic response of concrete gravity dams with dam–reservoir interaction. *Engineering Structures*, 21(4):306–315, 1999.

Soumya Gorai and Damodar Maity. Seismic performance evaluation of concrete gravity dams in finite-element framework. *Practice Periodical on Structural Design and Construction*, 27(1): 04021072, 2022. doi: 10.1061/(ASCE)SC.1943-5576.0000656.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022b.

J. Huang. Earthquake damage analysis of concrete gravity dams: Modeling and behavior under near-fault seismic excitations. *Journal of Earthquake Engineering*, 19(7):1037–1085, 2015.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890.

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008. doi: 10.1049/el:20080522.

Mojtaba Khoshnafas Anari, S. Mohammad Mirhosseini, Seyed Hamid Lajevardi, and Ehsanollah Zeighami. Effect of foundation stiffness on the fragility curves of a concrete gravity dam under far-field ground motions. *Results in Engineering*, 24:102962, 2024. doi: 10.1016/j.rineng.2024.102962.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2012. doi: 10.1109/TIP.2011.2175935.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, Zeyu Wang, Zhifeng Li, Xiu Li, Wei Liu, Dan Xu, Linfeng Zhang, and Qifeng Chen. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806, 2025.

B. Miquel and N. Bouaanani. Simplified evaluation of the vibration period and seismic response of gravity dam–water systems. *Engineering Structures*, 32(8):2488–2502, 2010.

S. Mridha and D. Maity. Experimental investigation on nonlinear dynamic response of concrete gravity dam-reservoir system. *Engineering Structures*, 80:289–297, 2014.

Shrey K. Shahi and Jack W. Baker. Nga-west2 models for ground motion directionality. *Earthquake Spectra*, 30(3):1285–1300, 2014.

Iarly Vanderlei da Silveira, Lineu José Pedroso, and Giuliano Santa Marotta. Study of the influence of the foundation and the reservoir on the dynamic response in a concrete gravity dam profile. *Revista IBRACON de Estruturas e Materiais*, 14(4):e14403, 2021. doi: 10.1590/S1983-41952021000400003.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023.

R. Tinawi, P. Léger, M. Leclerc, et al. Seismic safety of gravity dams: From shake table experiments to numerical analyses. *Journal of Structural Engineering*, 126(4):518–529, 2000.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23371–23385, 2022.

M. Wang, J. Chen, and W. Xiao. Experimental and numerical comparative study on gravity dam-reservoir coupling system. *KSCE Journal of Civil Engineering*, 22(10):3980–3987, 2018.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Ying Shan, Tien-Tsin Wong, and Chao Weng. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024a.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024b. doi: 10.1145/3696415.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

APPENDIX

APPENDIX CONTENTS

A Experimental System and Physical Similitude Design **16**

A.1 Large-Scale Shaking-Table System 16

A.2 Scaled Gravity-Dam Model and Model-Box System 16

A.3 Physical Similitude and Scaling Relations 20

A.4 Model Material Properties and Design Requirements 22

A.5 Boundary Conditions, Model Installation, and Numerical Characterization 22

A.6 Damage Accumulation and Local Repair Protocol 23

B Earthquake Inputs and Experimental Condition Design **24**

B.1 Ground-Motion Library 24

B.2 Ground-Motion Preprocessing 26

B.3 PGA Scaling and Multi-Axis Excitation 28

B.4 Experimental State Evolution and Loading History 29

B.5 Experimental Matrix 30

C Multimodal Data Acquisition System **35**

C.1 UAV-Based Multi-View Video Acquisition 35

C.2 Far-View Camera Configuration 35

C.3 Near-View Camera Configuration 36

C.4 Sensor Layout and Instrumentation System 37

C.5 Ground-Motion Preprocessing 38

C.6 PGA Scaling and Multi-Axis Excitation 38

C.7 Auxiliary Identification and Calibration Tests 39

C.8 Video–Sensor Temporal Synchronization 40

D Dataset Construction and Task Protocols **41**

D.1 From Raw Experimental Records to DamEQ-Bench 41

D.2 Dataset Statistics and Coverage 42

D.3 Standardized Prompt Construction 43

E Unified Evaluation Preprocessing and Fairness Protocol **44**

E.1 Generated-to-Reference Video Pairing 44

E.2 Spatial and Temporal Standardization 44

E.3 Global Camera-Motion Compensation 46

E.4 Structural ROI and Mask Extraction 47

E.5 Optical Flow and Motion Preprocessing 47

E.6 Acceleration-Only Sensor Alignment 49

F Visual Quality Metrics **50**

F.1 Subject Consistency 50

F.2 Official VBench Aesthetic Quality 51

F.3 Custom Aesthetic Presentation Quality 51

F.4 Temporal Flickering 52

F.5 Imaging Quality 53

F.6 Scene Coherence 54

G Response-Oriented Evaluation Metrics **54**

G.1 Structure-Level Displacement Difference (D-Structure) 56

G.2 Pixel-Level Edge Difference (DEdge-Pixel) 56

G.3 Segment-Level Structure Consistency (SAM-Structure) 58

G.4 Dynamic Time Warping Similarity (DTW-Sim) 59

G.5 Multi-Order Modal Consistency 60

G.6 Dynamic Response Consistency 61

G.7 Peak/Crest Response Consistency 62

G.8 Acceleration Amplification Consistency 63

H Baseline Generation and LoRA Adaptation Results **64**

H.1 Qualitative Results of Baseline Models 64

H.2 Qualitative Results of LoRA Adaptation 65

H.3 Quantitative Results on the Evaluated Subset 66

H.4 Metadata Verification 66

H.5 Limitations of the Scaled Physical Model 67

H.6 Limitations of Sensor-Based References 68

H.7 Generalization Scope 68

H.8 Experimental Cost, Data Scarcity, and Resource Reuse 69

A Experimental System and Physical Similitude Design

A.1 Large-Scale Shaking-Table System

The experimental system provides three-axis excitation with six-degree-of-freedom motion control. The shaking table has a platform size of $20\ \mathrm{m} \times 16\ \mathrm{m}$ and a maximum payload of $1350\ \mathrm{t}$. Under full-payload conditions, the nominal peak acceleration capacities are $1.5g$ in the horizontal directions and $2.0g$ in the vertical direction, with an effective operating frequency range of 0.1–$25\ \mathrm{Hz}$. The system is driven by 28 horizontal and 18 vertical actuators, enabling multi-axis excitation of the large-scale dam–foundation–reservoir model.

For each experimental condition, the processed earthquake record was converted into a command signal according to the prescribed PGA, excitation direction, and component ratio. The command signal was then executed by the shaking-table control system to excite the model box and the enclosed dam–foundation–reservoir system. Importantly, the commanded motion and the actual excitation experienced by the specimen were treated as distinct quantities. We therefore archived the target input, control command, and measured table/base acceleration separately. Figure A1 further illustrates the hierarchical control architecture used to convert the prescribed reference acceleration into the realized shaking-table response. The reference signal is processed through signal conditioning, multi-stage control, inverse compensation, and feedforward decoupling before producing the measured acceleration response.

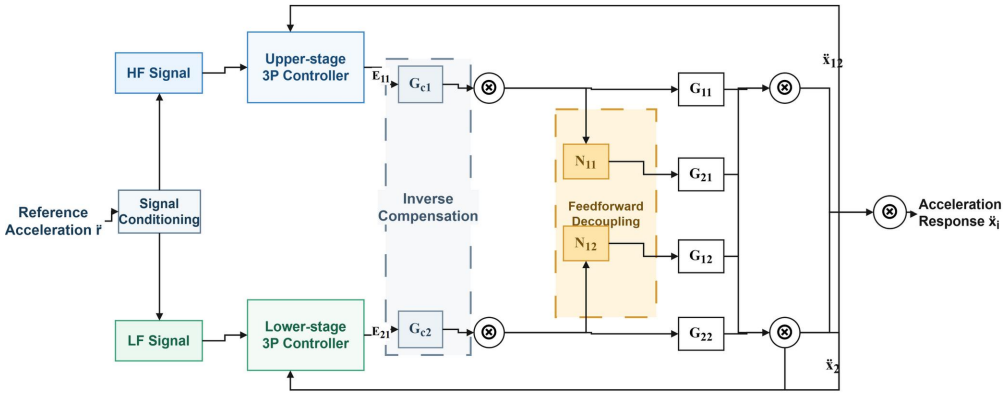


Figure A1: Control architecture of the large-scale shaking-table system.

All response-based analyses in DamEQ-Bench use the calibrated measured table or dam-base acceleration as the reference excitation whenever available. The nominal target PGA is retained as experimental metadata but is not assumed to be identical to the achieved PGA. This distinction is particularly important for high-payload and multi-axis tests, where servo-control performance and specimen–table interaction may introduce deviations between the prescribed and realized motions.

Table A1 summarizes the nominal facility specifications used to characterize the experimental platform. Condition-specific excitation amplitudes and measured input histories are reported separately with the corresponding loading records.

The actuator arrangement of the large-scale shaking-table system is illustrated in Fig. A2, showing the 28 horizontal and 18 vertical hydraulic actuators used to realize the multi-axis excitation.

A.2 Scaled Gravity-Dam Model and Model-Box System

The gravity-dam specimen was constructed according to a geometric scale of $1{:}50$. The experimental system retained the principal physical components required to reproduce the observable seismic response of a gravity dam, including the concrete dam body, the near-dam foundation region, the

Table A1: Nominal specifications of the large-scale seismic simulation shaking table.

Parameter	**Specification**
Control capability	Three-axis, six-degree-of-freedom (6-DOF)
Table size	$20\ \mathrm{m} \times 16\ \mathrm{m}$
Maximum payload	$1350\ \mathrm{t}$
Horizontal peak acceleration	$1.5g$ at full payload
Vertical peak acceleration	$2.0g$ at full payload
Effective frequency range	0.1–25 Hz
Actuator configuration	28 horizontal + 18 vertical actuators
Anchor-bolt configuration	357 M48 + 638 M36

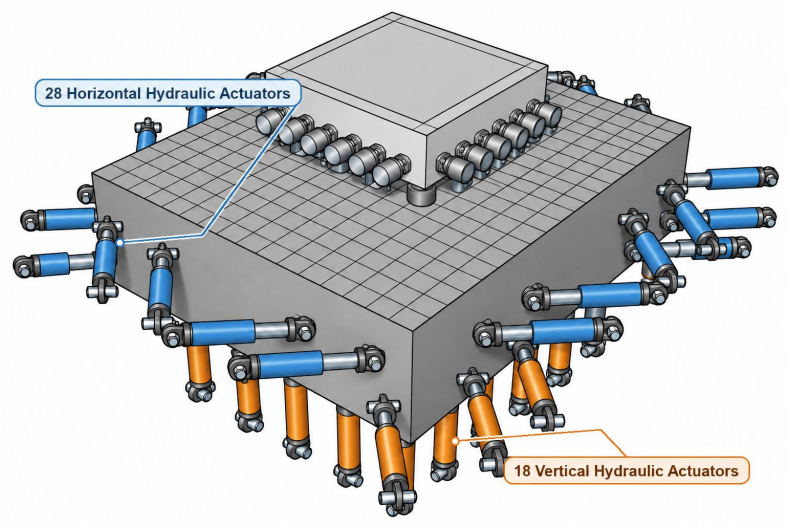


Figure A2: Actuator configuration of the large-scale shaking-table system, including 28 horizontal and 18 vertical hydraulic actuators.

reservoir domain, and the spillway. The scaled specimen was installed inside a large steel model box, which provided structural support and transmitted the shaking-table input to the dam–foundation–reservoir system. The overall configuration and principal components of the scaled gravity-dam model are shown in Fig. A3.

The adopted configuration had a dam height of approximately $4.20\ \mathrm{m}$, a transverse extent of approximately $5.68\ \mathrm{m}$, and a model width of approximately $2.68\ \mathrm{m}$. These principal dimensions are indicated in Fig. A3. The spillway was retained in the scaled geometry to preserve the principal geometric discontinuities of the prototype dam. The local spillway geometry consisted of a segment depth of $100\ \mathrm{mm}$, a segment width of $B_1 = 125\ \mathrm{mm}$, and an adjacent-segment spacing of $B_2 = 200\ \mathrm{mm}$. Here, B_1 and B_2 correspond to the geometric definitions illustrated in the spillway-detail panel of Fig. A3.

Because the finite dimensions of the experimental container may introduce artificial wave reflections, absorbing boundary treatments were incorporated into the model system. A nominally $120\ \mathrm{mm}$-thick absorbing layer was adopted. Rubber material was arranged along the bottom boundary, whereas extruded polystyrene (XPS) boards were installed along the lateral boundaries. The relative configuration of the concrete dam body and the absorbing boundaries is also illustrated in Fig. A3. These components were regarded as integral parts of the experimental boundary condition rather than as auxiliary background structures, since their mechanical properties directly affect wave transmission, reflection, and the overall dynamic response of the model.

The concrete dam body had a design mass of approximately $90.7\ \mathrm{t}$, while the rubber absorbing components contributed approximately $4.6\ \mathrm{t}$. The combined design mass of the dam body and rubber components was therefore approximately $95.3\ \mathrm{t}$. This value should not be interpreted as the total

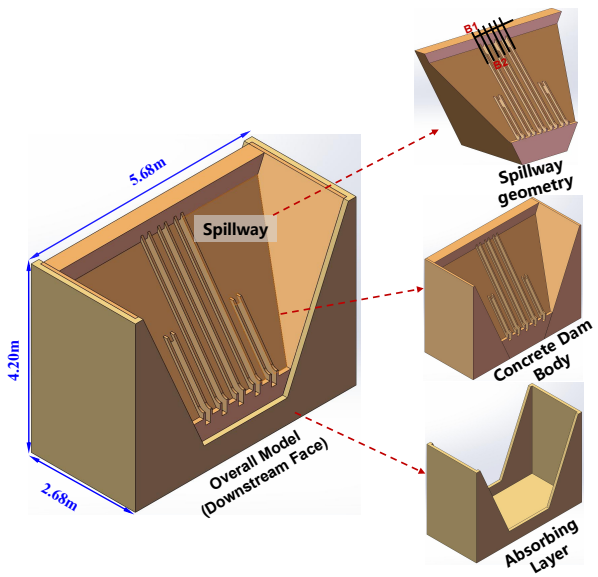


Figure A3: Overall configuration and component decomposition of the scaled gravity-dam model.

Table A2: First six numerical modal frequencies of the empty steel model-box system.

Mode	1	2	3	4	5	6
Frequency (Hz)	25.25	33.38	34.49	37.38	38.47	41.97

shaking-table payload, because it excludes the steel model box, reservoir water, instrumentation, connection components, and other auxiliary structures. Reservoir water was introduced according to the corresponding experimental stage, thereby forming distinct dry and impounded test configurations.

The scaled dam system was surrounded by a large steel model box with external dimensions of approximately $9.8\ \mathrm{m} \times 7.8\ \mathrm{m} \times 4.5\ \mathrm{m}$ and an assembled mass of approximately $80.02\ \mathrm{t}$. The three-dimensional configuration of the steel model box and its supporting frame is shown in Fig. A4.

Numerical modal analysis of the empty steel model-box system.

The bottom and side steel plates were both $10\ \mathrm{mm}$ thick, while the supporting ground beams had a height of $600\ \mathrm{mm}$. The columns were directly supported on the ground-beam system, with local stiffening provided near the column bases. Box columns were constructed using a combination of box sections and H-shaped steel members, while the principal ground beams employed channel sections, H-shaped steel, and built-up double-channel sections. The upper beams were primarily H-shaped steel members, and tubular members were used in the truss system.

The steel model box was mechanically connected to the shaking-table platform through the ground-beam system and table anchor bolts. This configuration provided a direct load-transfer path from the shaking table to the model box and subsequently to the dam–foundation–reservoir system. The model box therefore constituted an integral component of the experimental system and was considered when interpreting boundary effects and measured dynamic responses. The model-box anchorage used 135 M36 bolts and 70 M48 bolts.

To further characterize the dynamic properties of the supporting experimental system, an eigenvalue analysis was performed for the empty steel model-box system using MIDAS. The analysis was conducted without the gravity-dam model and reservoir water. The first six numerical modal frequencies were 25.25, 33.38, 34.49, 37.38, 38.47, and 41.97 Hz, respectively. The corresponding mode shapes are shown in Fig. A5.

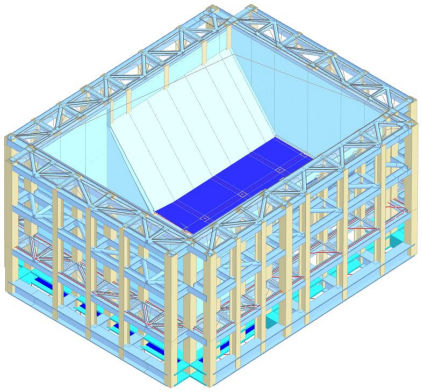


Figure A4: Three-dimensional configuration of the steel model box and supporting frame used to accommodate the scaled dam–foundation–reservoir system.

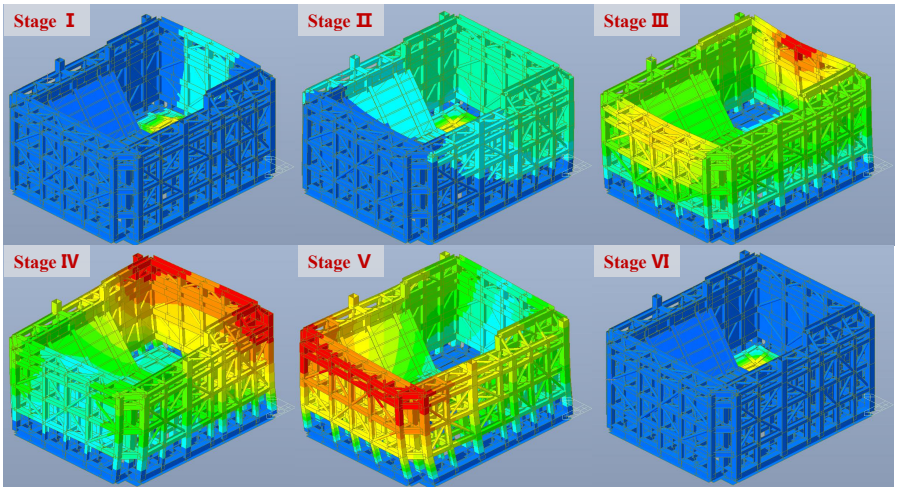


Figure A5: Numerical modal analysis of the empty steel model-box system.

Table A3 summarizes the principal geometric, mass, boundary, and installation parameters of the adopted experimental configuration. Unless otherwise stated, the reported dimensions and masses correspond to the revised design configuration and should be interpreted as design values rather than post-installation survey or weighing results.

Dimensions and masses extracted from preliminary design-stage drawings are retained only for internal design traceability and are not mixed with the adopted configuration reported in Table A3.

The physical similitude framework governing the scaled gravity-dam model is described separately in Section A.3, while the corresponding material-property requirements are summarized in Section A.4.

The simulated concrete was consequently designed to maintain a density close to that of the prototype concrete while substantially reducing its stiffness. The adopted target density was approximately $2400\ \mathrm{kg/m^3}$, and the target elastic modulus was approximately $400\ \mathrm{MPa}$. These properties were selected to satisfy the adopted similitude requirements while maintaining a material density close to that of the prototype concrete. The detailed mixture design and material characterization procedure are described subsequently in Section A.4.

Table A3: Principal geometric, mass, boundary, and installation parameters of the scaled gravity-dam experimental system.

Subsystem	Parameter	Specification
Gravity-dam model	Geometric scale	1:50
	Model type	Scaled concrete gravity dam
	Dam height	≈ 4.20 m
	Transverse extent	≈ 5.68 m
	Model width	≈ 2.68 m
	Spillway	Retained
	Gravity-dam model mass	≈ 93.66 t
	Reservoir water mass	≈ 32.05 t
Spillway detail	Segment depth	100 mm
	Segment width B_1	125 mm
	Segment spacing B_2	200 mm
Absorbing boundary	Nominal thickness	120 mm
	Bottom boundary	Rubber
	Lateral boundaries	XPS boards
Steel model box	External length	9.8 m
	External width	7.8 m
	External height	4.5 m
	Assembled mass	≈ 80.02 t
	Bottom-plate thickness	10 mm
	Side-plate thickness	10 mm
	Ground-beam height	600 mm
Structural components	Column support	Columns directly supported on ground beams
	Column section	Box sections and H-shaped steel members
	Ground-beam section	Channel, H-shaped, and built-up double-channel sections
	Beam section	H-shaped steel members
	Truss members	Tubular members
	Local reinforcement	Ground-beam stiffening near column bases
	Table connection	Ground beams anchored to the shaking-table platform
	Model-box anchorage	135 M36 bolts and 70 M48 bolts
Reservoir condition	Dry stage	Reservoir empty
	Impounded stage	2.3 m water depth
	Total impounded mass	≈ 205.73 t

Table A4: Target material properties of the simulated concrete used in the scaled gravity-dam model.

