[TINY PAPER] TEMPORAL REVERSAL ASYMMETRY: A PHYSICS-INSPIRED METRIC FOR EVALUATING WORLD MODELS

Corrected version, July 2026 — see Appendix M for changes from the workshop version

Kanpat Vesessook*
Department of Computer Science
Brown University
Providence, RI 02912, USA
kanpat_vesessook@brown.edu

Kevin Yang*
Department of Computer Science
Brown University
Providence, RI 02912, USA
kevin_c_yang@brown.edu

ABSTRACT

World models that understand physics should recognize the thermodynamic arrow of time: some processes look physically plausible when reversed, while others do not. We propose Temporal Reversal Asymmetry (TRA), comparing prediction loss on forward versus time-reversed videos. We evaluated four video models on 380 physics simulations and found that TRA for V-JEPA2 varies continuously with dissipation strength, peaking at intermediate values (restitution 0.5, damping 2.0) where energy loss remains visible throughout the video, while approaching zero for both genuinely reversible processes and motion that stops too quickly. This non-monotonic relationship suggests V-JEPA2 has learned to recognize irreversible processes by how they look as opposed to memorizing categories. In contrast, VideoMAE shows inverted TRA (negative asymmetry), suggesting that latent prediction captures temporal causality more faithfully than pixel reconstruction. Two further baselines (MVD, Hiera) show near-zero or inconsistent TRA, but evaluation limitations disclosed in Appendix A prevent conclusions about those objectives. This work offers a training-free, physics-grounded probe that complements existing world model benchmarks.

1 INTRODUCTION

The second law of thermodynamics establishes an arrow of time: entropy tends to increase, making dissipative physical processes irreversible. A glass can shatter but spontaneously reassembling is vanishingly improbable. In contrast, idealized conservative systems like a frictionless pendulum are time-symmetric. A video of a pendulum played backwards remains physically plausible.

World models trained on natural video should, in principle, learn this distinction. If a model truly understands physics, it should find time-reversed dissipative processes surprising or difficult to predict, while treating low-dissipation processes symmetrically. This intuition motivates our proposed metric: Temporal Reversal Asymmetry (TRA).

Prior work has evaluated world models on intuitive physics benchmarks like IntPhys (Riochet et al., 2020) and PHYRE (Bakhtin et al., 2019), which test whether models distinguish possible from impossible scenarios. TRA offers a complementary approach: rather than asking whether a model knows something is impossible, we ask whether its prediction errors reflect thermodynamic asymmetry.

Our contributions are:

1. We introduce TRA, a training-free metric that probes whether world models respect the thermodynamic arrow of time (Section 2).

*Equal contribution.

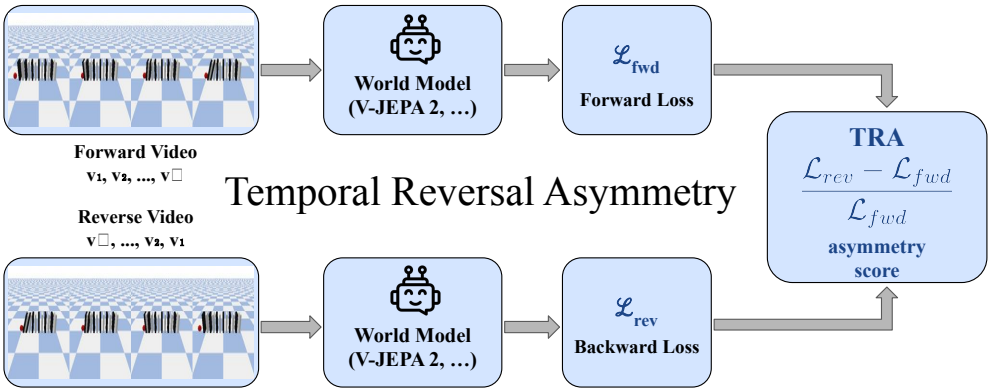

Figure 1: TRA computation: forward and time-reversed videos are passed through a world model; we compare their prediction losses.

