PHYTRAJ: MOTION AS A QUERY FOR VISUAL PHYSICS AUDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

A generated video can look convincing while an object's motion makes the event physically implausible. Detecting such failures calls for relating motion to its visual context. We introduce PHYTRAJ, which uses estimated object trajectories as queries into frozen video representations for visual physics auditing. Metadata-derived object names guide tracking; a trajectory encoder turns each motion history into a query that retrieves context from VideoMAE tokens. The resulting motion–video representations are contextualized across objects and aggregated into a clip-level violation score. We evaluate this approach on 2,334 human-labeled clips with available object tracks from one video generator, spanning six physical-rule families. Labels describe the behavior visible in the rendered clips, and object compositions are held out across splits. In an exploratory representation study on 353 test clips, the full trajectory–video model achieves 0.758 AUROC, compared with 0.683 for the video-only control. A complementary study on frozen fusion features finds that simple pooling remains competitive with structured aggregation. These findings support trajectory-conditioned video representations for detecting visible physical violations, with motion serving as an object-centered interface to visual context. The artifact is available at https://anonymous.4open.science/r/PhyTraj-B046/.

1 INTRODUCTION

A ball can be rendered with convincing texture and smooth frame-to-frame motion while passing through a solid barrier. The failure is not located in one frame alone, but in how an object moves within its visual context. As video generators are increasingly studied as candidate models of the physical world, detecting such failures becomes an evaluation problem that visual realism alone cannot solve (Bansal et al., 2025; Motamed et al., 2025). We study this problem as a representation question: *can explicit object motion help a model audit visible physical violations beyond pretrained video features?*

Object trajectories are a tempting interface for this question. Continuity, support, and collision response are naturally stated in terms of objects and their motion, and object-centric models have long used entities and relations to represent physical dynamics (Battaglia et al., 2016; Wu et al., 2023). For generated video auditing, however, trajectories are only visual estimates. They can be missing, noisy, or geometrically imperfect, and they do not by themselves encode the surfaces, contact cues, occlusions, or surrounding scene context that may make a motion plausible or implausible. Pretrained video tokens contain richer spatiotemporal context, but do not force the readout to organize that context around the motion of a particular object. The design problem is therefore how to let estimated motion ask for the visual evidence needed to interpret it.

We introduce PHYTRAJ, a trajectory-conditioned representation for *visual physics auditing*: predicting whether the behavior visible in a clip violates the corpus annotation rules. The central idea is to use motion as a query. Metadata-derived object names guide the upstream perception stage, which estimates object trajectories; the final classifier receives neither the generation prompt nor the physical-rule family. A trajectory encoder maps estimated positions, velocities, accelerations, and observation-validity indicators into object tokens. Each object token then queries frozen VideoMAE spatiotemporal tokens through cross-attention, retrieving visual context conditioned on that object's motion. The fused object representations are contextualized by a relation encoder before a clip-level

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.

score is produced. The resulting representation connects an object's estimated motion with the scene evidence used to assess its physical plausibility.

Existing work provides complementary ways to evaluate generated physics. VideoPhy, VideoPhy-2, and PhyGenBench evaluate physical commonsense through human judgments and automatic evaluators (Bansal et al., 2025; 2026; Meng et al., 2025). VAMP incorporates motion plausibility, and PDI-Bench uses object tracking and reconstructed geometry to compute consistency residuals (Wang et al., 2024; Wu et al., 2026). PHYTRAJ follows this line of work in treating the rendered video as the object of evaluation, but focuses on a learned representation that lets estimated object motion condition access to pretrained video features. The key comparison is therefore between video-only features and trajectory-conditioned video representations under a controlled visible-violation task.

Our evaluation uses 2,334 human-labeled clips retained after trajectory availability filtering and class balancing. The clips come from one generator under fixed-camera generation constraints and span six physical-rule families. Human labels refer to the events visible in the clips rather than the events the prompts intended to generate. Train, validation, calibration, and test splits hold out normalized object compositions; the final test contains 353 clips from 45 groups. This design tests generalization to new object compositions within the controlled corpus.

Our experiments examine the fusion representation, its motion inputs, and the readout used for prediction. In an exploratory representation study, full trajectory–video fusion reaches 0.758 AUROC versus 0.683 for the video-only control. We then compare position-only, image-plane, and temporally shuffled trajectories to characterize sensitivity to the motion input. A complementary study freezes the fusion features and compares eight readouts: mean/max linear pooling reaches 0.760 AUROC, close to the structured evidence head at 0.762. These results show that useful violation predictions are accessible through lightweight readouts of the fused representation. Comparisons with existing physics evaluators and a separate calibration analysis place these findings in the broader evaluation context. Section 6.1 explains the two study protocols and the exploratory status of the representation comparisons.

Our contributions are:

- a trajectory-conditioned video representation in which estimated object motion queries frozen spatiotemporal video tokens, producing contextual object representations for clip-level physics auditing;
- a controlled visible-violation setting with 2,334 human-labeled generated videos, six physical-rule families, metadata-derived object queries, and object-composition-held-out evaluation; and
- an empirical study of motion representations and readout complexity, comparing 3D and image-plane trajectories, temporal-order controls, and lightweight versus structured aggregation, with paired uncertainty estimates and probability-reliability analysis.

2 RELATED WORK

Physics evaluation for generated video. General video-quality suites measure perceptual and semantic properties (Huang et al., 2024). VideoPhy and PhyGenBench make physical commonsense an explicit evaluation target (Bansal et al., 2025; Meng et al., 2025); VideoPhy-2 and VideoScore2 extend this line with action-centric and multi-dimensional evaluator protocols (Bansal et al., 2026; He et al., 2025). Recent benchmarks broaden the phenomena, evidence types, and model families being tested (Gu et al., 2026; Upadhyay et al., 2026; Lin et al., 2026; Pothiraj et al., 2026; Zhang et al., 2026; Ba et al., 2026; Chow et al., 2025). These works often pair a generated clip with a prompt, rubric, question graph, or model-generated score. Our detector instead uses supervised human labels of the visible clip behavior and does not receive the prompt or rule family. We run three existing evaluator protocols on our test split as zero-shot transfer baselines, not as matched-training ablations.

Physics-IQ asks whether video generators capture physical principles across domains (Motamed et al., 2025), and other work studies broader physical law generalization or conservation-style diagnostics (Tragoudaras et al., 2026; Kang et al., 2025). Our object-composition split is narrower: it tests one form of compositional separation within the proposed dataset.

Object-centric motion and anomaly evidence. Entity-centric dynamics models motivate persistent objects and relations as useful units of physical reasoning (Battaglia et al., 2016; Wu et al., 2023). Our setting is a discriminative audit of an observed video, where monocular trajectories are estimated evidence rather than simulator truth. VideoMAE supplies dense visual context (Tong et al., 2022), while SAM 3 and DetAny3D support prompt-conditioned masks and estimated 3D states (Carion et al., 2025; Zhang et al., 2025); retaining both streams is important because tracks can fail under occlusion, camera ambiguity, or identity switches.

VAMP uses tracked motion to evaluate video generation (Wang et al., 2024). PDI-Bench is the closest object-centric 3D precedent: it audits generated videos with tracking, camera/depth estimates, and deterministic geometric residuals (Wu et al., 2026). PHYTRAJ studies a complementary approach: it learns a supervised video-and-track decision from human labels, while PDI-Bench uses hand-designed residual checks. Phys-AD studies physics-grounded anomaly detection in real-world industrial videos (Li et al., 2025), a different source-domain setting from the generated-video benchmark studied here.

3 PROBLEM FORMULATION

Let a video be $V = (I_1, \ldots, I_T)$, and let q denote the object-concept queries supplied to the promptable perception stage. Each clip has a binary label $y \in \{0, 1\}$: $y = 0$ denotes motion judged physically consistent under the annotation rule, and $y = 1$ denotes an observable violation. The learning objective is

$$f_\theta : (V, q) \mapsto \hat{p}(y = 1 \mid V, q). \tag{1}$$

The classifier receives neither the generation prompt nor the physical-rule category. The score supports threshold-free ranking, thresholded detection, and post-hoc calibration; AUROC is the primary final discrimination metric, with AP, balanced accuracy, NLL, Brier score, ECE, and risk–coverage behavior reported as complementary views.

The label is a judgment of the physical behavior visible in the rendered clip. It is not inferred from prompt text, file names, source metadata, or the generation intent. This distinction matters because a prompt can ask for a physically valid event while the video renders a violation, or conversely ask for an impossible event that is not visibly realized.

The object-query input q specifies which objects to track. In this corpus, it is a per-video list of object names drawn from source metadata and normalized to a 30-name vocabulary. These names configure the perception pipeline, giving the system access to metadata-specified object identity. The classifier then operates on the video features and estimated trajectories. Appendix A describes this input requirement.

Object-centric evidence. For up to K objects, the perception pipeline estimates a trajectory $\tau_k = (\mathbf{m}_{k,1}, \ldots, \mathbf{m}_{k,T})$, where

$$\mathbf{m}_{k,t} = [\mathbf{p}_{k,t}; \mathbf{v}_{k,t}; \mathbf{a}_{k,t}; r_{k,t}] \in \mathbb{R}^{10}. \tag{2}$$

Here $\mathbf{p}$, $\mathbf{v}$, and $\mathbf{a}$ are estimated 3D position, velocity, and acceleration, and r marks validity. These are visual estimates, not privileged simulator states; they can be noisy, missing, or camera-frame dependent. We therefore combine them with dense video tokens instead of treating reconstructed trajectories as ground truth.

Scope of violations. A clip is in scope when the relevant evidence is visible and belongs to one of the annotated rule families: gravity/projectile behavior, collision response, contact and support, friction, inertia/external-force consistency, or trajectory continuity. Hidden forces, unknown masses or materials, off-screen agents, and domain-specific exceptions are not assumed merely to rescue an implausible clip. Cases requiring such unobserved information fall outside the intended scope. The current study does not establish judgments about fluids, aerodynamics, deformable bodies, articulated biomechanics, or strongly moving-camera scenes.

