EVENTS, NOT FRAMES: DETECTING VIOLATIONS OF PHYSICAL EXPECTATION FROM A SINGLE VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Infants look longer when a box hangs in mid-air instead of falling (Needham & Baillargeon, 1993); detecting such violations of physical expectation is a hallmark of intuitive physics (Battaglia et al., 2013) and a long-standing goal for machines (Lake et al., 2017; Piloto et al., 2022). Current methods often use a matched plausible clip to interpret a prediction error (Piloto et al., 2022; Bordes et al., 2025), or annotated violations to train an evaluator (Motamed et al., 2025; Bansal et al., 2026). Infants need neither to judge a single event from ordinary experience alone. Here we show that event structure may help close this gap by separating the physical expectations that frame-level prediction errors conflate. To this end, our model, Physics Using Latent State-Events (`PULSE`), parses tracked 3D object states into events and asks, stage by stage, whether an object should be visible (Visibility), whether the interaction begins as it should (Onset), and whether the state it leads to is consistent (Outcome). Each of these checks asks the same question of every video, and plausible videos alone define the normal range against which an unmatched clip is judged. On three benchmarks, including `PhysEvent`, our new benchmark with stage-specific violations and ground-truth 3D states, `PULSE` achieves the strongest overall single-video detection and paired ranking among the evaluated continuous scorers. Each decision also names the event, check, and objects behind it. More broadly, this work points toward machines that turn experience with ordinary events into an understanding of what can and cannot happen.

1 INTRODUCTION

Recognizing that an observed event has violated a physical expectation is central to physical reasoning, in humans and machines alike (Battaglia et al., 2013; Lake et al., 2017). The problem has its roots in developmental psychology, where infants form expectations about hidden objects and unfolding interactions and look longer when those expectations are broken (Baillargeon et al., 1985; Needham & Baillargeon, 1993; Margoni et al., 2024). We study its computational counterpart, in which a machine learns physical expectations from ordinary experience and uses them to judge an individual video.

Two common kinds of system detect such violations today. Prediction-based violation of expectation (VoE) systems typically compare a clip against a matched plausible counterpart (Riochet et al., 2022; Piloto et al., 2022; Bordes et al., 2025), and evaluators trained on annotated violations judge a clip alone but need labels that experience alone does not provide (Motamed et al., 2025; Bansal et al., 2026). The comparison in the first kind is not incidental. The two clips share a scene and its events, so the error common to both cancels and only their difference is read. A single clip offers no such cancellation, and Figure 1 shows why its error is hard to read alone. Disappearance, penetration, and post-impact ejection produce errors that differ in magnitude and duration, so the same magnitude can signal a violation in one scene yet arise from plausible dynamics in another. