WORLD EMBEDDING BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical fidelity has received increasing attention in world models and video generation, yet how video representations encode physical information remains less understood. We introduce the World Embedding Benchmark, comprising 8,000 controlled simulation cases from 80 families spanning fluid mechanics, solid mechanics, dynamics, and optics & electromagnetism. Each case pairs a rendered video with simulation-derived physical annotations, supporting three complementary tasks: text-video retrieval, physical-property regression, and multiple-choice video-description pair classification. We use these tasks to distinguish cross-modal physical alignment from the recoverability of quantitative physical information. Evaluated pre-trained omnimodal embedding models show weak retrieval and near-chance within-family pair classification, while lightweight probes recover useful physical information from frozen video embeddings. Continual contrastive training with physics-specific video-text pairs improves retrieval and pair classification but degrades physical-property regression, revealing a trade-off between alignment and quantitative information recoverability. Finally, we use the embeddings to retrieve reference videos for retrieval-augmented generation with MiniMax-H3. Retrieved references improve the physical fidelity of generated videos, with stronger retrieval models yielding larger gains in our experiments. Together, these findings highlight the need to evaluate physical alignment and property recoverability jointly, and demonstrate the utility of physical representations for improving video generation.

1 INTRODUCTION

Physical fidelity is essential to the ambition of using video generative models as world models: a useful model of the physical world must capture how systems behave under different conditions, beyond producing visually plausible scenes (Bansal et al., 2025; Meng et al., 2025; Kang et al., 2025; Li et al., 2026a). Recent evaluations have made this challenge explicit, revealing substantial gaps between visual realism and adherence to physical commonsense in generated videos (Bansal et al., 2025; Meng et al., 2025; Zheng et al., 2025). This growing attention to physical fidelity raises a critical question about representation learning: *what information about the physical world is preserved in video embeddings, and how accessible is that information?* A representation may identify a scene as fluid flow or elastic deformation yet fail to distinguish changes in viscosity or stiffness. Such distinctions matter when embeddings are used to retrieve physical examples, estimate system properties, or provide reference evidence for generation (Peruzzo et al., 2025; Cheng et al., 2026).

Evaluating physical representations requires separating two capabilities: encoding information about physical quantities and aligning that information with language. A physical property may be recoverable by a trained probe even when embedding similarity cannot match a video to a description of that property. Conversely, successful matching of broad scene categories need not imply sensitivity to the parameters governing their behavior. Existing scientific simulation benchmarks provide rich resources for learning physical dynamics (Takamoto et al., 2022; Luo et al., 2024; Ohana et al., 2024a), but evaluating multimodal embeddings calls for paired video-language observations and quantitative physical targets within a shared, controlled setting. This motivates a benchmark that measures both cross-modal alignment and the recoverability of physical information in video representations.

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.

We introduce the **World Embedding Benchmark**, a comprehensive benchmark for evaluating physical information in video embeddings. It comprises 8,000 simulation cases from 80 families spanning fluid mechanics, solid mechanics, dynamics, and optics & electromagnetism. Each family contains 100 instances of a physical system with different parameter settings. We combine branch-specific simulation engines with analytical formulations and numerical integration where appropriate. These controlled configurations provide rendered videos together with their generating parameters and computed physical quantities, enabling evaluation beyond coarse scene semantics.

The benchmark supports three complementary tasks. **Text-video retrieval** measures cross-modal matching between physical descriptions and simulation videos. **Physical-property regression** measures whether lightweight probes can recover continuous physical quantities from frozen video embeddings. **Pair classification** asks models to select the correct video-description pair from a small set of candidates. Together, these tasks assess cross-modal physical alignment, quantitative information recoverability, and fine-grained physical discrimination. Using this benchmark, we investigate three research questions.

RQ1: Do current representation models encode physics in video representations? We evaluate pre-trained omnimodal embedding models and find a marked discrepancy across tasks: text-video retrieval and within-family pair classification remain close to their respective chance baselines, whereas lightweight regression probes recover useful physical information from frozen video representations. These results suggest that the evaluated embeddings retain signals predictive of physical properties, but that these signals are not readily accessible through cross-modal similarity. Directly prompted MLLMs perform substantially better on pair classification, suggesting that relevant physical capabilities exist but are not fully exposed in embedding space.

RQ2: Can contrastive adaptation improve physical alignment? We investigate whether adding physics-specific video-text pairs to continual contrastive training improves the physical alignment of pre-trained embeddings. Adaptation improves retrieval and pair classification, but reduces physical-property regression. This trade-off suggests that organizing embeddings for cross-modal matching can compromise the recoverability of quantitative physical information. Overall, these results suggest that contrastive training primarily reorganizes and aligns existing physical signals in the representation space, rather than substantially enhancing the underlying physical capabilities.

RQ3: Can physical representations improve video generation? We use omnimodal embedding models as retrievers to provide physically relevant reference videos for retrieval-augmented generation with MiniMax-H3 (MiniMax, 2026). Retrieved references improve the physical fidelity of generated videos, and embeddings with stronger physics retrieval performance yield larger improvements in our experiments. The results suggest that improving physical representations offers a practical route to supporting more faithful generation of physical scenes.

Our contributions are three-fold: (1) We introduce the World Embedding Benchmark, comprising 8,000 controlled simulation cases across 80 physical families and three complementary tasks for evaluating physical representations. (2) We identify a gap between physical information encoded by multimodal models and its cross-modal alignment in embedding space, and show that contrastive adaptation improves alignment while degrading physical-property regression. (3) We demonstrate the downstream utility of physical embeddings for retrieval-augmented video generation, connecting stronger physical retrieval performance to improved physical fidelity of generated videos. More broadly, these results point to physically grounded representations as a useful bridge between understanding physical systems and improving generative world models.

2 World Embedding Benchmark

The World Embedding Benchmark evaluates the extent to which video representations preserve information about physical systems and their underlying parameters. The benchmark is built from 8000 physics simulation cases from 80 families across multiple branches: Fluid Mechanics, Solid Mechanics, Dynamics, and Optics & Electromagnetism. Each family contains 100 cases, representing the same physical system under different parameter settings.

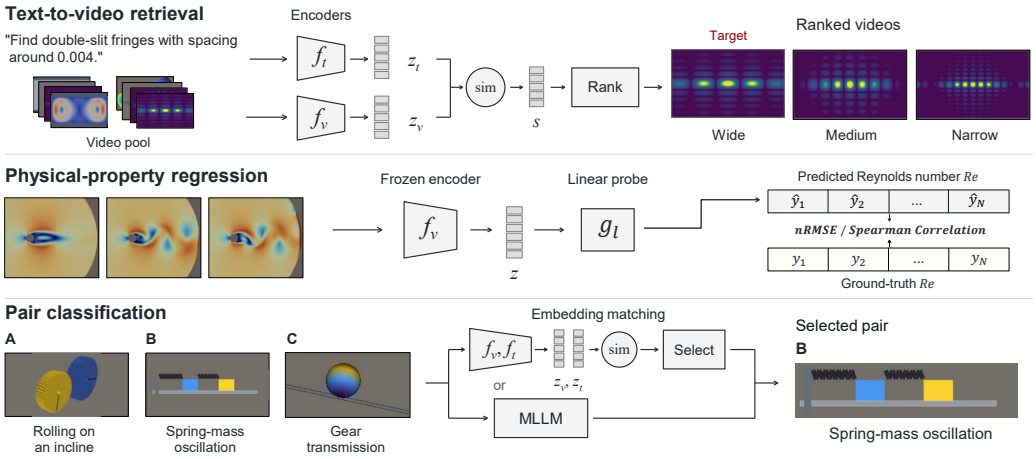


Figure 1: Overview of the three evaluation tasks: text-video retrieval, physical-property regression, and video-description pair classification.

The simulation cases support three complementary evaluation tasks: 1) **Text-video retrieval**, which evaluates the alignment between natural-language descriptions of physical systems and their corresponding video representations. 2) **Physical property regression**, which evaluates whether continuous quantities can be recovered from frozen video representations using lightweight regression models. 3) **Pair classification**, which evaluates whether models can distinguish the correct video-description correspondence from alternative candidates. Figure 1 summarizes the three evaluation tasks, which respectively assess cross-modal physical alignment, quantitative physical information recoverability, and fine-grained physical discrimination.

Table 1: Simulation families, case counts, and computation methods across physics branches.

Branch	Families	Cases	Computation
Fluid Mechanics	7	700	OpenFOAM
Solid Mechanics	27	2,700	DOLFINx
Dynamics	19	1,900	Chrono + Analytical / ODE integration
Optics & EM	27	2,700	Analytical + HCIPy + Meep (FDTD)
Total	80	8,000	–

2.1 Physics Simulation

We construct the physics simulation subset from parameterized, representative scene families across Fluid Mechanics, Solid Mechanics, Dynamics, and Optics & Electromagnetism. Each instance is defined by its governing model, constitutive law, geometry, material parameters, boundary conditions, and initial conditions. Physical responses are generated using OpenFOAM for fluid mechanics (OpenCFD, 2024; Weller et al., 1998), FEniCSx/DOLFINx for solid mechanics (Baratta et al., 2023), Project Chrono for rigid-body and multibody dynamics (Tasora et al., 2016), and HCIPy together with Mitsuba 3 for optics and MEEP for electromagnetics (Por et al., 2018; Jakob et al., 2022; Oskooi et al., 2010). The exported fields and trajectories are rendered into PNG frames with ParaView 6.0.1 (Ahrens et al., 2005) and encoded into MP4 clips with FFmpeg. The rendered clips therefore depict physically faithful responses with quantitative meaning, and generating evaluation data from validated simulators follows a growing line of scientific machine learning benchmarks (Takamoto et al., 2022; Ohana et al., 2024a; Luo et al., 2024). Simulation provides explicit access to the generating parameters and computed physical states, enabling controlled evaluation of physical information in video representations. Representative examples from the four physics branches are shown in Figures 2. Detailed formulations, computation methods, case parameters, and field outputs are provided in Appendix A.