Evaluation unit. Multiple generations from related prompts can share objects, layouts, and generator artifacts, so random clip splits can reward template memorization. We group clips by their normalized object composition—the set of queried object types in a clip—and assign each group

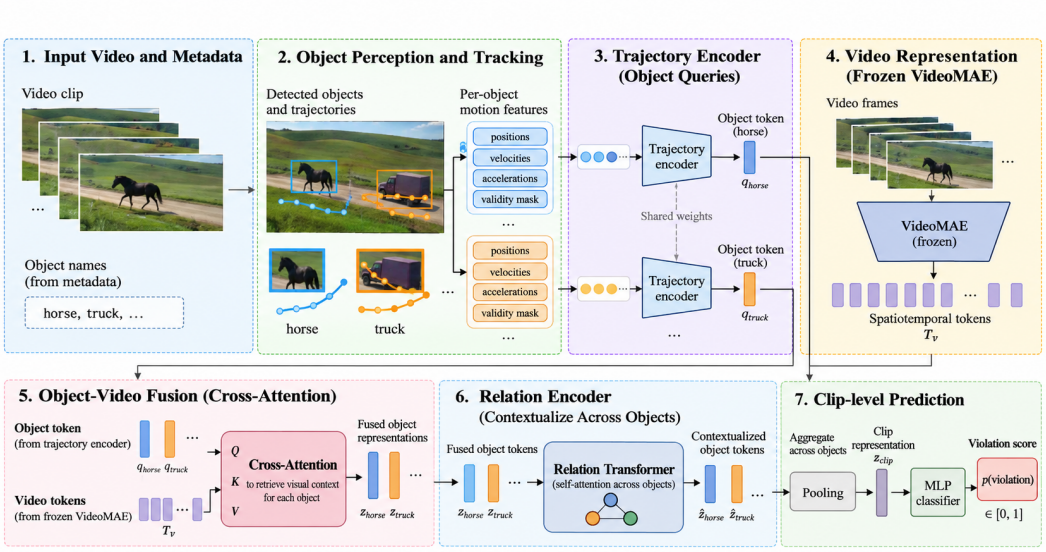


Figure 1: Motion as a query. Estimated object histories provide queries for reading frozen video features. Cross-attention and fusion produce object tokens, which exchange information before clip-level prediction. Object names supplied to perception come from metadata. The separate frozen-feature readout protocol is detailed in Appendix B.

to exactly one of train, validation, calibration, or test. This evaluates prediction on unseen object compositions within the corpus. Section 5 gives the split sizes, and Section 6.1 describes their experimental use.

4 PHYTRAJ: MOTION AS A QUERY

Physical plausibility depends on how an object moves within its visual context. For example, an abrupt stop may be consistent with a visible collision, while a similar motion without visible contact requires a different judgment. Estimated trajectories make object motion explicit, but compress away scene information and inherit errors from monocular perception. PHYTRAJ uses these estimates to organize access to pretrained video features: each trajectory becomes a query that retrieves visual context, forming an object representation that combines motion and video information. The representations are then contextualized across objects and aggregated into a clip-level violation score (Figure 1).

4.1 OBSERVING OBJECT MOTION

Given a video V and metadata-derived object queries q, the perception pipeline estimates up to $K = 10$ object tracks using SAM3 and DetAny3D (Section 5). At each of $T = 16$ sampled times, a track stores camera-frame position, finite-difference velocity and acceleration, and an observation-validity indicator:

$$\mathbf{m}_{k,t} = [\mathbf{p}_{k,t}; \mathbf{v}_{k,t}; \mathbf{a}_{k,t}; r_{k,t}] \in \mathbb{R}^{10}. \tag{3}$$

The object-validity mask has $v_k = 1$ when at least one observation of object k is valid. These states are visual estimates; the model uses the video stream to retain context beyond the estimated centers. The final classifier receives neither the generation prompt nor the physical-rule family; object names are used by the upstream perception stage. Appendix B details the input tensors and motion controls; Appendix C describes trajectory extraction.

4.2 TURNING TRAJECTORIES INTO QUERIES

A shared temporal Transformer encodes each object's sampled history. Let ϕ_m denote a learned state projection with layer normalization, GELU, and dropout, and let $\mathbf{e}_t$ be a learned temporal

embedding. We form

$$\mathbf{A}_k = E_{\text{traj}}\big((\phi_m(\mathbf{m}_{k,t}) + \mathbf{e}_t)_{t=1}^{T}; r_{k,1:T}\big)\,, \tag{4}$$

$$\mathbf{o}_k = v_k\,\mathrm{LN}\left(\frac{\sum_t r_{k,t}\mathbf{A}_{k,t}}{\max(1, \sum_t r_{k,t})}\right). \tag{5}$$

Invalid observations are masked in temporal attention and averaging. The resulting token summarizes an object's estimated motion history and serves as its query into the video representation. Fully missing tracks are excluded from object aggregation; the implementation handles all-missing masks before attention to avoid undefined outputs.

4.3 Retrieving Visual Context with Motion Queries

The video stream maps the sampled frames, resized to 224×224, to spatiotemporal patch tokens using frozen VideoMAE. A trainable projection ϕ_v, comprising a linear layer, layer normalization, and GELU, maps the tokens to the same dimension as the trajectory queries:

$$\mathbf{X} = E_{\text{vid}}(V), \qquad \mathbf{U} = \phi_v(\mathbf{X}) \in \mathbb{R}^{N\times d}. \tag{6}$$

Each trajectory token then reads the video through multi-head attention:

$$\mathbf{c}_k = \mathrm{MHA}(Q = \mathbf{o}_k, K = \mathbf{U}, V = \mathbf{U}), \tag{7}$$

$$\mathbf{z}_k = v_k F_{\text{fuse}}([\mathbf{o}_k; \mathrm{LN}(\mathbf{o}_k + \mathbf{c}_k)])\,. \tag{8}$$

Here MHA includes its learned query, key, and value projections, and F_{fuse} maps the concatenated features back to dimension d. The trajectory determines the query, while video tokens supply the context being read. The fused token retains both the explicit motion representation and a trajectory-conditioned summary of the video. This design allows scene information absent from object-center tracks to enter the decision; it does not require the reconstructed coordinates to specify a complete physical state. The attention operates over video tokens without a box-based spatial mask, allowing each motion query to access context throughout the clip.

4.4 Contextualizing Objects and Scoring the Clip

An object-relation Transformer lets the fused tokens exchange information under the object-validity mask:

$$\mathbf{H} = E_{\text{rel}}(\mathbf{z}_{1:K}; \mathbf{v}), \qquad \mathbf{h}_k \in \mathbb{R}^d. \tag{9}$$

This provides context across tracked objects before the video-level decision. It does not require an annotated interaction graph. During encoder training, a classifier pools valid tokens by their mean and maximum,

$$\mathbf{h}_{\text{clip}} = \left[\frac{\sum_k v_k\mathbf{h}_k}{\max(1, \sum_k v_k)}\;;\; \max_{k:v_k=1} \mathbf{h}_k\right], \qquad \hat{p} = \sigma(g_{\text{train}}(\mathbf{h}_{\text{clip}})), \tag{10}$$

where g_{train} is an MLP producing one logit. The prediction concerns a visible violation somewhere in the clip; the model has no object-level or time-localized supervision. In the frozen-feature study, alternative readouts use the same $\mathbf{H}$ and mask, with some also using the global projected video feature $\bar{\mathbf{u}} = N^{-1}\sum_n \mathbf{U}_n$.

4.5 Learning and Controlled Readout Evaluation

The encoder and its training classifier are fit under video-level supervision, with the VideoMAE backbone frozen. For a training minibatch $\mathcal{B}$, the implemented objective is binary cross-entropy with logits:

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} [y_i \log\sigma(s_i) + (1-y_i)\log(1-\sigma(s_i))]\,, \tag{11}$$

where s_i is the classifier logit. Checkpoints are selected by validation NLL. We then distinguish two experimental questions. Representation controls retrain the encoder for each motion-input variant and compare them under a shared readout family. Readout controls freeze a trained encoder, discard its training classifier, and fit eight specified heads on the same cached features, also using binary cross-entropy. This second comparison measures readout differences without changing the input representation. The full head definitions and cache protocol are in Appendix B; Section 6.1 details the two evaluation suites and their comparison protocols.

5 DIAGNOSTIC DATASET

Corpus construction. We build a controlled generated-video corpus for binary visible physical-consistency diagnosis. The six paper-level families are gravity and projectile motion, collision response, contact and support, friction, inertia and external-force consistency, and trajectory continuity; we exclude phenomena such as optics, fluids, heat, deformable bodies, ropes, smoke, and fireworks where object-center trajectories are unlikely to be sufficient evidence. Structured event templates specify the physical family, event stages, relevant objects, observability constraints, and plausible or implausible variants. A programmatic expansion over curated objects, materials, surfaces, environments, heights, distances, and camera angles is then verbalized into natural Chinese prompts by a language model while preserving the specified event and objects. The language model adds linguistic and scene diversity; it does not determine the final physical label.

All prompts require a fixed camera, one continuous shot, no cuts, no camera following or zoom, real-time motion, complete visibility of the main objects, and a simple background, and they avoid label-leaking phrases such as direct statements of "violate physics." After format, constraint, and duplicate checks, Seedance 2.0 generates 2,454 candidate videos. Appendix C details video properties, annotation, and filtering.

Human-observed labels and filtering. Prompt intent is metadata rather than ground truth: a prompt or `source_id` can intend plausible or violating behavior, but the video generator can fail to follow that intent or introduce new artifacts. Two annotators reviewed all 2,454 candidates for the physical behavior visible in the generated video, cross-validated the annotations, and jointly adjudicated ambiguous or disagreeing cases. The dataset records the final adjudicated label: 0 denotes physically plausible behavior and 1 denotes a physical violation. Inter-annotator agreement cannot be estimated from these final adjudicated labels alone.

The adjudicated candidate set is exactly balanced: 1,227 plausible clips and 1,227 violation clips. We retain a candidate only if sampled frames and tracking output exist and at least one queried object has a finite 3D center in at least one sampled frame. This removes 96 videos (36 label 0, 60 label 1), leaving 2,358 clips with 1,191 plausible and 1,167 violation clips. We then downsample 24 additional label-0 clips with random seed 42, yielding the final 2,334-video benchmark with 1,167 examples per class. The evaluation thus measures performance on clips with at least one available trajectory.

Object queries and trajectories. Each video is paired with normalized object queries used by the perception stage for open-vocabulary segmentation and tracking. In the 2,454 candidates, 450 videos have one query object and 2,004 have at least two; in the retained 2,334-video benchmark, 426 videos have one query object and 1,908 have at least two. The normalized query vocabulary contains 30 object names. Some raw object descriptions are merged for detection consistency, such as human roles to `person` and specific ball types to `ball`; 1,343 candidate examples have a normalized query list that does not exactly match the raw object field. Appendix C documents the query statistics and normalization examples.

For each retained video, the perception pipeline samples 16 frames, resizes them to 640×360, segments and tracks the normalized object queries with SAM3, and estimates camera-frame 3D object centers with DetAny3D. Model features are built from estimated center positions, finite-difference velocities and accelerations, and validity masks. Appendix C describes this preprocessing, including the interpretation of the estimated motion features.

Split and scope. All experiments use one fixed object-composition split over 290 normalized `object_combo` groups, with every clip sharing an object composition assigned to the same subset (Table 1). Validation selects checkpoints, calibration fits scalar temperatures, and test measures predictive performance on held-out object compositions. Section 6.1 explains the evaluation sequence for the frozen-feature and exploratory representation studies.

The benchmark is balanced overall but not within every physical family: 78.3% of gravity/projectile clips are violations, compared with 39.0% of trajectory-continuity clips (Table 2). These different family prevalences are relevant when interpreting aggregate metrics. The split evaluates new nor-

Table 1: Frozen object-composition split for training, model selection, temperature calibration, and final testing. The manifest hash and split report are recorded in Appendix D.

Split	Clips	Label 0	Label 1	Groups
Train	1,401	701	700	172
Validation	350	176	174	44
Calibration	230	115	115	29
Test	353	175	178	45
Total	2,334	1,167	1,167	290

Table 2: Composition of the 2,334 retained clips by paper-level physical family after canonical mapping. Label 0 is plausible and label 1 is a visible violation. The corpus is balanced overall but not within every family; Appendix C gives category-mapping caveats.

Physical family	Label 0	Label 1	Clips	Viol. %
Trajectory continuity	420	269	689	39.0
Contact / support	269	215	484	44.4
Gravity / projectile motion	89	322	411	78.3
Collision response	135	129	264	48.9
Friction / energy dissipation	142	120	262	45.8
Inertia / external-force consistency	112	112	224	50.0
Total	1,167	1,167	2,334	50.0

malized object compositions within this controlled generated-video distribution; Section E discusses family-correlated cues and evaluation across broader video sources.

6 Experiments

6.1 Protocol

All results use the frozen split in Table 1. We ask four questions. *RQ1*: does complete trajectory-conditioned fusion improve over video-only features? *RQ2*: how sensitive is performance to the motion representation? *RQ3*: once the fusion representation is fixed, does a structured readout improve over simple pooling? *RQ4*: what does the same test set reveal about transfer baselines and probability reliability?

The protocol uses two complementary suites. In the representation suite, each trajectory variant is evaluated with a linear readout over mean- and max-pooled object features (`meanmax_linear`). We compare these variants with a VideoMAE-only control with expanded trainable capacity. In the frozen-head suite, we first cache frozen VideoMAE features for all 2,334 videos, train fusion encoders with seeds 42–44, select each encoder by validation NLL, and export frozen fusion features. Each cache is then evaluated with eight lightweight heads and five head seeds (42–46): three mean/max pooling variants, one LogSumExp multiple-instance head, one prototype head, and three structured evidence readouts with object-level scoring and optional pairwise-object and global-video branches. The full three-branch readout is denoted `evidence`. Metrics average head seeds within each encoder before averaging the three encoder-level means. VideoMAE-only controls average three model seeds.

The primary ranking metric is raw AUROC. We also report AP, balanced accuracy, NLL, Brier score, ECE, and AURC where relevant. Checkpoints are selected on validation data, scalar temperature is fit on the separate calibration split, and the test split is not used to choose seeds, thresholds, learning rates, or checkpoints. Paired uncertainty intervals use 2,000 bootstrap resamples over the 45 held-out object-composition groups.

Evaluation sequence and interpretation. The eight-head suite was specified before its test evaluation. After examining those results, we added the representation controls in Section 6.2 and evaluated them on the same 353-clip test split. These supplemental comparisons are exploratory: their bootstrap intervals describe variation within the evaluated test sample and do not provide an independent confirmatory test. The `evidence` head is highlighted because it achieved the highest test AUROC among the eight specified heads; head-specific comparisons are therefore also exploratory. We report the complete head suite alongside it.

Table 3: Trajectory representation ablations on the 353-clip test split (mean across seeds; higher is better for the first three metrics). The full mean±standard-deviation tables and paired intervals appear in Appendix D.4. EC denotes the expanded-capacity video-only control; exact parameter parity is not established.

Method	AUROC ↑	AP ↑	BAcc ↑	NLL ↓	Brier ↓
EC VideoMAE	0.6834	0.6416	0.6340	0.6549	0.2287
Position-only 3D	0.7396	0.7133	0.6768	0.6024	0.2077
Shuffled 3D	0.7526	0.7184	0.6893	0.5925	0.2027
Image-plane 2D	0.7525	0.6841	0.6822	0.5973	0.2034
Full 3D trajectory	**0.7582**	**0.7210**	**0.7016**	**0.5864**	**0.1998**

Table 4: Separate supervised test suite (353 clips). VideoMAE-only averages three model seeds; fusion heads average five head seeds within each of three frozen fusion encoders. "No global" and "No pair" ablate branches of the `evidence` head. Encoder-seed AUROC standard deviations are in Appendix D.5.

Model / head	AUROC ↑	AP ↑	BAcc ↑	NLL ↓	Brier ↓	AURC ↓
VideoMAE-only	0.7173	0.6846	0.6596	0.6363	0.2200	0.2624
Fusion + `evidence`	**0.7622**	0.7156	0.6974	**0.5821**	**0.1980**	**0.1888**
`meanmax_linear`	0.7599	0.7240	0.7034	0.5859	0.1994	0.2073
`lse_mil`	0.7592	**0.7259**	0.7005	0.5843	0.1993	0.2037
`prototype`	0.7591	0.7225	0.7047	0.5845	0.1989	0.1982
No global	0.7570	0.7147	**0.7053**	0.5865	0.1994	0.2131
No pair	0.7564	0.7021	0.6977	0.5842	0.1991	0.1912
`meanmax_mlp`	0.7558	0.7129	0.6981	0.5963	0.2023	0.2040
`meanmax_global_mlp`	0.7538	0.7164	0.6948	0.5920	0.2022	0.2027

6.2 Trajectory Representation Questions

The representation suite addresses RQ1 by comparing Full 3D trajectory fusion against an expanded-capacity VideoMAE-only control. It addresses RQ2 under a shared trajectory readout by varying the motion stream through three controls: position-only 3D, image-plane 2D tracks, and a temporal-order shuffle applied after local velocity and acceleration have been computed. All trajectory variants in Table 3 use the same split and `meanmax_linear` head family. The frozen-feature study below instead compares alternative readouts on exported fusion features. Each study has its own video-only control and fusion estimate, reported in Tables 3 and 4, respectively.

RQ1: complete object-conditioned fusion. Full 3D outperforms the expanded-capacity VideoMAE control on every reported metric. The paired object-group differences are +0.0748 [0.0267, 0.1186] AUROC, +0.0794 [0.0223, 0.1345] AP, and +0.0677 [0.0187, 0.1129] balanced accuracy; NLL falls by 0.0685 [0.0216, 0.1107] and Brier by 0.0290 [0.0089, 0.0470]. These unadjusted intervals all favor Full 3D (Appendix D.4). This supports complete object-conditioned motion and video fusion as a useful representation in this controlled setting.

RQ2: motion representation choices. Position-only 3D reaches 0.7396 AUROC; adding velocity and acceleration raises the point estimate to 0.7582, with most paired intervals spanning zero. Image-plane 2D reaches 0.7525 AUROC, close to the full 3D representation: the paired AUROC difference is +0.0056 [-0.0103, 0.0212], while the AP difference favors full 3D (+0.0369 [0.0072, 0.0654]). The temporal-order control permutes complete tokens after local derivatives have been computed, preserving local motion descriptors; it reaches 0.7526 AUROC. Thus, the full representation has the highest point estimates, while simpler motion inputs also support competitive violation detection.

6.3 Readout Complexity on Frozen Fusion Features

RQ3 asks whether the cached fusion representation needs a structured evidence head. Table 4 shows the separate supervised comparison and all eight heads. Fusion plus `evidence` leads in raw AUROC, but three simple heads are within 0.0031; `lse_mil` leads in AP and removing the global branch improves default-threshold balanced accuracy. The fused features support competitive prediction across several readout families, including simple mean/max pooling. Encoder-seed variation is in Appendix D.5.

RQ3 answer and cross-suite check. The fusion-versus-video AUROC difference is positive for each aligned encoder seed (+0.0481, +0.0504, +0.0362). Its paired object-group difference is

Table 5: External evaluator transfer on the fixed 353-clip test set. The first three methods are zero-shot; supervised PHYTRAJ uses metadata-specified object queries. Inputs, training, and scoring differ; see Appendix D.6 for protocols and paired intervals.

Method	AUROC ↑	AP ↑	Kendall τ_b ↑	Spearman ρ ↑
VideoPhy-2	0.5524	0.5365	0.1129	0.1150
VideoScore2	0.5422	0.5280	0.0658	0.0754
PhyGenEval	0.5253	0.5104	0.0486	0.0516
PHYTRAJ	**0.7622**	**0.7156**	**0.3714**	**0.4542**

Table 6: Test probability metrics before → after scalar temperature scaling fitted on the separate 230-clip calibration split (lower is better). The full head suite is in Appendix D.5.

Model / head	NLL	Brier	ECE
VideoMAE-only	0.6363 → 0.6224	0.2200 → 0.2161	0.0857 → 0.0452
Fusion + `evidence`	0.5821 → 0.5832	0.1980 → 0.1983	0.0746 → 0.0727

+0.0449, with a 95% interval of [-0.0006, 0.0865]; AP and balanced accuracy also have zero-crossing intervals. Raw NLL and Brier differences are -0.0542 [-0.0933, -0.0120] and -0.0220 [-0.0397, -0.0032], respectively, providing secondary proper-score evidence under the same unadjusted analysis (Appendix D.5). Simple pooling and multiple-instance heads obtain AUROC point estimates close to the structured evidence variants.