Material property	Target value
Elastic modulus	400 MPa
Density	2400 kg/m^3

A.3 Physical Similitude and Scaling Relations

To establish a consistent correspondence between the scaled model and the prototype structure, all similitude ratios are defined as the ratio of a model quantity to its corresponding prototype quantity,

$$\lambda_q = \frac{q_m}{q_p}, \tag{12}$$

where q_m and q_p denote the model and prototype quantities, respectively. The adopted geometric scale is

$$\lambda_L = \frac{L_m}{L_p} = \frac{1}{50}. \tag{13}$$

Table A5: Physical similitude relations adopted for the 1:50 gravity-dam model. All ratios are defined as model-to-prototype quantities.

Physical quantity	**Scaling relation**	**Model/Prototype**
Length	λ_L	$1/50$
Displacement	$\lambda_u = \lambda_L$	$1/50$
Density	λ_ρ	1
Acceleration	λ_a	1
Time	$\lambda_t = \sqrt{\lambda_L/\lambda_a}$	0.1414
Frequency	$\lambda_f = 1/\lambda_t$	7.0711
Velocity	$\lambda_v = \lambda_L/\lambda_t$	0.1414
Stress	$\lambda_\sigma = \lambda_\rho \lambda_a \lambda_L$	$1/50$
Elastic modulus	$\lambda_E = \lambda_\sigma/\lambda_\varepsilon$	$1/50$
Material strength	λ_{str}	$1/50$
Strain	λ_ε	1
Mass	$\lambda_m = \lambda_\rho \lambda_L^3$	$1/125000$
Force	$\lambda_F = \lambda_\rho \lambda_L^3 \lambda_a$	$1/125000$

The model was designed with a density ratio of $\lambda_\rho = 1$ and an acceleration ratio of $\lambda_a = 1$. The latter preserves gravitational acceleration between the model and prototype systems and forms the basis of the adopted gravity–inertia similitude framework.

Because acceleration scales dimensionally as $a \sim L/t^2$,

$$\lambda_a = \frac{\lambda_L}{\lambda_t^2}, \tag{14}$$

which gives

$$\lambda_t = \sqrt{\frac{\lambda_L}{\lambda_a}} = \sqrt{\frac{1}{50}} \approx 0.1414. \tag{15}$$

Accordingly, the frequency and velocity scales are

$$\lambda_f = \frac{1}{\lambda_t} = \sqrt{50} \approx 7.0711, \qquad \lambda_v = \frac{\lambda_L}{\lambda_t} \approx 0.1414. \tag{16}$$

Under gravity–inertia similitude, the characteristic stress scale follows from $\sigma \sim \rho a L$,

$$\lambda_\sigma = \lambda_\rho \lambda_a \lambda_L = \frac{1}{50}. \tag{17}$$

Assuming strain similitude, $\lambda_\varepsilon = 1$, the elastic-modulus scale becomes

$$\lambda_E = \frac{\lambda_\sigma}{\lambda_\varepsilon} = \frac{1}{50}. \tag{18}$$

The same scaling requirement is adopted for the characteristic material strength. The mass and force scales are

$$\lambda_m = \lambda_\rho \lambda_L^3 = \frac{1}{125000}, \qquad \lambda_F = \lambda_\rho \lambda_L^3 \lambda_a = \frac{1}{125000}. \tag{19}$$

The resulting similitude relations are summarized in Table A5.

For a prototype acceleration history $a_p(t_p)$, the corresponding time-scaled model input is expressed as

Table A6: Target material properties of the simulated concrete used in the scaled gravity-dam model.

Material property	Target value
Elastic modulus	$400\ \mathrm{MPa}$
Density	$2400\ \mathrm{kg/m^3}$

$$a_m(t_m) = \lambda_a a_p \left(\frac{t_m}{\lambda_t} \right), \qquad t_m = \lambda_t t_p. \tag{20}$$

A.4 Model Material Properties and Design Requirements

The physical similitude framework in Section A.3 requires the model material to maintain a density close to that of the prototype concrete while substantially reducing its elastic modulus and characteristic strength according to the prescribed scaling relations.

Accordingly, the simulated concrete was designed with a target density of approximately $2400\ \mathrm{kg/m^3}$ and a target elastic modulus of approximately $400\ \mathrm{MPa}$. These target properties were selected to satisfy the adopted gravity–inertia similitude requirements while maintaining a practical material system for the large-scale physical model.

A.5 Boundary Conditions, Model Installation, and Numerical Characterization

The dynamic response of the scaled gravity-dam system is governed not only by the dam specimen itself, but also by the finite experimental boundaries, the supporting model box, the dam–foundation interface, and the mechanical connections within the supporting system. Accordingly, these components were treated as integral parts of the experimental configuration when interpreting the measured dynamic response.

The internal boundaries of the experimental system consisted of the near-dam foundation model, the bottom and lateral absorbing layers, and the reservoir domain described in Section A.2. The absorbing materials were introduced to reduce artificial wave reflections generated by the finite dimensions of the model container.

Mechanical installation and load-transfer path. The steel model box was mechanically connected to the shaking-table platform through the ground-beam system and table anchor bolts. The principal columns were supported directly on the ground beams, and local stiffening was introduced near the column bases to improve the local load-transfer behavior. The ground-beam height was approximately $600\ \mathrm{mm}$, providing sufficient space for the anchoring and connection system.

Representative connection details of the steel supporting frame are shown in Fig. A6. The supporting system contains both bolted column–beam connections and stiffened column–ground-beam connections. These connections form part of the load-transfer path through which the shaking-table excitation is transmitted from the supporting system to the model box and subsequently to the enclosed dam–foundation–reservoir system.

Numerical representation of the experimental system. Numerical models were developed to provide auxiliary physical references for interpreting the large-scale shaking-table experiment. As shown in Fig. A7, the numerical representations were constructed at several levels of structural complexity.

Specifically, Fig. A7(a) shows the finite-element model of the gravity-dam body, Fig. A7(b) represents the foundation and riverbank regions, Fig. A7(c) shows the coupled dam–foundation system, and Fig. A7(d) presents an integrated numerical model incorporating the surrounding supporting structures.

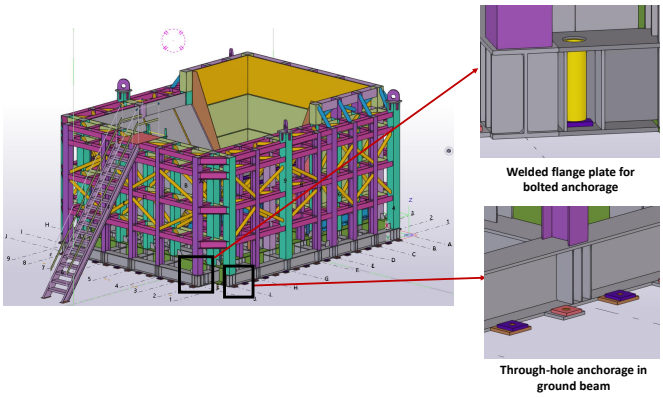


Figure A6: Steel supporting frame and representative connection details.

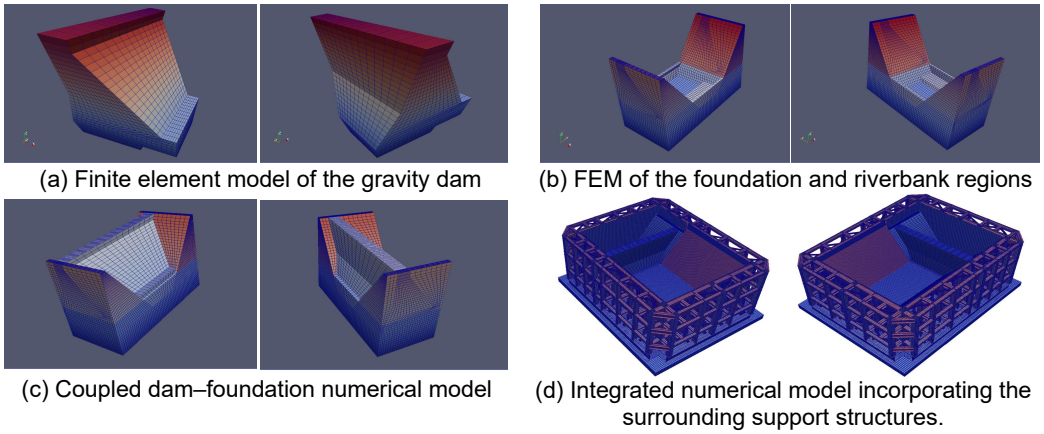


Figure A7: Numerical representations of the gravity-dam experimental system.

These models document the spatial extent and discretization of the numerical analysis domain and provide auxiliary information for interpreting boundary effects, structural dynamic characteristics, and response distributions.

Modal analysis and model-version consistency. Representative numerical mode shapes of the gravity-dam model are shown in Fig. A8. The first six numerical modes, denoted by ϕ_1^{FEM}–ϕ_6^{FEM}, illustrate the deformation patterns predicted by the corresponding finite-element model.

For comparison, the numerical dynamic characteristics of the adopted gravity-dam model were also evaluated through eigenvalue analysis. The first six numerical modal frequencies are summarized in Table A7. The fundamental frequency of the gravity-dam model is approximately 24.13 Hz, while the first six modal frequencies span approximately 24.13–52.46 Hz.

A.6 Damage Accumulation and Local Repair Protocol

The experimental program was organized as a sequential structural-state evolution process rather than as a collection of independent shaking-table tests. The principal stages consisted of dry-reservoir loading, impounded-reservoir loading, damage inspection and local repair, and post-repair impounded loading. Consequently, the structural state associated with a given test condition depends not only on the instantaneous seismic input, but also on the preceding loading history, accumulated damage, reservoir condition, and repair intervention.

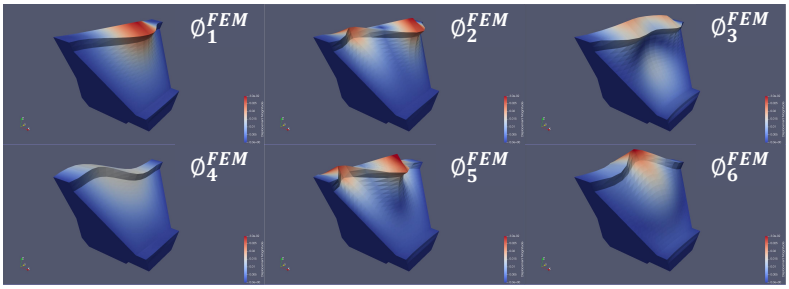


Figure A8: Representative first six numerical mode shapes of the gravity-dam model.

Table A7: First six numerical modal frequencies of the adopted gravity-dam model.

Mode	**1**	**2**	**3**	**4**	**5**	**6**
Frequency (Hz)	24.13	34.17	43.62	43.96	48.62	52.46

Local repair and structural-state transition. After the preceding seismic-loading and damage-inspection stages, local repair was carried out in the identified damaged region. A shape-memory-alloy fiber-reinforced cementitious material (SMA-FRC) was used as the local repair material. In DamEQ-Bench, SMA-FRC is recorded only as the repair-material type; its detailed material formulation, mixture design, and mechanical characterization are beyond the scope of the benchmark.

The repair procedure included reservoir drainage, identification and treatment of the locally damaged region, placement of the repair material, curing, and subsequent re-impoundment before the post-repair shaking-table tests. Following repair, the dam specimen was subjected to further seismic excitation to obtain dynamic-response records under the repaired structural state.

Representative views of the same general dam region before and after repair are shown in Fig. A9.

B Earthquake Inputs and Experimental Condition Design

B.1 Ground-Motion Library

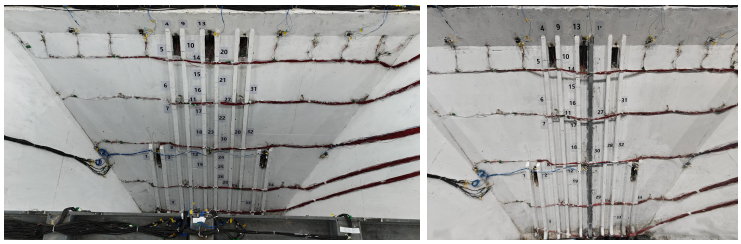
The ground-motion records used in the experimental program were obtained from the PEER Ground Motion Database and were subsequently organized and processed at the record level for shaking-table excitation. For each candidate ground motion, the database separately retains the event identifier, recording station, component definition, original waveform, preprocessing version, and the waveform used for shaking-table excitation.

Where record-level files are available, each processed input waveform is associated with its corresponding experimental condition through a unique file identifier. This record-level organization preserves the provenance of the seismic excitation and avoids treating different waveform versions from the same earthquake event as identical experimental inputs.

Record-level waveform statistics. Table B1 summarizes the valid numerical records extracted from the available ground-motion workbook. The quantities reported in this table describe the waveform files contained in the workbook and should be distinguished from both the original station metadata and the final shaking-table control records.

For a waveform sampled at time coordinates $\{t_i\}_{i=1}^{N}$, the nominal sampling interval is estimated from the time axis as

$$\Delta t = \text{median}\,(t_{i+1} - t_i)\,, \qquad f_s = \frac{1}{\Delta t}, \tag{21}$$



(a) Before Repair (b) After Repair

Figure A9: Comparison of the gravity-dam model before and after local repair.

Table B1: Summary of the ground-motion records used in DamEQ-Bench.

Ground-motion record	Samples	Sampling rate (Hz)	Duration (s)	|PGA| in X/Y/Z (g)
Chi-Chi (pulse-type)	18000	200	89.995	0.446910 / 0.357550 / 0.197293
Chi-Chi (non-pulse-type)	18000	200	89.995	0.596438 / 0.379659 / 0.276077
Taft	5424	100	54.230	0.180286 / 0.158885 / 0.110806
Artificial wave	3001	100	30.000	0.162131 / 0.159350 / 0.137266
Livermore–Eastman Kodak	18000	200	89.995	0.279879 / 0.081792 / 0.042328
Friuli–Forgaria Cornino	15001	200	75.000	0.306752 / 0.261463 / 0.098454
Chi-Chi (large-displacement)	18000	200	89.995	0.511683 / 0.370909 / 0.529778
Wave No. 1, original record	30000	100	299.990	0.007491 / 0.007915 / 0.005322

where f_s denotes the sampling rate inferred from the stored waveform. The record duration is defined as

$$T = t_N - t_1, \tag{22}$$

and the absolute peak ground acceleration of component $j \in \{X, Y, Z\}$ is computed as

$$\mathrm{PGA}_j = \max_t |a_j(t)| . \tag{23}$$

The durations in Table B1 refer to the waveforms as stored in the source workbook and do not necessarily correspond to the time windows displayed in the representative plots or to the final duration applied during shaking-table excitation.

Representative processed acceleration time histories of the Livermore and Friuli mainshock–aftershock sequences are shown in Fig. B1. These examples illustrate the temporal characteristics of selected natural earthquake records used during the experimental design. The plotted time windows are provided for qualitative visualization and should not be interpreted as the full source-record durations reported in Table B1.

An artificial ground motion was also considered to provide an excitation with spectral characteristics consistent with the prescribed target spectrum. Figure B2 compares the response spectra of the three components of the artificial wave with the target spectrum. The comparison provides a direct reference for evaluating the spectral compatibility of the artificial excitation over the period range considered in its design.

Ground-motion sources and dominant-frequency metadata. In addition to the waveform-level statistics, the available experimental documentation provides recording-station information and dominant-frequency values for a subset of the selected earthquake records. These values are summarized in Table C3.

The dominant-frequency values characterize prominent spectral components of the corresponding earthquake records. They should not be interpreted as single-frequency shaking-table inputs, and they are distinct from both the natural frequencies of the gravity-dam model and the temporal sampling frequency of the video recordings.

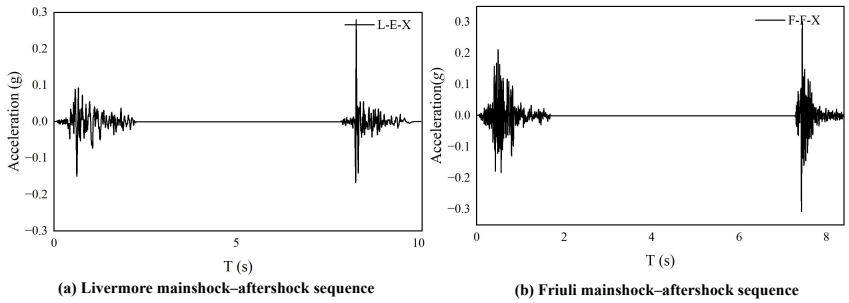


Figure B1: Representative processed acceleration time histories of selected mainshock–aftershock earthquake sequences: (a) Livermore–Eastman Kodak and (b) Friuli–Forgaria Cornino.

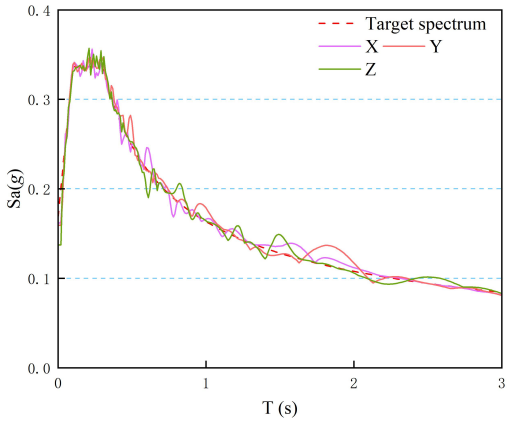


Figure B2: Comparison between the target response spectrum and the response spectra of the X-, Y-, and Z-components of the artificial ground motion.

In the original ground-motion documentation, "p" denotes the pulse-type Chi-Chi record, "d" denotes the long-period large-displacement record, and the superscript * denotes a mainshock–aftershock sequence.

B.2 Ground-Motion Preprocessing

The purpose of ground-motion preprocessing is to convert the source earthquake records into numerically stable and device-compatible input signals while preserving the principal temporal and spectral characteristics of the original motions. The preprocessing stage is applied independently to each available acceleration component before the PGA scaling and multi-axis combination described in Section C.6.

Let $a_j^{\mathrm{raw}}(t)$ denote the raw acceleration record of component $j \in \{X, Y, Z\}$. The preprocessing pipeline is represented as

$$\tilde{a}_j(t) = \mathcal{H}_j \left[\mathcal{R}_j \left(a_j^{\mathrm{raw}}(t) - b_j(t) \right) \right], \tag{24}$$

where $b_j(t)$ denotes the estimated baseline component, $\mathcal{R}_j(\cdot)$ denotes the resampling operation including anti-aliasing treatment when required, and $\mathcal{H}_j(\cdot)$ denotes the recorded device-compatible filtering operation. The resulting waveform $\tilde{a}_j(t)$ is referred to as the *processed ground-motion component* and serves as the input to the subsequent amplitude-scaling stage.

Table B2: Recording stations and dominant frequencies of the ground-motion records.

Ground motion	Station	X (Hz)	Y (Hz)	Z (Hz)
Chi-Chi	TCU052	9.64	9.84	13.12
Chi-Chi(p)	TCU074	4.77	4.07	1.29
Chi-Chi(d)	TCU068	3.73	0.99	2.68
Taft	Taft Lincoln School	26.88	12.20	39.08
Livermore*	Eastman Kodak	9.44	11.93	10.34
Friuli*	Forgaria Cornino	23.97	22.66	33.39
Artificial wave	N/A	44.71	33.08	29.80

Input consistency checks. Before signal processing, each waveform is checked for time-axis monotonicity, missing or duplicated samples, unit consistency, and component identity. The stored time vector is retained rather than reconstructed solely from the nominal sampling rate. Acceleration units are converted to a consistent representation before subsequent scaling, and the correspondence between the waveform axes and the experimental X, Y, and Z directions is preserved explicitly in the metadata.

Baseline correction. Long-duration earthquake records may contain small offsets or low-frequency baseline drift that can accumulate into unrealistic velocity or displacement when integrated. For each component, the baseline estimate $b_j(t)$ is therefore removed prior to amplitude scaling,

$$a_j^{\mathrm{bc}}(t) = a_j^{\mathrm{raw}}(t) - b_j(t), \tag{25}$$

where $a_j^{\mathrm{bc}}(t)$ denotes the baseline-corrected acceleration. The specific baseline model used for a loading file is treated as part of the preprocessing configuration and should be retained together with the processed waveform.

Resampling and anti-aliasing. When the source waveform sampling rate differs from that required by the loading or storage pipeline, the baseline-corrected signal is resampled through the operator $\mathcal{R}_j$. Let the original and target sampling rates be f_s^{src} and f_s^{dst}, respectively. For down-sampling, frequency content above the Nyquist frequency of the target sequence must be suppressed before sample-rate reduction,

$$f_{\mathrm{Nyq}}^{\mathrm{dst}} = \frac{f_s^{\mathrm{dst}}}{2}. \tag{26}$$

Frequency-band conditioning. Where filtering is required to remove numerical drift, suppress out-of-band content, or satisfy shaking-table operating constraints, the filtered waveform is written as

$$a_j^{\mathrm{f}}(t) = \mathcal{H}_j \left[a_j^{\mathrm{rs}}(t) \right], \tag{27}$$

where $a_j^{\mathrm{rs}}(t)$ is the resampled signal and $\mathcal{H}_j$ denotes the corresponding filter operator.