2. We show V-JEPA2 exhibits physically appropriate TRA while VideoMAE shows an inverted pattern, and that TRA tracks visible irreversibility by peaking at intermediate physical parameters (Section 3).

2 METHOD

2.1 TEMPORAL REVERSAL ASYMMETRY

Given a video $\mathbf{v} = (v_1, v_2, \ldots, v_T)$ and its time-reversed version $\tilde{\mathbf{v}} = (v_T, v_{T-1}, \ldots, v_1)$, we define TRA as:

$$\text{TRA}(\mathbf{v}) = \frac{\mathcal{L}(\tilde{\mathbf{v}}) - \mathcal{L}(\mathbf{v})}{\mathcal{L}(\mathbf{v})} \tag{1}$$

where $\mathcal{L}(\cdot)$ is the model's prediction loss. A TRA of zero indicates the model treats forward and reversed videos identically. Positive TRA means reversed videos are harder to predict; negative TRA means forward videos are harder.

For a world model with physically grounded temporal understanding, we expect TRA ≈ 0 for low-dissipation processes (elastic collisions, pendulums) and TRA > 0 for dissipative processes (falling objects, toppling dominos). Figure 1 illustrates this computation.

2.2 MODELS

We evaluate four self-supervised video models: **V-JEPA2** (Assran et al., 2025), a Joint Embedding Predictive Architecture (ViT-Large, 303M params) that predicts latent representations with 3D rotary position embeddings; **VideoMAE** (Tong et al., 2022), a masked autoencoder (ViT-Base, 94M params) that reconstructs pixel patches; **MVD** (Wang et al., 2023), a teacher-student distillation model (ViT-Base, 94M params) that learns to match features from a pretrained teacher; and **Hiera** (Ryali et al., 2023), a hierarchical vision transformer (79M params) with multi-scale mask units. For all models, we compute prediction loss using sliding windows: given C context frames, the model predicts representations/pixels for remaining frames.

2.3 VIDEO DATASET

We generate 380 physics simulations using PyBullet (Coumans & Bai, 2016), shown in Figure 2: 160 videos across four scenarios (*bouncing ball*, *pendulum* = low-dissipation; *falling objects*, *dominos* = dissipative) plus 220 videos with varying restitution/damping for continuous probing. Each video: 48 frames at 256×256. Bouncing ball uses restitution 0.9 (near-elastic); pendulum uses zero damping.

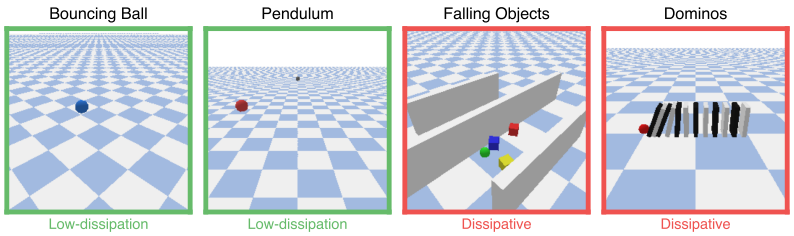

Figure 2: Physics simulation scenarios. Low-dissipation processes (green) look plausible when time-reversed; dissipative processes (red) do not; reversed falling objects would spontaneously jump upward.

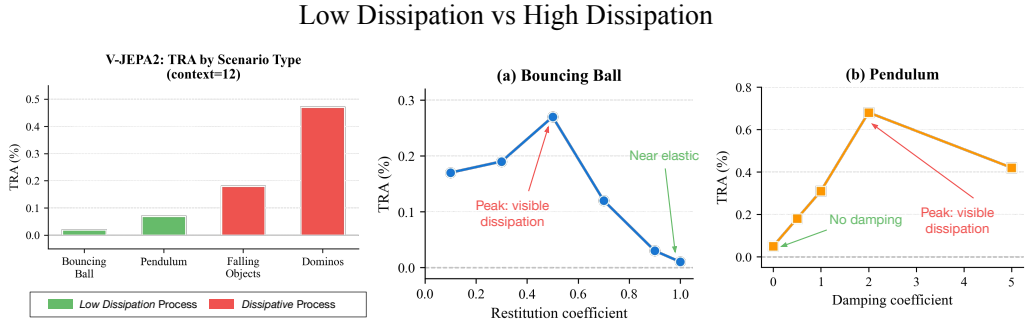

Figure 3: V-JEPA2 TRA reveals graded sensitivity to dissipation. **Middle + Right:** TRA varies continuously with physical parameters, peaking where energy loss remains visible throughout the video duration. **Left:** Discrete scenarios validate this pattern—low-dissipation processes (green) cluster near zero while dissipative ones (red) show elevated TRA.