6.4 EXTERNAL EVALUATOR TRANSFER AND PROBABILITY RELIABILITY

RQ4 examines how existing physics evaluators transfer to this benchmark and how reliable the supervised models' probabilities are. Under the reported zero-shot protocols, three physics evaluators rank the same 353 held-out clips less effectively (Table 5). These are transfer baselines rather than matched ablations: the evaluators are zero-shot, whereas PHYTRAJ is trained on this benchmark; prompt/caption access, target task, score definition, and input protocol also differ. The AUROC gaps to PHYTRAJ are large and their paired intervals exclude zero (Appendix D.6). They characterize the evaluators' transfer to this task under their respective input and training protocols.

The paired AUROC gaps are +0.2098 [0.1387, 0.2756] versus VideoPhy-2, +0.2200 [0.1563, 0.2870] versus VideoScore2, and +0.2369 [0.1680, 0.3047] versus the single-backbone PhyGenEval-style evaluator. The AP intervals and score protocols are in Appendix D.6.

Head branches and diagnostic signals. For the cached heads, ablating the pair or global branch yields small AUROC differences whose intervals cross zero. Object scores, pair scores, and branch weights should therefore be read as diagnostic signals rather than validated explanations (Appendix D.5).

Probability reliability. Temperature scaling improves the VideoMAE-only probability metrics but slightly worsens `evidence`'s mean NLL and Brier on test. Fusion still has the best raw and calibrated NLL/Brier, whereas VideoMAE-only has the lowest calibrated ECE (Table 6). Ranking quality and probability calibration therefore provide complementary views of performance.

Selective risk follows the same broad pattern: the lower-is-better AURC is 0.1888 for fusion plus `evidence` and 0.2624 for VideoMAE-only (Table 4), indicating lower aggregate selective risk in this evaluation.

Across the two studies, trajectory-conditioned fusion improves the AUROC point estimates over the respective video-only controls, and several lightweight readouts perform competitively on the fused features. Appendix Figures 2 and 3 pair two labeled clips with source frames and estimated camera-frame paths as qualitative pipeline diagnostics.

7 CONCLUSION

PHYTRAJ uses estimated object trajectories to query frozen video features for visual physics auditing. On 2,334 human-labeled clips, the exploratory representation study reaches 0.758 AUROC on 353 object-composition-held-out test clips, versus 0.683 for a video-only control, while frozen-feature experiments show that simple pooling remains competitive with structured aggregation.

Reproducibility Statement

Appendix D specifies the split and manifest hash, training and model-selection protocols, seed hierarchy, calibration procedure, and group-bootstrap design. The accompanying materials contain the model implementation, dataset metadata, split manifest, and aggregate experiment results. These support inspection of the experimental design, but do not provide all trained checkpoints and per-example outputs needed to regenerate every numerical comparison. Appendix D.7 identifies the corresponding source versions and describes these reproducibility limits.

AI Use Statement

In this work, an LLM is utilized to revise the paper written by the authors. We also use LLM to assist academic figure drafting and revisions.

References

Zhongjie Ba, Shengwang Xu, Peng Cheng, Jinyang Zou, Ting Yu, Zhibo Wang, and Zhan Qin. PhyCheck: Fine-grained evidence-grounded dataset for physical law understanding in video-llms, 2026. URL https://arxiv.org/abs/2608.02150.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fce2d8a485746f76aac7b5650db2679d-Abstract-Conference.html.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c02f6a1d5c55e16db50d339dad905b4d-Abstract-Conference.html.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1612.00222.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Zhaolong Du, Shasha Mao, Xuequan Lu, Mengnan Qi, Yimeng Zhang, Jing Gu, and Licheng Jiao. Rethinking multiple-instance learning from feature space to probability space. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/463a91da3c832bd28912cd0d1b8d9974-Abstract-Conference.html.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. PhyWorldBench: A comprehensive evaluation

of physical realism in text-to-video models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79d86433c2acd12b6fa98553435d226e-Abstract-Conference.html. Oral presentation.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. VideoScore2: Think before you score in generative video evaluation, 2025. URL https://arxiv.org/abs/2509.22799.

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*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28991–29017. PMLR, 2025. URL https://proceedings.mlr.press/v267/kang25g.html.

Wenqiao Li, Yao Gu, Xintao Chen, Xiaohao Xu, Ming Hu, Xiaonan Huang, and Yingna Wu. Towards visual discrimination and reasoning of real-world physical dynamics: Physics-grounded anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30409–30419, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_Towards_Visual_Discrimination_and_Reasoning_of_Real-World_Physical_Dynamics_Physics-Grounded_CVPR_2025_paper.html.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y. Lu, Enfu Nan, Hokin Deng, Edmund Yeh, Sarah Ostadabbas, Yun Fu, Jennifer Dy, Pu Zhao, and Yanzhi Wang. PhyGround: Benchmarking physical reasoning in generative world models, 2026. URL https://arxiv.org/abs/2605.10806.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 2025. URL https://proceedings.mlr.press/v267/meng25c.html.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos?, 2025. URL https://arxiv.org/abs/2501.09038.

Atin Pothiraj, Jaemin Cho, Yue Zhang, Elias Stengel-Eskin, and Mohit Bansal. Physics question scene graph: Fine-grained evaluation of physical plausibility in text-to-video generation. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.25306.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.12602.

Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Evaluating newtonian mechanics in video generative models with real physical systems. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2504.02918.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Pranay Boreddy, Shruti Satya Narayana, Yunhao Ba, Alex Wong, Celso M. de Melo, and Achuta Kadambi. WorldBench: Disambiguating physics for diagnostic evaluation of world models, 2026. URL https://arxiv.org/abs/2601.21282.

Zihan Wang, Songlin Li, Lingyan Hao, Xinyu Hu, and Bowen Song. What you see is what matters: A novel visual and physics-based metric for evaluating video generation quality, 2024. URL https://arxiv.org/abs/2411.13609.

Jiaxin Wu, Yihao Pi, Yinling Zhang, Yuheng Li, and Xueyan Zou. Quantitative video world model evaluation for geometric-consistency, 2026. URL https://arxiv.org/abs/2605.15185.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.05861.

Hanxue Zhang, Haoran Jiang, Qingsong Yao, Yanan Sun, Renrui Zhang, Hao Zhao, Hongyang Li, Hongzi Zhu, and Zetong Yang. Detect anything 3d in the wild, 2025. URL https://arxiv.org/abs/2504.07958. ICCV 2025.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. Physion-Eval: Evaluating physical realism in generated video via human reasoning, 2026. URL https://arxiv.org/abs/2603.19607.

A Background Details for Task Scope

This appendix gives additional scope details for the visual physics-auditing task and situates the benchmark relative to neighboring evaluator protocols.

Visible-event labels versus prompt intent. The task deliberately separates the rendered event from the prompt's intended event. A valid prompt can yield a visually invalid clip, and an invalid prompt can yield a clip where the impossible event is not actually visible. The classifier therefore receives neither prompt text nor physical-rule family. Labels are meant to reflect visible evidence such as motion, contact, support, continuity, and energy dissipation.

Examples of in-scope violations. Representative violations include a projectile accelerating against the apparent gravity direction, an object changing velocity without visible interaction, rigid bodies passing through one another or failing to respond at contact, a discontinuous trajectory jump, an unsupported object remaining static, or motion failing to dissipate under a frictional interaction. Human or animal actors may appear in continuity, contact, or gravity examples, but the current study does not claim to evaluate biomechanics.

Object-query inputs. The corpus uses metadata-associated object names normalized to a 30-name vocabulary. The system therefore receives object identity through the perception interface even though the final classifier does not receive the full prompt. For example, human roles are merged into `person` and specific ball types into `ball`. The evaluation assumes that these object queries are supplied; automatic query selection for videos without source object metadata is outside its scope. Query normalization is consequently part of the input protocol, not a learned component of the classifier.

Related evaluation boundaries. VideoPhy, PhyGenBench, VideoPhy-2, VideoScore2, PhysBench, Physics-IQ, and related suites are closest in problem motivation, but their input contracts differ from ours: some use prompts, rule text, questions, rubrics, or model-derived physical-consistency scores. PDI-Bench and VAMP are closest to the object-motion direction. PDI-Bench uses deterministic geometric residuals over object-centric 3D estimates; VAMP uses tracked motion scores for video generation evaluation. These approaches motivate a complementary comparison between learned motion–video representations and explicit geometric or motion-based scores under a common input protocol.

B Method Details

This appendix supplements the trajectory-query mechanism in Section 4 with implementation details, the frozen-cache protocol, and the complete readout definitions.

B.1 Input Tensors and Motion Controls

For a clip V, the video stream samples $T = 16$ frames at resolution 224×224 and feeds them to a frozen VideoMAE backbone:

$$\mathbf{X} = E_{\text{vid}}(V) \in \mathbb{R}^{N \times d_v}. \tag{12}$$

These tokens are computed once per example and reused throughout the cached-head experiments.

For object k at sampled frame t, the motion tensor stores

$$\mathbf{m}_{k,t} = [\mathbf{p}_{k,t}; \mathbf{v}_{k,t}; \mathbf{a}_{k,t}; r_{k,t}] \in \mathbb{R}^{10}, \tag{13}$$

where $\mathbf{p}, \mathbf{v}, \mathbf{a} \in \mathbb{R}^3$ are monocular camera-frame position, finite-difference velocity, and finite-difference acceleration, and $r_{k,t}$ is a valid-observation indicator. Positions, velocities, and accelerations are clipped before they are passed to the network. The object mask marks an object valid if at least one sampled frame contains a valid observation. Fully missing tracks are masked to avoid NaNs and are zeroed by the object-validity mask.

Trajectory extraction uses 16 uniformly sampled frames resized to 640×360. We compute finite-difference velocities and accelerations using a frame rate of 3.2 FPS. These features describe local changes in estimated object centers and inherit the monocular coordinate scale, so we treat them as motion descriptors rather than calibrated physical measurements (Appendix C).