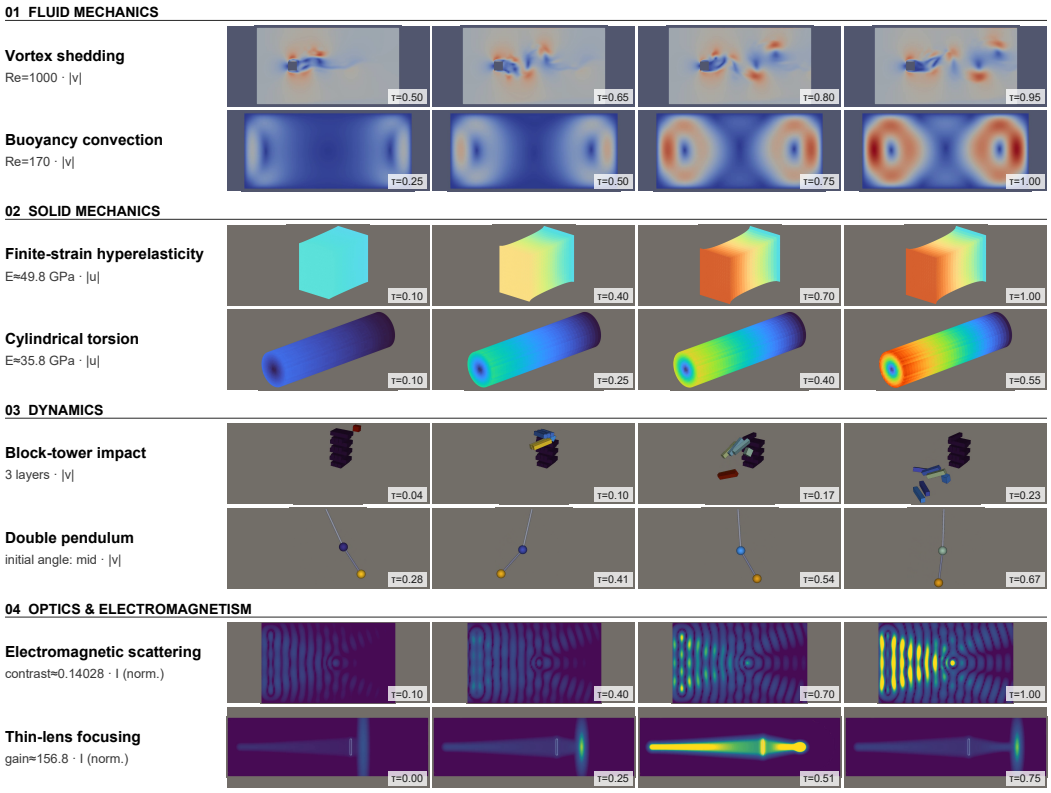


Figure 2: Example video frames of representative scene families of the 4 physics branches, illustrating different stages of each simulation sequence. The full dataset contains videos from 80 families.

Quality control and self-checks To improve data quality and reduce duplication, we apply automated checks during dataset construction. These checks flag failed or non-convergent runs, NaN or infinite values, and unexpected physical responses. Video checks examine file properties and use sampled-frame statistics to identify blank, nearly static, or flickering videos. We also check case IDs and parameter settings for duplicates.

2.1.1 Fluid Mechanics

The fluid dynamics subset contains 700 videos across seven families instantiated from 15 OpenFOAM configurations (OpenCFD, 2024; Weller et al., 1998). The configurations cover internal and external flows, cavity and buoyancy-driven flows, jets, non-Newtonian flow, and passive-scalar transport, including lid-driven cavity and backward-facing step flow (Ghia et al., 1982; Armaly et al., 1983). The case parameters include Reynolds and Péclet numbers, velocity, viscosity, geometry, scalar diffusivity, buoyancy parameters, and rheological parameters. Outputs include velocity, temperature, and transported-scalar fields. Further details are provided in Appendix A.2.1.

2.1.2 Solid Mechanics

The solid mechanics subset contains 2,700 videos across 27 families generated with FEniCSx/DOLFINx (Baratta et al., 2023). The families cover linear elasticity, finite-strain hyperelasticity, geometrically nonlinear deformation, indentation, stress concentration, and history-dependent response under tension, compression, bending, and torsion. The cases vary in material properties, geometry, and loading conditions, and the released videos primarily visualize displacement magnitude on the deformed geometry. Detailed formulations are provided in Appendix A.2.2.

2.1.3 DYNAMICS

The dynamics subset contains 1,900 videos across 19 families combining Project Chrono (Tasora et al., 2016) with Analytical / ODE integration. The families cover rigid-body motion, collision and contact, articulated mechanisms, vibration, wave propagation, rolling and sliding, rotating machinery, and structural response. The case parameters include geometry, mass, gravity, initial conditions, restitution and friction coefficients, stiffness and damping, and excitation conditions. Outputs include trajectories and family-specific quantities such as velocity, displacement, angular speed, stress, equivalent plastic strain, and damage. Detailed formulations are provided in Appendix A.2.3.

2.1.4 OPTICS & ELECTROMAGNETISM

The optics and electromagnetism subset contains 2,700 videos across 27 families. Nineteen optics families cover reflection and refraction, diffraction, interference, focusing, Gaussian-beam propagation, dispersion, and polarization, using analytical relations or HCIPy (Por et al., 2018), with Mitsuba 3 (Jakob et al., 2022) used as an auxiliary backend for selected geometric-optics configurations. The remaining eight families use Meep FDTD (Oskooi et al., 2010) for electromagnetic scattering, waveguides, photonic-crystal structures, gratings, dipole radiation, evanescent fields, and dispersive plasmonic systems. Detailed formulations and field definitions are provided in Appendix A.2.4.

2.2 BENCHMARK FORMULATION

Retrieval We formulate physics-aware retrieval as bidirectional text-video retrieval within each physics branch. We evaluate both text-to-video and video-to-text retrieval over the full branch corpus and report their mean Recall@K. For example, the Solid Mechanics corpus contains 2,700 paired videos and queries from 27 families. For each family, we define a natural-language template describing the physical system, with slots for case-specific physical conditions, parameters, and derived quantities. These slots are instantiated from the simulation metadata, so queries within a family share similar wording but differ in the physical properties and numerical values associated with each case. See Figure 5 for an example of a natural-language template instantiated with metadata values.

Regression To evaluate whether quantitative physical information is recoverable from video representations, we define ten family-specific physical-property regression tasks. Each task associates one simulation family with a continuous scalar target, such as gravity, Reynolds number, wavelength, or elastic modulus. Given frozen video embeddings, we fit a linear regression probe to predict the target values $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_N]$ from video representations alone. We report normalized root mean squared error (nRMSE) as the primary metric and Spearman's rank correlation ρ as a complementary measure of whether the predicted values preserve the ordering of the ground-truth property. Details of the regression targets and evaluation protocol are provided in Appendix B.2.

Pair Classification Pair classification provides a lower-cost, controlled test of whether models can distinguish the correct video-text correspondence from alternatives. We sample 10 anchors per family and pair each with a *within-family* negative and a *cross-family* negative, yielding two binary settings. The former tests fine-grained sensitivity to physical properties and numerical values, while the latter tests discrimination across physical systems. For embedding models, we compare the anchor-positive similarity against the corresponding anchor-negative similarity. For MLLMs, we directly prompt the model to answer each binary multiple-choice question, rather than evaluating its embeddings. We report accuracy for both settings.

3 EXPERIMENTS

3.1 MAIN RESULTS

We evaluate three classes of models with different representation paradigms. **Embedding models** include Video-CLIP-XL (Wang et al., 2024), Qwen3-VL-Embedding (2B and 8B) (Li et al., 2026b), Omni-Embed-Nemotron 3B (Xu et al., 2025), and LCO-Embedding-Omni (3B and 7B) (Xiao et al., 2026). These models provide both video and text embeddings and are therefore evaluated on all

Table 2: Performance on the World Embedding Benchmark. Retrieval is evaluated globally within each physics branch. Regression reports the macro-average nRMSE across 10 physical-property regression tasks. Pair classification reports within-family and cross-family accuracy. **Bold** denotes best across all models, underline denotes second-best; * denotes best in embedding models.

Model	Modality	Retrieval R@10 ↑				Regression nRMSE ↓	Pair Classification Acc. ↑	
		Fluid	Solid	Dynamics	Optics & EM	Avg. (×100)	Within-Fam.	Cross-Fam.
Embedding Models								
Video-CLIP-XL	T, V	3.3	1.4	2.3	0.6	**0.81**	48.1	66.1
Qwen3-VL-Embedding-2B	T, I, V	3.9	1.6	7.3	2.3	1.76	51.5	80.8
Qwen3-VL-Embedding-8B	T, I, V	6.6	1.9	7.4	3.1	1.80	53.1	86.2
Omni-Embed-Nemotron-3B	T, I, V, A	3.1	1.4	3.1	0.7	1.41	52.5	71.2
LCO-Embedding-3B	T, I, V, A	3.9	1.5	5.1	0.7	2.47	51.5	68.5
LCO-Embedding-3B-2605	T, I, V, A	4.6	1.5	4.5	1.2	1.85	50.5	72.9
LCO-Embedding-7B	T, I, V, A	4.9	1.6	5.7	1.9	2.02	46.8	76.4
Physics-adapted Embedding Models								
LCO-Embedding-3B + Physics Adaptation	T, I, V, A	**18.6**	8.0	16.5	8.9	3.95	53.5	98.2
LCO-Embedding-7B + Physics Adaptation	T, I, V, A	17.9	**9.4**	**21.0**	**9.4**	2.89	59.6*	**98.4**
JEPA Models								
V-JEPA2 ViT-L	V	—	—	—	—	0.94	—	—
V-JEPA2 ViT-g	V	—	—	—	—	1.02	—	—
MLLMs								
Qwen3.5-27B	T, I, V	—	—	—	—	—	**61.3**	92.0
Qwen3.5-35B-A3B	T, I, V	—	—	—	—	—	58.8	91.2
Qwen3.6-27B	T, I, V	—	—	—	—	—	60.2	91.9
Qwen3.6-35B-A3B	T, I, V	—	—	—	—	—	59.0	90.9

three benchmark tasks. **JEPA models** are represented by V-JEPA2 (Assran et al., 2025). As V-JEPA2 produces video representations without a corresponding text embedding space, we evaluate it only on physical-property regression using frozen video features. **MLLMs** include Qwen3.5 (Qwen Team, 2026a) and Qwen3.6 (Qwen Team, 2026c;b), each with 27B and 35B-A3B variants. Rather than extracting embeddings, we directly prompt these models to answer the pair-classification questions, providing a generative-model reference for physical video-text discrimination.