3 Results

3.1 TRA Tracks Visible Irreversibility

To test whether TRA captures graded irreversibility, we varied physical parameters: restitution (0.1–1.0) for bouncing balls and damping (0.0–5.0) for pendulums. Figure 3 (right) reveals a non-monotonic pattern: TRA peaks at intermediate values. The bouncing ball shows maximum TRA at restitution 0.5 (+0.27%), falling to near zero at the elastic extreme (restitution 1.0, +0.01%) and reduced at restitution 0.1 (+0.17%), where motion stops within a few frames. The pendulum peaks at damping 2.0 (+0.68%).

This reflects that TRA measures *visible* irreversibility: at high dissipation, motion stops quickly leaving a symmetric stationary scene; at low dissipation, the process is genuinely reversible. At restitution 0.5, the ball bounces throughout all 48 frames with visibly decreasing amplitude; time-reversing this creates an implausible video of a ball spontaneously gaining energy. V-JEPA2 responds to visual evidence of ongoing energy dissipation rather than abstract physical categories.

3.2 V-JEPA2 Exhibits Physically Appropriate TRA

These findings validate across discrete physics scenarios (Figure 3, left; Table 1). V-JEPA2 shows a clear distinction between low-dissipation and dissipative processes: low-dissipation processes have near-zero TRA (+0.04%), an order of magnitude smaller than dissipative processes (+0.32%); the group difference is highly significant ($p < 0.001$).

The bouncing ball (restitution 0.9, near the elastic extreme) shows essentially zero TRA (+0.02%, $p = 0.33$). In contrast, dominos shows the strongest signal (+0.47%), and falling objects shows robust positive TRA (+0.18%)— both using lower restitution values in the regime where TRA is

Table 1: Model comparison at context=8. V-JEPA2 shows physically appropriate TRA; VideoMAE is inverted; random baseline confirms learning is required. †MVD and ‡Hiera rows are retained from the workshop version but are not valid evidence about their objectives due to evaluation limitations (Appendix A, Appendix M).

Model	Objective	Low-dissip.	Dissipative	Diff.	p-value
V-JEPA2	Latent pred.	+0.03%	+0.22%	+0.20%	<0.001***
VideoMAE	Pixel recon.	−0.07%	−0.28%	−0.21%	<0.001***
MVD†	Distillation	+0.00%	+0.00%	+0.00%	0.74 ns
Hiera‡	Hier. MAE	+0.19%	−0.09%	−0.29%	0.22 ns
Random	None	≈0%	≈0%	≈0%	0.56 ns

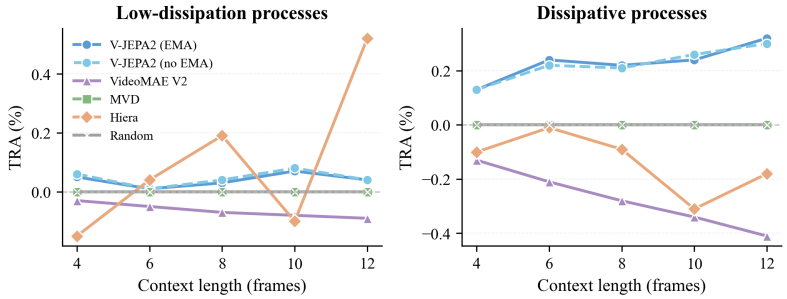

Figure 4: TRA across context lengths for all models. V-JEPA2 (both EMA and no-EMA variants) shows consistent positive TRA for dissipative processes; VideoMAE shows inverted (negative) TRA; MVD and Random show near-zero TRA; Hiera shows inconsistent results.

elevated. A randomly initialized V-JEPA2 shows negligible TRA for all conditions (<0.01%), confirming that meaningful TRA sensitivity emerges from learned representations rather than model architecture. The effect is consistent across context lengths 6–12.