Representation controls retrain the fusion encoder under a shared linear mean/max decision head. The position-only control keeps $\mathbf{p}_{k,t}$, object identity, validity, and missingness while zeroing $\mathbf{v}_{k,t}$ and $\mathbf{a}_{k,t}$. The 2D control replaces camera-frame centers and derivatives with normalized image-plane box centers and their finite differences. The temporal-order control permutes the full $(\mathbf{p}, \mathbf{v}, \mathbf{a}, r)$ tokens after derivatives have been computed; it disrupts global order while preserving local motion values, so it is not a no-dynamics intervention. The VideoMAE-only control increases the capacity of the trainable visual classifier while omitting the trajectory stream. This comparison evaluates the complete fusion model against an expanded-capacity visual baseline; it does not isolate the motion-query mechanism (Appendix D.4).

B.2 Fusion Encoder

Section 4.3 gives the query and fusion equations. The motion projection ϕ_m is a linear layer followed by layer normalization, GELU, and dropout. Its output is added to a learned temporal embedding and processed by a two-layer, four-head trajectory Transformer with GELU feed-forward blocks and dropout. Invalid observations are masked in attention and temporal pooling. The pooled token is layer normalized and multiplied by the object-validity mask.

The video projection ϕ_v applies a linear layer, layer normalization, and GELU to each frozen VideoMAE token. Multi-head attention then uses the trajectory tokens as queries and the projected video tokens $\mathbf{U}$ as both keys and values, with its own internal learned projections. Equations 7–8 include the residual connection, layer normalization, and fusion module. The latter comprises a linear map from $2d$ to d, layer normalization, GELU, and dropout. Fused tokens are zeroed for invalid objects before a two-layer, four-head relation Transformer processes them under the object mask. Its outputs are normalized and masked again.

The exported cache contains contextual object tokens $\mathbf{H}$, the object-validity mask, the global projected video feature $\bar{\mathbf{u}} = N^{-1}\sum_n \mathbf{U}_n$, and the original motion tensors and observation masks. During encoder fitting, masked mean/max pooling feeds an MLP with a $2d$-to-d linear map, layer normalization, GELU, dropout, and a final scalar linear output. This training classifier is discarded before the cached-head experiments.

In the frozen-feature suite, the fusion encoder and cached heads use unweighted binary cross-entropy with logits and shuffled training minibatches. Training follows the empirical class distribution, without class-balanced resampling. Checkpoint selection uses validation NLL.

B.3 Frozen-Cache Protocol

The formal protocol uses a fixed object-composition-held-out manifest with 1,401 train, 350 validation, 230 calibration, and 353 test examples. The split holds out object compositions within one generator's output distribution; prompt and scene templates can be shared across subsets.

The training sequence is:

1. cache frozen VideoMAE tokens for all 2,334 examples;
2. fit the fusion encoder on train for seeds 42, 43, and 44, selecting the checkpoint with the lowest validation NLL;
3. export one frozen fusion cache for each encoder seed;
4. fit each cached head on train, selecting on validation, using head seeds 42–46; and
5. for the original fixed-head suite, fit a scalar temperature on the calibration split after checkpoint selection and report locked test metrics.

The original head suite did not use test labels for training, checkpoints, thresholds, or calibration. All eight frozen heads are reported. Highlighting the highest-AUROC head after inspecting the fixed head matrix is therefore an exploratory presentation choice rather than a pre-specified primary contrast.

B.4 Cached Classification Heads

Let $\mathbf{H} \in \mathbb{R}^{K\times d}$ be the frozen object tokens and $\mathbf{v} \in \{0,1\}^K$ the object mask. The formal head matrix contains eight heads: `meanmax_linear`, `meanmax_mlp`, `meanmax_global_mlp`, `lse_mil`, `prototype`, `evidence`, `evidence_no_pair`, and `evidence_no_global`. Additional experimental interfaces exist in the implementation, but they are not part of the frozen final comparison.

Mean/max heads. The simplest heads pool object tokens with a masked mean and masked max:

$$\phi_{\mathrm{mm}}(\mathbf{H},\mathbf{v}) = \left[\frac{\sum_k v_k \mathbf{h}_k}{\max(1,\sum_k v_k)}\ ;\ \max_{k:v_k=1} \mathbf{h}_k\right]. \tag{14}$$

`meanmax_linear` applies a linear classifier to this vector, while `meanmax_mlp` applies a small LayerNorm-GELU-dropout MLP. `meanmax_global_mlp` concatenates the global video feature $\bar{\mathbf{u}}$ to the same mean/max vector.

Multiple-instance and prototype heads. `lse_mil` scores each object independently and aggregates valid instance logits with length-normalized log-sum-exp,

$$s_{\mathrm{lse}} = \tau \log \sum_{k:v_k=1} \exp(s_k/\tau) - \tau \log \max(1, \sum_k v_k). \tag{15}$$

The normalization prevents the score from changing only because an identical bag is duplicated. `prototype` projects each object token, compares it with learned positive and negative prototype sets by cosine similarity, takes the strongest positive and negative matches, and aggregates the resulting object scores with the same normalized LSE rule. These prototypes are learned evidence centers, not annotated physical-law categories.

Evidence head and ablations. The evidence head is an aggregation head with three optional branches. The unary branch is the prototype MIL scorer described above. The pair branch enumerates unordered valid object pairs (i, j) and scores

$$\psi_{ij} = [|\mathbf{h}_i - \mathbf{h}_j|; \mathbf{h}_i \odot \mathbf{h}_j], \tag{16}$$

then compares a projected pair feature with learned positive and negative pair prototypes and aggregates pair scores by normalized LSE. The global branch is a small MLP over $\bar{\mathbf{u}}$. Available branch logits are combined by learned nonnegative weights obtained from `softplus` parameters:

$$s_{\text{evi}} = \frac{\sum_b a_b w_b s_b}{\max(\epsilon, \sum_b a_b w_b)}, \tag{17}$$

where a_b indicates whether branch b is available for the sample. The no-pair ablation disables the pair branch, and the no-global ablation disables the global branch. The object scores, pair scores, and branch weights are useful diagnostic signals, but the current experiments do not validate them as causal explanations or recovered physical rules.

C Dataset Construction Details

This section gives the full dataset accounting for Section 5.

Generation and annotation. The corpus is built from structured event templates that specify physical family, event stages, objects, observability constraints, and plausible or implausible variants. Programmatic template expansion supplies object, material, surface, environment, height, distance, and camera-angle variables from a curated asset pool. A language model verbalizes these structured instances into natural Chinese prompts while preserving the specified event, objects, and constraints; prompt intent remains metadata and is not the final label. After format checks, constraint checks, and duplicate-prompt checks, Seedance 2.0 produces 2,454 candidate clips. The `v4` and `v5` metadata identifiers denote generation batches within this process. The two example videos in Figure 2 each contain 121 frames at 1280×720 pixels and 24 FPS (5.04 seconds); these properties describe the examples, not a verified corpus-wide distribution.

Two annotators independently reviewed all 2,454 candidates for the physical behavior visible in the generated video, then cross-validated annotations and jointly reviewed ambiguous or disagreeing cases to produce the final label. The released metadata stores only the final adjudicated label, where 0 denotes physically plausible behavior and 1 denotes a physical violation. These adjudicated labels define the prediction target; they do not support an estimate of initial inter-annotator agreement.

Table 7: Candidate filtering and balancing. The first exclusion step removes clips without usable 3D trajectories; the second step is class balancing only.

Stage	Clips	Label 0	Label 1
Adjudicated candidates	2,454	1,227	1,227
After 3D-trajectory availability filter	2,358	1,191	1,167
Final balanced benchmark	2,334	1,167	1,167

The trajectory-availability filter removes 96 candidates (36 plausible and 60 violating). Class balancing then removes 24 plausible clips with seed 42. Because the first filter depends on perception success, the retained set may underrepresent difficult-to-track events; the reported results are conditional on at least one usable object trajectory.

Object queries and trajectory extraction. Each candidate video is paired with normalized object queries for open-vocabulary segmentation and tracking. In the 2,454-candidate set, 450 videos have one query object and 2,004 have at least two query objects. In the retained 2,334-video benchmark, 426 videos have one query object and 1,908 have at least two query objects. The normalized query vocabulary contains 30 object names. Some raw object descriptions are merged for detection consistency: for example, human roles are normalized to `person`, and specific ball types are merged to `ball`. In the candidate metadata, 1,343 examples have a normalized query list that does not exactly match the raw object field. These queries supply object identity to perception and determine the object-composition groups used for splitting.

For each retained video, the perception pipeline samples 16 frames, resizes them to 640×360, segments and tracks normalized object queries with SAM3, and estimates camera-frame 3D object centers with DetAny3D. The trajectory features used by the model are built from estimated center positions, finite-difference velocities and accelerations, and validity masks. Velocity and acceleration are computed by finite differences using a 3.2 FPS time base, corresponding approximately to 16 samples over a five-second clip. The estimated positions and their derivatives inherit the uncertain monocular spatial scale, which limits their interpretation in physical units.



Figure 2: Representative source frames from two generated clips. `video_0868` (top) is an adjudicated violation from validation; `video_1521` (bottom) is plausible and in test. Columns show source frames 30, 75, and 112 (zero-based), at 1.25, 3.125, and approximately 4.667 seconds. Labels describe the complete clips, rather than individual frames. Source videos are included at companion-code commit `f062e76`; the corresponding estimated trajectories are shown in Figure 3.

Physical families. Table 2 in the main text reports the six analysis families used in the paper. These are semantic families after canonical mapping, not raw metadata values: the source `physics_category` field contains multiple English and Chinese strings, and `category_key` also contains scene strata such as `living` and `living_vehicle`. The table should therefore be read as the paper-level family distribution rather than as a direct dump of one raw field.

Overall class balance does not imply balance within these families. Differences in family prevalence and queried objects can support predictive shortcuts, so aggregate discrimination alone does not establish category-independent physical reasoning. Per-family and track-quality comparisons have not been evaluated with uncertainty intervals.

Object-composition-held-out split. All formal experiments use one frozen split. Examples are grouped by normalized object composition, and every clip with the same `object_combo` is assigned to the same subset. This prevents the same normalized object pair or set from appearing in both train and test, yielding 290 object-composition groups.

The subset sizes, label counts, and group counts are reported in Table 1 in the main paper.