Preservation of the time axis. The source-record time axis, any similitude-scaled time axis, and the final loading-file time axis are stored as distinct quantities. If the prototype record is temporally compressed according to the $1{:}50$ similitude relation, the corresponding physical transformation is governed by

$$t_m = \lambda_t t_p, \qquad \lambda_t = \sqrt{\lambda_L} \approx 0.1414, \tag{28}$$

under the adopted acceleration similitude $\lambda_a = 1$.

Quality control of the processed waveform. After preprocessing, the resulting signal is checked against the source record to ensure that the intended earthquake-motion characteristics have not been inadvertently altered. At minimum, the following record-level quantities are retained or recomputed when available: sample count, sampling rate, duration, component-wise absolute PGA, and basic waveform provenance.

B.3 PGA Scaling and Multi-Axis Excitation

After the preprocessing procedure described in Section C.5, the processed acceleration components are scaled according to the prescribed PGA level, excitation direction, and inter-component amplitude ratio of each experimental condition. Let $\tilde{a}_j(t)$ denote the processed acceleration component along axis $j \in \{X, Y, Z\}$, and define its absolute peak amplitude as

$$P_j = \max_t |\tilde{a}_j(t)| \,. \tag{29}$$

For experimental conditions in which the target PGA of each active axis is specified explicitly, the commanded component is constructed as

$$a_j^{\mathrm{cmd}}(t) = \begin{cases} \dfrac{p\, r_j}{P_j}\, \tilde{a}_j(t), & j \in \mathcal{D}, \\ 0, & j \notin \mathcal{D}, \end{cases} \tag{30}$$

where p denotes the reference target PGA, r_j denotes the prescribed amplitude ratio of component j, and $\mathcal{D}$ is the set of actively excited axes. With this definition,

$$\max_t \left|a_j^{\mathrm{cmd}}(t)\right| = p\, r_j, \qquad j \in \mathcal{D}. \tag{31}$$

Excitation-direction encoding. Three principal excitation configurations were used in the experimental program:

$$\begin{aligned} \mathrm{X}: &\quad \mathcal{D} = \{X\}, \\ \mathrm{XZ}: &\quad \mathcal{D} = \{X, Z\}, \\ \mathrm{XYZ}: &\quad \mathcal{D} = \{X, Y, Z\}. \end{aligned} \tag{32}$$

Components outside the active-axis set are assigned zero command input. For example, an XZ condition with the ratio $X{:}Z = 1{:}0.667$ is represented by

$$(r_X, r_Z) = (1, 0.667), \tag{33}$$

such that the corresponding target component PGAs are

$$\mathrm{PGA}_X^{\mathrm{target}} = p, \qquad \mathrm{PGA}_Z^{\mathrm{target}} = 0.667p. \tag{34}$$

Similarly, an XYZ condition with the ratio $X{:}Y{:}Z = 1{:}1{:}0.667$ uses

$$(r_X, r_Y, r_Z) = (1, 1, 0.667). \tag{35}$$

The excitation-direction and ratio conventions used in the experimental metadata are summarized in Table C4.

Table B3: Encoding of representative excitation directions and component-amplitude ratios.

Direction	Active axes	Ratio label	Target ratio
X	$\{X\}$	Single	$X = 1$
XZ	$\{X, Z\}$	R1-1	X:Z = 1:1
XZ	$\{X, Z\}$	R1-0.667	X:Z = 1:0.667
XYZ	$\{X, Y, Z\}$	R1-1-1	X:Y:Z = 1:1:1
XYZ	$\{X, Y, Z\}$	R1-1-0.667	X:Y:Z = 1:1:0.667
Multi-axis	Record dependent	OriginalRatio	Original component ratio retained

Preservation of the original component ratio. A separate scaling rule is used for conditions labeled *OriginalRatio*. In these cases, independently normalizing each component would alter the relative amplitude structure of the original multi-component earthquake record. Instead, a single common scaling factor α is applied to all active components,

$$a_j^{\mathrm{cmd}}(t) = \alpha\, \tilde{a}_j(t), \qquad j \in \mathcal{D}. \tag{36}$$

The scaling factor is defined from a documented reference PGA criterion,

$$\alpha = \frac{p}{P_{\mathrm{ref}}}, \tag{37}$$

where P_{ref} denotes the reference peak quantity used for the corresponding loading file.

Because the same α is applied to all active components:

$$\frac{\max_t |a_i^{\mathrm{cmd}}(t)|}{\max_t |a_j^{\mathrm{cmd}}(t)|} = \frac{P_i}{P_j}, \tag{38}$$

Target and measured PGA. The PGA values specified above define the commanded excitation rather than the acceleration necessarily realized on the shaking table. For each test, the nominal target PGA is therefore stored separately from the measured table or dam-base PGA. The latter is computed from the calibrated reference acceleration channel as

$$\mathrm{PGA}_j^{\mathrm{meas}} = \max_t \left| a_j^{\mathrm{meas}}(t) \right|. \tag{39}$$

Acceleration units and gravitational constant. The acceleration representation used in each processing stage is retained explicitly. If acceleration is stored in units of g, conversion to SI units follows

$$a^{(\mathrm{SI})}(t) = g_0\, a^{(g)}(t), \tag{40}$$

where g_0 is the standard gravitational acceleration adopted by the corresponding processing configuration.

B.4 Experimental State Evolution and Loading History

The shaking-table experiment was conducted as a sequential state-evolution program rather than as a collection of independent tests. The complete experimental process was organized into four stages: S1, dry-reservoir loading; S2, impounded-reservoir loading; S3, damage inspection and local repair; and S4, post-repair impounded-reservoir loading. The overall state-transition protocol is illustrated in Fig. B3.

Among these four stages, S1, S2, and S4 correspond to the three structural states represented in DamEQ-Bench, whereas S3 is an intervention stage used to transform the damaged specimen into the repaired configuration. S3 is therefore not treated as an additional video-generation condition.

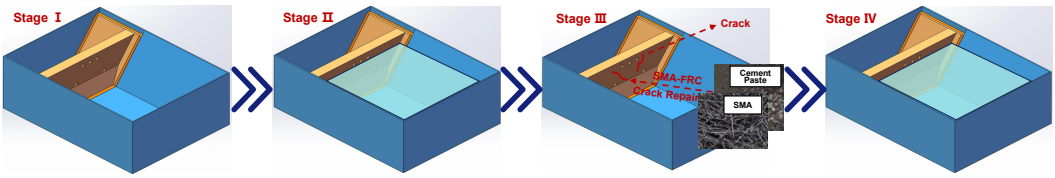


Figure B3: Experimental state evolution and repair-transition protocol.

Table B4: Structural states and their roles in the sequential experimental program.

Stage	Reservoir state	Structural condition	Benchmark role
S1	Dry	Initial / pre-damage state	Generation state
S2	Impounded	Progressively damaged state	Generation state
S3	Drained during intervention	Damage inspection and local repair	State transition
S4	Impounded	Post-repair, history-dependent state	Generation state

The transition between these stages changes several physical factors simultaneously. Reservoir impoundment modifies the hydrodynamic loading, fluid–structure interaction, and effective boundary conditions of the dam.

Importantly, the stage labels describe the observed structural state at a given point in the loading sequence rather than statistically independent experimental groups.

For an experimental event e_k, we conceptually represent its structural state as

$$\mathcal{S}_k = \mathcal{F}\left(\mathcal{S}_{k-1}, \mathbf{a}_k, \mathcal{R}_k, \mathcal{I}_k\right), \tag{41}$$

where $\mathcal{S}_{k-1}$ denotes the structural state inherited from the preceding loading history, $\mathbf{a}_k$ denotes the current seismic excitation, $\mathcal{R}_k$ represents the reservoir condition, and $\mathcal{I}_k$ denotes any intervention such as local repair.

B.5 Experimental Matrix

The final DamEQ-Bench experimental matrix contains 198 shaking-table conditions distributed across three benchmark states: the dry-reservoir stage (S1), the impounded-reservoir stage (S2), and the post-repair impounded stage (S4). The intermediate repair stage (S3) is treated as a state-transition operation and is therefore not included as an independent generation condition.

As summarized in Table B5, S1 contains 70 conditions, including 66 earthquake-excitation cases and four white-noise tests; S2 contains 63 conditions, including 57 earthquake cases and six white-noise tests; and S4 contains 65 conditions, including 56 earthquake cases and nine white-noise tests. The resulting benchmark therefore contains 179 earthquake-excitation conditions and 19 white-noise identification conditions, for a total of 198 experiments.

The experimental matrix spans seven waveform labels: Chi-Chi, Chi-Chi(d), Artificial wave, Friuli, Taft, Wave No. 1, and white noise. Across all three stages, the corresponding condition counts are 36, 34, 25, 35, 44, 5, and 19, respectively. The excitation-direction distribution consists of 59 X-axis conditions, 58 XZ conditions, and 81 XYZ conditions. White-noise identification tests use a nominal PGA of $0.07g$, whereas the earthquake-excitation conditions span nominal PGA levels from $0.10g$ to $0.80g$.

The three condition tables below follow the stage order S1 $\rightarrow$ S2 $\rightarrow$ S4. For readability, the condition number shown in each table is a continuous stage-local index.

Each benchmark condition is associated, where available, with the ground-motion identifier, target PGA, excitation direction, component ratio, structural state, processed input file, measured table/base input, video record, and sensor data.

Table B5: Distribution of the 198 experimental conditions across the three benchmark states.

Stage	Structural–reservoir state	Earthquake	White noise	Total
S1	Dry reservoir	66	4	70
S2	Impounded reservoir	57	6	63
S4	Post-repair impounded reservoir	56	9	65
Total	Three benchmark states	**179**	**19**	**198**

Table B6: Experimental conditions for the dry-reservoir stage (S1).

Condition	PGA/g	Ground motion	Direction	Component ratio
1	0.07	White noise	X+Y+Z	–
2	0.10	Chi-Chi	X	–
3	0.10	Chi-Chi(d)	X	–
4	0.10	Artificial wave	X	–
5	0.10	Friuli*	X	–
6	0.10	Taft	X	–
7	0.10	Chi-Chi	X+Z	1:1
8	0.10	Chi-Chi(d)	X+Z	1:1
9	0.10	Artificial wave	X+Z	1:1
10	0.10	Friuli*	X+Z	1:1
11	0.10	Taft	X+Z	1:2/3
12	0.10	Chi-Chi	X+Z	OriginalRatio
13	0.10	Friuli*	X+Z	OriginalRatio
14	0.10	Taft	X+Z	OriginalRatio
15	0.10	Chi-Chi	X+Y+Z	1:1:1
16	0.10	Chi-Chi(d)	X+Y+Z	1:1:1
17	0.10	Artificial wave	X+Y+Z	1:1:1
18	0.10	Friuli*	X+Y+Z	1:1:1
19	0.10	Taft	X+Y+Z	1:1:2/3
20	0.10	Chi-Chi	X+Y+Z	OriginalRatio
21	0.10	Friuli*	X+Y+Z	OriginalRatio
22	0.10	Taft	X+Y+Z	OriginalRatio
23	0.10	Wave No. 1	X+Y+Z	OriginalRatio
24	0.07	White noise	X+Y+Z	–
25	0.20	Chi-Chi	X	–
26	0.20	Chi-Chi(d)	X	–
27	0.20	Artificial wave	X	–
28	0.20	Friuli*	X	–
29	0.20	Taft	X	–
30	0.20	Chi-Chi	X+Z	1:1
31	0.20	Chi-Chi(d)	X+Z	1:1
32	0.20	Artificial wave	X+Z	1:1
33	0.20	Friuli*	X+Z	1:1
34	0.20	Taft	X+Z	1:2/3
35	0.20	Chi-Chi	X+Z	OriginalRatio
36	0.20	Friuli*	X+Z	OriginalRatio
37	0.20	Taft	X+Z	OriginalRatio
38	0.20	Chi-Chi	X+Y+Z	1:1:1
39	0.20	Chi-Chi(d)	X+Y+Z	1:1:1
40	0.20	Artificial wave	X+Y+Z	1:1:1
41	0.20	Friuli*	X+Y+Z	1:1:1
42	0.20	Taft	X+Y+Z	1:1:2/3
43	0.20	Chi-Chi	X+Y+Z	OriginalRatio
44	0.20	Friuli*	X+Y+Z	OriginalRatio

Continued on next page

Table B6 continued

Condition	PGA/g	Ground motion	Direction	Component ratio
45	0.20	Taft	X+Y+Z	OriginalRatio
46	0.07	White noise	X+Y+Z	–
47	0.30	Chi-Chi	X	–
48	0.30	Chi-Chi(d)	X	–
49	0.30	Artificial wave	X	–
50	0.30	Friuli*	X	–
51	0.30	Taft	X	–
52	0.30	Chi-Chi	X+Z	1:1
53	0.30	Chi-Chi(d)	X+Z	1:1
54	0.30	Artificial wave	X+Z	1:1
55	0.30	Friuli*	X+Z	1:1
56	0.30	Taft	X+Z	1:2/3
57	0.30	Chi-Chi	X+Z	OriginalRatio
58	0.30	Friuli*	X+Z	OriginalRatio
59	0.30	Taft	X+Z	OriginalRatio
60	0.30	Chi-Chi	X+Y+Z	1:1:1
61	0.30	Chi-Chi(d)	X+Y+Z	1:1:1
62	0.30	Artificial wave	X+Y+Z	1:1:1
63	0.30	Friuli*	X+Y+Z	1:1:1
64	0.30	Taft	X+Y+Z	1:1:2/3
65	0.30	Chi-Chi	X+Y+Z	OriginalRatio
66	0.30	Friuli*	X+Y+Z	OriginalRatio
67	0.30	Taft	X+Y+Z	OriginalRatio
68	0.07	White noise	X+Y+Z	–
69	0.20	Wave No. 1	X+Y+Z	OriginalRatio
70	0.30	Wave No. 1	X+Y+Z	OriginalRatio

Table B7: Experimental conditions for the impounded-reservoir stage (S2).

Condition	PGA/g	Ground motion	Direction	Component ratio
1	0.07	White noise	X+Y+Z	–
2	0.10	Chi-Chi	X	–
3	0.10	Chi-Chi(d)	X	–
4	0.10	Artificial wave	X	–
5	0.10	Friuli*	X	–
6	0.10	Taft	X	–
7	0.10	Chi-Chi	X+Z	1:1
8	0.10	Chi-Chi(d)	X+Z	1:1
9	0.10	Artificial wave	X+Z	1:1
10	0.10	Friuli*	X+Z	1:1
11	0.10	Taft	X+Z	1:2/3
12	0.10	Chi-Chi	X+Z	OriginalRatio
13	0.10	Friuli*	X+Z	OriginalRatio
14	0.10	Taft	X+Z	OriginalRatio
15	0.10	Chi-Chi	X+Y+Z	1:1:1
16	0.10	Chi-Chi(d)	X+Y+Z	1:1:1
17	0.10	Artificial wave	X+Y+Z	1:1:1
18	0.10	Friuli*	X+Y+Z	1:1:1
19	0.10	Taft	X+Y+Z	1:1:2/3
20	0.10	Chi-Chi	X+Y+Z	OriginalRatio
21	0.10	Friuli*	X+Y+Z	OriginalRatio
22	0.10	Taft	X+Y+Z	OriginalRatio

Continued on next page

Table B7 continued

Condition	PGA/g	Ground motion	Direction	Component ratio
23	0.10	Wave No. 1	X+Y+Z	OriginalRatio
24	0.07	White noise	X+Y+Z	–
25	0.20	Chi-Chi	X	–
26	0.20	Chi-Chi(d)	X	–
27	0.20	Artificial wave	X	–
28	0.20	Friuli*	X	–
29	0.20	Taft	X	–
30	0.20	Chi-Chi	X+Z	1:1
31	0.20	Chi-Chi(d)	X+Z	1:1
32	0.20	Artificial wave	X+Z	1:1
33	0.20	Friuli*	X+Z	1:1
34	0.20	Taft	X+Z	1:2/3
35	0.20	Chi-Chi	X+Z	OriginalRatio
36	0.20	Friuli*	X+Z	OriginalRatio
37	0.20	Taft	X+Z	OriginalRatio
38	0.20	Chi-Chi	X+Y+Z	1:1:1
39	0.20	Chi-Chi(d)	X+Y+Z	1:1:1
40	0.20	Artificial wave	X+Y+Z	1:1:1
41	0.20	Friuli*	X+Y+Z	1:1:1
42	0.20	Taft	X+Y+Z	1:1:2/3
43	0.20	Chi-Chi	X+Y+Z	OriginalRatio
44	0.20	Friuli*	X+Y+Z	OriginalRatio
45	0.20	Taft	X+Y+Z	OriginalRatio
46	0.07	White noise	X+Y+Z	–
47	0.30	Chi-Chi	X	–
48	0.30	Chi-Chi(d)	X	–
49	0.30	Artificial wave	X	–
50	0.30	Friuli*	X	–
51	0.30	Taft	X	–
52	0.30	Chi-Chi	X+Z	1:1
53	0.30	Chi-Chi(d)	X+Z	1:1
54	0.30	Taft	X+Z	1:2/3
55	0.30	Chi-Chi(d)	X+Y+Z	1:1:1
56	0.30	Taft	X+Y+Z	1:1:2/3
57	0.07	White noise	X+Y+Z	–
58	0.40	Chi-Chi(d)	X	–
59	0.40	Taft	X	–
60	0.07	White noise	X+Y+Z	–
61	0.50	Chi-Chi(d)	X	–
62	0.50	Taft	X	–
63	0.07	White noise	X+Y+Z	–

Table B8: Experimental conditions for the post-repair impounded-reservoir stage (S4).

Condition	PGA/g	Ground motion	Direction	Component ratio
1	0.07	White noise	X+Y+Z	–
2	0.10	Chi-Chi	X	–
3	0.10	Chi-Chi(d)	X	–
4	0.10	Artificial wave	X	–
5	0.10	Friuli*	X	–
6	0.10	Taft	X	–
7	0.10	Chi-Chi	X+Z	1:1

Continued on next page

Table B8 continued

Condition	PGA/g	Ground motion	Direction	Component ratio
8	0.10	Chi-Chi(d)	X+Z	1:1
9	0.10	Artificial wave	X+Z	1:1
10	0.10	Friuli*	X+Z	1:1
11	0.10	Taft	X+Z	1:2/3
12	0.10	Chi-Chi	X+Y+Z	1:1:1
13	0.10	Chi-Chi(d)	X+Y+Z	1:1:1
14	0.10	Artificial wave	X+Y+Z	1:1:1
15	0.10	Friuli*	X+Y+Z	1:1:1
16	0.10	Taft	X+Y+Z	1:1:2/3
17	0.07	White noise	X+Y+Z	–
18	0.20	Chi-Chi	X	–
19	0.20	Chi-Chi(d)	X	–
20	0.20	Artificial wave	X	–
21	0.20	Friuli*	X	–
22	0.20	Taft	X	–
23	0.20	Chi-Chi	X+Z	1:1
24	0.20	Chi-Chi(d)	X+Z	1:1
25	0.20	Artificial wave	X+Z	1:1
26	0.20	Friuli*	X+Z	1:1
27	0.20	Taft	X+Z	1:2/3
28	0.20	Chi-Chi	X+Y+Z	1:1:1
29	0.20	Chi-Chi(d)	X+Y+Z	1:1:1
30	0.20	Artificial wave	X+Y+Z	1:1:1
31	0.20	Friuli*	X+Y+Z	1:1:1
32	0.20	Taft	X+Y+Z	1:1:2/3
33	0.07	White noise	X+Y+Z	–
34	0.30	Chi-Chi	X	–
35	0.30	Chi-Chi(d)	X	–
36	0.30	Artificial wave	X	–
37	0.30	Friuli*	X	–
38	0.30	Taft	X	–
39	0.30	Chi-Chi	X+Z	1:1
40	0.30	Chi-Chi(d)	X+Z	1:1
41	0.30	Artificial wave	X+Z	1:1
42	0.30	Friuli*	X+Z	1:1
43	0.30	Taft	X+Z	1:2/3
44	0.30	Chi-Chi	X+Y+Z	1:1:1
45	0.30	Chi-Chi(d)	X+Y+Z	1:1:1
46	0.30	Artificial wave	X+Y+Z	1:1:1
47	0.30	Friuli*	X+Y+Z	1:1:1
48	0.30	Taft	X+Y+Z	1:1:2/3
49	0.07	White noise	X+Y+Z	–
50	0.40	Chi-Chi(d)	X	–
51	0.40	Taft	X	–
52	0.07	White noise	X+Y+Z	–
53	0.50	Chi-Chi(d)	X	–
54	0.50	Taft	X	–
55	0.07	White noise	X+Y+Z	–
56	0.60	Chi-Chi(d)	X	–
57	0.60	Taft	X	–
58	0.07	White noise	X+Y+Z	–
59	0.70	Chi-Chi(d)	X	–
60	0.70	Taft	X	–
61	0.07	White noise	X+Y+Z	–
62	0.80	Chi-Chi(d)	X	–

Continued on next page

Table C1: Principal DJI Mavic 4 Pro platform specifications relevant to the seismic video acquisition system.