3.3 VideoMAE Shows Inverted TRA

Remarkably, VideoMAE exhibits the opposite pattern from V-JEPA2. All TRA values are negative, meaning the model finds forward videos harder to predict than reversed ones. At context length 12, low-dissipation processes show TRA of -0.09% while dissipative processes show -0.41% ($p < 0.001$), the inverse of what physics would predict. This pattern strengthens with longer context (from -0.10% at context 4 to -0.40% at context 14), suggesting a systematic bias rather than noise.

4 Discussion

Why does VideoMAE show inverted TRA? We hypothesize pixel reconstruction encourages learning low-level statistical regularities (motion blur, lighting gradients) that are easier to extrapolate in reverse, without corresponding to physical understanding. JEPA's latent prediction may filter out such artifacts.

What do the MVD and Hiera rows show? Post-publication auditing revealed that neither row supports conclusions about its training objective. The public MVD checkpoint contains encoder weights only, so our evaluation decoder was effectively untrained — near-zero TRA is exactly what an untrained decoder produces, as the random baseline confirms. The Hiera evaluator relies on randomly sampled mask units rather than future-frame prediction, and the reported numbers additionally predate mask-RNG pinning, so forward and reversed runs used different masks. We retain both rows for continuity with the workshop version, annotated accordingly (Appendix M).

Limitations. Our study uses synthetic physics simulations. On Something-Something V2 (real human-performed actions), human action timing confounds TRA; however, temporally cropping to isolate physics-only segments recovers the expected pattern (Appendix L).

5 Conclusion

We introduced Temporal Reversal Asymmetry (TRA) as a metric for probing whether world models respect the thermodynamic arrow of time. V-JEPA2 exhibits the predicted pattern — positive TRA that grades continuously with visible dissipation — while VideoMAE shows a robust inverted pattern. This latent-prediction versus pixel-reconstruction contrast suggests the training objective shapes whether a model encodes temporal causality; broader claims across architectures await valid MVD and Hiera evaluations (Appendix M). TRA offers a training-free, physics-grounded evaluation approach for world models.

References

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.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning, 2019.

Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning, 2016.

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.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles, 2023.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning, 2023.

APPENDIX

A IMPLEMENTATION DETAILS

V-JEPA2. We use the official ViT-Large checkpoint (303M parameters) with 3D rotary position embeddings (RoPE). Prediction loss is computed using the EMA target encoder with layer normalization applied to targets. We use a sliding window approach: window size of 16 frames with stride 2. For each window, we provide C context frames and compute L1 loss between predicted and target latent representations for the remaining frames.

VideoMAE. We use the ViT-Base checkpoint pretrained on Kinetics-400 with decoder depth 4. We compute pixel reconstruction loss using the same sliding window approach. The model reconstructs masked patches given visible context patches. *Correction:* the workshop version labeled this model "VideoMAE V2"; the evaluated checkpoint is VideoMAE (v1) ViT-B, Kinetics-400, 800-epoch pretraining, from the MCG-NJU `VideoMAE` model zoo. The citation has been corrected accordingly.

MVD. We use the ViT-Base checkpoint (94M parameters) pretrained on Kinetics-400 with ViT-Base teacher. Our released code evaluates a checkpoint-compatible MVD-style baseline using the VideoMAE architecture wrapper and normalized patch targets, rather than vendoring the full official MVD teacher-feature pipeline. *Correction:* the public checkpoint (`mvd_b_from_b_ckpt_399.pth`) contains encoder weights only; all decoder parameters in our wrapper remained randomly initialized. The MVD numbers in this paper therefore reflect an untrained decoder and are not evidence about the MVD objective.