Train fits weights; validation selects checkpoints by NLL; calibration fits thresholds and scalar temperatures. The eight-head suite preceded its first test evaluation; later representation controls reused the inspected test, so their intervals are descriptive. This split holds out object compositions, not prompts, scenes, generator batches, or source domains. Object-composition separation also does not exclude near-duplicate clips or semantically similar prompt variants across subsets.

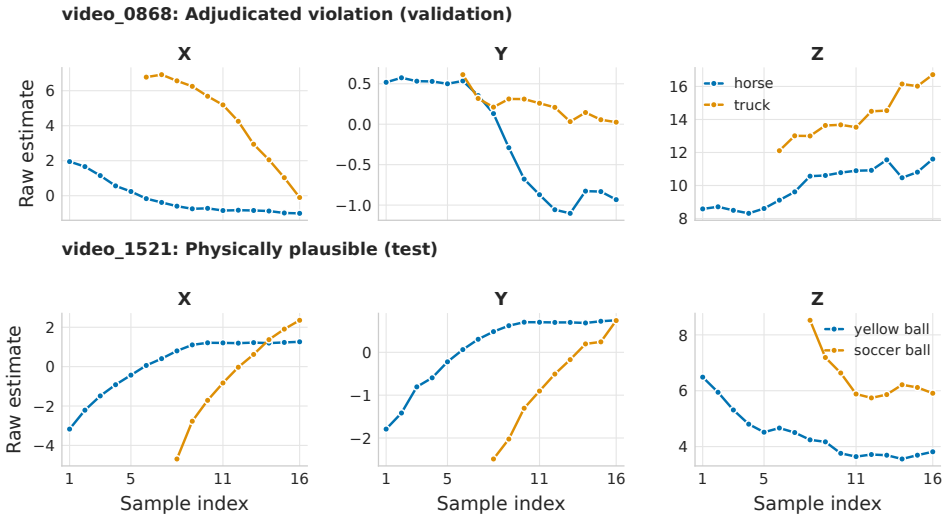


Figure 3: Stored camera-frame object-center components for the clips in Figure 2. Each point is a valid SAM3/DetAny3D record; lines connect available samples without smoothing or extrapolation. These monocular camera-frame estimates have unverified scale and are not calibrated world trajectories. The companion score file lists 0.7278 for `video_0868` and 0.1885 for `video_1521`; these reported mean predicted-violation scores average 3 fusion-encoder seeds (42–44) and 5 evidence-head seeds (42–46), using the selected checkpoint for each pair. These scores illustrate model outputs on the two clips and are not object-level explanations.

D Experimental Details and Additional Results

This appendix specifies the experimental protocols and expands the results for PhyTraj. It distinguishes the fixed-head study from the exploratory representation controls, and describes data and implementation availability.

D.1 Experimental Configuration

Split report's `manifest_sha256` identifier:

```
4cf843f6f552a7a23e9f757b2a67ca3b
2f1d9ab8d1e117dee934ac44ddf05946
```

Bytewise SHA-256 of `data/splits/manifest.json` at code commit `57e78c4`:

```
f13a8e8445af5e1a5f7dfbde7efa8c06
e80c099b2239a19cc46f3db490cb4cf8
```

D.2 Interpretation of the Main Comparisons

Table 9 summarizes the experimental comparisons and their statistical interpretation.

D.3 Training and Evaluation Details

Fusion encoder. The frozen fusion encoders were fit on train with VideoMAE frozen, cached VideoMAE tokens when available, hidden dimension 256, maximum 10 objects, 16 sampled frames, two-layer trajectory and relation Transformer encoders, four attention heads, dropout 0.2, learning rate 10^{-5}, batch size 16, weight decay 3×10^{-4}, and 200 epochs. Checkpoints are selected by validation NLL, not by the last epoch. Seeds 42, 43, and 44 reached best validation NLL at epochs 50, 57, and 51, respectively.

Table 8: Shared configuration and evaluation sequence.

Component	Configuration
Split manifest	2,334 examples; train/val/calibration/test = 1,401/350/230/353; object-composition groups do not cross splits. The split report's `manifest_sha256` field begins `4cf843f6`; the bytewise file hash is recorded separately below.
VideoMAE cache	Frozen VideoMAE tokens, 16 frames, 224×224, feature dimension 768, generated once for all examples.
Fusion encoders	Seeds 42/43/44; fit on train, with checkpoint selected by lowest validation NLL; exported caches contain object tokens, object masks, global features, motion, and motion masks.
Cached heads	Eight formal heads: meanmax linear, meanmax MLP, meanmax global MLP, LSE-MIL, prototype, evidence, evidence without pair branch, and evidence without global branch. Head seeds are 42–46 for each fusion cache.
Supplemental representation ablations	Added after the first final-test inspection; results provided at code commit `83d2f8979d701f82212b181d2befac9b4a7b389d`. The controls are expanded-capacity VideoMAE-only, position-only 3D, image-plane 2D trajectory, temporal-order shuffled 3D, and Full 3D trajectory. Trajectory variants use a fixed `meanmax_linear` head with three encoder seeds and five head seeds per encoder. These controls provide descriptive reused-test evidence.
Calibration/test	Temperature is fit only on the calibration split. The original eight-head suite was fixed before its first test evaluation; the supplemental controls reuse the inspected test and are descriptive.

Cached heads. Cached-head training uses the frozen fusion caches and does not update VideoMAE or the fusion encoder. The formal comparison uses head seeds 42–46, batch size 128, weight decay 10^{-4}, hidden dimension 128, and checkpoint selection by validation NLL. Learning rates and epoch budgets differ by head according to the frozen development protocol (Table 10). The linear and LogSumExp heads received a 500-epoch cap after a training-budget check on validation; other heads used a 100-epoch cap. The selected weights still come from the lowest validation NLL, not necessarily the final epoch.

VideoMAE-only control. The visual-only model uses the same frozen manifest and VideoMAE feature cache. Weights are fit on train with learning rate 10^{-4}, batch size 32, weight decay 3×10^{-4}, and early-stopping patience 60; validation NLL selects the checkpoint. Formal seeds 42, 43, and 44 stopped at epochs 85, 86, and 88, with best checkpoints at epochs 25, 26, and 28. An earlier development run at 10^{-3} is not part of the formal comparison.

Metrics and uncertainty. The primary ranking metric is raw AUROC. Companion metrics are AP, balanced accuracy, raw and calibrated NLL, raw and calibrated Brier score, ECE, and selective-risk/risk-coverage diagnostics where available. Test-comparison uncertainty uses the object-composition group bootstrap over the 45 test groups rather than treating individual clips as independent.

D.4 Supplemental Representation-Ablation Protocol

The representation controls were added after inspection of the test results from the frozen-feature study and evaluated on the same 353 clips. Their confidence intervals are therefore exploratory and do not account for test-informed experimental design. The following specifies the interventions, seeds, metrics, and group-bootstrap procedure.

Seed hierarchy. Each trajectory representation uses encoder seeds 42, 43, and 44. For each encoder, the `meanmax_linear` head is trained with head seeds 42–46. Metrics first average the five head seeds within the same encoder cache, then average the three encoder-level values; the reported standard deviation is the population standard deviation of these three encoder-level means. The

Table 9: Findings and interpretation of the main experimental comparisons.

Comparison / question	Findings and interpretation
Fusion versus video-only (frozen-feature suite)	On the fixed object-composition test, Fusion+evidence has a raw AUROC difference over VideoMAE-only of +0.0449 on average across aligned encoder seeds, with all three seed-wise differences positive (+0.0481, +0.0504, +0.0362). The 2,000-sample group bootstrap CI is $[-0.0006, 0.0865]$, so the AUROC result is a consistent point estimate rather than a 95% significant gain. Raw NLL improves by -0.0542 with CI $[-0.0933, -0.0120]$ and raw Brier by -0.0220 with CI $[-0.0397, -0.0032]$, supporting favorable secondary proper-score evidence under the same unadjusted bootstrap analysis. The evidence head was highlighted after comparing all eight fixed heads on test, so this interval is conditional on that choice.
Pairwise and global readout branches	On final test, evidence has the best raw AUROC/NLL/Brier among cached heads, but its margins over `evidence_no_global` and `evidence_no_pair` are small: +0.0053 and +0.0058 AUROC, respectively, and all major 95% CIs cross zero. The global and pair branches therefore provide directionally positive point estimates, not independently validated components.
Fusion versus video-only (representation suite)	In the exploratory representation study, trajectory variants use a `meanmax_linear` readout. Full 3D exceeds the expanded-capacity VideoMAE control by +0.0748 AUROC, +0.0794 AP, and +0.0677 balanced accuracy, with CIs $[0.0267, 0.1186]$, $[0.0223, 0.1345]$, and $[0.0187, 0.1129]$. Raw NLL improves by -0.0685 with CI $[-0.1107, -0.0216]$ and raw Brier by -0.0290 with CI $[-0.0470, -0.0089]$. This expanded-capacity VideoMAE row uses a different protocol from the earlier VideoMAE-only baseline in the main fusion comparison; the two AUROC values (0.6834 and 0.7173) are therefore not a capacity monotonicity result.
Full 3D versus simpler motion inputs	Full 3D improves over position-only 3D on all five point estimates, with the balanced-accuracy interval excluding zero (+0.0249, CI $[0.0008, 0.0481]$); the AUROC, AP, NLL, and Brier intervals cross zero. Full 3D also improves over image-plane 2D on all five point estimates, with AP excluding zero (+0.0369, CI $[0.0072, 0.0654]$). These results support a consistent Full-3D advantage pattern, without broad statistical separation on every metric.
Sensitivity to global temporal order	Temporal-order shuffle preserves each token's correctly computed position, velocity, acceleration, and validity while permuting token order. Full 3D has slightly favorable point estimates against this control, but all reported intervals cross zero. The comparison is consistent with a small benefit from global order, without ruling out the contribution of token-local motion quantities.
Lightweight readout performance	Prototype, linear mean/max, LSE-MIL, and global mean/max MLP are competitive cached-head baselines. Linear mean/max reaches 0.7599 test AUROC versus 0.7622 for the evidence head; these close point estimates do not establish statistical equivalence.
Interpretation of internal scores	Object scores, pair scores, and branch weights are inspectable model internals. Without object/time evidence labels or counterfactual audits, they are diagnostic signals rather than causal explanations or recovered physical rules.
Generalization measured by the split	The split is object-composition-held-out. It does not measure generalization across prompt templates, generators, or source domains.

expanded-capacity VideoMAE-only control uses three model seeds. Bootstrap intervals resample the 45 held-out `object_combo` groups 2,000 times with bootstrap seed 20260925.