The main results are summarized in Table 2. We make three observations. First, current pre-trained embedding models remain weak at physics-aware cross-modal alignment: retrieval performance stays low across all four physics branches, and within-family pair classification remains close to random chance, although cross-family discrimination is considerably easier. Second, directly prompted MLLMs perform substantially better on pair classification, reaching around 91–92% cross-family accuracy and 59–61% within-family accuracy. This suggests that stronger multimodal generative models already possess meaningful capabilities for understanding physical video-text correspondences, whereas current embedding models do not expose comparable capabilities through similarity matching, motivating us to study whether such capabilities can be better activated for representation learning. Third, V-JEPA2, despite being trained without text alignment, achieves strong physical-property regression performance with frozen video features. This suggests that predictive video representation learning itself can preserve substantial quantitative physical information, highlighting a distinction between modeling physical information in the representation and aligning it with language for similarity matching.

3.2 Enhancement of Physics Capabilities

We next ask whether physics adaptation can improve embedding model performance on our benchmark. Existing embedding models are rarely trained explicitly to capture physical information. Among the two dominant paradigms for multimodal retrieval, prior work has identified important mechanistic differences between CLIP and MLLM-based embedding models (Xiao et al., 2025; 2026). First, CLIP models exhibit a scaling plateau on complex tasks requiring spatial and compositional cross-modal alignment, where simply scaling model size or data provides limited gains. Second, MLLMs acquire richer multimodal knowledge through generative pretraining, as the LLM decoder learns to leverage multimodal information in a shared representation space; lightweight contrastive learning can then activate this latent alignment for similarity matching. Third, a generation-representation scaling pattern emerges: stronger generative models tend to provide higher representation upper bounds after contrastive activation.

These observations motivate us to focus on the MLLM-based paradigm and ask whether physics capabilities already present in generative models can be released through physics-aware contrastive learning. Our setup therefore calls for a backbone with strong world knowledge, a validated contrastive activation recipe that permits injecting physics-aware examples while retaining general-domain data, and omnimodal support for video and audio, which also enables downstream retrieval-augmented video generation.

We adopt the training recipe of LCO-Embedding (Xiao et al., 2026), an omnimodal embedding model trained with lightweight contrastive activation that achieves competitive performance across text, image, video, and audio embedding benchmarks.

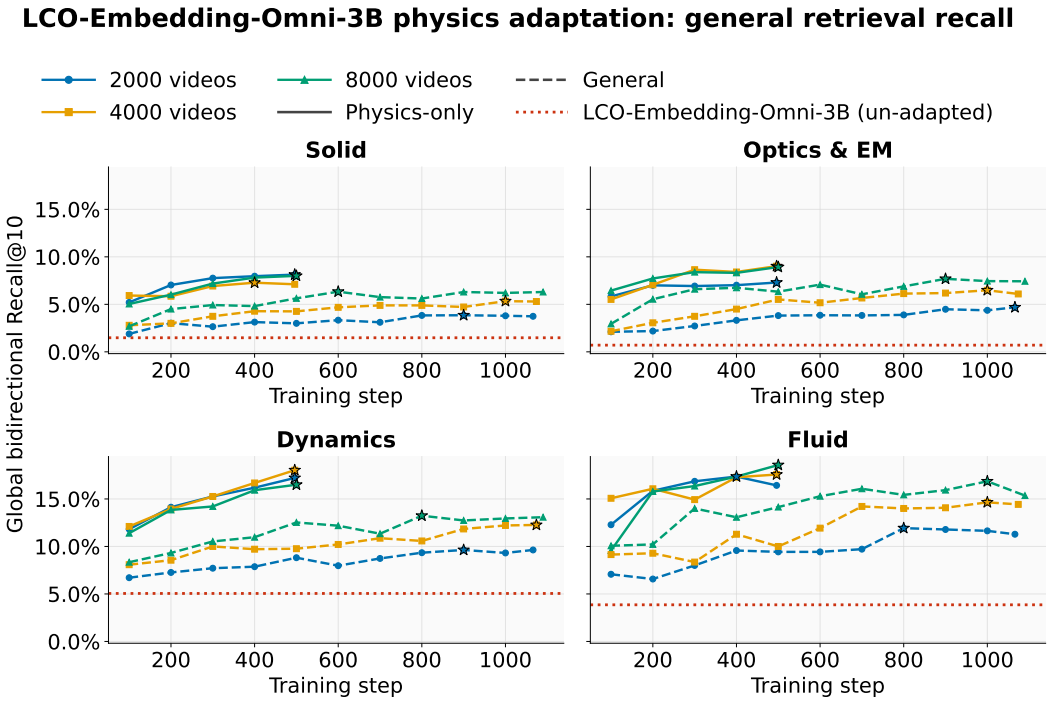


Figure 3: Retrieval performance (y-axis) across training steps (x-axis) and training dataset settings (different lines).

Experiment Setups We construct the training set using the same physics simulation engines while ensuring that physics quantities take different numerical values from those in the benchmark. In total, we generate 8,000 text–video pairs by simulating 100 instantiations for each of the 80 physics families. Videos are rendered from the simulation results, with captions expressing the corresponding physics quantities in natural language following the benchmark templates. For each anchor, we sample one hard negative from the same scene family.

We conduct two training settings. **Physics-only Training** uses only physics video-text pairs, with nested subsets of $\{2k, 4k, 8k\}$ examples. Given the small dataset size and high visual similarity within physics families, we use a batch size of 64 and train for $\{16, 8, 4\}$ epochs, respectively, to keep the number of optimization steps approximately constant. **General Training** mixes the same physics subsets with 271k general video-text pairs from the `LCO-Embedding-Omni-3B-2605` collection, yielding $\{273k, 275k, 279k\}$ examples. We initially hypothesize that general-domain data can stabilize training and prevent capability collapse, although later results provide evidence against this hypothesis. We use a batch size of 512, and following Xiao et al. (2026), we train all three settings for 2 epochs; their total training steps differ only slightly because physics data constitute a small fraction of the mixture.

We initialize from `LCO-Embedding-Omni-3B` and `LCO-Embedding-Omni-7B`. We find that initializing directly from `Qwen2.5-Omni` (which is itself the base model of LCO-Embedding)

leads to frequent loss spikes when physics data are introduced. We attribute this to the anisotropic representation space of the generative model, whereas LCO-Embedding has already been calibrated for similarity-based representation learning.

Results and Analyses Figure 3 summarizes retrieval performance of models across training settings and steps. We make two observations: First, contrastive learning substantially improves physics-aware video retrieval for MLLM-based embedding models. Although recall@10 remains well below that of general-domain video-text retrieval (Assadi et al., 2026), it clearly surpasses the near-chance performance of the base model in Table 2. Second, physics-aware retrieval requires prolonged optimization. Retrieval performance continues to improve throughout training and shows no signs of converging within our allocated training resources. For the physics-only training with 2000 examples, this results in an unusually high number of 16 epochs. Moreover, mixing general-domain video-text data slows down the training. Both behaviors contrast with general-domain MLLM-based embedding model training, where ≤ 2 epochs is typically sufficient (Xiao et al., 2026).

Figure 4 compares checkpoints trained from 3B and 7B backbones to test whether physics retrieval scales with backbone capability. Consistent with Xiao et al. (2026; 2025), stronger generative backbones yield stronger retrieval performance after identical contrastive training, supporting the generation-representation scaling law in physics representation learning. Figure 8 displays the retrieval performance of 7B models throughout the physics adaptation training.

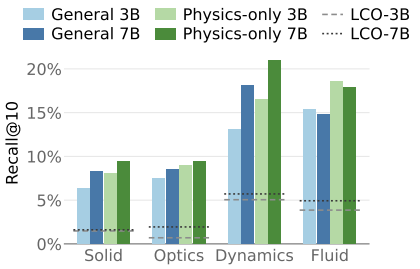


Figure 4: Retrieval performance of 3B and 7B backbones after physics adaptation across the four physics branches.

Regression performance in Figure 6, however, calls for a more nuanced interpretation. Across the 10 regression tasks, models continually trained on in-domain physics data mostly, if not consistently, underperform `LCO-Embedding-Omni`. We provide two possible interpretations. First, this observation is broadly consistent with Xiao et al. (2026), which argues that contrastive learning primarily activates and calibrates MLLM capabilities rather than enhancing the underlying capabilities themselves. While in-domain physics training improves retrieval by better calibrating the similarity-matching space, a linear regression probe instead measures whether specific information is encoded and linearly accessible in the embeddings. As a side effect of this representation calibration, the model's ability to encode certain physics properties may therefore degrade. Second, physics-property regression may benefit from training objectives beyond contrastive learning, such as explicit representation regression losses, similar to objectives used to optimize semantic textual similarity in text embedding models.

3.3 Retrieval-Augmented Video Generation

Having attained stronger video-text alignment in the physics representation space, we analyze whether such physics-aware representation can enhance video generation. We consider retrieval-augmented video generation. We equip MiniMax-H3 (MiniMax, 2026), a state-of-the-art open-weight video generation model, with multimodal embedding models serving as retrievers. We leverage the external video database built in (Cheng et al., 2026), and use the `ref2va` version of Minimax-H3 (MiniMax, 2026), which inherently has certain capabilities of grounding on external contexts.

We benchmark the performance on PhyGenBench (Meng et al., 2025), and take a closer look at two physics branches related to World Embedding Benchmark. The results are displayed in Table 3. As a frontier open-weight model, MiniMax-H3 without RAG already achieves strong PhyGenBench performance. When forced to refer to potentially suboptimal videos, average performance slightly degrades from the baseline, shown in the VideoCLIP-XL (Wang et al., 2024) column. The original LCO-Embedding-Omni-3B (Xiao et al., 2026) enhances the performance over the no-RAG baseline. The LCO variant adapted for physics-aware video retrieval in this work further improves performance, with the *general + physics* variant attaining the best average score of 0.66 and the physics-only variant reaching 0.65.

Table 3: Retrieval-augmented video generation with MiniMax-H3 using top-1 retrieved references.