Hiera. We use the Hiera-Base checkpoint (79M parameters) pretrained with MAE on Kinetics-400. Hiera uses a hierarchical architecture with multi-scale mask units. Because the public Hiera MAE interface exposes a mask ratio rather than a future-frame context mask, our released code uses a mask-ratio proxy matched to each context length: this measures masked-region inpainting, not future prediction. *Correction:* the reported Hiera numbers were additionally produced before the mask RNG was pinned, so forward and reversed evaluations sampled different random masks; the reported values contain mask-sampling noise.

Random baseline. We initialize V-JEPA2 with random weights (no pretraining) to verify that TRA sensitivity requires learned representations.

B DATASET GENERATION

All videos are generated using PyBullet physics engine at 256×256 resolution with 48 frames per video.

Bouncing ball. A sphere is dropped onto a plane. Default restitution coefficient: 0.9 (near-elastic). For continuous probing, we vary restitution from 0.1 to 1.0.

Pendulum. A rigid body swings from a fixed pivot point. Default damping: 0.0 (no energy loss). For continuous probing, we vary damping from 0.0 to 5.0.

Falling objects. Multiple rigid bodies fall and collide, settling into a pile. *Correction:* the workshop version stated restitution 0.3; the generator does not set a restitution coefficient for this scene (PyBullet's default contact parameters apply).

Dominos. A row of rectangular blocks topples sequentially after the first is pushed. *Correction:* the workshop version stated restitution 0.2; as above, no restitution coefficient is set for this scene.

C V-JEPA2 RESULTS ACROSS CONTEXT LENGTHS

Table 2 shows V-JEPA2 TRA across all tested context lengths. The effect is robust from context 6–12, with context 12 showing the strongest separation.

Table 2: V-JEPA2 (EMA) TRA by context length.

Context	Low-dissip.	Dissipative	Difference	p-value
4	+0.05%	+0.13%	+0.08%	0.037*
6	+0.01%	+0.24%	+0.23%	<0.001***
8	+0.03%	+0.22%	+0.20%	<0.001***
10	+0.07%	+0.24%	+0.17%	<0.001***
12	+0.04%	+0.32%	+0.28%	<0.001***
14	+0.09%	−0.11%	−0.21%	<0.001***

D V-JEPA2 NO-EMA RESULTS

Table 3 shows results using the online encoder instead of the EMA target encoder. The pattern is consistent, confirming robustness.

Table 3: V-JEPA2 (no EMA) TRA by context length.

Context	Low-dissip.	Dissipative	Difference	p-value
4	+0.06%	+0.13%	+0.07%	0.050 ns
6	+0.01%	+0.22%	+0.21%	<0.001***
8	+0.04%	+0.21%	+0.17%	<0.001***
10	+0.08%	+0.26%	+0.18%	<0.001***
12	+0.04%	+0.30%	+0.26%	<0.001***
14	+0.09%	−0.11%	−0.20%	<0.001***

E VIDEOMAE FULL RESULTS

Table 4 shows VideoMAE TRA across all context lengths. All values are negative, and the inverted pattern strengthens with longer context.

Table 4: VideoMAE TRA by context length. All values negative (inverted pattern).

Context	Low-dissip.	Dissipative	Difference	p-value
4	−0.03%	−0.13%	−0.10%	<0.001***
6	−0.05%	−0.21%	−0.16%	<0.001***
8	−0.07%	−0.28%	−0.21%	<0.001***
10	−0.08%	−0.34%	−0.26%	<0.001***
12	−0.09%	−0.41%	−0.32%	<0.001***
14	−0.10%	−0.50%	−0.40%	<0.001***

F RANDOM BASELINE RESULTS

Table 5 shows results for randomly initialized V-JEPA2 (no pretraining). TRA values are negligible for all conditions: mean |TRA| is below 0.004% and no single video exceeds 0.017%, one to two orders of magnitude below the trained-model effects. While some p-values reach nominal significance (and, because within-group variance is correspondingly tiny, standardized effect sizes are not uniformly small — a correction to the workshop version, which claimed Cohen's $d < 0.05$), the absolute magnitudes confirm that meaningful temporal asymmetry sensitivity requires learned representations.