Interventions. The Full 3D representation uses $(P_t, V_t, A_t, \mathrm{valid}_t)$. The position-only control preserves 3D positions, object identity, validity masks, time order, and missingness patterns, but zeros the velocity and acceleration channels: $(P_t, 0, 0, \mathrm{valid}_t)$. The image-plane 2D control uses normalized detection-box centers and finite-difference velocity and acceleration in image coordinates. The temporal-order shuffle first computes complete motion tokens $(P_t, V_t, A_t, \mathrm{valid}_t)$, then

Table 10: Documented cached-head training settings. Batch size 128, weight decay 10^{-4}, and five head seeds are shared across rows. Epochs are caps, not selected checkpoint epochs.

Head	Learning rate	Epoch cap
`meanmax_linear`	3×10^{-4}	500
`meanmax_mlp`	10^{-3}	100
`meanmax_global_mlp`	10^{-4}	100
`lse_mil`	3×10^{-4}	500
`prototype`	10^{-3}	100
`evidence`	10^{-3}	100
`evidence_no_pair`	10^{-3}	100
`evidence_no_global`	10^{-3}	100

permutes the token order. It therefore breaks global temporal ordering while preserving token-local motion quantities; it is not a "no dynamics" control.

Expanded-capacity video control. The VideoMAE-only control increases the capacity of its trainable visual MLP with the aim of approximating the Full 3D model's trainable parameter budget. Exact parameter parity is not established, so this comparison does not fully control for capacity. Moreover, the fusion model jointly introduces object queries, motion-conditioned retrieval, and inter-object context. Its gain therefore characterizes the complete model rather than any single component.

Table 11: Discrimination and threshold metrics for Table 3, shown as mean±standard deviation. For trajectory models, five head seeds are averaged within each of three encoder seeds; the displayed standard deviation is across encoder-level means. The VideoMAE-only row summarizes three model seeds.

Method	AUROC ↑	AP ↑	BAcc ↑
EC VideoMAE	0.6834 ± 0.0121	0.6416 ± 0.0053	0.6340 ± 0.0198
Position-only 3D	0.7396 ± 0.0069	0.7133 ± 0.0065	0.6768 ± 0.0068
Shuffled 3D	0.7526 ± 0.0118	0.7184 ± 0.0141	0.6893 ± 0.0165
Image-plane 2D	0.7525 ± 0.0037	0.6841 ± 0.0018	0.6822 ± 0.0033
Full 3D trajectory	0.7582 ± 0.0036	0.7210 ± 0.0088	0.7016 ± 0.0049

Table 12: Companion proper-score metrics for the representation controls, shown as mean±standard deviation under the same seed hierarchy as Table 11.

Method	NLL ↓	Brier ↓
EC VideoMAE	0.6549 ± 0.0064	0.2287 ± 0.0032
Position-only 3D	0.6024 ± 0.0070	0.2077 ± 0.0032
Shuffled 3D	0.5925 ± 0.0108	0.2027 ± 0.0046
Image-plane 2D	0.5973 ± 0.0019	0.2034 ± 0.0006
Full 3D trajectory	0.5864 ± 0.0045	0.1998 ± 0.0017

D.5 Frozen Fusion Head Suite

Table 15 reports all eight pre-frozen cached heads on the same frozen fusion features. The `evidence` head ranks first on raw AUROC, raw NLL, raw Brier, and AURC, but the margins over other fusion heads are small. `lse_mil` has the highest AP, `evidence_no_global` has the highest default-threshold balanced accuracy, and `meanmax_linear` has the smallest encoder-seed AUROC standard deviation. The evidence head is therefore the best member of a fixed diagnostic suite, without establishing the head design as an independently necessary architectural component.

The ablations of the internal `evidence` branches are conservative diagnostics. The full head's AUROC point estimate is higher than `evidence_no_global` and `evidence_no_pair`, but the paired bootstrap intervals in Table 16 cross zero for the reported metrics. Object scores, pair scores, and branch weights are therefore diagnostic signals rather than validated object-level or temporal explanations.

Table 13: Descriptive reused-test bootstrap intervals for the supplemental representation ablations. Differences are Full 3D minus the comparison model; negative NLL/Brier differences favor Full 3D.

Comparison	Metric	Difference [95% CI]
Full 3D vs EC VideoMAE	ΔAUROC	+0.0748 [0.0267, 0.1186]
Full 3D vs EC VideoMAE	ΔAP	+0.0794 [0.0223, 0.1345]
Full 3D vs EC VideoMAE	ΔBAcc	+0.0677 [0.0187, 0.1129]
Full 3D vs EC VideoMAE	ΔNLL	-0.0685 [-0.1107, -0.0216]
Full 3D vs EC VideoMAE	ΔBrier	-0.0290 [-0.0470, -0.0089]
Full 3D vs position-only 3D	ΔAUROC	+0.0185 [-0.0059, 0.0440]
Full 3D vs position-only 3D	ΔAP	+0.0077 [-0.0208, 0.0386]
Full 3D vs position-only 3D	ΔBAcc	+0.0249 [0.0008, 0.0481]
Full 3D vs position-only 3D	ΔNLL	-0.0160 [-0.0413, 0.0091]
Full 3D vs position-only 3D	ΔBrier	-0.0080 [-0.0187, 0.0023]
Full 3D vs temporal-order shuffled 3D	ΔAUROC	+0.0056 [-0.0115, 0.0219]
Full 3D vs temporal-order shuffled 3D	ΔAP	+0.0025 [-0.0180, 0.0238]
Full 3D vs temporal-order shuffled 3D	ΔBAcc	+0.0124 [-0.0085, 0.0323]
Full 3D vs temporal-order shuffled 3D	ΔNLL	-0.0061 [-0.0236, 0.0116]
Full 3D vs temporal-order shuffled 3D	ΔBrier	-0.0030 [-0.0106, 0.0048]
Full 3D vs image-plane 2D	ΔAUROC	+0.0056 [-0.0103, 0.0212]
Full 3D vs image-plane 2D	ΔAP	+0.0369 [0.0072, 0.0654]
Full 3D vs image-plane 2D	ΔBAcc	+0.0194 [-0.0024, 0.0402]
Full 3D vs image-plane 2D	ΔNLL	-0.0109 [-0.0261, 0.0038]
Full 3D vs image-plane 2D	ΔBrier	-0.0037 [-0.0092, 0.0016]

Table 14: Final test comparison between the pure-video baseline and object / trajectory fusion. Metrics are means over three VideoMAE-only seeds or over three fusion encoder seeds after averaging five head seeds per encoder.

Model	AUROC ↑	AP ↑	BAcc ↑	NLL ↓	Brier ↓	AURC ↓
VideoMAE-only	0.7173	0.6846	0.6596	0.6363	0.2200	0.2624
Fusion + `evidence`	**0.7622**	**0.7156**	**0.6974**	**0.5821**	**0.1980**	**0.1888**

Table 15: Final test results for cached heads on frozen fusion features. Each row averages five head seeds within each encoder, then averages over encoder seeds 42–44. "Enc. SD" is the standard deviation of the three encoder-level AUROC means, not a confidence interval.

Head	AUROC ↑	Enc. SD	AP ↑	BAcc ↑	NLL ↓	Brier ↓	AURC ↓
`evidence`	**0.7622**	0.0056	0.7156	0.6974	**0.5821**	**0.1980**	**0.1888**
`meanmax_linear`	0.7599	**0.0038**	0.7240	0.7034	0.5859	0.1994	0.2073
`lse_mil`	0.7592	0.0069	**0.7259**	0.7005	0.5843	0.1993	0.2037
`prototype`	0.7591	0.0052	0.7225	0.7047	0.5845	0.1989	0.1982
`evidence_no_global`	0.7570	0.0051	0.7147	**0.7053**	0.5865	0.1994	0.2131
`evidence_no_pair`	0.7564	0.0072	0.7021	0.6977	0.5842	0.1991	0.1912
`meanmax_mlp`	0.7558	0.0091	0.7129	0.6981	0.5963	0.2023	0.2040
`meanmax_global_mlp`	0.7538	0.0109	0.7164	0.6948	0.5920	0.2022	0.2027

Table 16: Paired group bootstrap on the 353-video test split. Differences are `evidence` minus the baseline; lower is better for NLL and Brier.

Baseline	Metric	Difference [95% CI]
VideoMAE-only	ΔAUROC	+0.0449 [-0.0006, 0.0865]
VideoMAE-only	ΔAP	+0.0311 [-0.0138, 0.0751]
VideoMAE-only	ΔBAcc	+0.0378 [-0.0130, 0.0843]
VideoMAE-only	ΔNLL	**-0.0542 [-0.0933, -0.0120]**
VideoMAE-only	ΔBrier	**-0.0220 [-0.0397, -0.0032]**
`evidence_no_global`	ΔAUROC	+0.0053 [-0.0073, 0.0173]
`evidence_no_global`	ΔNLL	-0.0043 [-0.0134, 0.0040]
`evidence_no_global`	ΔBrier	-0.0014 [-0.0049, 0.0019]
`evidence_no_pair`	ΔAUROC	+0.0058 [-0.0035, 0.0150]
`evidence_no_pair`	ΔNLL	-0.0021 [-0.0078, 0.0037]
`evidence_no_pair`	ΔBrier	-0.0011 [-0.0037, 0.0016]

Table 17 reports probability metrics before and after calibration-only temperature scaling. Temperature scaling improves VideoMAE-only NLL, Brier, and ECE, but the calibrated baseline remains worse than raw or calibrated fusion on NLL and Brier. For `evidence`, calibration slightly worsens mean NLL and Brier (0.5821 to 0.5832, and 0.1980 to 0.1983), suggesting a calibration shift between the 230-video calibration split and the 353-video test split. Fusion with `evidence` has the

lowest mean raw and calibrated NLL and Brier score, while VideoMAE-only has the lowest calibrated ECE; the proper-score gains do not imply uniform dominance on every calibration diagnostic.