Parameter	Specification
UAV platform	DJI Mavic 4 Pro
Main imaging sensor	$4/3$-inch CMOS Hasselblad camera
Maximum still-image resolution	100 MP
Maximum main-camera video mode	6K/60 fps HDR
High-frame-rate video mode	up to 4K/120 fps
Gimbal system	Multi-axis stabilized Infinity Gimbal
Gimbal viewing capability	Up to 70° upward / 360° rotation
Maximum nominal flight time	51 min
Aircraft mass	approximately 1063 g

Table B8 continued

Condition	PGA/g	Ground motion	Direction	Component ratio
63	0.80	Taft	X	–
64	0.07	White noise	X+Y+Z	–
65	0.30	Wave No. 1	X+Y+Z	OriginalRatio

The asterisk in Friuli* follows the source ground-motion naming convention introduced in Section B.1.

C Multimodal Data Acquisition System

C.1 UAV-Based Multi-View Video Acquisition

The seismic-response videos were acquired using a DJI Mavic 4 Pro unmanned aerial vehicle (UAV), which enabled flexible observation of the gravity-dam specimen from multiple viewpoints without mechanically coupling the imaging system to the shaking table or its supporting structures. Compared with cameras rigidly mounted on the shaking-table system, the UAV configuration reduces direct transmission of table vibration to the camera body and allows the upstream, downstream, lateral, oblique, and top regions of the dam to be observed from different viewing directions.

The use of a UAV, however, does not imply an inertially fixed imaging coordinate system. Residual hovering drift, attitude correction, gimbal motion, and automatic imaging adjustments may introduce global image motion that is unrelated to the structural deformation of the dam. These effects are therefore treated explicitly during preprocessing and physical-response extraction rather than being assumed negligible.

The DJI Mavic 4 Pro employs a multi-camera imaging system. Its Hasselblad main camera uses a $4/3$-inch CMOS sensor and supports video capture up to 6K/60 fps HDR and 4K/120 fps, while the integrated gimbal provides multi-axis stabilization and a wide range of viewing orientations. Table C1 summarizes the principal platform capabilities relevant to the present video-acquisition task.

The actual recording configuration is determined from the corresponding video file and acquisition log. For reproducibility, the condition-level video metadata retain, where available, the image resolution, frame rate, codec, focal-length setting, exposure parameters, gimbal state, recording duration, and camera viewpoint.

The overall UAV-based multi-view acquisition configuration is illustrated in the main paper.

C.2 Far-View Camera Configuration

The far-view video configuration was designed to provide global observations of the gravity-dam specimen from complementary viewing directions. The far-view set consists of four lateral

viewpoints—East, South, West, and North—together with a top-view configuration. These directional labels follow the site-specific experimental coordinate convention and are retained as categorical viewpoint metadata throughout the benchmark.

The four lateral views primarily capture the overall structural outline, the relative motion between the dam body and its supporting foundation, and the surrounding experimental context. Because each lateral camera observes the structure from a different projection direction, the same physical motion may appear with different image-space amplitudes and orientations. The top-view configuration provides an additional projection of the motion field and is particularly useful for observing plan-view displacement and global motion patterns that may be weakly represented in the lateral views.

The far-view recordings are intended primarily for global structural observation rather than fine-scale crack or local-surface inspection. Accordingly, they are used to support tasks such as structural-region tracking, global displacement estimation, temporal-consistency analysis, and cross-view comparison of the generated seismic response.

Importantly, the image coordinate system of a recorded video is not assumed to coincide directly with the physical X, Y, and Z axes of the shaking-table experiment. A horizontal or vertical image-plane displacement may contain contributions from multiple physical motion components, depending on the UAV position, camera orientation, gimbal attitude, and perspective projection. Therefore, viewpoint labels such as East, West, South, North, and Top are preserved explicitly, while any transformation from image coordinates to physical structural coordinates requires corresponding camera-geometry or calibration information.

For a far-view recording $V^{(v)}$, where $v \in \{\mathrm{E}, \mathrm{S}, \mathrm{W}, \mathrm{N}, \mathrm{Top}\}$ denotes the viewpoint, the observed image-space motion can be represented conceptually as

$$\mathbf{u}^{(v)}(t) = \Pi^{(v)}\left(\mathbf{x}(t)\right) + \mathbf{c}^{(v)}(t), \tag{42}$$

where $\mathbf{x}(t)$ denotes the physical structural motion, $\Pi^{(v)}(\cdot)$ represents the viewpoint-dependent projection into the image plane, and $\mathbf{c}^{(v)}(t)$ denotes residual camera motion caused by UAV drift, gimbal adjustment, or stabilization behavior.

C.3 Near-View Camera Configuration

The near-view configuration was designed to complement the global observations provided by the far-view recordings in Section C.2. Two near-view labels were used: *Upstream* and *Downstream*, corresponding to observations of the upstream and downstream dam surfaces, respectively. These recordings provide higher spatial detail of local surface regions, dam boundaries, geometric discontinuities, and visually observable repaired areas.

Unlike the far-view configuration, the near-view recordings are not required to contain the complete dam body within every frame. Their primary purpose is to preserve local structural appearance and motion information at a finer image scale. Depending on the UAV position and camera field of view, individual near-view videos may emphasize only a portion of the dam surface or a specific region of engineering interest. This property makes the near-view data particularly useful for local surface tracking, structural-boundary preservation, repair-region observation, and fine-scale visual-consistency evaluation.

The near-view recordings should not, however, be assumed to provide complete correspondence with the sensor layout. In particular, a single video is not guaranteed to simultaneously cover all sensor elevations or all structural regions used in the response analysis. Sensor–video comparisons are therefore performed only when the corresponding structural region, measurement elevation, and observation direction are sufficiently supported by the actual video coverage.

Together with the five far-view labels (East, South, West, North, and Top), the two near-view labels (Upstream and Downstream) define seven observation categories in the benchmark:

$$\mathcal{V} = \{\mathrm{East}, \mathrm{South}, \mathrm{West}, \mathrm{North}, \mathrm{Top}, \mathrm{Upstream}, \mathrm{Downstream}\}. \tag{43}$$

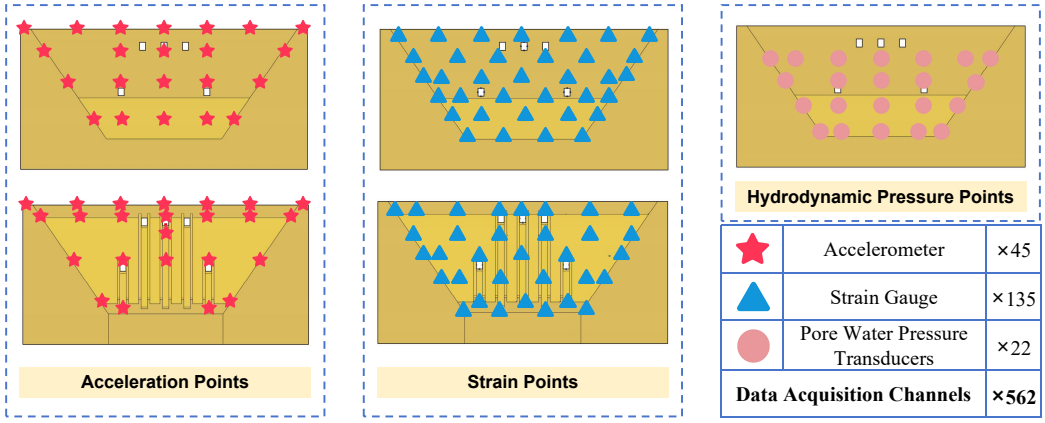


Figure C1: Adopted measurement layout of the gravity-dam experimental system.

C.4 Sensor Layout and Instrumentation System

The gravity-dam experiment employed a distributed instrumentation system to record structural acceleration, strain, and hydrodynamic pressure at multiple locations of the dam–foundation–reservoir system. Because the instrumentation configuration evolved during experimental preparation and loading, the physical sensor layout, acquisition-channel assignment, and recorded signal labels are treated as related but distinct pieces of metadata.

For the present benchmark documentation, the adopted sensor-count reference is the finalized measurement-layout drawing shown in Fig. C1. According to this configuration, the experimental system includes 45 accelerometers, 135 strain gauges, and 22 dynamic water-pressure measurement points.

Spatial distribution of the measurement system. The measurement points were distributed over the upstream and downstream dam surfaces, selected elevations of the dam body, the surrounding foundation and model-box system, and the reservoir-facing regions. Acceleration sensors were arranged to characterize the global and height-dependent dynamic response of the structure, while the strain measurements provided denser local coverage of deformation-sensitive regions. Dynamic water-pressure sensors were distributed primarily along the reservoir-facing portions of the system to record hydrodynamic response during the impounded stages.

Physical sensors, acquisition channels, and signal labels. A unique distinction is maintained among (i) the physical measurement location, (ii) the installed sensor, (iii) the acquisition-hardware channel, and (iv) the recorded signal label.

Each recorded signal is therefore identified, whenever available, using a hierarchical hardware address consisting of

$$\mathcal{C} = (\text{cabinet}, \text{unit}, \text{channel}), \tag{44}$$

together with the corresponding physical location, measurement direction, recorded label, engineering unit, calibration configuration, and applicable loading interval.

The available wiring records contain the channel groups summarized in Table C2.

Across the first seven acceleration-related acquisition units, the wiring records contain 156 channel entries and 153 distinct A-series directional labels.

Table C2: Channel composition identified from the available acquisition-wiring records.

Acquisition location	Recorded entries	Recorded label / role
Cabinet 5, Units 1–5	120	Acceleration records with A-series labels
Cabinet 6, Unit 1	24	Acceleration records with A-series labels
Cabinet 6, Unit 2	12	A-series records including replacement wiring and reused labels
Cabinet 6, Unit 3	22	W-series records; selected channels correspond to side-wall acceleration

Table C3: Recording stations and source-reported dominant frequencies of selected earthquake records.

Ground motion	Station	X (Hz)	Y (Hz)	Z (Hz)
Chi-Chi	TCU052	9.64	9.84	13.12
Chi-Chi(p)	TCU074	4.77	4.07	1.29
Chi-Chi(d)	TCU068	3.73	0.99	2.68
Taft	Taft Lincoln School	26.88	12.20	39.08
Livermore*	Eastman Kodak	9.44	11.93	10.34
Friuli*	Forgaria Cornino	23.97	22.66	33.39
Artificial wave	N/A	44.71	33.08	29.80

In the original ground-motion documentation, "p" denotes the pulse-type Chi-Chi record, "d" denotes the long-period large-displacement record, and the superscript * denotes a mainshock–aftershock sequence.

C.5 Ground-Motion Preprocessing

The purpose of ground-motion preprocessing is to convert the source earthquake records into numerically stable and device-compatible input signals while preserving the principal temporal and spectral characteristics of the original motions. The preprocessing stage is applied independently to each available acceleration component before the PGA scaling and multi-axis combination described in Section C.6.

Let $a_j^{\mathrm{raw}}(t)$ denote the raw acceleration record of component $j \in \{X, Y, Z\}$. The preprocessing pipeline is represented as

$$\tilde{a}_j(t) = \mathcal{H}_j \left[\mathcal{R}_j \left(a_j^{\mathrm{raw}}(t) - b_j(t) \right) \right], \tag{45}$$

where $b_j(t)$ denotes the estimated baseline component, $\mathcal{R}_j(\cdot)$ denotes the resampling operation including anti-aliasing treatment when required, and $\mathcal{H}_j(\cdot)$ denotes the recorded device-compatible filtering operation. The resulting waveform $\tilde{a}_j(t)$ is referred to as the *processed ground-motion component* and serves as the input to the subsequent amplitude-scaling stage.

C.6 PGA Scaling and Multi-Axis Excitation

After the preprocessing procedure described in Section C.5, the processed acceleration components are scaled according to the prescribed PGA level, excitation direction, and inter-component amplitude ratio of each experimental condition. Let $\tilde{a}_j(t)$ denote the processed acceleration component along axis $j \in \{X, Y, Z\}$, and define its absolute peak amplitude as

$$P_j = \max_t |\tilde{a}_j(t)| . \tag{46}$$

For experimental conditions in which the target PGA of each active axis is specified explicitly, the commanded component is constructed as

$$a_j^{\mathrm{cmd}}(t) = \begin{cases} \dfrac{p\, r_j}{P_j}\, \tilde{a}_j(t), & j \in \mathcal{D}, \\ 0, & j \notin \mathcal{D}, \end{cases} \tag{47}$$

Table C4: Encoding of representative excitation directions and component-amplitude ratios used in the shaking-table experiments.

Direction	Active axes	Ratio label	Target ratio
X	$\{X\}$	Single	$X = 1$
XZ	$\{X, Z\}$	R1-1	X:Z = 1:1
XZ	$\{X, Z\}$	R1-0.667	X:Z = 1:0.667
XYZ	$\{X, Y, Z\}$	R1-1-1	X:Y:Z = 1:1:1
XYZ	$\{X, Y, Z\}$	R1-1-0.667	X:Y:Z = 1:1:0.667
Multi-axis	Record dependent	OriginalRatio	Original component ratio retained

where p denotes the reference target PGA, r_j denotes the prescribed amplitude ratio of component j, and $\mathcal{D}$ is the set of actively excited axes. With this definition,

$$\max_t \left|a_j^{\mathrm{cmd}}(t)\right| = p\, r_j, \qquad j \in \mathcal{D}. \tag{48}$$

Thus, the temporal waveform of each processed component is preserved while its amplitude is adjusted to the target axis-specific PGA.

Excitation-direction encoding. Three principal excitation configurations were used in the experimental program:

$$\begin{aligned} \mathrm{X}: &\quad \mathcal{D} = \{X\}, \\ \mathrm{XZ}: &\quad \mathcal{D} = \{X, Z\}, \\ \mathrm{XYZ}: &\quad \mathcal{D} = \{X, Y, Z\}. \end{aligned} \tag{49}$$

Components outside the active-axis set are assigned zero command input. For example, an XZ condition with the ratio X:Z = 1:0.667 is represented by

$$(r_X, r_Z) = (1, 0.667), \tag{50}$$

such that the corresponding target component PGAs are

$$\mathrm{PGA}_X^{\mathrm{target}} = p, \qquad \mathrm{PGA}_Z^{\mathrm{target}} = 0.667p. \tag{51}$$

Similarly, an XYZ condition with the ratio X:Y:Z = 1:1:0.667 uses

$$(r_X, r_Y, r_Z) = (1, 1, 0.667). \tag{52}$$

The excitation-direction and ratio conventions used in the experimental metadata are summarized in Table C4.

C.7 Auxiliary Identification and Calibration Tests

In addition to the formal earthquake-loading conditions, the experimental program included white-noise identification, system-calibration, and commissioning tests to support dynamic characterization and data-quality control. Among the 198 experimental conditions, 19 white-noise tests were used primarily for stage-dependent modal and frequency-response identification, while calibration records were used to verify channel identity, sensor orientation, timing consistency, excitation transfer, and sensor–video synchronization. A representative auxiliary frequency-response analysis is shown in Fig. C2, where the box-top and dam-top responses are compared in both the time and frequency domains. The dam-top response exhibits a pronounced amplification around 20 Hz, whereas the box-top response shows no comparably strong localized amplification, providing auxiliary evidence of the dominant dynamic-response range of the coupled dam system. These auxiliary results are used for system characterization and quality control rather than as experimental modal ground

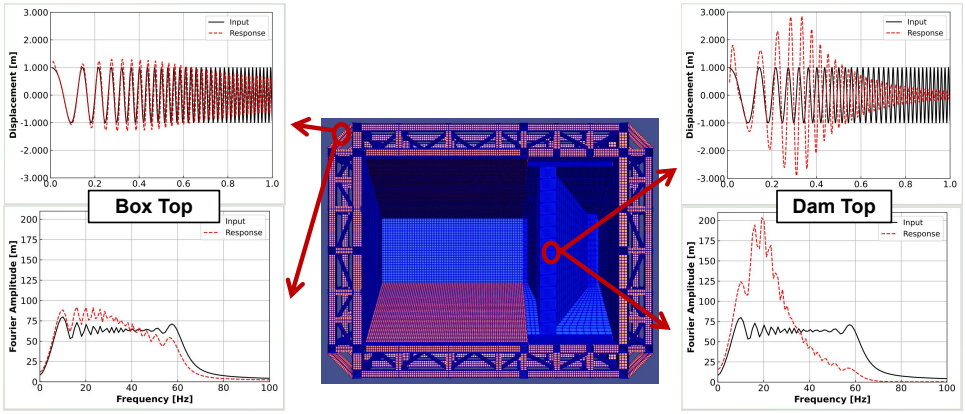


Figure C2: Auxiliary frequency-response analysis at the box-top and dam-top locations.

truth. Accordingly, inclusion in the experimental matrix is distinguished from metric-specific eligibility, and a condition is used for a particular video- or physics-based evaluation only when the required sensor, video, synchronization, and signal-quality information is available.

C.8 Video–Sensor Temporal Synchronization

Reliable temporal alignment between the UAV videos and the sensor measurements is required for response-based physical evaluation. In particular, comparisons involving peak-response timing, phase consistency, acceleration amplification, or time-history correlation require that video-derived and sensor-derived responses refer to the same physical interval of the shaking-table event.

For each video–sensor pair, the relationship between the video time t_v and the sensor time t_s is represented by an affine mapping,

$$t_s = \alpha\, t_v + \beta, \tag{53}$$

where β denotes the temporal offset between the two acquisition systems and α denotes a clock-scale correction. In the absence of independent evidence of clock-rate drift, α is fixed to unity and only the offset β is estimated. A non-unit value of α is introduced only when supported by the acquisition records or by multiple independent temporal correspondences over the same recording.

Offset estimation from measured responses. When direct hardware synchronization is unavailable, the temporal offset may be estimated from physically corresponding measured signals. Let $x_s(t)$ denote a sensor-derived reference response and $x_v(t)$ a video-derived response extracted from a spatially corresponding region. After restricting the comparison to a physically plausible search interval $\mathcal{B}$, the offset can be estimated from the normalized cross-correlation,

$$\hat{\beta} = \arg\max_{\beta \in \mathcal{B}} \rho\left[x_s(t), x_v(t-\beta)\right], \tag{54}$$

where $\rho(\cdot,\cdot)$ denotes the normalized correlation coefficient.

The search interval $\mathcal{B}$ is specified before optimization using the available acquisition records and approximate event timing.

Clock-scale correction. For sufficiently long recordings, a constant temporal offset may be insufficient if the video and sensor acquisition clocks exhibit measurable relative drift. If two or more independently identifiable temporal correspondences are available, the affine model in Eq. (60) may be used to estimate both α and β.

Given matched temporal landmarks $\{(t_{v,k}, t_{s,k})\}_{k=1}^{K}$, the synchronization residual for landmark k is

$$e_k = t_{s,k} - (\alpha t_{v,k} + \beta)\,, \tag{55}$$

and the residual synchronization error can be summarized as

$$\varepsilon_{\mathrm{sync}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} e_k^2}. \tag{56}$$

D Dataset Construction and Task Protocols

D.1 From Raw Experimental Records to DamEQ-Bench

DamEQ-Bench was derived from synchronized shaking-table experiments by integrating raw condition records, multi-view UAV videos, sensor metadata, and experimental logs within a unified data-processing framework. Rather than treating these sources independently, we organized them at the event level so that each video sequence could be consistently associated with its corresponding excitation condition, structural state, viewpoint, and experimental record. The raw data were subsequently subjected to integrity checking, visual-quality screening, condition–viewpoint alignment, and deterministic spatiotemporal standardization before being converted into task-specific benchmark samples.

Condition and viewpoint matching. After integrity checking, each candidate video is associated with an experimental condition and a viewpoint label using multiple sources of evidence. These may include the recorded condition board, original file information, recording continuity, neighboring files in the acquisition sequence, experimental logs, known loading order, and visible observation geometry.

For a candidate video v_i, the resulting association can be represented as

$$v_i \longrightarrow (c_i,\, q_i,\, s_i)\,, \tag{57}$$

where c_i denotes the experimental condition, q_i denotes the viewpoint category, and s_i denotes the matching status or confidence supported by the available evidence.

I2V and V2V sample construction. The event-level video observations are subsequently converted into task-specific benchmark samples.