Table 5: Random baseline (untrained V-JEPA2). TRA values are negligible (<0.01%) for all conditions. Some p-values reach significance due to large sample size, but effect sizes are negligible.

Context	Low-dissip.	Dissipative	Difference	p-value
4	+0.00%	+0.00%	+0.00%	0.02*
6	+0.00%	+0.00%	+0.00%	0.05*
8	+0.00%	+0.00%	+0.00%	0.56 ns
10	+0.00%	+0.00%	+0.00%	0.02*
12	+0.00%	+0.00%	+0.00%	<0.001***
14	+0.00%	+0.00%	+0.00%	<0.001***

G MVD Full Results

Table 6 shows MVD (Masked Video Distillation) TRA across all context lengths. TRA values are essentially zero, indicating that teacher-student distillation does not encode temporal asymmetry.

Table 6: MVD TRA by context length. All values near zero.

Context	Low-dissip.	Dissipative	Difference	p-value
4	+0.00%	+0.00%	+0.00%	0.006**
6	+0.00%	+0.00%	+0.00%	0.047*
8	+0.00%	+0.00%	+0.00%	0.74 ns
10	+0.00%	−0.00%	−0.00%	0.07 ns
12	+0.00%	−0.00%	−0.00%	0.015*
14	+0.00%	+0.00%	−0.00%	0.97 ns

H Hiera Full Results

Table 7 shows Hiera (Hierarchical MAE) TRA across all context lengths. Results are inconsistent across context lengths and mostly not statistically significant, suggesting Hiera does not reliably encode temporal asymmetry.

Table 7: Hiera TRA by context length. Results inconsistent and mostly not significant.

Context	Low-dissip.	Dissipative	Difference	p-value
4	−0.15%	−0.10%	+0.05%	0.80 ns
6	+0.04%	−0.01%	−0.06%	0.77 ns
8	+0.19%	−0.09%	−0.29%	0.22 ns
10	−0.10%	−0.31%*	−0.21%	0.32 ns
12	+0.52%**	−0.18%	−0.70%	0.019*
14	−0.01%	+0.06%	+0.07%	0.84 ns

I Per-Scenario Breakdown

Table 8 shows detailed per-scenario results for V-JEPA2 (EMA) at context length 12.

Note. The released PyBullet generator used for these JSONs did not introduce seed-dependent variation in the domino setup: all 40 domino videos are identical, which is why the t-statistic is infinite (zero within-scenario variance). The dominos row is therefore a deterministic scenario check (effectively $n = 1$), not 40 independent samples, and every group-level test that includes dominos contains 40 copies of one measurement. Appendix M reports a sensitivity analysis excluding dominos. The repository includes a randomized domino generator for reruns while preserving the legacy setting for exact reproduction of the reported tables.

Table 8: V-JEPA2 TRA by individual scenario (context=12).

Scenario	Type	TRA	t-stat	p-value
Bouncing ball	Low-dissip.	+0.02%	0.99	0.33 ns
Pendulum	Low-dissip.	+0.07%	3.29	0.002**
Falling objects	Dissipative	+0.18%	4.10	<0.001***
Dominos	Dissipative	+0.47%	∞	<0.001***

J Continuous Probe Data

Table 9 shows TRA as a function of physical parameters. Both scenarios exhibit non-monotonic patterns: TRA peaks at intermediate values where energy dissipation is visually apparent throughout the video, approaching zero at both extremes (perfectly elastic/undamped, or motion stops too quickly to observe dissipation).

Table 9: TRA vs. physical parameters (context=12). Peak values in bold.