	No RAG	VideoCLIP-XL	LCO-Omni-3B	LCO-Omni-3B + Ft. (general+physics)	LCO-Omni-3B + Ft. (physics-only)
Optics	0.67	0.63	0.65	0.68	0.65
Thermal	0.53	0.54	0.61	0.63	0.64
Average	0.60	0.59	0.63	**0.66**	<u>0.65</u>

4 Related Work

4.1 Physics Simulation

Numerical benchmarks evaluate surrogate models on parameterized physical systems (Takamoto et al., 2022; Bonnet et al., 2022; Gupta & Brandstetter, 2023; Toshev et al., 2023; Luo et al., 2024; Ohana et al., 2024b), while visual benchmarks study causal reasoning and event prediction in simulated environments (Riochet et al., 2020; Yi et al., 2020; Bear et al., 2021; Chen et al., 2025). Existing work therefore focuses mainly on either numerical fields or object-level dynamics. Our benchmark instead renders simulations across four physics domains as videos paired with physical quantities, enabling retrieval and physical property regression on video representations.

4.2 Physics-aware Video Generation and World Modeling

Video generation evaluation has expanded from visual quality and text alignment (Liu et al., 2023; 2024; Huang et al., 2024) to physical plausibility and world-modeling capabilities (Bansal et al., 2025; Meng et al., 2025; Zheng et al., 2025; Duan et al., 2025; Li et al., 2026a). Related methods improve physical faithfulness through prompt refinement (Xue et al., 2025), physics-oriented training (Wang et al., 2025), or retrieved video references (Cheng et al., 2026). In contrast, we examine the physical information encoded in pretrained video representations and its utility for retrieval, regression, and retrieval-augmented generation.

4.3 Omnimodal Representation Learning

Omnimodal representation learning has progressed from image-text alignment (Radford et al., 2021) to unified embedding spaces across modalities (Girdhar et al., 2023; Zhu et al., 2024). Recent MLLM-based approaches obtain general-purpose embeddings through text-only contrastive learning (Jiang et al., 2024), instruction-conditioned training (Jiang et al., 2025), or language-centric contrastive refinement (Xiao et al., 2026). Recent models support videos and visual documents (Meng et al., 2026; Li et al., 2026b), interleaved inputs (Zhou et al., 2026), and audio or composed multimodal queries (Huynh et al., 2026). Huang et al. (2026) has expanded the evaluation accordingly, covering a broader range of modalities and reveals persistent modality bias and asymmetric cross-modal retrieval. Complementing this broad semantic evaluation, our benchmark focuses on whether video embeddings preserve fine-grained physical information, as measured by physical-description alignment and quantitative property recovery.

5 Conclusion

We introduced the World Embedding Benchmark to evaluate how video representations capture and align physical information. Across 8,000 controlled simulation cases, we find that pre-trained omnimodal embeddings preserve recoverable physical signals but exhibit weak cross-modal physical alignment. Physics-specific contrastive adaptation improves retrieval and pair classification, but degrades physical-property regression, revealing a trade-off between alignment and quantitative information recoverability. Finally, physically relevant retrieval improves the fidelity of generated videos, with stronger retrieval models yielding larger gains. These results highlight the importance of jointly evaluating physical alignment and information recoverability when developing representations for physically grounded world models.

AI DISCLOSURE

In this work, we used generative AI tools to assist with data preprocessing and the development of portions of the code implementing our method. We also used generative AI tools to assist with the development of code for producing figures and charts used to present the experimental results. In addition, generative AI tools were used to proofread and revise the manuscript to improve its clarity, readability, and presentation.

We did not use AI in the conceptualization process, the formulation of the research questions, and experiment designs.

All AI-assisted outputs incorporated into the work were reviewed by the authors. The authors take responsibility for the final content of the paper, including its methodology, experiments, code, analyses, claims, and artifacts.

REFERENCES

James Ahrens, Berk Geveci, and Charles Law. ParaView: An end-user tool for large-data visualization. In Charles D. Hansen and Christopher R. Johnson (eds.), *Visualization Handbook*, pp. 717–731. Elsevier Inc., Burlington, MA, USA, 2005. ISBN 9780123875822.

B. F. Armaly, F. Durst, J. C. F. Pereira, and B. Schönung. Experimental and theoretical investigation of backward-facing step flow. *Journal of Fluid Mechanics*, 127:473–496, 1983. doi: 10.1017/S0022112083002839.

Adnan El Assadi, Roman Solomatin, Isaac Chung, Chenghao Xiao, Deep Shah, Manan Dey, Shriya Sudhakar, Zacharie Bugaud, Wissam Siblini, Ayush Sunil Munot, Yashwanth Devavarapu, Rakshitha Ireddi, Michelle Yang, Márton Kardos, Niklas Muennighoff, and Kenneth Enevoldsen. Mveb: Massive video embedding benchmark, 2026. URL https://arxiv.org/abs/2606.14958.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Igor A. Baratta, Joseph P. Dean, Jørgen S. Dokken, Michal Habera, Jack S. Hale, Chris N. Richardson, Marie E. Rognes, Matthew W. Scroggs, Nathan Sime, and Garth N. Wells. DOLFINx: The next generation FEniCS problem solving environment. https://doi.org/10.5281/zenodo.10447666, 2023.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Fan-Yun Sun, Fei-Fei Li, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d09bf41544a3365a46c9077ebb5e35c3-Abstract-round1.html.

Florent Bonnet, Jocelyn Mazari, Paola Cinnella, and Patrick Gallinari. Airfrans: High fidelity computational fluid dynamics dataset for approximating reynolds-averaged navier–stokes solutions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23463–23478. Curran Associates, Inc., 2022. doi: 10.52202/068431-1705. URL

https://proceedings.neurips.cc/paper_files/paper/2022/file/94ab7b23a345f93333eac8748a66c763-Paper-Datasets_and_Benchmarks.pdf.

Zhenfang Chen, Shilong Dong, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Compositional physical reasoning of objects and events from videos. *IEEE transactions on pattern analysis and machine intelligence*, 47(9):7689–7703, 2025.

Kexu Cheng, Zicheng Liu, Mingju Gao, Chunhe Song, and Hao Tang. Physrag: Enhancing physics-awareness in video generation via retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2606.26916.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 27713–27724, 2025.

U. Ghia, K. N. Ghia, and C. T. Shin. High-Re solutions for incompressible flow using the Navier–Stokes equations and a multigrid method. *Journal of Computational Physics*, 48(3):387–411, 1982. doi: 10.1016/0021-9991(82)90058-4.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15180–15190. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01457. URL https://doi.org/10.1109/CVPR52729.2023.01457.

Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=dPSTDbGtBY.

Haohang Huang, Xuan Lu, Mingyi Su, Xuan Zhang, Ziyan Jiang, Ping Nie, Kai Zou, Tomas Pfister, Wenhu Chen, Wei Zhang, Xiaoyu Shen, and Rui Meng. MMEB-v3: Measuring the performance gaps of omni-modality embedding models. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=EgNjgS4yOh.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Chuong Huynh, Manh Luong, and Abhinav Shrivastava. Efficient and high-fidelity omni modality retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8770–8780, June 2026.

Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 Renderer. https://mitsuba-renderer.org, 2022. Version 3.8.0.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: universal embeddings with multimodal large language models. *CoRR*, abs/2407.12580, 2024. doi: 10.48550/ARXIV.2407.12580. URL https://doi.org/10.48550/arXiv.2407.12580.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TE0KOzWYAF.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DLlVjZQ7vD.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026b. URL https://arxiv.org/abs/2601.04720.

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, 2024. URL https://arxiv.org/abs/2310.11440.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=yWpY5I3XyX.

Yining Luo, Yingfa Chen, and Zhen Zhang. Cfdbench: A large-scale benchmark for machine learning methods in fluid dynamics, 2024. URL https://arxiv.org/abs/2310.05963.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dIjMswSzgF.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Raghuveer Thirukovalluru, Xuan Zhang, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=TpU38jbKIJ.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities. https://www.minimax.io/blog/minimax-h3, 2026.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024a. Curran Associates Inc. ISBN 9798331314385.

Ruben Ohana, Michael McCabe, Lucas Thibaut Meyer, Rudy Morel, Fruzsina Julia Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond Buschman Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich Kerswell, Suryanarayana Maddu, Jonah M. Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Holden Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=00Sx577BT3.

OpenCFD. OpenFOAM v2406: The open source cfd toolbox. https://www.openfoam.com/, 2024.

Ardavan F. Oskooi, David Roundy, Mihai Ibanescu, Peter Bermel, J.D. Joannopoulos, and Steven G. Johnson. Meep: A flexible free-software package for electromagnetic simulations by the fdtd method. *Computer Physics Communications*, 181(3):687–702, 2010. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2009.11.008. URL https://www.sciencedirect.com/science/article/pii/S001046550900383X.

Elia Peruzzo, Dejia Xu, Xingqian Xu, Humphrey Shi, and Nicu Sebe. Ragme: Retrieval augmented video generation for enhanced motion realism. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ICMR '25, pp. 1081–1090, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718779. doi: 10.1145/3731715.3733417. URL https://doi.org/10.1145/3731715.3733417.

E. H. Por, S. Y. Haffert, V. M. Radhakrishnan, D. S. Doelman, M. Van Kooten, and S. P. Bos. High Contrast Imaging for Python (HCIPy): an open-source adaptive optics and coronagraph simulator. In *Adaptive Optics Systems VI*, volume 10703 of *Proc. SPIE*, 2018. doi: 10.1117/12.2314407.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning, 2020. URL https://arxiv.org/abs/1803.07616.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Alessandro Tasora, Radu Serban, Hammad Mazhar, Arman Pazouki, Daniel Melanz, Jonathan Fleischmann, Michael Taylor, Hiroyuki Sugiyama, and Dan Negrut. Chrono: An open source multiphysics dynamics engine. In *High Performance Computing in Science and Engineering (HPCSE 2015)*, volume 9611 of *Lecture Notes in Computer Science*, pp. 19–49. Springer, Cham, 2016. doi: 10.1007/978-3-319-40361-8_2.

Artur Toshev, Gianluca Galletti, Fabian Fritz, Stefan Adami, and Nikolaus Adams. Lagrangebench: A lagrangian fluid mechanics benchmarking suite. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64857–64884. Curran Associates, Inc., 2023. doi: 10.52202/075280-2830. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ccac3b120c7dc86d45f56830732b62be-Paper-Datasets_and_Benchmarks.pdf.

Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. Videoclip-xl: Advancing long description understanding for video clip models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16061–16075, 2024.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. WISA: World simulator assistant for physics-aware text-to-video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4jWuS5hye1.

Henry G. Weller, Gavin Tabor, Hrvoje Jasak, and Christer Fureby. A tensorial approach to computational continuum mechanics using object-oriented techniques. *Computers in Physics*, 12(6): 620–631, 1998. doi: 10.1063/1.168744.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22187–22198. IEEE, 2025.

Chenghao Xiao, Hou Pong Ken Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. *Advances in Neural Information Processing Systems*, 38:158370–158401, 2026.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*, 2025.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18826–18836. IEEE, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: CoLlision Events for Video REpresentation and Reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. Wemm-embedding: Wechat multi-modal embedding technical report, 2026. URL https://arxiv.org/abs/2608.24053.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QmZKc7UZCy.

A Dataset and Generation Details

A.1 Dataset Overview

The physics simulation dataset contains 8,000 cases organized into 80 families across four branches: Fluid Mechanics, Solid Mechanics, Dynamics, and Optics and Electromagnetism. Each family contains 100 cases. Cases within a family share the same physical scenario while varying selected material properties, geometric parameters, initial or boundary conditions, loading conditions, or excitation parameters.

Each case contains its simulation configuration, computed physical response, associated physical quantities, and rendered video. These records provide the source data for the retrieval, physical-property regression, and pair-classification tasks described in Appendix B.

A.2 Physical Formulations and Simulation

The simulation families use established simulation backends, analytical evaluation, or numerical ODE integration according to the underlying physical formulation. In this appendix, *Analytical / ODE integration* denotes analytical evaluation or numerical time integration implemented directly from the specified formulation. The formulations below summarize the equations used by the generation pipelines. Table 4 maps each family to its formulation, computation method, case parameters, and field output.

A.2.1 Fluid Mechanics

Governing formulations The fluid families use the incompressible mass and momentum equations,

$$\begin{aligned} \nabla\cdot\mathbf{u} &= 0, \\ \frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} &= -\frac{1}{\rho}\nabla p + \nabla\cdot(2\nu_{\text{eff}}\mathbf{D}) + \mathbf{f}, \\ \mathbf{D} &= \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right). \end{aligned} \tag{1}$$

Here, ν_{eff} is the kinematic viscosity used by the selected laminar, turbulence, or rheological configuration.

Family organization The fluid subset contains seven benchmark families instantiated from fifteen lower-level simulation configurations. The family organization groups configurations according to their dominant physical mechanism instead of their geometry. These families cover internal flows, external flows, transport phenomena, buoyancy-driven flows, and non-Newtonian rheology.

Family-specific relations Buoyancy-driven configurations couple thermal transport to momentum through the Boussinesq approximation,

$$\begin{aligned} \frac{\partial T}{\partial t} + \mathbf{u}\cdot\nabla T &= \alpha\nabla^2 T, \\ \mathbf{f}_b &= \mathbf{g}\,\beta(T - T_0). \end{aligned} \tag{2}$$

The non-Newtonian configurations use a generalized-Newtonian stress relation,

$$\boldsymbol{\tau} = 2\rho\nu_{\text{eff}}(\dot{\gamma})\mathbf{D}, \qquad \dot{\gamma} = \sqrt{2\mathbf{D}:\mathbf{D}}, \tag{3}$$

where ν_{eff} follows the PowerLaw or HerschelBulkley transport model selected by the OpenFOAM case. The corresponding case configuration specifies K, n, τ_0, ν_0, $\nu_{\min}$, and $\nu_{\max}$ as applicable. Passive-scalar transport follows

$$\frac{\partial c}{\partial t} + \mathbf{u}\cdot\nabla c = D\nabla^2 c. \tag{4}$$

Computation and visualization Fluid cases are solved with OpenFOAM using the finite-volume method. The principal field outputs are velocity magnitude and, for the corresponding families, temperature or transported-scalar fields. Video frames follow simulation time.

A.2.2 SOLID MECHANICS

Governing formulations The linear-elastic families solve static equilibrium with the small-strain isotropic constitutive relation,

$$\begin{aligned} \nabla\cdot\boldsymbol{\sigma} + \mathbf{b} &= \mathbf{0}, \\ \boldsymbol{\varepsilon} &= \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right), \\ \boldsymbol{\sigma} &= \lambda\,\mathrm{tr}(\boldsymbol{\varepsilon})\mathbf{I} + 2\mu\boldsymbol{\varepsilon}. \end{aligned} \tag{5}$$

Finite-strain families use the compressible Neo-Hookean energy implemented by the solver,

$$\begin{aligned} \mathbf{F} &= \mathbf{I} + \nabla\mathbf{u}, & J &= \det\mathbf{F}, \quad \mathbf{C} = \mathbf{F}^T\mathbf{F}, \\ \psi(\mathbf{F}) &= \frac{\mu}{2}\left[\mathrm{tr}(\mathbf{C}) - d - 2\ln J\right] + \frac{\kappa}{2}(\ln J)^2, & & \\ \mathrm{Div}\,\mathbf{P} &= \mathbf{0}, & \mathbf{P} &= \frac{\partial\psi}{\partial\mathbf{F}}. \end{aligned} \tag{6}$$

In equation 6, κ denotes the volumetric coefficient multiplying $(\ln J)^2$. The geometrically nonlinear buckling family uses the same finite-strain energy form together with an initial geometric imperfection and compressive loading.

History-dependent response Families carrying J2-related identifiers use the internal-variable update implemented by the current DOLFINx pipeline,

$$\begin{aligned} \varepsilon_{\mathrm{eq}} &= \sqrt{\frac{2}{3}\,\mathrm{dev}(\boldsymbol{\varepsilon}) : \mathrm{dev}(\boldsymbol{\varepsilon})}, \\ p^{k+1} &= p^k + c_p\left\langle \varepsilon_{\mathrm{eq}} - \left(c_y\varepsilon_y + c_H\frac{H}{E}p^k\right)\right\rangle_+, \\ \mu_{\mathrm{eff}} &= \frac{\mu}{1+\beta p}. \end{aligned} \tag{7}$$

The coefficients c_p, c_y, and c_H follow the corresponding 2D or 3D implementation and the selected loading family. The equation 7 is the implemented history-dependent constitutive update used to generate these cases.

Computation and visualization Solid-mechanics cases are solved with DOLFINx using finite-element discretization. The released videos primarily visualize displacement magnitude on the deformed geometry. Frame progression follows successive loading states.

A.2.3 DYNAMICS

Rigid-body and lumped-parameter dynamics Chrono-based families use constrained Newton-Euler dynamics,

$$\begin{aligned} m_i\dot{\mathbf{v}}_i &= \sum\mathbf{F}_i, \\ \mathbf{I}_i\dot{\boldsymbol{\omega}}_i + \boldsymbol{\omega}_i\times(\mathbf{I}_i\boldsymbol{\omega}_i) &= \sum\boldsymbol{\tau}_i, \\ \boldsymbol{\Phi}(\mathbf{q},t) &= \mathbf{0}. \end{aligned} \tag{8}$$

Rigid collision uses an impulse-based contact update. The normal component satisfies

$$m_i\left(\mathbf{v}_i^+ - \mathbf{v}_i^-\right) = \pm J_n\mathbf{n}, \qquad \mathbf{n}\cdot\mathbf{v}_{\mathrm{rel}}^+ = -e\,\mathbf{n}\cdot\mathbf{v}_{\mathrm{rel}}^-. \tag{9}$$

The implementation additionally accounts for eccentric contact and the associated rotational response, together with a bounded tangential friction impulse when friction is enabled.

The mass-spring-damper family integrates

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{C}\dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{f}(t), \tag{10}$$

and the rolling/sliding family uses

$$m\dot{\mathbf{v}} = \sum \mathbf{F}, \qquad I\dot{\omega} = \sum \tau, \qquad v_{\mathrm{rel}} = v - R\omega. \tag{11}$$

Drop-weight impact is integrated using a spring-damper contact force coupled to an elastic plate degree of freedom,

$$\delta = \left[z_{\mathrm{imp}} + r - z_{\mathrm{plate}}\right]_+, \tag{12a}$$

$$F_c = \left[k_n\delta + c_n\left(v_{\mathrm{imp}} - v_{\mathrm{plate}}\right)\right]_+, \tag{12b}$$

$$m_{\mathrm{plate}}\ddot{q} + c_{\mathrm{plate}}\dot{q} + k_{\mathrm{plate}}q = F_c. \tag{12c}$$

Mechanism relations The crank-slider family uses geometric closure,

$$\theta(t) = \theta_0 + \omega t, \qquad x_s = r\cos\theta + \sqrt{L^2 - r^2\sin^2\theta}. \tag{13}$$

The gear family uses tooth-count kinematics,

$$\omega_j = -\frac{N_i}{N_j}\omega_i, \qquad \theta_j - \theta_{j,0} = -\frac{N_i}{N_j}(\theta_i - \theta_{i,0}). \tag{14}$$

The articulated-chain generator uses a damped, driven angular response,

$$\theta_j(t) = \theta_{0,j}e^{-\zeta_j t}\cos(\omega_j t + \phi_j) + A_d r_d^j \sin(\Omega t - j\Delta\phi) + c_j\theta_{j-1}(t), \tag{15}$$

with ω_j determined from gravity and the mean upstream link length in the implementation.