Table 17: Final-test probability metrics, raw and after calibration-only temperature scaling. Lower is better for every column. ECE is a descriptive binned estimate, not a distribution-shift guarantee.

Model/head	Raw NLL	Cal NLL	Raw Brier	Cal Brier	Raw ECE	Cal ECE
VideoMAE-only	0.6363	0.6224	0.2200	0.2161	0.0857	**0.0452**
`evidence`	**0.5821**	**0.5832**	**0.1980**	**0.1983**	0.0746	0.0727
`meanmax_linear`	0.5859	0.5880	0.1994	0.2006	0.0736	0.0709
`lse_mil`	0.5843	0.5879	0.1993	0.2007	0.0675	0.0676
`prototype`	0.5845	0.5872	0.1989	0.1999	0.0628	0.0653
`evidence_no_global`	0.5865	0.5879	0.1994	0.2001	**0.0612**	0.0614
`evidence_no_pair`	0.5842	0.5861	0.1991	0.1996	0.0622	0.0617
`meanmax_mlp`	0.5963	0.5867	0.2023	0.2002	0.0775	0.0614
`meanmax_global_mlp`	0.5920	0.5896	0.2022	0.2016	0.0693	0.0642

D.6 External Baseline Protocols

Table 18 records the exact identity and score used for the external evaluator comparison summarized in the main text. All rows use the same fixed test split, but the comparison is not parameter-matched and not an equal-input ablation. The proposed model is trained on the benchmark training split; the external evaluators are zero-shot transfer baselines.

Table 18: External baseline protocols. "Training on this dataset" refers to using the proposed train split for model fitting or fine-tuning. Larger ranking scores always mean a stronger prediction of physical violation.

Method	Training on this dataset	Input used at inference	Main ranking score
VideoPhy-2 AutoEval (PC)	No; zero-shot	Video only; main run uses 16 frames	Reversed PC score
VideoScore2	No; zero-shot	Video plus cleaned English generation caption; 3.2 FPS, sampling seed 42	Reversed expected physical-consistency score
PhyGenEval-style 7B (single-backbone)	No; zero-shot	Video, generation caption, and physical-rule family; 16 frames	Reversed unrounded three-stage compliance mean
PhyTraj Fusion + `evidence`	Yes; supervised on train split	Video plus metadata-specified object queries used by the perception stage	Raw predicted violation score; metrics averaged across seeds

Table 19: External evaluator transfer on the fixed test split. All external evaluators are zero-shot on the proposed dataset, while PhyTraj is supervised on the training split. The PhyTraj row highlights the highest-AUROC head after evaluating the fixed eight-head suite. Inputs and scoring protocols differ.

Method	AUROC ↑	AP ↑	Kendall τ_b ↑	Spearman ρ ↑
VideoPhy-2 AutoEval (zero-shot, PC)	0.5524	0.5365	0.1129	0.1150
VideoScore2 (zero-shot)	0.5422	0.5280	0.0658	0.0754
PhyGenEval-style 7B (single-backbone)	0.5253	0.5104	0.0486	0.0516
PhyTraj Fusion + `evidence`	**0.7622**	**0.7156**	**0.3714**	**0.4542**

Conditional on these score definitions and the test-highlighted `evidence` head, the paired object-group AUROC differences are +0.2098 over VideoPhy-2 (95% CI [0.1387, 0.2756]), +0.2200 over VideoScore2 ([0.1563, 0.2870]), and +0.2369 over the single-backbone PhyGenEval-style evaluator ([0.1680, 0.3047]). The corresponding AP differences are +0.1791 ([0.1238, 0.2402]), +0.1877 ([0.1216, 0.2533]), and +0.2052 ([0.1418, 0.2687]). Pairwise intervals among the three external evaluators include zero. Because their inputs and training differ, these intervals describe transfer performance rather than an isolated architecture effect.

VideoPhy-2. The main row uses the Physical Commonsense task with 16 frames to approximately match the temporal frame budget of PhyTraj. VideoPhy-2 scores physical plausibility on a 1–5 scale, so violation ranking uses the reversed score. The official 32-frame setting gives AUROC 0.5468 and AP 0.5322 on the same test split, close to the 16-frame result of 0.5524/0.5365. The 32-minus-16 paired group-bootstrap intervals for AUROC, AP, and balanced accuracy all cross zero; therefore the 16-frame setting is not established as better than the official 32-frame setting. The fixed-threshold diagnostic uses the validation-selected rule $\mathrm{PC_score} < 4$. VideoPhy-2's ordinal 1–5 score is not a probability; any NLL/Brier/ECE comparison requires a calibration-split Platt mapping and is post-processing.

VideoScore2. The main run uses 3.2 FPS with the cleaned English generation caption, sampling, temperature 0.7, and seed 42. For approximately five-second clips, this provides a temporal sampling budget comparable to 16 frames, without ensuring an identical number of model-consumed frames for every clip. The pre-specified score is the expected physical-consistency score; hard-score and standalone soft-score variants are sensitivity analyses. Only one generation seed was run, so sampling variability is unmeasured.

PhyGenEval-style single-backbone adaptation. This baseline is a PhyGenEval-style three-stage prompt adaptation using one frozen `LLaVA-Interleave 7B` checkpoint (`lmms-lab/llava-next-interleave-qwen-7b-dpo`) for all stages. It is not the official multi-model PhyGenEval-Open ensemble and does not use GPT, CLIP-FlanT5, CLIP retrieval, or InternVideo2.

The reported main score reverses the unrounded native three-stage compliance mean. We treat this score-based comparison as exploratory because its specification before test inspection is not established. Alternative score definitions are post-hoc sensitivity analyses. The self-reported violation probability gives the same near-random conclusion (AUROC 0.5317, AP 0.5178), so it is reported only as a score-definition sensitivity. Stage-specific scores and alternative thresholds are test-set diagnostics, not replacements for the reported main score.

Category and error diagnostics. The individual reports include category-level diagnostics for eight `category_key` values, but `living` and `living_vehicle` are scene strata rather than standalone physical laws. Per-law behavior is therefore defined by the standardized `physics_category` labels. Category sample sizes are small (29–81 clips), so these results are error-analysis diagnostics. Current diagnostics indicate that PHYTRAJ is stronger on friction, inertia, and continuity, while gravity and contact remain weak; this characterizes a boundary of the method rather than a separate significance claim.

D.7 DATA AND IMPLEMENTATION AVAILABILITY

The accompanying code includes the model implementation, candidate metadata, normalized object queries, final labels, and split manifest. The original baseline results are provided at commit `57e78c4`; the dataset description and representation-ablation results are provided at `83d2f89`. Example videos, stored detections and trajectories, and aggregate example scores are included at `f062e76`.

The metadata retains adjudicated labels rather than individual annotator decisions or adjudication histories. It identifies candidate and retained clips, but does not distinguish trajectory-filtered from class-balancing exclusions on a per-clip basis. Complete generation settings and template revisions, the raw-to-normalized query procedure and any manual edits, and the mapping to the six analysis families are not included. These omissions limit reconstruction of corpus construction and analysis of annotation or selection effects.

The implementation specifies the feature construction and training objectives, but complete run configurations, trained checkpoints, and per-example predictions with group and seed identifiers are unavailable for some comparisons. In particular, the expanded-capacity baseline's exact architecture and parameter count, the representation-suite bootstrap outputs, and the seed-specific predictions underlying the example scores are not included. The stored detections do not include their correspondence to source-video frames. Complete specifications of backbone and perception-model revisions, preprocessing thresholds, track association, and coordinate conventions are also unavailable. These materials support implementation-level inspection, but not complete numerical reproduction of all reported results.

D.8 INTERPRETATION OF DIAGNOSTIC SCORES

The evidence head exposes unary object scores, pair scores, branch availability, and branch weights. These quantities are inspectable diagnostic signals trained with clip-level supervision. Their correspondence to the objects or time intervals responsible for a violation has not been validated with localized annotations or interventions. They therefore support model inspection, rather than causal explanation or recovery of physical laws. Generalization across generators and source domains is outside the present evaluation.

E DISCUSSION AND LIMITATIONS

What the representation offers. PHYTRAJ connects explicit object motion with the scene context available in pretrained video features. The representation study favors the complete fusion stack over its video-only control, and the frozen-feature study shows that lightweight readouts remain competitive with structured aggregation. Together, these findings motivate motion-conditioned access to visual context as a useful design direction for visible-violation detection. The comparisons evaluate the complete stack, including metadata-specified object queries and inter-object context. They do not isolate the querying mechanism alone. Full 3D improves AP over image-plane 2D,

while their paired AUROC interval spans zero; token shuffling preserves local motion descriptors. These distinctions guide the interpretation of the representation controls.

Evaluation scope. The corpus contains trajectory-available clips from one generator with fixed-camera prompts, and the split holds out object compositions. Query selection, tracking success, family prevalences, and shared prompt templates can still provide correlated cues, so aggregate performance does not isolate physical reasoning from these potential shortcuts. Performance over all attempted clips, moving cameras, and unseen generators is unmeasured. Object and pair scores are diagnostic model outputs, with object/time annotations and intervention tests needed to validate localization or explanation. Likewise, the calibration and selective-risk results describe this test distribution; a human-review operating point would require a separately validated risk threshold. Appendix D.7 describes the scope of the available data and implementation.

Next experimental questions. As detailed in Section 6.1, the representation controls reuse the test split examined in the frozen-feature study. An independent confirmation set would test whether their observed gains persist. On that set, matched-input comparisons with non-motion object-conditioned video readouts, trajectory-only prediction, and late fusion would distinguish the role of motion-conditioned querying. Query-only, validity-only, track-quality, and depth-perturbation controls would further separate motion evidence from preprocessing cues. Broader generator and camera settings would then test where this representation is most useful.

F Additional Related Work

Decision heads and calibration. Attention-based multiple-instance learning motivates bag-level decisions from instance evidence (Ilse et al., 2018), and recent probability-space MIL work highlights instability in feature-space selection (Du et al., 2025). These ideas motivate comparing mean/max, LogSumExp, prototype, and evidence-aggregation heads on identical frozen features. Temperature scaling provides a complementary way to assess the probabilities used for decisions (Guo et al., 2017). Our study connects these two questions: how the readout uses trajectory-conditioned video features, and how well the resulting probabilities reflect observed labels.