For the image-to-video (I2V) task, a valid observation provides an initial frame together with the standardized experimental conditioning information, while the subsequent real video provides the reference dynamic sequence. A generic I2V sample is therefore represented as

$$\mathcal{D}_{\mathrm{I2V}} = \left\{I^0, \mathbf{c}, V^{\mathrm{gt}}\right\}, \tag{58}$$

where I^0 is the experiment-specific initial frame, $\mathbf{c}$ contains the corresponding seismic and state metadata, and V^{gt} is the real experimental video.

For the video-to-video (V2V) task, an observed initial sequence of length L is used as the visual conditioning input,

$$\mathcal{D}_{\mathrm{V2V}} = \left\{V^{0:L-1}, \mathbf{c}, V^{L:T-1}\right\}, \tag{59}$$

Table D1: Condition-level composition of the final 198-condition experimental matrix.

Stage	Earthquake	White noise	Total
S1: Dry reservoir	66	4	70
S2: Impounded reservoir	57	6	63
S4: Post-repair impounded	56	9	65
Total	**179**	**19**	**198**

Table D2: Condition-level distribution of excitation records.

Excitation label	Number of conditions
Chi-Chi	36
Chi-Chi(d)	34
Artificial wave	25
Friuli	35
Taft	44
Wave No. 1	5
White noise	19
Total	**198**

where $V^{0:L-1}$ denotes the observed conditioning frames and $V^{L:T-1}$ denotes the corresponding future experimental sequence.

D.2 Dataset Statistics and Coverage

Dataset statistics are reported at multiple levels to distinguish physical shaking-table conditions from video observations and task-specific benchmark samples. This distinction is essential because a single experimental condition may contain multiple valid UAV recordings from different viewpoints, whereas another condition may contain only partial or no valid video coverage.

We therefore distinguish three basic statistical units:

$$\mathcal{N} = \{N_{\text{condition}}, N_{\text{video}}, N_{\text{sample}}\}, \tag{60}$$

where $N_{\text{condition}}$ denotes the number of physical shaking-table conditions, N_{video} denotes the number of valid view-specific video observations, and N_{sample} denotes the number of task-specific I2V or V2V samples derived from those observations.

Condition-level composition. The 198 experimental conditions are distributed across the three benchmark states defined in Section B.4. Stage S1 contains 70 dry-reservoir conditions, Stage S2 contains 63 impounded-reservoir conditions, and Stage S4 contains 65 post-repair impounded conditions. Among all conditions, 179 correspond to earthquake excitation and 19 correspond to white-noise identification tests.

The experimental matrix contains seven waveform labels. Their condition-level distribution is summarized in Table D2.

The excitation-direction distribution consists of 59 single-axis X conditions, 58 biaxial XZ conditions, and 81 triaxial XYZ conditions:

$$N_X = 59, \qquad N_{XZ} = 58, \qquad N_{XYZ} = 81. \tag{61}$$

White-noise identification conditions use a nominal PGA of $0.07g$, whereas the earthquake-loading conditions span nominal PGA levels from $0.10g$ to $0.80g$. The distribution across PGA levels is intentionally non-uniform because stronger excitation was introduced progressively as the experimental program transitioned from low-damage to severe-damage regimes.

Task-specific and modality-specific availability. The number of valid videos is not used directly as the number of I2V, V2V, or physics-evaluable samples. Task eligibility is determined independently according to the input-length, temporal-coverage, viewpoint, sensor-availability, and synchronization requirements of the corresponding protocol.

Thus,

$$N_{\mathrm{I2V}}, \qquad N_{\mathrm{V2V}}, \qquad N_{\mathrm{sensor}} \tag{62}$$

are reported separately in the final release statistics.

D.3 Standardized Prompt Construction

To ensure that different video-generation models receive semantically consistent experimental information, DamEQ-Bench uses a standardized prompt-construction protocol. Prompt content is generated deterministically from the released condition metadata rather than written manually for individual samples or models.

The canonical field order is

$$\begin{aligned} \text{task/object} &\to \text{ground motion} \to \text{PGA} \\ &\to \text{excitation direction} \to \text{reservoir/structural state} \\ &\to \text{component ratio} \to \text{viewpoint} \\ &\to \text{generation duration}. \end{aligned} \tag{63}$$

This fixed ordering reduces prompt variation unrelated to the physical condition and improves reproducibility across baseline models.

Condition information included in the prompt. The standardized prompt is constructed from experimentally defined metadata rather than from the observed future response. The principal conditioning fields include the ground-motion label, nominal target PGA, excitation direction, reservoir state, structural stage, component-scaling rule, observation viewpoint, and requested output duration.

I2V prompt template. For the I2V task, the canonical English template is

```
 Generate the gravity-dam shaking-table response from
the provided initial image.  Ground motion:  {wave}.
Target PGA: {pga_g} g.  Excitation axes:  {direction}.
Reservoir state:  {reservoir_state}.  Structural stage:
{structural_stage}.  Component scaling:  {ratio_description}.
Observation:  {viewpoint}.  Continue the same dam, reservoir
boundaries, experimental background, and camera viewpoint
for {duration_s} seconds under these conditions.
```

The phrase "continue the same dam" is used to emphasize preservation of the experiment-specific geometry and appearance provided by the initial image.

V2V prompt template. For the V2V task, the canonical English template is

```
 Continue the gravity-dam shaking-table video after
the provided 12-frame prefix.  Ground motion:  {wave}.
Target PGA: {pga_g} g.  Excitation axes:  {direction}.
Reservoir state:  {reservoir_state}.  Structural stage:
{structural_stage}.  Component scaling:  {ratio_description}.
Observation:  {viewpoint}.  Generate the subsequent response
for {prediction_duration_s} seconds while retaining the same
observed system and viewpoint.
```

E UNIFIED EVALUATION PREPROCESSING AND FAIRNESS PROTOCOL

E.1 GENERATED-TO-REFERENCE VIDEO PAIRING

For all models, generated videos are deterministically paired with the corresponding experimental references using the benchmark sample metadata, and all methods are evaluated over the same predefined prediction interval. Conditioning content, including the initial image for I2V and the 12-frame prefix for V2V, is excluded from metric computation. Generated outputs are not temporally shifted, stretched, looped, or otherwise post-processed to improve agreement with the reference; incomplete or invalid outputs are instead recorded explicitly according to the common evaluation protocol. This unified preprocessing ensures that all visual, temporal, and physics-based metrics are computed under identical evaluation conditions.

E.2 SPATIAL AND TEMPORAL STANDARDIZATION

To ensure fair comparison across video-generation models, DamEQ-Bench applies a unified spatial and temporal standardization procedure before metric computation. The same preprocessing definition is applied to the generated and real reference videos associated with a given benchmark sample.

Spatial standardization is designed to preserve the structural field of view and geometric consistency, whereas temporal standardization is defined with respect to physical time rather than frame count alone.

Spatial standardization. Let an original video frame be denoted by $I_t \in \mathbb{R}^{H\times W\times 3}$. The standardized frame is obtained through a fixed crop-and-resize operator

$$\tilde{I}_t = \mathcal{R}_{\mathrm{sp}}\left[\mathcal{C}\left(I_t\right)\right], \tag{64}$$

where $\mathcal{C}(\cdot)$ denotes the predefined spatial crop and $\mathcal{R}_{\mathrm{sp}}(\cdot)$ denotes aspect-ratio-preserving rescaling to the resolution required by the corresponding evaluation protocol.

Metric-specific spatial resolutions. A single spatial resolution is not required for all evaluation tasks. RGB- or perception-oriented metrics may operate on a standardized evaluation resolution, whereas subpixel displacement extraction, edge tracking, or local structural-motion analysis may require a higher resolution to preserve small-amplitude structural response.

We therefore distinguish

$$(H_{\mathrm{rgb}}, W_{\mathrm{rgb}}) \tag{65}$$

for appearance-oriented evaluation from

$$(H_{\mathrm{motion}}, W_{\mathrm{motion}}) \tag{66}$$

for motion-sensitive physical analysis.

Coordinate mapping after crop and resize. Spatial preprocessing changes the numerical value of image-space displacement and must therefore be included when converting tracked motion back to the source-image or physical coordinate system.

Let the crop origin in the source image be (x_0, y_0) and let the spatial scaling factors be

$$s_x = \frac{W_{\mathrm{out}}}{W_{\mathrm{crop}}}, \qquad s_y = \frac{H_{\mathrm{out}}}{H_{\mathrm{crop}}}. \tag{67}$$

A source-image coordinate (x, y) is mapped to the standardized image as

$$\begin{bmatrix}\tilde{x}\\ \tilde{y}\end{bmatrix} = \begin{bmatrix}s_x(x-x_0)\\ s_y(y-y_0)\end{bmatrix}. \tag{68}$$

Correspondingly, an image-space displacement transforms as

$$\begin{bmatrix}\Delta\tilde{x}\\ \Delta\tilde{y}\end{bmatrix} = \begin{bmatrix}s_x & 0\\ 0 & s_y\end{bmatrix}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}. \tag{69}$$

Reference regions and masks. Metric-specific masks, such as the dam structural region, background stabilization region, or valid evaluation region, are transformed using the same crop-and-resize geometry as the corresponding video.

Masks are not independently stretched or redefined after observing model outputs. This prevents spatial preprocessing from changing the effective evaluation region between models.

Temporal standardization in physical time. Temporal comparison is defined in seconds rather than by directly matching frame indices.

Let the verified source-frame timestamps of a video be

$$\mathcal{T}_{\mathrm{src}} = \{\tau_0, \tau_1, \ldots, \tau_{M-1}\}. \tag{70}$$

For an evaluation rate f_{eval}, the target time grid is

$$t_n = t_0 + \frac{n}{f_{\mathrm{eval}}}, \qquad n = 0, \ldots, N-1, \tag{71}$$

where t_0 is the beginning of the canonical prediction interval.

Each target time t_n is mapped to the corresponding source-frame timestamp or neighboring source frames according to the predefined temporal resampling rule.

Thus, temporal standardization is represented as

$$\tilde{I}(t_n) = \mathcal{R}_{\mathrm{t}}\left(I(\tau_k), t_n\right), \tag{72}$$

where $\mathcal{R}_{\mathrm{t}}$ denotes the fixed temporal sampling or interpolation operator.

Source-index provenance. For every standardized frame, the evaluator records the source frame index or source-frame pair used to construct it.

If nearest-frame sampling is used,

$$k_n = \arg\min_k \left|\tau_k - t_n\right|, \tag{73}$$

Different native frame rates. Generated videos may have native frame rates different from those of the reference videos. Such differences are handled through explicit temporal resampling rather than by modifying the FPS metadata while leaving the frame sequence unchanged.

and the physical duration of the generated sequence is preserved during evaluation.

Variable-frame-rate recordings. When reliable frame timestamps are available, they take precedence over the nominal container frame rate. This is particularly important for variable-frame-rate or irregularly sampled recordings.

A nominal value such as 30 fps is therefore not assumed to imply

$$\tau_{k+1} - \tau_k = \frac{1}{30} \tag{74}$$

for every frame unless the timestamp sequence confirms uniform temporal spacing.

Fairness across models. All spatial and temporal preprocessing parameters are determined by the benchmark protocol and frozen before evaluating the compared generation models. They are not optimized independently for a particular model output.

The same reference crop, target time interval, temporal sampling rule, evaluation resolution, and metric-specific preprocessing are therefore applied to all models evaluated on the same benchmark sample.

E.3 Global Camera-Motion Compensation

Global camera-motion compensation is applied before motion-sensitive evaluation to reduce residual UAV drift while preserving the physical response of the gravity-dam system. The stabilization procedure follows the background-based protocol and is applied independently to the real and generated videos.

For a video sequence V, the frame-wise camera transformation is estimated only from the predefined stationary-background regions,

$$\mathbf{T}_t = \mathcal{F}\left(\mathcal{B}_t, \mathcal{B}_{\mathrm{ref}}\right), \tag{75}$$

where $\mathcal{B}_t$ denotes the verified background features at frame t, $\mathcal{B}_{\mathrm{ref}}$ denotes the corresponding reference background, and $\mathcal{F}(\cdot)$ denotes the robust transformation estimator defined by the fixed stabilization protocol.

The real video and generated video are processed separately:

$$\mathbf{T}_t^{\mathrm{real}} \neq \mathbf{T}_t^{\mathrm{gen}} \quad \text{in general}, \tag{76}$$

because each sequence may contain different residual camera motion. However, both transformation sequences are estimated using the same background-region definition, transformation hierarchy, and quality-control criteria.

Preservation of physical structural motion. Let the observed image-space motion be decomposed conceptually as

$$\mathbf{u}_{\mathrm{obs}}(t) = \mathbf{u}_{\mathrm{cam}}(t) + \mathbf{u}_{\mathrm{str}}(t) + \boldsymbol{\epsilon}(t), \tag{77}$$

where $\mathbf{u}_{\mathrm{cam}}$ denotes apparent motion introduced by the UAV, $\mathbf{u}_{\mathrm{str}}$ denotes the projected physical response of the experimental system, and $\boldsymbol{\epsilon}$ denotes residual tracking and imaging error.

The stabilization stage is designed to estimate and remove $\mathbf{u}_{\mathrm{cam}}$ from independent background regions while retaining $\mathbf{u}_{\mathrm{str}}$. Relative motion among the dam, foundation, model box, and other physically coupled components is therefore not treated automatically as camera motion.

Valid evaluation region after warping. Spatial warping may introduce invalid pixels near the image boundaries. These regions are not included in paired metric computation.

For frame t, let Ω_t^{real} and Ω_t^{gen} denote the valid image domains after stabilization. The paired evaluation mask is defined as

$$\Omega_t^{\mathrm{eval}} = \Omega_t^{\mathrm{real}} \cap \Omega_t^{\mathrm{gen}} \cap \Omega^{\mathrm{ROI}}, \tag{78}$$

where Ω^{ROI} denotes the predefined metric-specific evaluation region.

Pixels introduced by constant padding, reflection padding, replicated borders, or other warp-boundary filling operations are excluded from metric computation. This prevents artificial border content from affecting image-quality or motion scores.

Stabilization quality control. The stabilization quality is evaluated independently for the real and generated sequences using the criteria. These include background residual motion, robust-fitting inlier ratio, spatial coverage of the background control features, and temporal continuity of the estimated transformation.

For diagnostic purposes, the residual background displacement after stabilization may be summarized as

$$e_{\mathrm{bg}} = \frac{1}{T} \sum_{t=1}^{T} e_{\mathrm{bg},t}, \tag{79}$$

where $e_{\mathrm{bg},t}$ is the frame-level background residual.

E.4 Structural ROI and Mask Extraction

All structure-dependent metrics are evaluated within a predefined gravity-dam region of interest (ROI) rather than over the complete image. The structural ROI and segmentation prompts are defined from the real initial reference image and are shared across all evaluated models for the same benchmark sample.

Initial ROI definition. Let Ω_{ROI} denote the predefined spatial region containing the structural portion relevant to evaluation. The ROI is determined before model evaluation using the real conditioning image and the corresponding viewpoint protocol.

The same ROI definition is applied to the real reference and all generated videos associated with the sample, subject to the common spatial transformations.

Regions outside the physical structure of interest are not automatically included in structure-based metrics merely because they fall inside the image bounding box.

Promptable structural segmentation. A fixed version of SAM 2 is used as the primary promptable video segmentation tool for extracting and propagating the gravity-dam mask through time. The segmentation configuration is frozen before benchmark evaluation and includes the model checkpoint, prompt coordinates, confidence settings, propagation strategy, and review rules.

For video frame I_t, the resulting structural mask is denoted by

$$M_t = \mathcal{S}_{\mathrm{SAM2}}\left(I_t, \mathcal{P}, \Theta_{\mathrm{seg}}\right), \tag{80}$$

where $\mathcal{P}$ denotes the predefined prompt set and Θ_{seg} denotes the frozen segmentation configuration.

The prompt locations are established from the initial real reference frame and are not manually re-positioned for individual generated results to improve segmentation quality.

E.5 Optical Flow and Motion Preprocessing

Motion-sensitive evaluation uses complementary dense and sparse tracking strategies because global structural response and local boundary motion have different observational requirements.

Dense optical flow is used primarily to characterize global motion over extended structural regions, whereas sparse feature tracking is used for localized structural edges and high-confidence point trajectories.

Dense optical flow. Global image-space motion is estimated using a fixed Farnebäck dense optical-flow configuration. For two consecutive stabilized frames, I_t and I_{t+1}, the dense flow field is written as

$$\mathbf{F}_t(x, y) = \begin{bmatrix} u_t(x, y) \\ v_t(x, y) \end{bmatrix}, \tag{81}$$

where u_t and v_t denote the horizontal and vertical image-space displacements between the two frames.

Block-wise robust aggregation. To reduce sensitivity to local texture noise and isolated flow outliers, the dense flow may be aggregated over fixed spatial blocks.

Under the current protocol, a nominal 8×8-pixel block is used at the designated motion-evaluation resolution. For block $\mathcal{B}_q$, a robust representative displacement is computed as

$$\bar{\mathbf{F}}_{t,q} = \mathrm{RobustAgg}\left\{\mathbf{F}_t(x,y) \mid (x,y) \in \mathcal{B}_q \cap M_t^{\mathrm{eval}}\right\}, \tag{82}$$

where the aggregation operator is fixed across all evaluated models.

Sparse feature tracking. Local structural boundaries and selected high-texture regions are tracked using feature points and Lucas–Kanade optical flow.

For feature k, the tracked image location at time t is

$$\mathbf{x}_{k,t} = \begin{bmatrix} x_{k,t} \\ y_{k,t} \end{bmatrix}, \tag{83}$$

and its displacement relative to the initial location is

$$\mathbf{d}_k^0(t) = \mathbf{x}_{k,t} - \mathbf{x}_{k,0}. \tag{84}$$

The corresponding incremental displacement is

$$\Delta\mathbf{d}_k(t) = \mathbf{x}_{k,t} - \mathbf{x}_{k,t-1}. \tag{85}$$

Forward–backward reliability check. If a point is tracked from frame t to $t+1$ and then backward to frame t, the forward–backward residual is

$$e_{k,t}^{\mathrm{FB}} = \left\|\mathbf{x}_{k,t} - \mathbf{x}_{k,t}^{\mathrm{back}}\right\|_2. \tag{86}$$

Points with residuals above the frozen threshold are rejected from motion-sensitive evaluation.

Points are also rejected when they leave the valid structural ROI, enter an occluded region, cross an invalid warp boundary, or fail the corresponding feature-quality criterion.

Multi-scale displacement normalization. When optical flow or feature tracking is performed at multiple image scales, displacements are converted to a common source-image pixel unit before any cross-scale aggregation.

Let scale l have spatial scaling factor s_l relative to the common reference resolution. A displacement estimated at that scale, $\mathbf{d}^{(l)}$, is converted as

$$\mathbf{d}^{\mathrm{ref},(l)} = \frac{\mathbf{d}^{(l)}}{s_l}. \tag{87}$$

Only after this normalization are multi-scale observations eligible for fusion.

If multiple scales are combined, the fused displacement is represented as

$$\mathbf{d}^{\mathrm{fused}} = \sum_{l=1}^{L} w_l\, \mathbf{d}^{\mathrm{ref},(l)}, \qquad \sum_{l=1}^{L} w_l = 1, \tag{88}$$

where the scale weights w_l are fixed by the released evaluation configuration and are identical across models.

E.6 Acceleration-Only Sensor Alignment

The sensor-based branch of DamEQ-Bench uses measured acceleration as the principal experimental reference. The benchmark does not assume that ground-truth structural displacement or velocity measurements are available for every evaluated condition.

Sensor preprocessing is performed before any video–sensor comparison and follows the documented acquisition-channel mappings.

The processing sequence is

$$\begin{aligned}\text{raw hardware channel} &\rightarrow \text{event-specific channel mapping} \rightarrow \text{range/calibration verification} \\ &\rightarrow \text{axis correction} \rightarrow \text{coordinate interpretation} \rightarrow \text{unit conversion} \\ &\rightarrow \text{temporal alignment} \rightarrow \text{common-band filtering}.\end{aligned} \tag{89}$$

Video-derived physical displacement. Let $u_i^{\mathrm{pix}}(t)$ denote the stabilized image-space displacement estimated at structural location i.

When reliable local geometric calibration is available, the image-space motion is transformed into physical displacement along the corresponding structural direction,

$$u_i^{\mathrm{phys}}(t) = \mathcal{G}_i\left[u_i^{\mathrm{pix}}(t)\right], \tag{90}$$

where $\mathcal{G}_i(\cdot)$ represents the location- and viewpoint-dependent geometric mapping.

For a locally linear approximation,

$$u_i^{\mathrm{phys}}(t) = \kappa_i u_i^{\mathrm{pix}}(t), \tag{91}$$

where κ_i is the verified local physical scale.

Displacement smoothing and numerical differentiation. Because second-order differentiation strongly amplifies image-tracking noise, physical displacement is processed with a fixed smoothing procedure before acceleration is estimated.