Structural-response relations The transient cantilever family uses the Euler-Bernoulli tip displacement and first-mode frequency,

$$\delta_s = \frac{FL^3}{3EI}, \qquad f_1 = \frac{1.875^2}{2\pi}\sqrt{\frac{EI}{\rho AL^4}}. \tag{16}$$

The harmonic-response family uses

$$A(\omega) = \frac{F_0}{\sqrt{(k - m_{\mathrm{eff}}\omega^2)^2 + (c_{\mathrm{eff}}\omega)^2}}. \tag{17}$$

The dataset identifier `reissnerShellModal` currently uses a thin-plate modal generator,

$$D = \frac{Et^3}{12(1-\nu^2)}, \qquad w_{mn}(x,y,t) = A_{mn}\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b}\sin(2\pi f_{mn}t)e^{-0.12t}. \tag{18}$$

Longitudinal stress-wave propagation follows

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, \qquad c = \sqrt{\frac{E}{\rho}}. \tag{19}$$

The unbalanced-rotor response uses the rotating-imbalance amplitude relation

$$r = \frac{\Omega}{\Omega_c}, \qquad A(r) = e\,\frac{r^2}{\sqrt{(1-r^2)^2 + (2\zeta r)^2}}. \tag{20}$$

Parameterized structural responses The `j2PlasticBeam`, `lemaitreDamageBar`, and `rubberBlockCompression` families use explicit parameterized response relations defined in their generation pipelines. Their displacement, internal-variable, and deformation fields are given by equation 21a–equation 21c.

$$u_{\mathrm{beam}}(x,t) = a(t)\phi(x), \qquad \bar{\varepsilon}^p(x,t) = \left(1 - \frac{x}{L}\right)^2 p(t). \tag{21a}$$

$$D(x,t) = d_{\max}(t)\exp\left[-\frac{(x - L/2)^2}{2w^2}\right], \qquad u_{\mathrm{damage}}(x,t) = u_0\sin(2\pi f t)\frac{x}{L}\left[1 - 0.55D(x,t)\right]. \tag{21b}$$

$$x' = x\,[1 + 0.24c(t)]\,, \qquad y' = y\,[1 + 0.24c(t)]\,, \qquad z' = z\,[1 - c(t)]\,. \tag{21c}$$

Here, $a(t)$ and $p(t)$ specify the prescribed beam-response histories, $d_{\max}(t)$ controls the evolution of the localized damage field, and $c(t)$ specifies the compression history of the rubber-block family. The spatial functions $\phi(x)$ and the Gaussian profile in equation 21b determine the corresponding field distributions.

Computation and visualization Dynamics combines Chrono and Analytical / ODE integration of the specified equations. The displayed quantities depend on the family and include velocity magnitude, displacement, angular speed, axial stress, equivalent plastic strain, and scalar damage. Frames represent physical time.

A.2.4 Optics and Electromagnetism

Geometric optics Reflection, paraxial imaging, and refraction use

$$\mathbf{d}_r = \mathbf{d}_i - 2(\mathbf{d}_i \cdot \mathbf{n})\mathbf{n}, \tag{22}$$

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}, \tag{23}$$

and

$$n_1 \sin\theta_1 = n_2 \sin\theta_2, \qquad \theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right), \quad n_1 > n_2. \tag{24}$$

Interface reflection uses the Fresnel coefficients,

$$\begin{aligned} r_s &= \frac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t}, \\ r_p &= \frac{n_2\cos\theta_i - n_1\cos\theta_t}{n_2\cos\theta_i + n_1\cos\theta_t}, \qquad R = |r|^2. \end{aligned} \tag{25}$$

Wave and polarization optics HCIPy families propagate a scalar optical field,

$$U_z = \mathcal{P}_z\{U_0\}, \qquad I(x,y) = |U_z(x,y)|^2. \tag{26}$$

Analytical interference families use

$$d\sin\theta_m = m\lambda, \qquad \Delta\phi_{\text{film}} = \frac{4\pi n t\cos\theta_t}{\lambda}. \tag{27}$$

Gaussian-beam propagation follows

$$w(z) = w_0\sqrt{1 + \left(\frac{z}{z_R}\right)^2}, \qquad z_R = \frac{\pi w_0^2}{\lambda}. \tag{28}$$

Polarization families use

$$I = I_0\cos^2\theta, \qquad \mathbf{E}_{\text{out}} = \mathbf{R}(\theta)\begin{bmatrix}1 & 0\\ 0 & e^{i\delta}\end{bmatrix}\mathbf{R}(-\theta)\mathbf{E}_{\text{in}}. \tag{29}$$

Electromagnetism The Meep families solve Maxwell's equations in the time domain,

$$\nabla\times\mathbf{E} = -\mu\frac{\partial\mathbf{H}}{\partial t}, \qquad \nabla\times\mathbf{H} = \mathbf{J} + \epsilon\frac{\partial\mathbf{E}}{\partial t}. \tag{30}$$

The plasmonic-dimer family additionally uses the configured Drude-type dispersive material relation,

$$\epsilon(\omega) = \epsilon_\infty - \frac{\omega_p^2}{\omega^2 + i\gamma\omega}. \tag{31}$$

Computation and visualization Optical families use Analytical / ODE integration or scalar wave propagation with HCIPy. Electromagnetic families use Meep FDTD. Optical field outputs are rendered as irradiance. Meep frames use

$$Q = (\mathrm{Re}\, E_x)^2 + (\mathrm{Re}\, E_y)^2 + (\mathrm{Re}\, E_z)^2 + 0.15(\mathrm{Re}\, H_z)^2, \tag{32}$$

followed by per-case normalization over the sampled frames. Optical sequences may follow a parameter scan; Meep sequences follow simulation time.

Table 4: Summary of simulation families, their physical formulations, computation methods, case parameters, and field outputs.

Family	Formulation	Computation	Case parameters	Field output
Fluid Mechanics				
`buoyancy`	equation 1, equation 2	OpenFOAM (FVM)	cavity dimensions; temperature difference; reference temperature; kinematic viscosity; thermal diffusivity; gravity	$\|\mathbf{U}\|$
`cavity`	equation 1	OpenFOAM (FVM)	cavity dimensions; lid velocity; kinematic viscosity; Reynolds number Re	$\|\mathbf{U}\|$
`external_aero`	equation 1	OpenFOAM (FVM)	airfoil chord; thickness scale; free-stream velocity; kinematic viscosity; Reynolds number Re	$\|\mathbf{U}\|$
`external_wake`	equation 1	OpenFOAM (FVM)	obstacle diameter or width; free-stream velocity; kinematic viscosity; Reynolds number Re	$\|\mathbf{U}\|$
`internal_step`	equation 1, equation 3	OpenFOAM (FVM)	channel dimensions; inlet velocity; Reynolds number Re; reference viscosity ν_0; viscosity bounds $\nu_{\min}, \nu_{\max}$; consistency K; exponent n; yield stress τ_0	$\|\mathbf{U}\|$
`jet_axi`	equation 1	OpenFOAM (FVM)	jet radius; outer radius; jet velocity; kinematic viscosity; Reynolds number Re	$\|\mathbf{U}\|$
`scalar_transport`	equation 1, equation 4	OpenFOAM (FVM)	domain dimensions; carrier velocity; scalar diffusivity; Péclet number Pe; injection position and duration	T (transported scalar)
Solid Mechanics				
`uniaxialTension`	equation 5	DOLFINx (FEM)	L_x; L_y; neck position; neck width; Young's modulus E; Poisson ratio ν; load magnitude	$\|\mathbf{u}\|$
`uniaxial Compression`	equation 5	DOLFINx (FEM)	L_x; L_y; Young's modulus E; Poisson ratio ν; load magnitude; loading direction	$\|\mathbf{u}\|$
`cantilever Localization`	equation 5	DOLFINx (FEM)	L_x; L_y; localization-band position; Young's modulus E; Poisson ratio ν; load magnitude; tip force direction	$\|\mathbf{u}\|$
`plateWithHole`	equation 5	DOLFINx (FEM)	hole radius; hole center x; hole center y; L_x; L_y; Young's modulus E; Poisson ratio ν; load magnitude	$\|\mathbf{u}\|$
`pureShear`	equation 5	DOLFINx (FEM)	L_x; L_y; Young's modulus E; Poisson ratio ν; load magnitude; shear ratio	$\|\mathbf{u}\|$
`threePointBending`	equation 5	DOLFINx (FEM)	support span; loading width; L_x; L_y; Young's modulus E; Poisson ratio ν; load magnitude; center-drop ratio	$\|\mathbf{u}\|$
`torsionShear`	equation 5	DOLFINx (FEM)	twist-axis position; warping coefficient; L_x; L_y; Young's modulus E; Poisson ratio ν; load magnitude; twist magnitude	$\|\mathbf{u}\|$
`biaxialTension`	equation 5	DOLFINx (FEM)	L_x; L_y; biaxial loading ratio; Young's modulus E; Poisson ratio ν; load magnitude	$\|\mathbf{u}\|$
`contactPatch Compression`	equation 5	DOLFINx (FEM)	patch center; patch half-width; L_x; L_y; Young's modulus E; Poisson ratio ν; load magnitude; patch-load magnitude	$\|\mathbf{u}\|$
`cubeCompression3D`	equation 5	DOLFINx (FEM)	L_x; L_y; L_z; Young's modulus E; Poisson ratio ν; load magnitude; compression direction	$\|\mathbf{u}\|$
`longCuboidBending 3D`	equation 5	DOLFINx (FEM)	L_x; L_y; L_z; Young's modulus E; Poisson ratio ν; load magnitude	$\|\mathbf{u}\|$
`cylinderTorsion3D`	equation 5	DOLFINx (FEM)	cylinder radius R; L_x; L_y; L_z; Young's modulus E; Poisson ratio ν; twist angle; load magnitude	$\|\mathbf{u}\|$
`sphereIndentation 3D`	equation 5	DOLFINx (FEM)	sphere radius R; L_x; L_y; L_z; Young's modulus E; Poisson ratio ν; indentation-depth ratio; load magnitude	$\|\mathbf{u}\|$
`plateWithHole3D`	equation 5	DOLFINx (FEM)	hole radius; hole center x; hole center y; hole center z; L_x; L_y; L_z; Young's modulus E	$\|\mathbf{u}\|$