Bouncing Ball			Pendulum		
Restitution	TRA	Note	Damping	TRA	Note
0.1	+0.17%	Stops quickly	0.0	+0.05%	No damping
0.3	+0.19%		0.5	+0.18%	
0.5	**+0.27%**	Peak	1.0	+0.31%	
0.7	+0.12%		2.0	**+0.68%**	Peak
0.9	+0.03%		5.0	+0.42%	Stops quickly
1.0	+0.01%	Elastic			

K Context Length 14 Anomaly

At context length 14 (out of 16 total frames in each window), V-JEPA2's TRA pattern inverts: dissipative processes show *negative* TRA (-0.11%) while low-dissipation processes remain slightly positive (+0.09%). We hypothesize that this anomaly occurs because:

1. With only 2 frames left to predict, the model's task at hand changes significantly
2. Short prediction horizons may not provide enough temporal extent to distinguish dissipative dynamics
3. The prediction task becomes more about single-frame extrapolation than sequence modeling

We exclude context=14 from our main analysis but report it here for completeness. The consistent pattern across context lengths 6–12 suggests this anomaly reflects a boundary condition rather than a fundamental limitation of TRA.

L Real-World Video: Something-Something V2

To test generalization beyond synthetic physics, we evaluated TRA on Something-Something V2 (SSv2), a dataset of human-performed actions. We selected 250 "reversible" actions (spinning, rolling) and 250 "irreversible" actions (dropping, pouring, tearing); after decode failures, 496 videos enter the full-video analysis and 499 the cropped analysis.

Initial results. V-JEPA2 shows *negative* TRA for *all* SSv2 actions (group means between -0.59% and -0.76%), with no significant difference between reversible and irreversible categories at any context length ($p \geq 0.20$ everywhere; $p = 0.58$ at the context length of 8 used below). This contradicts the pattern observed on synthetic physics.

Figure 5: SSv2 temporal cropping. The first half of videos (red border) shows human hand manipulation; the second half (green border) shows physics-dominated motion after release. Cropping to the second half isolates the physics signal from human action timing.

Root cause: human action initiation. We hypothesized that human action initiation (hands entering the frame to manipulate objects) creates a temporal asymmetry unrelated to physics. In forward videos, predicting *when* a hand will appear and initiate action is difficult. In reversed videos, the action "undoes" itself predictably. This human-induced asymmetry may dominate over the smaller physics-based asymmetry.

Temporal cropping experiment. To isolate physics from human agency, we temporally cropped videos using a simple heuristic: the **second half** of SSv2 videos typically contains physics-dominated motion after the hand releases the object (Figure 5). This is an approximate method; the exact transition from manipulation to physics varies across videos. However, it provides a proof-of-concept that the physics signal exists when human action is reduced.

Results are shown in Table 10. When using only the second half, a significant difference emerges: irreversible actions show near-zero TRA (-0.04%, ns) while reversible actions remain negative (-0.29%, $p < 0.001$). The difference is significant ($p = 0.0008$).

Table 10: SSv2 cropped experiment (V-JEPA2, context=8). Cropping to second half reveals physics signal.

Crop	Reversible	Irreversible	Diff.	p-value
First half (hands)	-0.30%***	-0.35%***	-0.05%	0.51 ns
Second half (physics)	-0.29%***	-0.04% ns	+0.25%	**0.0008*** **
Middle 50%	-0.25%***	-0.16%***	+0.09%	0.21 ns

This confirms that TRA measures physics understanding when physical dynamics dominate, but human action timing confounds the signal in videos with visible agents. Future work could develop more sophisticated methods to segment human manipulation from physics-dominated phases, such as hand detection or motion-based segmentation. Alternatively, datasets with minimal human presence (surveillance footage, nature videos, robotic manipulation) may provide cleaner signals for evaluating physics understanding.

M Errata and Changes from the Workshop Version

This corrected version follows a post-publication audit in which every reported number was independently recomputed from the released per-video results and the evaluation code was re-examined. The released result files and audit are available in the accompanying repository. Changes from the version presented at the workshop:

1. **MVD row invalidated.** The public MVD checkpoint contains encoder weights only; the evaluation decoder was randomly initialized (Appendix A). All MVD-based claims, including the discussion of distillation objectives, are withdrawn. The row is retained, annotated, for continuity.
2. **Hiera row downgraded.** The Hiera evaluator measures random-mask inpainting rather than future prediction, and the reported numbers predate mask-RNG pinning, so forward and reversed runs used different random masks. Hiera-based claims are withdrawn.
3. **Conclusion softened.** "Training objective matters more than architecture" is reduced to the supported two-model contrast between latent prediction (V-JEPA2) and pixel reconstruction (VideoMAE).
4. **Dominos pseudo-replication disclosed.** All 40 domino videos are identical (Appendix I). Table 11 reports the group comparison with dominos excluded (dissipative = falling objects only, $n = 40$, vs. low-dissipation, $n = 80$): the V-JEPA2 separation survives at context 4–8 and 12 but not at context 10; the VideoMAE inverted pattern is unaffected at every context length. A full rerun on a corrected dataset with randomized dominos confirms the separation at every context 4–12 (Table 12).
5. **Loss function.** The workshop version's Appendix A said MSE; all released results use L1 prediction loss.
6. **Physics parameters.** The workshop version claimed restitution 0.3 (falling objects) and 0.2 (dominos); neither scene sets a restitution coefficient (Appendix B).
7. **Model identification corrected.** The pixel-reconstruction model was labeled "VideoMAE V2" throughout the workshop version; the evaluated checkpoint is in fact VideoMAE (v1) ViT-B (Kinetics-400, 800-epoch pretraining, MCG-NJU model zoo), and the reference has been corrected to the VideoMAE paper. No numbers change; the inverted-TRA finding attaches to VideoMAE v1.
8. **Statistical corrections.** Table 3, context 4: $p = 0.050$ (ns), previously misreported as 0.048*. Section 3.2 no longer describes the low-dissipation group as "not significant" (its one-sample $p = 0.004$; Table 8 itself marks the pendulum row **). Appendix F replaces the incorrect "Cohen's $d < 0.05$" claim with absolute magnitudes. Appendix L replaces "$p > 0.5$" with the exact range ($p \geq 0.20$) and reports achieved sample sizes.

Table 11: Sensitivity analysis: low-dissipation ($n = 80$) vs. dissipative excluding dominos (falling objects only, $n = 40$). Independent t-test on per-video TRA, matching the main analysis.

	V-JEPA2 (EMA)		VideoMAE	
Context	Diff.	p-value	Diff.	p-value
4	+0.34%	<0.001***	−0.15%	<0.001***
6	+0.35%	<0.001***	−0.23%	<0.001***
8	+0.24%	<0.001***	−0.30%	<0.001***
10	+0.01%	0.78 ns	−0.38%	<0.001***
12	+0.13%	<0.001***	−0.46%	<0.001***
14	−0.20%	<0.001***	−0.52%	<0.001***

Corrected-dataset rerun. Beyond the exclusion analysis above, we regenerated the discrete benchmark with a randomized domino generator (40 distinct seeds; the other three scenarios reuse the released seed offsets) and re-evaluated the main models under the released protocol ($w = 16$, $s = 2$, L1 loss). Table 12 reports the dissipative vs. low-dissipation comparison ($n = 80$ per

group, as in the main analysis). The headline pattern is unchanged: V-JEPA2's separation is positive and significant at every context 4–12 — including context 10, where the dominos-excluded sensitivity analysis was inconclusive — and reverses at context 14, as in the released results; VideoMAE remains inverted at every context length. The randomized dominos now contribute genuine between-video variance (per-scenario one-sample $t = 9.04$ at context 12, mean TRA $+0.25\%$), and the randomly initialized control stays at $|\text{TRA}| \leq 0.01\%$ throughout. Full per-model results are released alongside the code.

Table 12: Corrected-dataset rerun (randomized dominos): low-dissipation ($n = 80$) vs. dissipative ($n = 80$), independent t-test on per-video TRA, matching the main analysis.

	V-JEPA2 (EMA)		VideoMAE	
Context	Diff.	p-value	Diff.	p-value
4	+0.12%	0.003**	−0.12%	<0.001***
6	+0.21%	<0.001***	−0.17%	<0.001***
8	+0.17%	<0.001***	−0.24%	<0.001***
10	+0.12%	<0.001***	−0.29%	<0.001***
12	+0.17%	<0.001***	−0.33%	<0.001***
14	−0.22%	<0.001***	−0.40%	<0.001***