Let

$$u_i(t) = \mathcal{H}_u\left[u_i^{\mathrm{phys}}(t)\right] \tag{92}$$

denote the displacement after applying the frozen displacement-processing filter $\mathcal{H}_u$.

For a uniformly sampled evaluation grid with interval Δt, the video-derived acceleration can be estimated using the central second difference

$$a_i^v(t_n) = \frac{u_i(t_{n+1}) - 2u_i(t_n) + u_i(t_{n-1})}{(\Delta t)^2}. \tag{93}$$

More generally,

$$a_i^v(t) = \frac{d^2 u_i(t)}{dt^2}. \tag{94}$$

Absolute versus relative response. The physical interpretation of the video-derived acceleration depends on the reference used for displacement extraction.

If the stabilized displacement is measured with respect to an inertially stationary background and the camera geometry is sufficiently calibrated, the resulting displacement corresponds conceptually

to absolute structural motion, and its second derivative is compared with the corresponding absolute sensor acceleration.

If instead the visual displacement is measured relative to the dam base, let

$$u_i^{\mathrm{rel}}(t) = u_i^{\mathrm{abs}}(t) - u_0(t), \tag{95}$$

where $u_0(t)$ denotes the base displacement.

The corresponding relative acceleration is

$$a_i^{\mathrm{rel}}(t) = \frac{d^2 u_i^{\mathrm{rel}}(t)}{dt^2} = a_i^{\mathrm{abs}}(t) - a_0(t). \tag{96}$$

Therefore, absolute acceleration can be recovered from a reliable base-relative video displacement through

$$a_i^{\mathrm{abs}}(t) = \frac{d^2 u_i^{\mathrm{rel}}(t)}{dt^2} + a_0(t), \tag{97}$$

provided that the base acceleration $a_0(t)$ is available on the corresponding axis and is correctly synchronized.

Alternatively, the video-derived relative acceleration can be compared directly with the relative sensor response

$$a_{i,s}^{\mathrm{rel}}(t) = a_i^s(t) - a_0^s(t). \tag{98}$$

F Visual Quality Metrics

The six visual-quality dimensions evaluate whether a generated seismic video remains visually observable, structurally recognizable, temporally stable, and consistent with the appearance of the corresponding real experiment. These metrics are intentionally separated from the physics-oriented metrics introduced later: a visually high-quality video is not necessarily a physically correct seismic-response video.

Let the luminance range be normalized to $[0, 1]$, and let the common evaluation interval contain N frames. For scalar image or video statistics for which neither a larger nor a smaller raw value is intrinsically preferable, we evaluate similarity to the real reference rather than rewarding the raw magnitude itself. For a generated statistic x_g and reference statistic x_r, we define

$$Q_{\mathrm{ref}}(x_g, x_r; \sigma_x) = \exp\left(-\frac{|x_g - x_r|}{\sigma_x + \epsilon}\right), \tag{99}$$

F.1 Subject Consistency

Subject Consistency evaluates whether the gravity-dam specimen remains visible and recognizable throughout generation and whether its structural appearance remains compatible with the real experimental reference.

The metric consists of three components: Object Persistence, Identity Continuity, and Structural Appearance Consistency.

Object persistence. Let M_t^g denote the valid generated structural mask at frame t. The visibility score is

$$R_{\mathrm{vis}} = \frac{1}{N}\sum_{t=1}^{N} \mathbb{I}\left(M_t^g \text{ is valid and sufficiently visible}\right). \tag{100}$$

The visibility criterion and minimum valid structural area are frozen before evaluation.

Identity continuity. A fixed visual feature encoder $\phi(\cdot)$ is applied to the structural ROI. The generated structure is compared with the experiment-specific conditioning appearance rather than with an unrelated generic object:

$$S_{\mathrm{id}} = \frac{1}{N}\sum_{t=1}^{N} \mathrm{CosSim}\left[\phi(I_t^g \odot M_t^g), \phi(I^0 \odot M^0)\right]. \tag{101}$$

This term detects identity drift such as material-texture replacement, unrealistic color changes, disappearance of characteristic boundaries, or progressive corruption of the dam appearance.

Structural appearance consistency. To avoid penalizing physically valid structural translation, structural appearance is compared after the common camera-motion compensation and metric-specific motion compensation:

$$S_{\mathrm{struct}} = \frac{1}{N}\sum_{t=1}^{N} \mathrm{SSIM}\left(I_t^{g,\mathrm{mc}}, I_t^{r,\mathrm{mc}}\right)_{\Omega_t^{\mathrm{struct}}}, \tag{102}$$

where $\Omega_t^{\mathrm{struct}}$ denotes the valid common structural region.

The final Subject Consistency score is

$$S_{\mathrm{sub}} = 0.35R_{\mathrm{vis}} + 0.25S_{\mathrm{id}} + 0.40S_{\mathrm{struct}}. \tag{103}$$

All three components are mapped to $[0, 1]$, and higher values indicate better visual subject consistency.

F.2 Official VBench Aesthetic Quality

Aesthetic Quality is evaluated using the official VBench `aesthetic_quality` implementation without redefining its internal sub-components.

For each frame I_t, VBench extracts a normalized CLIP ViT-L/14 image embedding and applies the released LAION aesthetic predictor. Let $h_{\mathrm{aes}}(\cdot)$ denote the frozen aesthetic predictor and $e_{\mathrm{CLIP}}(\cdot)$ denote the normalized CLIP image embedding. The video-level score can be represented as

$$S_{\text{VB-Aes}} = \frac{1}{N}\sum_{t=1}^{N} \frac{h_{\mathrm{aes}}\left(e_{\mathrm{CLIP}}(I_t)\right)}{10}. \tag{104}$$

Higher values indicate higher predicted visual aesthetic quality.

This metric evaluates general perceptual presentation rather than experimental or physical correctness. In particular, a visually polished generated video may receive a high aesthetic score while exhibiting incorrect structural motion.

F.3 Custom Aesthetic Presentation Quality

Complementary to the learned VBench aesthetic score, we retain an interpretable reference-based presentation metric derived from the visual-quality protocol of our previous benchmark.

The metric evaluates four experimentally relevant appearance statistics: Color Richness, Color Distribution, Brightness Distribution, and Composition Consistency. Unlike Section F.2, this is a custom DamEQ-Bench metric and is not part of the official VBench implementation.

Color richness. Color richness is characterized using the frame-level color-dispersion statistic

$$x_{\text{color}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \left\| \mathbf{C}(i) - \bar{\mathbf{C}} \right\|_2^2 . \tag{105}$$

Rather than rewarding larger color variance, we compare the generated statistic with the real experimental reference:

$$q_{\text{color}} = Q_{\text{ref}} \left(x^g_{\text{color}}, x^r_{\text{color}}; \sigma_{\text{color}} \right) . \tag{106}$$

Color distribution. Color-distribution consistency is evaluated using a fixed histogram representation $\mathbf{h}$ and a frozen histogram distance D_{hist}:

$$q_{\text{dist}} = \exp \left[- \frac{D_{\text{hist}}(\mathbf{h}^g, \mathbf{h}^r)}{\sigma_{\text{hist}} + \epsilon} \right] . \tag{107}$$

Brightness distribution. Frame luminance statistics are computed within the predefined valid evaluation region, and their agreement with the real reference is measured as

$$q_{\text{lum}} = Q_{\text{ref}} \left(x^g_{\text{lum}}, x^r_{\text{lum}}; \sigma_{\text{lum}} \right) . \tag{108}$$

Composition consistency. Composition is characterized primarily by the relative structural occupancy and location in the frame. For example,

$$R_{\text{occ}} = \frac{|M_t^{\text{struct}}|}{|\Omega_t^{\text{valid}}|}, \tag{109}$$

and the generated occupancy statistics are compared with the corresponding real reference.

The final score is

$$S_{\text{Aes-Custom}} = 0.25 q_{\text{color}} + 0.25 q_{\text{dist}} + 0.25 q_{\text{lum}} + 0.25 q_{\text{comp}}. \tag{110}$$

Higher values indicate closer agreement with the visual presentation of the real experiment.

F.4 Temporal Flickering

Temporal Flickering measures visually undesirable frame-to-frame instability after accounting for the legitimate motion of the seismic response. It is designed to penalize texture popping, brightness oscillation, structural-mask jitter, and perceptual discontinuity rather than the physical vibration itself.

Let $\mathcal{W}_{t \to t+1}$ denote the frozen motion-compensation warp obtained from reliable flow and let Ω_t^{nc} denote the valid non-occluded region. The motion-compensated residual is

$$e_t = \frac{1}{|\Omega_t^{\text{nc}}|} \sum_{\mathbf{x} \in \Omega_t^{\text{nc}}} \left| I_{t+1}(\mathbf{x}) - \mathcal{W}_{t \to t+1} \left[I_t \right] (\mathbf{x}) \right| . \tag{111}$$

Four complementary terms are retained.

Segmentation consistency. q_{seg} measures undesirable frame-to-frame instability of the structural mask after accounting for expected structural motion.

Residual consistency. The mean motion-compensated residual is compared with that of the real video:

$$q_{\mathrm{res}} = Q_{\mathrm{ref}}\left(\bar{e}^{g}, \bar{e}^{r}; \sigma_{\mathrm{res}}\right). \tag{112}$$

Temporal residual stability. The temporal variation of the residual sequence is compared with the real reference:

$$q_{\mathrm{temp}} = Q_{\mathrm{ref}}\left(\mathrm{Std}(e^{g}), \mathrm{Std}(e^{r}); \sigma_{\mathrm{temp}}\right). \tag{113}$$

Perceptual stability. A fixed perceptual feature distance between motion-compensated neighboring frames provides the fourth component q_{perc}.

The final Temporal Flickering score is

$$S_{\mathrm{flick}} = 0.15q_{\mathrm{seg}} + 0.35q_{\mathrm{res}} + 0.30q_{\mathrm{temp}} + 0.20q_{\mathrm{perc}}. \tag{114}$$

Higher values indicate lower undesirable flicker.

F.5 Imaging Quality

Imaging Quality evaluates whether the generated frames retain sufficient visual clarity for structural observation. The metric contains four components: sharpness, contrast, noise consistency, and effective spatial detail.

Sharpness. Sharpness is measured over the fixed valid evaluation region using the mean gradient magnitude,

$$x_{\mathrm{sharp}} = \frac{1}{|\Omega|}\sum_{i\in\Omega}\left\|\nabla I(i)\right\|_2. \tag{115}$$

The generated value is compared with the real reference through Q_{ref} rather than rewarding arbitrarily strong sharpening.

Contrast. Luminance contrast is characterized by

$$x_{\mathrm{contrast}} = \sqrt{\frac{1}{|\Omega|}\sum_{i\in\Omega}\left(Y(i)-\bar{Y}\right)^2}. \tag{116}$$

Noise consistency. Noise is estimated from high-frequency residual energy in regions that exclude strong structural edges,

$$x_{\mathrm{noise}} = \frac{1}{|\Omega_{\mathrm{flat}}|}\sum_{i\in\Omega_{\mathrm{flat}}}\left|I(i) - G_{\sigma}(I)(i)\right|, \tag{117}$$

where G_{σ} denotes the frozen low-pass operator.

Effective detail. Effective detail is characterized from the proportion of spatial spectral energy retained within a predefined informative frequency band $\mathcal{F}_d$:

$$x_{\mathrm{detail}} = \frac{\sum_{f\in\mathcal{F}_d}|\mathcal{F}\{I\}(f)|^2}{\sum_{f}|\mathcal{F}\{I\}(f)|^2 + \epsilon}. \tag{118}$$

The final Imaging Quality score is

$$S_{\text{img}} = 0.30q_{\text{sharp}} + 0.25q_{\text{contrast}} + 0.25q_{\text{noise}} + 0.20q_{\text{detail}}. \tag{119}$$

Each component measures similarity to the real reference, and higher values indicate better imaging quality.

F.6 Scene Coherence

Scene Coherence evaluates whether the generated dam, reservoir, model box, instrumentation, and surrounding background remain visually compatible with the experiment-specific reference scene.

The metric retains three interpretable scene statistics from the previous benchmark design: Scene Complexity, Edge Distribution, and Texture Distribution. Unlike the earlier raw-statistic formulation, DamEQ-Bench evaluates agreement with the real reference rather than rewarding larger statistical values.

Scene complexity. Scene complexity is characterized by a fixed-grid appearance entropy,

$$H_{\text{scene}} = -\sum_{k} p_k \log(p_k + \epsilon), \tag{120}$$

and converted to a reference-similarity score

$$q_{\text{complexity}} = Q_{\text{ref}}\left(H^g_{\text{scene}}, H^r_{\text{scene}}; \sigma_H\right). \tag{121}$$

Edge distribution. The normalized edge density is

$$R_{\text{edge}} = \frac{1}{|\Omega|}\sum_{i\in\Omega} \mathbb{I}\left(\|\nabla I(i)\| > \tau_{\text{edge}}\right), \tag{122}$$

and the score q_{edge} measures agreement with the corresponding reference statistic.

Texture distribution. Local texture descriptors are extracted over a fixed spatial grid. Let $\mathbf{h}_{\text{tex}}$ denote the resulting normalized texture distribution. Its consistency with the real scene is defined as

$$q_{\text{texture}} = \exp\left[-\frac{D_{\text{tex}}(\mathbf{h}^g_{\text{tex}}, \mathbf{h}^r_{\text{tex}})}{\sigma_{\text{tex}} + \epsilon}\right]. \tag{123}$$

The final Scene Coherence score is

$$S_{\text{scene}} = 0.40q_{\text{complexity}} + 0.35q_{\text{edge}} + 0.25q_{\text{texture}}. \tag{124}$$

Higher values indicate stronger agreement with the real experimental scene.

A generated scene is therefore not rewarded merely for containing more edges, higher entropy, or richer texture. For example, hallucinated cracks, additional buildings, artificial water-surface patterns, or background objects may increase raw image complexity while decreasing the reference-based Scene Coherence score.

G Response-Oriented Evaluation Metrics

The eight response-oriented metrics evaluate whether the visible dynamics of generated videos are consistent with experimentally observed seismic responses. They comprise four physical consistency metrics inherited and refined from Earthquake-Bench—covering structure-level displacement, pixel-level edge motion, segment-level structure consistency, and temporal warping—and four

Table F1: Summary of the six general-purpose visual-quality metrics.

Metric	Primary target	Reference type
Subject Consistency	Structure identity and persistence	Real initial/reference video
VBench Aesthetic Quality	General perceptual aesthetics	Official learned predictor
Custom Aesthetic Presentation	Color, luminance, composition	Real reference statistics
Temporal Flickering	Undesirable temporal artifacts	Motion-compensated real video
Imaging Quality	Sharpness, contrast, noise, detail	Real reference statistics
Scene Coherence	Experimental-scene consistency	Real reference statistics

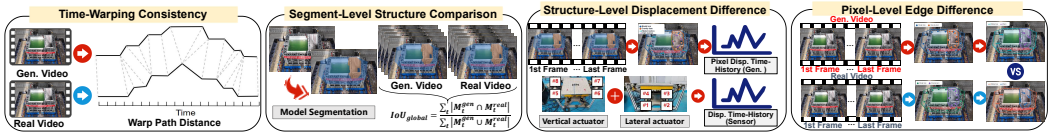


Figure G1: Calculation framework of the four physical consistency metrics.

newly introduced engineering response metrics targeting modal behavior, dynamic response histories, peak/crest characteristics, and height-dependent acceleration amplification. The engineering response metrics are introduced in the main paper, while Fig. G1 illustrates the calculation framework of the four physical consistency metrics.

All response comparisons follow the observation, synchronization, calibration, spatial-standardization, and metric-eligibility rules defined in Section E. These metrics quantify *observable response consistency*; they do not imply that complete stress, strain, internal-force, or damage fields can be reconstructed from RGB video alone.

Let $x_r(t)$ and $x_g(t)$ denote the reference and generated response trajectories over an observation interval T_{obs} containing N valid samples.

Common correlation, error, and amplitude terms. When both trajectories contain sufficient non-zero response energy, the correlation similarity is defined as

$$S_{\text{corr}} = \frac{1 + \rho(x_g, x_r)}{2}, \tag{125}$$

where $\rho(\cdot, \cdot)$ is the Pearson correlation coefficient.

The normalized root-mean-square error is

$$E_{\text{rms}} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(x_g(t_n) - x_r(t_n)\right)^2}}{\text{RMS}(x_r) + \delta_x}, \tag{126}$$

where δ_x is a quantity-specific noise-floor term expressed in the same physical or image-space unit as x_r. It is determined from the reference measurement uncertainty rather than introduced as an arbitrarily small dimensionless constant.

For a peak-amplitude operator

$$A(x) = \max_t |x(t)|, \tag{127}$$

the normalized amplitude error and corresponding similarity are

$$E_A = \frac{|A(x_g) - A(x_r)|}{A(x_r) + \delta_A}, \qquad S_A = \exp(-E_A), \tag{128}$$

where δ_A represents the corresponding observable amplitude floor.

Common frequency-domain terms. Auto- and cross-power spectral densities are estimated using the same segment length, overlap, window function, and common trustworthy frequency band for the real and generated responses.

The magnitude-squared coherence is

$$\gamma^2(f) = \frac{|P_{gr}(f)|^2}{P_{gg}(f)P_{rr}(f) + \epsilon}, \tag{129}$$

and the band-aggregated coherence score is

$$S_{\text{coh}} = \frac{\sum_{f\in\mathcal{B}} w_f \gamma^2(f)}{\sum_{f\in\mathcal{B}} w_f}, \tag{130}$$

where $\mathcal{B}$ denotes the valid comparison band and w_f is a frozen reference-derived spectral weight.

The cross-spectral phase difference is

$$\Delta\phi(f) = \text{wrap}\left[\arg P_{gr}(f)\right], \tag{131}$$

from which the phase-consistency score is defined as

$$S_{\text{phase}} = \exp\left[-\frac{\sum_{f\in\mathcal{B}} w_f |\Delta\phi(f)|}{\tau_\phi \sum_{f\in\mathcal{B}} w_f}\right], \tag{132}$$

where τ_ϕ is frozen before test evaluation.

The valid band and spectral weights are determined from the real response, measurement noise, and observable bandwidth and are shared by all models.

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

D-Structure evaluates the global visible displacement response of the dam. Unlike the rigid-body setting of the original Earthquake-Bench, the gravity-dam specimen is not assumed to undergo strictly rigid motion. A robust global trajectory is therefore extracted from reliable motion within the structural ROI, while local trajectories may additionally be retained to reveal deformation patterns that would otherwise cancel under spatial averaging.

Let $x_r(t)$ and $x_g(t)$ denote directionally comparable structure-level displacement trajectories extracted from the real and generated videos, respectively. The D-Structure score is

$$S_D = 0.40S_{\text{corr}} + 0.30S_{\text{phase}} + 0.20S_{\text{coh}} + 0.10S_A. \tag{133}$$

The component weights follow the response-oriented design inherited from Earthquake-Bench.

G.2 Pixel-Level Edge Difference (DEdge-Pixel)

DEdge-Pixel evaluates local structural-motion consistency along representative visible boundaries. Feature locations are initialized from the real reference image and propagated through the corresponding real and generated sequences using the common Lucas–Kanade tracking protocol.

For feature point i,

$$\mathbf{d}_i(t) = \mathbf{p}_i(t) - \mathbf{p}_i(0) \tag{134}$$

denotes its displacement relative to the initial image position.

Shape similarity. For reference-motion scale

$$r_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{d}_i^r(t_n)\|_2^2}, \tag{135}$$

the normalized displacement-field error is

$$E_{\text{shape}} = \frac{1}{N_p N}\sum_{i=1}^{N_p}\sum_{n=1}^{N} \frac{\|\mathbf{d}_i^g(t_n) - \mathbf{d}_i^r(t_n)\|_2}{r_i + \delta_d}, \tag{136}$$

and

$$S_{\text{shape}} = \exp(-E_{\text{shape}}). \tag{137}$$

Temporal increment similarity. Let

$$\Delta\mathbf{d}_i(t_n) = \mathbf{d}_i(t_n) - \mathbf{d}_i(t_{n-1}). \tag{138}$$

The temporal term compares generated and real displacement increments,

$$E_{\text{temporal}} = \frac{1}{N_p(N-1)}\sum_{i,n} \frac{\|\Delta\mathbf{d}_i^g(t_n) - \Delta\mathbf{d}_i^r(t_n)\|_2}{r_{\Delta,i} + \delta_\Delta}, \tag{139}$$

with

$$S_{\text{temporal}} = \exp(-E_{\text{temporal}}). \tag{140}$$

Amplitude similarity. For

$$A_i^{(\cdot)} = \max_t \|\mathbf{d}_i^{(\cdot)}(t)\|_2, \tag{141}$$

we define

$$E_{\text{amp}} = \frac{1}{N_p}\sum_{i=1}^{N_p} \frac{|A_i^g - A_i^r|}{A_i^r + \delta_A}, \qquad S_{\text{amplitude}} = \exp(-E_{\text{amp}}). \tag{142}$$

Tracking quality. The data-quality term retains the interpretable structure used in the original Earthquake-Bench implementation,

$$S_{\text{quality}} = 0.50R_{\text{track}} + 0.30R_{\text{edge}} + 0.20R_{\text{spatial}}, \tag{143}$$

where R_{track} is the valid tracking ratio, R_{edge} measures structural-edge coverage, and R_{spatial} quantifies the spatial distribution of the retained points.