Continued on next page

Family	Formulation	Computation	Case parameters	Field output
notchedBarTension 3D	equation 5	DOLFINx (FEM)	notch-depth ratio; notch-width ratio; notch position; L_x; L_y; L_z; Young's modulus E; Poisson ratio ν	$\|\mathbf{u}\|$
neoHookean	equation 6	DOLFINx (nonlinear FEM)	L_x; L_y; shear modulus μ; volumetric coefficient κ; load magnitude	$\|\mathbf{u}\|$
buckling Snapthrough	equation 6	DOLFINx (nonlinear FEM)	L_x; L_y; initial imperfection; Young's modulus E; Poisson ratio ν; load magnitude; critical-load parameter	$\|\mathbf{u}\|$
near Incompressible Block	equation 6	DOLFINx (nonlinear FEM)	L_x; L_y; shear modulus μ; volumetric coefficient κ; near-incompressibility bias; load magnitude	$\|\mathbf{u}\|$
hyperelasticNeo HookeanCuboid3D	equation 6	DOLFINx (nonlinear FEM)	L_x; L_y; L_z; shear modulus μ; volumetric coefficient κ; load magnitude	$\|\mathbf{u}\|$
hyperelasticNeo Hookean3D	equation 6	DOLFINx (nonlinear FEM)	cylinder radius R; L_x; L_y; L_z; shear modulus μ; volumetric coefficient κ; load magnitude	$\|\mathbf{u}\|$
hyperelasticNeo HookeanSphere3D	equation 6	DOLFINx (nonlinear FEM)	sphere radius R; L_x; L_y; L_z; shear modulus μ; volumetric coefficient κ; load magnitude	$\|\mathbf{u}\|$
yieldHardening	equation 7	DOLFINx (nonlinear FEM)	L_x; L_y; neck position; neck width; Young's modulus E; Poisson ratio ν; yield stress σ_y; hardening modulus H	$\|\mathbf{u}\|$
loadUnload Hysteresis	equation 7	DOLFINx (nonlinear FEM)	L_x; L_y; neck position; neck width; Young's modulus E; Poisson ratio ν; yield stress σ_y; hardening modulus H	$\|\mathbf{u}\|$
necking	equation 7	DOLFINx (nonlinear FEM)	L_x; L_y; neck position; neck width; Young's modulus E; Poisson ratio ν; yield stress σ_y; hardening modulus H	$\|\mathbf{u}\|$
plasticJ2Cylinder Tension3D	equation 7	DOLFINx (nonlinear FEM)	cylinder radius R; L_x; L_y; L_z; Young's modulus E; Poisson ratio ν; yield stress σ_y; hardening modulus H	$\|\mathbf{u}\|$
plasticJ2Notched Bar3D	equation 7	DOLFINx (nonlinear FEM)	notch-depth ratio; notch-width ratio; notch position; L_x; L_y; L_z; Young's modulus E; Poisson ratio ν	$\|\mathbf{u}\|$
plasticJ2Large Strain3D	equation 7	DOLFINx (nonlinear FEM)	L_x; L_y; L_z; Young's modulus E; Poisson ratio ν; yield stress σ_y; hardening modulus H; load magnitude	$\|\mathbf{u}\|$
Dynamics				
pendulum	equation 8	Chrono	link length; bob radius; mass; initial angle; initial angular velocity; gravity g; damping	$\|\mathbf{v}\|$ on bob
rigidCollision	equation 9	Analytical / ODE integration	body shape; body dimensions; body offset; mass; restitution coefficient e; friction coefficient; initial velocity; impact offset	$\|\mathbf{v}\|$ on bodies
sphereDrop3D	equation 8	Chrono	sphere radius R; drop height; mass; restitution coefficient e; friction coefficient; initial velocity; gravity g; ground contact	$\|\mathbf{v}\|$ on sphere
stackStability	equation 8	Chrono	block count; block dimensions; stack offset; mass; friction coefficient; restitution coefficient e; disturbance velocity; gravity g	$\|\mathbf{v}\|$ on blocks
blockTower3D	equation 8	Chrono	tower height; block dimensions; impactor dimensions; mass; friction coefficient; restitution coefficient e; impact velocity; gravity g	$\|\mathbf{v}\|$ on bodies
dropWeightImpact SMC	equation 12	Analytical / ODE integration	plate size and thickness; impactor radius and mass; contact stiffness k_n; contact damping c_n; plate stiffness, damping, and mass; drop height; initial velocity; gravity g	plate displacement $\|\mathbf{u}\|$
flexibleCrank Slider	equation 13	Analytical / ODE integration	crank radius; connecting-rod length; slider-axis direction; initial phase; angular speed; rod-deflection amplitude	rod deflection $u(s,t)$
gearPair3D	equation 14	Analytical / ODE integration	gear count; gear radii; tooth counts; shaft locations; mesh clearance; initial phase; motor-speed profile	angular speed $\|\omega\|$
cantileverBeam Transient	equation 16	Analytical / ODE integration	beam length L; width b; height h; Young's modulus E; density ρ; tip load F; load-removal time	beam displacement u
harmonicResponse FRF	equation 17	Analytical / ODE integration	frequency sweep; reference frequency f_1; force amplitude F_0; damping coefficient; drive frequency; beam length L	displacement u
reissnerShell Modal	equation 18	Analytical / ODE integration	plate dimensions a, b; thickness t; Young's modulus E; Poisson ratio ν; density ρ; mode indices (m,n); mode amplitude	modal displacement w
stressWave Propagation	equation 19	Analytical / ODE integration	rod length L; diameter; Young's modulus E; density ρ; pulse amplitude; rise and fall times	axial stress σ_{xx}

Continued on next page

Family	Formulation	Computation	Case parameters	Field output
`rubberBlock Compression`	equation 21	Analytical / ODE integration	block dimensions; mesh resolution; compression ratio; loading duration; effective stiffness	displacement-derived scalar u_{interp}
`j2PlasticBeam`	equation 21	Analytical / ODE integration	beam length L; element count; maximum tip deflection; residual tip deflection; maximum equivalent plastic strain	equivalent plastic strain $\bar{\varepsilon}^p$ / displacement u
`lemaitreDamageBar`	equation 21	Analytical / ODE integration	bar length L; node count; notch width; displacement amplitude; cycle frequency; damage-start cycle; failure cycle	damage D
`massSpringDamper`	equation 10	Analytical / ODE integration	number of degrees of freedom; mass spacing; cart count; mass; spring stiffness; damping; initial displacement; initial velocity	displacement x_i
`rollingSliding Transition`	equation 11	Analytical / ODE integration	roller radius; roller width; track length; friction coefficient; rolling friction; inertia factor; initial translational speed; initial angular speed	local surface speed $\|\mathbf{v}_s\|$
`unbalancedRotor3D`	equation 20	Analytical / ODE integration	shaft length; rotor radius; rotor mass; critical speed; spin-speed profile; damping ratio ζ; eccentricity; axis height	lateral displacement $u_\perp$
`articulatedChain`	equation 15	Analytical / ODE integration	link count; link lengths; initial joint angles; damping; base-drive amplitude and frequency; phase lag; terminal mass	link / joint speed $\|\mathbf{v}\|$
Optics & EM				
`planeMirror Reflection`	equation 22	Analytical / ODE integration	incidence angle; mirror reflectance; screen distance; beam divergence	$I(x, y)$ (irradiance)
`concaveMirror Focus`	equation 23	Analytical / ODE integration	focal length; aperture radius; object distance; detector distance	$I(x, y)$ (irradiance)
`dielectricBlock Refraction`	equation 24	Analytical / ODE integration	incidence angle; refractive index n; block thickness; screen distance	$I(x, y)$ (irradiance)
`totalInternal Reflection`	equation 24	Analytical / ODE integration	incidence angle; critical angle; core index n_1; cladding index n_2	$I(x, y)$ (irradiance)
`fresnelBrewster`	equation 25	Analytical / ODE integration	incidence angle; Brewster angle; refractive index n_2; polarization	$I(x, y)$ (irradiance)
`beamSplitterPlate`	equation 25	Analytical / ODE integration	incidence angle; reflectance R; transmittance T; plate thickness	$I(x, y)$ (irradiance)
`singleSlit Fraunhofer`	equation 26	HCIPy	slit width a; wavelength λ; screen distance; source offset	$I(x, y)$ (irradiance)
`doubleSlit Interference`	equation 26	HCIPy	slit width a; slit separation d; wavelength λ; relative phase ϕ	$I(x, y)$ (irradiance)
`circularAperture Airy`	equation 26	HCIPy	aperture diameter D; wavelength λ; screen distance; aperture scale	$I(x, y)$ (irradiance)
`knifeEdgeFresnel`	equation 26	HCIPy	edge position; wavelength λ; propagation distance z; beam width	$I(x, y)$ (irradiance)
`diffraction Grating`	equation 27	Analytical / ODE integration	grating pitch d; slit count N; wavelength λ; screen distance	$I(x, y)$ (irradiance)
`thinFilm Interference`	equation 27	Analytical / ODE integration	film thickness t; film index n; incidence angle; wavelength band	$I(x, y)$ (irradiance)
`gaussianBeam Propagation`	equation 28	Analytical / ODE integration	beam waist w_0; wavelength λ; propagation distance z; waist position	$I(x, y)$ (irradiance)
`prismDispersion`	equation 24	Analytical / ODE integration	prism apex angle; wavelength-dependent refractive indices; incidence angle; prism length	$I(x, y)$ (irradiance)
`thinLensFocus`	equation 23	Analytical / ODE integration	focal length; aperture radius; object distance; detector distance	$I(x, y)$ (irradiance)
`glassSphere Caustic`	equation 24	Analytical / ODE integration	sphere radius R; refractive index n; detector distance; source distance	$I(x, y)$ (irradiance)
`cornerCube Retroreflector`	equation 22	Analytical / ODE integration	incidence angle; face-angle error; aperture radius; return-screen distance	$I(x, y)$ (irradiance)
`polarizerMalus`	equation 29	Analytical / ODE integration	analyzer angle; extinction ratio; input polarization; polarization mode	$I(x, y)$ (irradiance)
`waveplate Polarization`	equation 29	Analytical / ODE integration	retardance δ; fast-axis angle; input polarization; analyzer angle	$I(x, y)$ (irradiance)
`meepCylinder Scattering`	equation 30	Meep (FDTD)	cylinder radius R; refractive index n; source wavelength λ	normalized $Q(x, y, t)$
`meepWaveguideMode`	equation 30	Meep (FDTD)	core width; core refractive index; source wavelength λ	normalized $Q(x, y, t)$
`meepPhotonic CrystalDefect`	equation 30	Meep (FDTD)	lattice pitch; rod radius; defect scale; rod refractive index; source wavelength λ	normalized $Q(x, y, t)$
`meepPhotonic CrystalWaveguide`	equation 30	Meep (FDTD)	lattice pitch; rod radius; rod refractive index; line-defect width; source wavelength λ	normalized $Q(x, y, t)$

Continued on next page

Family	Formulation	Computation	Case parameters	Field output
`meepGrating Diffraction`	equation 30	Meep (FDTD)	grating period; duty cycle; bar refractive index; source wavelength λ	normalized $Q(x, y, t)$
`meepPlasmonic Dimer`	equation 30, equation 31	Meep (FDTD)	particle radius; gap width; Drude frequency; Drude damping; source wavelength λ	normalized $Q(x, y, t)$
`meepDipole Radiation`	equation 30	Meep (FDTD)	source wavelength λ; dipole orientation	normalized $Q(x, y, t)$
`meepEvanescentTIR`	equation 30	Meep (FDTD)	incidence angle; high-index medium n_1; low-index medium n_2; source wavelength λ	normalized $Q(x, y, t)$

B Evaluation Protocols

B.1 Retrieval

Queries construction The retrieval evaluation contains one primary query for each simulation case. Each simulation case is associated with one primary retrieval query generated deterministically from a family-specific natural-language template. The template combines a description of the physical scenario, the underlying physical mechanism, the visualized quantity, case-specific physical conditions, and quantitative diagnostics. Case-dependent slots are instantiated from the simulation record and derived diagnostics. Figure 5 illustrates the resulting structure using a representative Dynamics example.