The final score is

$$S_{\text{edge}} = 0.35S_{\text{shape}} + 0.25S_{\text{temporal}} + 0.20S_{\text{amplitude}} + 0.20S_{\text{quality}}. \tag{144}$$

Multi-scale displacements are first converted back to the same reference pixel coordinate system according to Section E.5 before fusion.

G.3 Segment-Level Structure Consistency (SAM-Structure)

SAM-Structure evaluates the visible structural geometry using the fixed SAM 2 protocol defined in Section E.4.

For generated and real masks M_t^g and M_t^r, the frame-level Intersection over Union is

$$\mathrm{IoU}_t = \frac{|M_t^g \cap M_t^r|}{|M_t^g \cup M_t^r|}. \tag{145}$$

For completeness, coverage and precision may additionally be reported as

$$R_t = \frac{|M_t^g \cap M_t^r|}{|M_t^r|}, \qquad P_t = \frac{|M_t^g \cap M_t^r|}{|M_t^g|}. \tag{146}$$

These quantities correspond to recall and precision rather than to two directional definitions of IoU. Their harmonic mean is equivalent to a Dice/F1-type overlap score and is not treated as independent IoU evidence.

Spatial alignment. The spatial term is the mean frame-level IoU,

$$S_{\mathrm{spatial}} = \frac{1}{N}\sum_{t=1}^{N} \mathrm{IoU}_t\,. \tag{147}$$

Temporal consistency. For each sequence, define its adjacent-frame mask-overlap trajectory

$$J_t^{(\cdot)} = \mathrm{IoU}\left(M_t^{(\cdot)}, M_{t-1}^{(\cdot)}\right). \tag{148}$$

The temporal score compares the evolution of these two trajectories,

$$S_{\mathrm{temporal}} = \exp\left[-\frac{\frac{1}{N-1}\sum_{t=2}^{N}|J_t^g - J_t^r|}{\tau_J}\right], \tag{149}$$

where τ_J is frozen from the development data.

Mask-area stability. Let

$$q_t^{(\cdot)} = \frac{|M_t^{(\cdot)}| - |M_{t-1}^{(\cdot)}|}{|M_{t-1}^{(\cdot)}| + \epsilon}. \tag{150}$$

The stability term is

$$S_{\mathrm{stability}} = \exp\left[-\frac{\frac{1}{N-1}\sum_{t=2}^{N}|q_t^g - q_t^r|}{\tau_q}\right]. \tag{151}$$

This formulation evaluates whether the generated mask changes similarly to the real structural mask rather than requiring the structure to remain artificially static.

Boundary alignment. Let ∂M denote the structural boundary, and let $D_{\mathrm{sym}}(\partial M_t^g, \partial M_t^r)$ denote a symmetric mean boundary distance. Normalizing by the reference dam height H_{dam} in the same standardized image coordinates gives

$$e_{\mathrm{bnd}} = \frac{1}{N}\sum_{t=1}^{N}\frac{D_{\mathrm{sym}}(\partial M_t^g, \partial M_t^r)}{H_{\mathrm{dam}} + \epsilon}, \tag{152}$$

and

$$S_{\mathrm{alignment}} = \exp\left(-\frac{e_{\mathrm{bnd}}}{\tau_b}\right). \tag{153}$$

The final SAM-Structure score is

$$S_{\mathrm{SAM}} = 0.40S_{\mathrm{spatial}} + 0.30S_{\mathrm{temporal}} + 0.20S_{\mathrm{stability}} + 0.10S_{\mathrm{alignment}}. \tag{154}$$

G.4 Dynamic Time Warping Similarity (DTW-Sim)

DTW-Sim evaluates similarity of the observed motion pattern while allowing a limited amount of local temporal mismatch. It is complementary to, rather than a replacement for, strict physical-time evaluation.

At each time step, a motion feature vector is constructed from the standardized structural response. The feature may contain multi-scale displacement or optical-flow magnitude, directional encoding, and frame-to-frame change terms.

Feature normalization uses statistics frozen from the training/reference data:

$$\hat{\mathbf{f}} = \frac{\mathbf{f} - \boldsymbol{\mu}_{\mathrm{ref}}}{\boldsymbol{\sigma}_{\mathrm{ref}} + \boldsymbol{\epsilon}}. \tag{155}$$

The generated and real sequences are not normalized independently to unit variance, because such a procedure would remove genuine response-amplitude differences.

For generated feature $\hat{\mathbf{f}}_i^g$ and reference feature $\hat{\mathbf{f}}_j^r$, the local cost is

$$d(i,j) = \frac{1}{D_f}\left\|\hat{\mathbf{f}}_i^g - \hat{\mathbf{f}}_j^r\right\|_2, \tag{156}$$

where D_f is the feature dimension.

The accumulated cost satisfies

$$C(i,j) = d(i,j) + \min\left\{C(i-1,j), C(i,j-1), C(i-1,j-1)\right\}. \tag{157}$$

The admissible path is restricted by a frozen temporal window,

$$|i-j| \le w_{\mathrm{DTW}}, \tag{158}$$

preventing physically implausible unrestricted time warping.

For optimal path

$$\mathcal{W} = \{(i_k, j_k)\}_{k=1}^{K}, \tag{159}$$

the normalized distance is

$$D_{\mathrm{DTW}} = \frac{1}{K}\sum_{(i,j)\in\mathcal{W}} d(i,j). \tag{160}$$

The distance-similarity term is

$$S_{\text{dist}} = \exp\left(-\frac{D_{\text{DTW}}}{\tau_D}\right). \tag{161}$$

Rank similarity. A rank correlation is computed over the aligned response sequence and mapped to $[0, 1]$:

$$S_{\text{rank}} = \frac{1 + \tau_{\text{K}}}{2}, \tag{162}$$

where τ_{K} denotes Kendall's rank correlation.

Warping-path alignment. Let N_{nd} denote the number of non-diagonal transitions in the optimal path. The path-alignment score is

$$S_{\text{align}} = 1 - \frac{N_{\text{nd}}}{\max(K - 1, 1)}. \tag{163}$$

Large deviations from the diagonal therefore reduce the score even when the final DTW distance remains small.

Local dynamic consistency. Local velocities are computed on the original unwarped time axes. Differences are evaluated only for aligned path positions at which both sequences advance in time. Repeated DTW indices are not treated as new physical time samples.

Let $\mathcal{K}_d$ denote the set of valid jointly advancing path steps. Then

$$E_{\text{local}} = \frac{1}{|\mathcal{K}_d|} \sum_{k \in \mathcal{K}_d} \frac{|v_g(i_k) - v_r(j_k)|}{v_{\text{scale}} + \delta_v}, \tag{164}$$

and

$$S_{\text{local}} = \exp(-E_{\text{local}}). \tag{165}$$

The final score is

$$S_{\text{DTW}} = 0.30 S_{\text{dist}} + 0.25 S_{\text{rank}} + 0.25 S_{\text{align}} + 0.20 S_{\text{local}}. \tag{166}$$

G.5 Multi-Order Modal Consistency

Multi-Order Modal Consistency evaluates whether the generated video reproduces experimentally observable modal frequencies and spatial mode shapes of the gravity dam.

A set of N_p spatially corresponding structural points is selected along the dam height and other informative structural regions. Their displacement responses are assembled as

$$\mathbf{u}(t) = \begin{bmatrix} u_1(t) & u_2(t) & \cdots & u_{N_p}(t) \end{bmatrix}^{\mathsf{T}}. \tag{167}$$

The cross-power spectral density matrix is estimated using the frozen spectral-processing protocol,

$$\mathbf{S}_{uu}(f) = \mathbf{U}(f)\mathbf{\Sigma}(f)\mathbf{U}^{H}(f), \tag{168}$$

where $\mathbf{\Sigma}(f)$ contains the singular values and the columns of $\mathbf{U}(f)$ provide the corresponding singular vectors.

Frequency Domain Decomposition (FDD) is used to identify candidate response modes from the spectral matrix (Brincker et al., 2001). However, earthquake excitation is generally colored and non-stationary, and strong excitation may alter the structural state. Spectral peaks are therefore treated as modal candidates rather than automatically interpreted as structural natural frequencies.

Candidate modes must additionally satisfy the reference-data eligibility criteria, including spectral-peak stability, spatial-mode consistency, sufficient observation duration, adequate signal-to-noise ratio, and, where available, supporting white-noise or sensor evidence.

Mode matching. Generated and reference modes are matched one-to-one using both frequency proximity and spatial mode-shape correspondence rather than by independently ranking spectral peaks by amplitude.

For reference mode $\boldsymbol{\phi}_{k,r}$ and generated mode $\boldsymbol{\phi}_{k,g}$, the Modal Assurance Criterion is

$$\mathrm{MAC}_k = \frac{\left|\boldsymbol{\phi}_{k,r}^H \boldsymbol{\phi}_{k,g}\right|^2}{\left(\boldsymbol{\phi}_{k,r}^H \boldsymbol{\phi}_{k,r}\right)\left(\boldsymbol{\phi}_{k,g}^H \boldsymbol{\phi}_{k,g}\right)}. \tag{169}$$

For real-valued mode shapes, the conjugate transpose reduces to the ordinary transpose.

Frequency agreement is defined as

$$S_{f,k} = \exp\left[-\frac{|f_{k,g} - f_{k,r}|}{\tau_{f,k}}\right], \tag{170}$$

where $\tau_{f,k}$ is determined from the reference spectral resolution and mode-identification uncertainty and is frozen before model evaluation.

For the first three identifiable modes, the modal score is

$$S_{\mathrm{Modal3}} = \sum_{k=1}^{3} w_k \left(0.70\,\mathrm{MAC}_k + 0.30 S_{f,k}\right), \tag{171}$$

with

$$(w_1, w_2, w_3) = (0.50, 0.30, 0.20). \tag{172}$$

G.6 Dynamic Response Consistency

Dynamic Response Consistency evaluates the complete response history using time-domain correlation and normalized error together with frequency-domain coherence and phase agreement.

Two reference variants are distinguished: *Dyn-GT*, which compares directly against an experimentally derived reference response, and *Dyn-ODE*, which compares against a separately validated learned continuous-time reference model.

If a Neural ODE reference model is used, its latent state satisfies

$$\dot{\mathbf{s}}(t) = f_\theta\left(\mathbf{s}(t), \mathbf{a}_0(t), \mathbf{c}\right), \tag{173}$$

with output

$$R_{\mathrm{ODE}}(t) = h_\theta\left(\mathbf{s}(t)\right). \tag{174}$$

This follows the continuous-time formulation of Neural ODEs (Chen et al., 2018); the neural network parameterizes the state derivative and the continuous trajectory is obtained through numerical

integration. The architecture itself does not guarantee compliance with the true structural equations of motion.

For either reference R_{ref}, the dynamic-consistency score is

$$S_{\mathrm{Dyn}} = 0.35S_{\mathrm{corr}} + 0.30\exp(-E_{\mathrm{rms}}) + 0.20S_{\mathrm{coh}} + 0.15S_{\mathrm{phase}}. \tag{175}$$

G.7 Peak/Crest Response Consistency

Peak/Crest Response Consistency evaluates whether the generated response reproduces the maximum response magnitude, its occurrence time, and the prominence of that peak relative to the overall vibration level.

For response $D(t_n)$,

$$A = \max_n |D(t_n)|, \tag{176}$$

$$t^{\mathrm{peak}} = \min\left\{t_n : |D(t_n)| = A\right\}, \tag{177}$$

where the earliest occurrence is selected when several samples attain the same maximum magnitude.

The crest factor is

$$CF = \frac{A}{\sqrt{\frac{1}{N}\sum_{n=1}^{N} D^2(t_n)} + \delta_{\mathrm{RMS}}}. \tag{178}$$

The normalized amplitude error is

$$E_A = \frac{|A_g - A_r|}{A_r + \delta_A}, \tag{179}$$

while the peak-time error used for scoring is

$$E_T = \frac{|t_g^{\mathrm{peak}} - t_r^{\mathrm{peak}}|}{\tau_T}, \tag{180}$$

where τ_T is a fixed temporal tolerance derived from the benchmark sampling and validation protocol rather than optimized per model.

The absolute peak-time difference

$$\Delta t_{\mathrm{peak}} = |t_g^{\mathrm{peak}} - t_r^{\mathrm{peak}}| \tag{181}$$

is additionally reported in seconds.

The crest-factor error is

$$E_{\mathrm{CF}} = \frac{|CF_g - CF_r|}{CF_r + \delta_{\mathrm{CF}}}. \tag{182}$$

The final score is

$$S_{\mathrm{Peak}} = 0.45\exp(-E_A) + 0.35\exp(-E_T) + 0.20\exp(-E_{\mathrm{CF}}). \tag{183}$$

G.8 Acceleration Amplification Consistency

Acceleration Amplification Consistency (AAC) evaluates whether the generated response reproduces the magnitude and spatial variation of acceleration amplification along the dam height.

Three physically corresponding dam locations are selected from valid lower-, middle-, and upper-elevation measurements. They are not defined by dividing the image into three equal-height regions.

Let $a_0(t)$ denote the common measured table or dam-base reference acceleration, and let $a_i^s(t)$ and $a_i^v(t)$ denote the sensor-measured and video-derived absolute accelerations at elevation i, respectively, after unit conversion, directional mapping, synchronization, and common-band filtering.

For $i \in \{1, 2, 3\}$, the sensor and video amplification factors are

$$AF_i^s = \frac{\max_t |a_i^s(t)|}{\max_t |a_0(t)| + \delta_0}, \qquad AF_i^v = \frac{\max_t |a_i^v(t)|}{\max_t |a_0(t)| + \delta_0}. \tag{184}$$

AAC is unavailable when the reference input acceleration is below the validated measurement threshold.

Amplification-level consistency. The mean normalized amplification error is

$$E_{\text{amp}} = \frac{1}{3} \sum_{i=1}^{3} \frac{|AF_i^v - AF_i^s|}{AF_i^s + \delta_{\text{AF}}}, \tag{185}$$

with

$$S_{\text{amp}} = \exp(-E_{\text{amp}}). \tag{186}$$

Height-wise gradient consistency. Define the adjacent amplification increments as

$$G_{12}^{(\cdot)} = AF_2^{(\cdot)} - AF_1^{(\cdot)}, \qquad G_{23}^{(\cdot)} = AF_3^{(\cdot)} - AF_2^{(\cdot)}. \tag{187}$$

The normalized gradient error is

$$E_{\text{grad}} = \frac{1}{2} \sum_{(i,j) \in \{(1,2),(2,3)\}} \frac{|G_{ij}^v - G_{ij}^s|}{|G_{ij}^s| + \delta_G}, \tag{188}$$

and

$$S_{\text{grad}} = \exp(-E_{\text{grad}}). \tag{189}$$

The uncertainty term δ_G prevents the relative error from becoming unbounded when the experimentally measured amplification difference between two elevations approaches zero.

If the vertical spacing between the selected measurement locations is unequal, the benchmark additionally reports the physical gradient

$$\frac{AF_j - AF_i}{z_j - z_i}. \tag{190}$$

No monotonic increase of amplification with elevation is assumed.

Table G1: Summary of the eight response-oriented metrics in DamEQ-Bench.

Metric	Primary observation	Main components	Reference requirement
D-Structure	Global structural displacement	Corr./Phase/Coherence/Amplitude	Real video
SAM-Structure	Structural masks	Spatial/Temporal/Stability/Boundary	Real video
DTW-Sim	Motion-feature sequence	Distance/Rank/Path/Local dynamics	Real video
DEdge-Pixel	Local edge trajectories	Shape/Temporal/Amplitude/Quality	Real video
ODE-Response	Dynamic response history	Corr./NRMSE/Coherence/Phase	Valid response trajectory
Peak-Response	Extreme response characteristics	Amplitude/Timing/Crest factor	Valid response trajectory
AF-Profile	Height-wise acceleration response	Level/Gradient/Profile/Timing	Video–sensor calibrated subset
MAC-Modal	Modal frequencies and mode shapes	MAC/Frequency consistency	Modal-eligible real response

Spatial-profile consistency. Let

$$\mathbf{AF}^s = [AF_1^s, AF_2^s, AF_3^s]^\mathsf{T}, \qquad \mathbf{AF}^v = [AF_1^v, AF_2^v, AF_3^v]^\mathsf{T}. \tag{191}$$

When the reference profile contains sufficient spatial variation,

$$S_{\text{profile}} = \frac{1 + \rho(\mathbf{AF}^v, \mathbf{AF}^s)}{2}. \tag{192}$$

Because only three elevations are used, the raw amplification vectors are always reported together with this correlation-based profile term.

Peak-time consistency. For each elevation, let t_i^s and t_i^v denote the occurrence times of the maximum absolute acceleration. The mean normalized timing error is

$$E_{\text{time}} = \frac{1}{3} \sum_{i=1}^{3} \frac{|t_i^v - t_i^s|}{\tau_T}, \tag{193}$$

and

$$S_{\text{time}} = \exp(-E_{\text{time}}). \tag{194}$$

The final AAC score is

$$S_{\text{AAC}} = 0.40 S_{\text{amp}} + 0.25 S_{\text{grad}} + 0.20 S_{\text{profile}} + 0.15 S_{\text{time}}. \tag{195}$$

The four components correspond respectively to absolute amplification level, height-wise variation, spatial profile, and peak-response timing.

H Baseline Generation and LoRA Adaptation Results

H.1 Qualitative Results of Baseline Models

Figure H1 presents representative generation results from the evaluated baseline models under the DamEQ-Bench protocol. For each model, temporally ordered frames are shown to illustrate the evolution of the generated seismic-response video. The comparison covers representative I2V and V2V models, including both open-source and commercial systems, together with the StaticRepeat controls. These examples provide a direct qualitative view of each model's ability to preserve the dam, reservoir, supporting structure, and surrounding experimental scene while producing temporally coherent structural motion. The results also reveal typical failure modes of general-purpose video generators, including structural deformation, appearance drift, excessive blur, scene reconstruction artifacts, and unrealistic temporal changes.

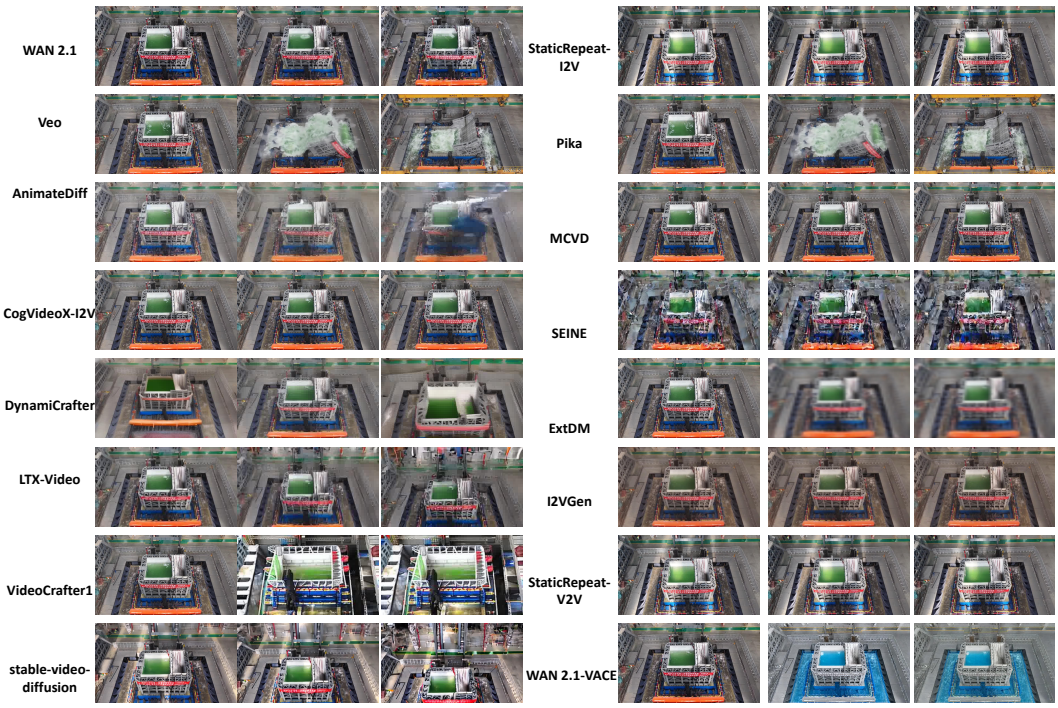


Figure H1: Qualitative Comparison of Baseline Video Generation Models.