Each query is associated with one designated target video from the same simulation case. During dataset export, duplicate query texts are checked and disambiguated when necessary to preserve a unique primary query for each released case.

Hard-negative sampling The training and adaptation datasets additionally provide five hard-negative cases for each anchor. For an anchor case, the candidate negatives are the other 99 cases in the same family. Five distinct cases are sampled with a fixed random seed, and the corresponding query texts and videos are kept aligned. The negative assignments are therefore deterministic for a given dataset revision. These training negatives are stored separately from the retrieval evaluation candidate set.

Figure 5: A template-generated retrieval query example `articulatedChain_S300` from the training set. Colors indicate the functional components of the query, while underlining marks case-specific values instantiated from the simulation record or derived diagnostics.

Representative query from the released Dynamics data `articulatedChain_S300`

Find the 2D dynamics video of 3 link articulated pendulum chain, showing a chain of hinged rigid links swinging under gravity or a root drive. Physically, joint constraints transmit motion along the chain, creating phase lag, folding, or free-swing response. The color map represents link and joint speed magnitude along the chain, so speed colors show which parts move fastest. Use the variant with drive mode set to free release, link count set to short, terminal mass set to low; Earth-scale gravity acts downward; no separate ground plane is part of the visible setup. To identify the exact video, look for the mechanism rotates through about 25.03 degrees, linked bodies reach a relative angle near 25.55 degrees, the fastest visible motion reaches about 2.251 m/s, and the recorded motion covers roughly 6.57 s.

Scenario / visible response · Physical mechanism · Visualized quantity · Case-specific conditions · Quantitative diagnostics

B.2 Physical Property Regression

Task definition The physical-property regression benchmark evaluates whether continuous physical quantities can be recovered from frozen video representations. Each selected simulation family defines an independent regression task with one scalar target quantity and physical unit. The re-

gression model receives only the video representation as input; no textual query or description is provided. The ten regression families and their target quantities are summarized in Table 5.

Representation probe Given a video v_i, the evaluated encoder produces a frozen representation $\mathbf{z}_i$. A linear regression probe is fitted independently for each family to predict the corresponding physical quantity,

$$\hat{y}_i = f_\theta(\mathbf{z}_i). \tag{33}$$

The video encoder remains fixed throughout probe training, so the probe measures the accessibility of the target quantity in the learned representation.

Evaluation metrics. For each regression task, we compare the predicted values $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_N]$ with the ground-truth values $\mathbf{y} = [y_1, \ldots, y_N]$. We report normalized root mean squared error (nRMSE) as the primary metric and Spearman's rank correlation ρ as a complementary measure of whether the predicted values preserve the ordering of the underlying physical quantity.

Table 5: Continuous physical quantities used in the regression benchmark.

Family	Target	Range
`pendulum`	gravity g	2–18 $\mathrm{m\,s^{-2}}$
`rigidCollision`	mass m_2	2–18 kg
`sphereDrop3D`	restitution e	0.35–0.90
`dye2d`	Péclet number Pe	30–1200
`flowpastcylinder2d`	Reynolds number Re	80–500
`totalInternalReflection`	refractive index n	1.44–1.78
`doubleSlitInterference`	wavelength λ	450–700 nm
`cylinderTorsion3D`	shear modulus G	2–10 GPa
`longCuboidBending3D`	Young's modulus E	5–50 GPa
`hyperelasticNeoHookeanCuboid3D`	shear modulus μ	5–25 MPa

C Regression and Pair Classification Results after Training

Figure 6 and Figure 7 display results of regression and pair classification across training steps.

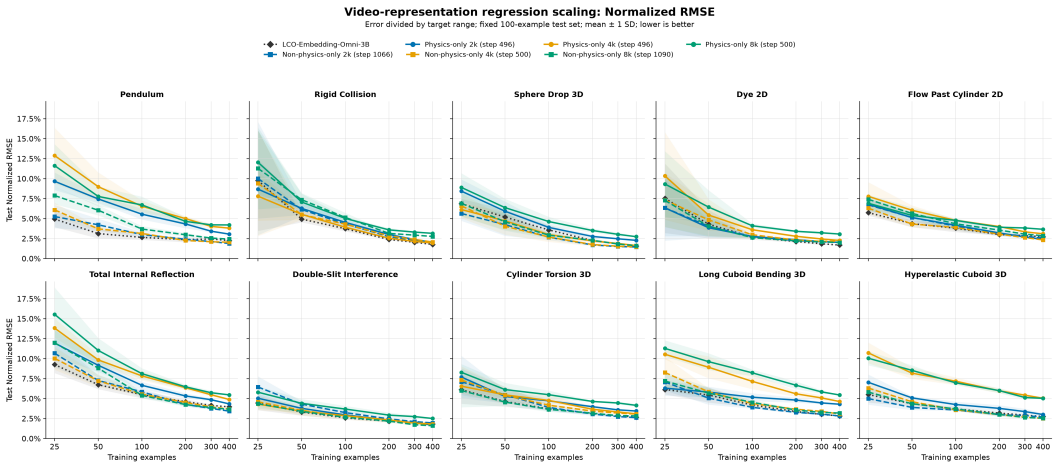


Figure 6: Linear regression performance (y-axis) of different contrastively trained checkpoints (different lines), across different numbers of probe examples (x-axis).

D Regression and Pair Classification Results after Training

Figure 8 shows the training dynamics from the 7B LCO variant, showing consistent behaviors with the 3B experiments in the main paper.

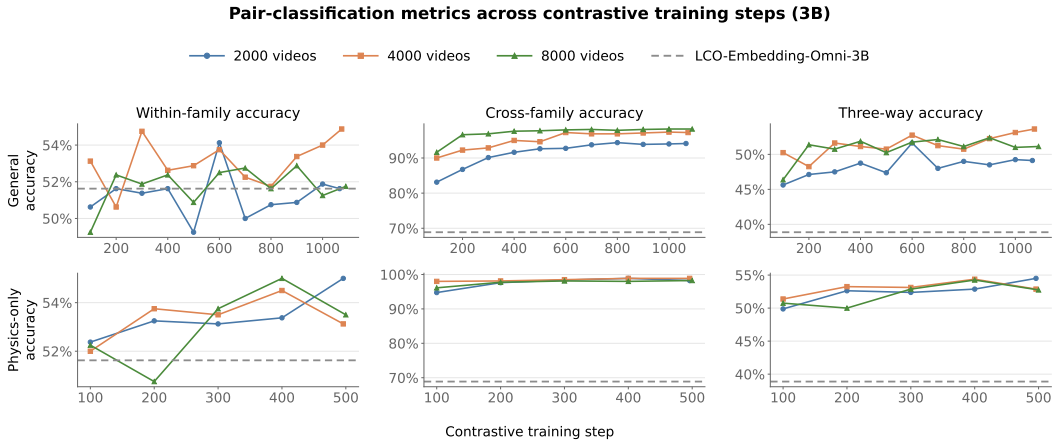


Figure 7: Within-family and cross-family pair classification performance across training steps.

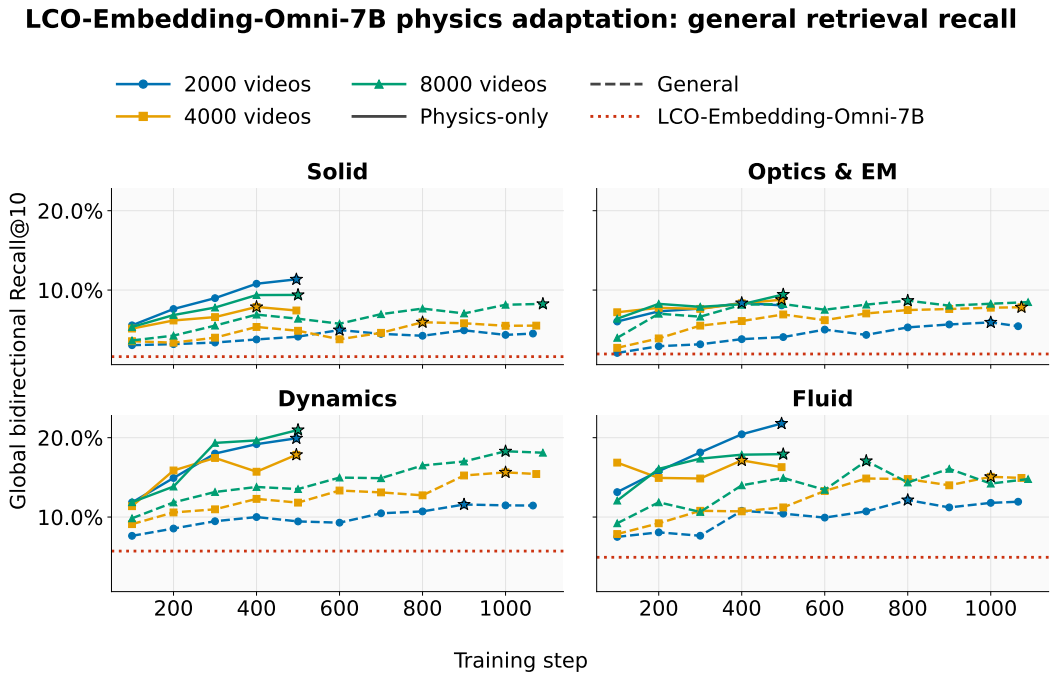


Figure 8: Retrieval performance (y-axis) across training steps (x-axis) and training dataset settings (different lines).