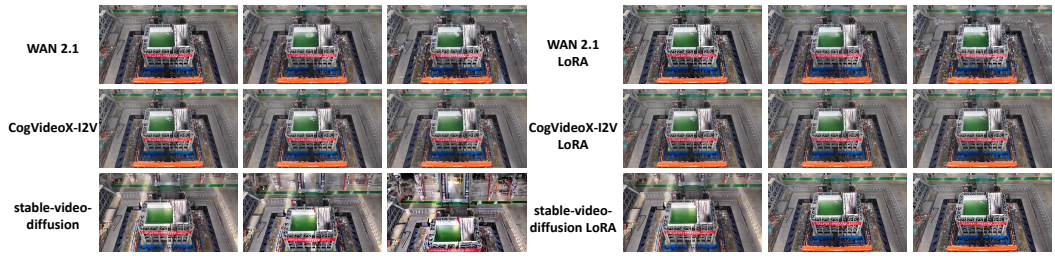


Figure H2: Qualitative Comparison Before and After LoRA Adaptation.

H.2 Qualitative Results of LoRA Adaptation

Figure H2 compares representative outputs before and after LoRA-based domain adaptation for selected video-generation backbones. The adapted models are evaluated using the same initial experimental observations and seismic-generation protocol as their corresponding base models, allowing the effect of domain adaptation to be examined directly. Qualitatively, LoRA adaptation improves the preservation of the gravity-dam experimental scene and reduces severe structural or background distortions in several cases, while maintaining the overall visual identity of the original experiment. These comparisons indicate that parameter-efficient adaptation can help general-purpose video priors better accommodate the visual characteristics of large-scale shaking-table experiments, although the generated dynamics must still be assessed quantitatively using the response-oriented metrics introduced earlier.

LoRA configuration. The principal LoRA configurations used for the three adapted backbones are summarized in Table H1.

Base versus LoRA comparison. For each adapted backbone, the effect of LoRA is evaluated by comparing the base model and its adapted counterpart using the same response-oriented evaluation protocol. For metric m, the score difference is defined as

Table H1: LoRA adaptation configurations for representative I2V models on DamEQ-Bench.

Backbone	Rank	α	Learning Rate	Training Steps	Target Modules	Adaptation Type
WAN2.1	16	16	1×10^{-4}	1000	`q, k, v, o`	LoRA
CogVideoX-I2V	32	32	1×10^{-4}	1500	`to_q, to_k, to_v, to_out.0`	LoRA
Stable-Video	16	16	1×10^{-4}	1000	`to_q, to_k, to_v, to_out.0`	LoRA

Table H2: Paired comparison between the base and LoRA-adapted I2V models on the eight response-oriented metrics.

Backbone	Setting	D-Structure	SAM-Structure	DTW-Sim	DEdge-Pixel	ODE-Response	Peak-Response	AF-Profile	MAC-Modal
WAN2.1	Base	0.1182	0.9109	0.6996	0.6423	0.4688	0.5649	0.3274	0.5718
	LoRA	0.1183	0.9048	0.6848	0.6490	0.4485	0.5745	0.3312	0.5842
	Δ	+0.0001	-0.0061	-0.0148	+0.0067	-0.0203	+0.0096	+0.0038	+0.0124
CogVideoX-I2V	Base	0.1278	0.4704	0.6163	0.4681	0.3270	0.3474	0.5567	0.5715
	LoRA	0.1222	0.4161	0.6131	0.5094	0.3314	0.3584	0.4235	0.5203
	Δ	-0.0056	-0.0543	-0.0032	+0.0413	+0.0044	+0.0110	-0.1332	-0.0512
Stable-Video	Base	0.1200	0.6110	0.6450	0.5894	0.4414	0.5282	0.3525	0.5134
	LoRA	0.1228	0.4894	0.6921	0.4038	0.4696	0.5652	0.4800	0.4705
	Δ	+0.0028	-0.1216	+0.0471	-0.1856	+0.0282	+0.0370	+0.1275	-0.0429

Notes. Δ denotes the difference between the LoRA-adapted and base models, computed as LoRA minus Base. All normalized scores are higher-is-better.

$$\Delta S_m = S_m^{\mathrm{LoRA}} - S_m^{\mathrm{Base}}, \tag{196}$$

where a positive value indicates a higher normalized similarity score after adaptation.

Figure H3 further illustrates the metric-wise changes before and after LoRA adaptation across the three representative I2V backbones, highlighting the different effects of adaptation on physical consistency and engineering response metrics.

The paired results indicate that the effect of LoRA adaptation is metric-dependent. WAN2.1 shows relatively small changes across most response dimensions, whereas CogVideoX-I2V exhibits improved edge-motion, dynamic-response, and peak/crest consistency but reduced modal and acceleration-amplification consistency. Stable-Video shows improvements in DTW-Sim, Neural ODE Dynamic Consistency, Peak/Crest Response Consistency, and Acceleration Amplification Consistency, accompanied by decreases in several structure-oriented metrics. These results suggest that domain adaptation does not uniformly improve all aspects of seismic-response generation.

H.3 Quantitative Results on the Evaluated Subset

The quantitative evaluation comprises six general-purpose visual-quality metrics and eight response-oriented metrics. The two metric groups are reported separately to distinguish perceptual video quality from structural and dynamic response consistency. The current tables correspond to the formally evaluated subset available in the benchmark result files rather than the complete set of requested generations.

Visual-quality results. The six visual dimensions are Subject Consistency, VBench Aesthetic Quality (VB-Aes), Custom Aesthetic Presentation Quality (Aes-Custom), Temporal Flickering, Imaging Quality, and Scene Coherence.

Response-oriented results. The eight response-oriented dimensions are D-Structure, DEdge-Pixel, SAM-Structure, DTW-Sim, Multi-Order Modal Consistency, Dynamic Response Consistency, Peak/Crest Response Consistency, and Acceleration Amplification Consistency.

H.4 Metadata Verification

All benchmark metadata were cross-checked against the available experimental execution records before dataset construction and evaluation. The verified fields include experimental stage, original condition identifier, ground-motion record, target and measured PGA, excitation direction, component ratio, input-frequency annotation, and viewpoint. Experimental loading records and source

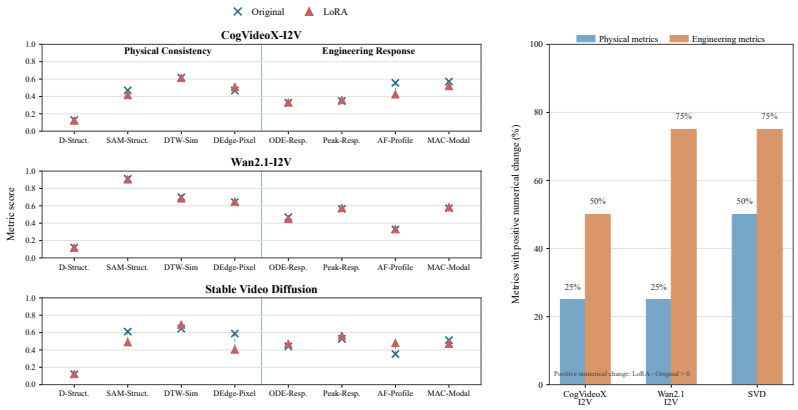


Figure H3: Metric-wise Comparison Before and After LoRA Adaptation.

Table H3: Visual-quality results on the currently evaluated DamEQ-Bench subset.

Task	Model	Subject Consistency	VBench Aesthetic Quality	Custom Aesthetic Presentation	Temporal Flickering	Imaging Quality	Scene Coherence
I2V	AnimateDiff-SparseCtrl	0.8153	0.6759	0.6409	0.9153	0.7884	0.3817
I2V	CogVideoX-I2V	0.6442	0.6708	0.6254	0.8261	0.7127	0.4405
I2V	CogVideoX-I2V-LoRA	0.6276	0.6788	0.6615	0.8371	0.7068	0.4372
I2V	DynamiCrafter	0.5782	0.6018	0.6159	0.8335	0.8131	0.3626
I2V	I2VGen-XL	0.5208	0.4921	0.6019	0.8248	0.7630	0.3458
I2V	LTX-Video-2B-distilled	0.7670	0.6494	0.6337	0.9135	0.8107	0.3807
I2V	VideoCrafter1-I2V	0.7554	0.7285	0.6844	0.8726	0.8287	0.4012
I2V	WAN2.1	0.8562	0.6746	0.6505	0.9405	0.7864	0.3973
I2V	WAN2.1-LoRA	0.8404	0.6824	0.6584	0.9343	0.7815	0.4022
I2V	Stable-Video	0.6661	0.6298	0.6244	0.8610	0.8113	0.3700
I2V	Stable-Video-LoRA	0.6054	0.7014	0.6903	0.7866	0.6752	0.4497
I2V	StaticRepeat-I2V	0.9973	0.6767	0.6491	0.9992	0.7749	0.3961
I2V	Kling†	0.8865	0.7332	0.6568	0.9537	0.7614	0.4100
I2V	Seedance†	0.9349	0.7153	0.6513	0.9753	0.7618	0.4041
I2V	Veo†	0.7599	0.7353	0.6589	0.9177	0.7632	0.4099
V2V	ExtDM	0.8180	0.5583	0.5894	0.9292	0.7753	0.3605
V2V	Pika†	0.9749	0.7255	0.6885	0.9896	0.8213	0.4140
V2V	MCVD	0.7153	0.6014	0.6437	0.7922	0.8327	0.3679
V2V	SEINE	0.5649	0.6850	0.6727	0.7551	0.8091	0.4048
V2V	WAN2.1-VACE-1.3B	0.8725	0.6894	0.6742	0.9405	0.7648	0.3987
V2V	StaticRepeat-V2V	0.9900	0.6746	0.6504	0.9964	0.7757	0.3962

Notes. All normalized scores are higher-is-better. Models marked with † are commercial models.

acquisition files were treated as primary evidence, while filenames were used only for organization and were not regarded as physical ground truth. Missing, unknown, zero, and not-applicable values were retained as distinct states, and conflicts among metadata sources were preserved for review rather than silently overwritten. Normalized filenames were generated only after metadata verification, and the released metadata retain the original identifiers and provenance needed to trace subsequent corrections.

H.5 Limitations of the Scaled Physical Model

DamEQ-Bench is constructed from a controlled 1:50 physical model of a gravity dam. Although the large-scale shaking-table experiment provides repeatable excitation, synchronized observations, and physically grounded structural responses, the scaled specimen should not be interpreted as a complete physical replica of a prototype-scale dam. The adopted similitude framework preserves the dominant scaling relations relevant to the experiment, but not all multi-physics effects can be matched simultaneously.

Table H4: Response-oriented results on the currently evaluated DamEQ-Bench subset.

Task	Model	D-Structure	DEdge-Pixel	SAM-Structure	DTW-Sim	Multi-Order Modal Consist.	Neural ODE Dynamic Consist.	Peak/Crest Response Consist.	Acceleration Amplification Consist.
I2V	AnimateDiff-SparseCtrl	0.1230	0.6213	0.7880	0.6889	0.4730	0.5583	0.5554	0.3695
I2V	CogVideoX-I2V	0.1278	0.4681	0.4704	0.6163	0.5715	0.3270	0.3474	0.5567
I2V	CogVideoX-I2V-LoRA	0.1222	0.5094	0.4161	0.6131	0.5203	0.3314	0.3584	0.4235
I2V	DynamiCrafter	0.1206	0.6201	0.5622	0.6428	0.5378	0.3804	0.4961	0.4031
I2V	I2VGen-XL	0.1289	0.5514	0.4907	0.5545	0.4560	0.4438	0.3975	0.3874
I2V	LTX-Video-2B-distilled	0.1246	0.6386	0.7919	0.7143	0.5850	0.4724	0.5959	0.3577
I2V	VideoCrafter1-I2V	0.1262	0.6012	0.5822	0.6690	0.4865	0.5879	0.6038	0.3695
I2V	WAN2.1	0.1182	0.6423	0.9109	0.6996	0.5718	0.4688	0.5649	0.3274
I2V	WAN2.1-I2V-LoRA	0.1183	0.6490	0.9048	0.6848	0.5842	0.4485	0.5745	0.3312
I2V	Stable-Video	0.1200	0.5894	0.6110	0.6450	0.5134	0.4414	0.5282	0.3525
I2V	Stable-Video-LoRA	0.1228	0.4038	0.4894	0.6921	0.4705	0.4696	0.5652	0.4800
I2V	StaticRepeat-I2V	0.1321	0.6547	0.9673	0.6948	0.5266	0.5039	0.5299	0.2442
I2V	Kling†	0.1338	0.6564	0.5092	0.6679	0.7616	0.2470	0.4298	0.3352
V2V	Pika†	0.0978	0.8225	0.5337	0.7467	0.7001	0.5887	0.6901	0.6275
I2V	Seedance†	0.1797	0.5164	0.5094	0.7437	0.6218	0.5019	0.5349	0.3853
I2V	Veo†	0.1147	0.5147	0.4968	0.6159	0.5654	0.4494	0.4329	0.3049
V2V	ExtDM	0.1248	0.6342	0.9279	0.6819	0.4513	0.4109	0.4873	0.3356
V2V	MCVD	0.1238	0.4842	0.6940	0.7698	0.4242	0.3874	0.4516	0.4883
V2V	SEINE	0.1289	0.3849	0.5057	0.7205	0.4973	0.4296	0.4890	0.4584
V2V	WAN2.1-VACE-1.3B	0.1134	0.6375	0.9108	0.6774	0.5812	0.5318	0.5661	0.2229
V2V	StaticRepeat-V2V	0.1342	0.7217	0.9691	0.6934	0.5960	0.5217	0.5367	0.2254

Notes. The eight dimensions follow the metric terminology used in the benchmark overview. All normalized scores are higher-is-better. "–" denotes an unavailable result. Models marked with † are commercial models.

Scale-dependent differences may remain in material constitutive behavior, fracture and damage evolution, damping, fluid–structure interaction, and boundary or foundation effects. Consequently, visible cracking, local deformation, damping characteristics, and reservoir-induced responses in the experiment should be interpreted as properties of the controlled experimental system rather than direct quantitative predictions of the prototype structure.

The post-repair stage is likewise used as a distinct experimentally observed structural condition for benchmarking state-dependent video generation. The SMA-FRC repair interface and its local mechanical behavior are model-scale realizations and should not be interpreted as direct evidence of prototype-scale repair efficiency. Therefore, DamEQ-Bench primarily evaluates whether generation models reproduce the experimentally observed seismic-response domain under different structural and reservoir states, rather than whether they provide full-scale structural-response prediction.

H.6 Limitations of Sensor-Based References

The current DamEQ-Bench evaluation primarily uses acceleration measurements as the sensor-based physical reference. Although strain gauges and dynamic water-pressure sensors were also deployed in the experiment, these modalities are not yet included in the benchmark evaluation. Consequently, the current response-oriented metrics should not be interpreted as direct validation of strain fields, stress states, hydrodynamic pressures, or complete structural damage states.

Video-derived acceleration is obtained from temporal differentiation of estimated displacement trajectories and is therefore sensitive to tracking noise, residual camera motion, image quality, frame rate, and the common trustworthy frequency band. Temporal interpolation cannot recover physical high-frequency content that was not reliably captured in the original video. In addition, accelerometers provide only sparse measurements at instrumented locations; agreement at these locations does not imply agreement over all structural degrees of freedom.

Accordingly, the sensor measurements in DamEQ-Bench are treated as localized physical references for selected observable response quantities rather than complete measurements of the dam state. Incorporating strain and hydrodynamic-pressure measurements into the multimodal evaluation framework remains an important direction for future extensions.

H.7 Generalization Scope

DamEQ-Bench contains diverse ground motions, PGA levels, excitation directions, reservoir states, and viewpoints, but these variations are primarily obtained from repeated testing of the same scaled gravity-dam specimen. Therefore, the benchmark mainly evaluates *within-system condition gener-*

alization, rather than generalization across independent dam geometries, material systems, foundations, or experimental facilities.

Performance on held-out ground motions or unseen combinations of PGA, direction, reservoir state, and viewpoint should consequently be interpreted only within the explicitly defined DamEQ-Bench holdout protocol. Such results do not establish structure-independent generalization to arbitrary gravity dams or prototype-scale field environments.

The experimental stages are also history dependent because they arise sequentially from cumulative loading, reservoir-state changes, damage evolution, and repair. Differences among dry, impounded, and post-repair stages therefore cannot always be attributed to a single isolated factor. Matched-condition comparisons can reduce some confounding effects, but do not provide complete causal identification.

Future extensions to multiple specimens, dam geometries, material systems, scale ratios, reservoir configurations, foundations, and experimental facilities will be required to evaluate true cross-structure generalization. The current DamEQ-Bench should therefore be regarded as a controlled benchmark for condition-aware seismic video generation within a well-defined experimental domain.

H.8 Experimental Cost, Data Scarcity, and Resource Reuse

DamEQ-Bench is a large-scale physical experimentation benchmark rather than an internet-scale video dataset. Each recording is generated from controlled shaking-table experiments using a physical gravity-dam specimen, prescribed seismic excitation, reservoir-state control, multi-view UAV acquisition, and extensive instrumentation. Documented costs include approximately RMB 0.85 million for the custom excitation/model-box system and RMB 0.60 million for construction and implementation of the gravity-dam experimental system, exceeding RMB 1.44 million in total. The experiments further involved seven UAV systems, dedicated batteries and storage, approximately six days of large shaking-table operation, and substantial lifting, instrumentation, reservoir operation, technical, storage, and computational resources. Since some resources were provided through existing institutional infrastructure, these figures distinguish documented expenditure from replacement-value estimates intended to characterize the scale of the physical experimentation rather than audited cash expenditure.

DamEQ-Bench therefore emphasizes **information density grounded in physical experiments** rather than simply maximizing the number of videos. Each usable sample corresponds to a clearly defined loading condition and is associated with structured information including seismic input, PGA, excitation direction, reservoir and structural state, viewpoint, loading history, and sensor-supported response. For large-scale seismic experiments, dataset size cannot be increased indefinitely through repeated recording, because strong excitation progressively alters the physical state of the specimen through reservoir interaction, cumulative damage, and repair. The objective of DamEQ-Bench is therefore not to maximize sample count, but to provide traceable, interpretable, and reusable physical observations for the standardized evaluation of seismic video-generation models.

Another important value of DamEQ-Bench lies in the reuse of costly physical experimental evidence. New generative models can be compared against the same experimental references without repeating a large-scale shaking-table campaign for every algorithmic study. More broadly, the significance of DamEQ-Bench extends beyond video generation itself: it supports the modeling, visualization, and understanding of gravity-dam seismic responses, and provides a foundation for future research on structural monitoring, seismic-response assessment, damage identification, and intelligent infrastructure management. Its ultimate engineering objective is to deepen our understanding of gravity-dam behavior under earthquakes and to support the development of safer and more resilient critical infrastructure.

Table H5: Representative cost and resource requirements associated with the construction of DamEQ-Bench.

Cost category	**Cost description**	**Approx. cost**
Custom excitation/model-box system	Procurement of the high-frequency excitation model-box system	RMB 852,100
Gravity-dam experimental construction	Concrete and barite materials, formwork, absorbing layers, temporary works, repair, pumping, demolition, wiring, labor, and tax	RMB 596,905
Seven UAV systems	Seven DJI Mavic 4 Pro 512GB Creator configurations, with three batteries per platform	RMB 167,916
External video storage media	Fourteen 512-GB memory cards, two per UAV	RMB 11,200
UAV auxiliary equipment	Card readers, spare propellers, cables, transport cases, and related accessories	RMB 10,000
Large shaking-table utilization	Six experimental days, including facility occupation and operation of hydraulic, control, and cooling infrastructure	RMB 300,000
Shaking-table energy and cooling	Electricity and cooling demand over the six-day experimental campaign	RMB 60,000
Lifting and rigging support	Installation, repositioning, and removal of the dam/model-box system	RMB 60,000
Sensor and data-acquisition infrastructure	Accelerometers, strain measurement, dynamic pressure sensing, and multi-channel data acquisition equipment	RMB 1,200,000
Calibration and metrology	Sensor calibration, channel verification, and measurement-quality checks	RMB 50,000
Reservoir operation	Water filling, drainage, pumps, hoses, and associated utilities	RMB 20,000
Technical personnel	Facility operators, instrumentation engineers, UAV operators, safety personnel, and experimental support	RMB 180,000
Safety and consumables	Protective equipment, temporary barriers, minor materials, and expendables	RMB 20,000
Data storage and backup	High-capacity storage and backup for raw, processed, and generated video data	RMB 35,000
Computational evaluation	GPU inference, commercial-model/API evaluation, metric computation, and data-processing resources	RMB 100,000
Facility coordination and logistics	Experiment scheduling, preparation, internal logistics, and supporting operations	RMB 50,000
Additional transport	Material transport, loading/unloading, and urgent logistics	RMB 30,000
Overall resource envelope		**∼RMB 3,753,000**
Equivalent in USD		**∼$521,250**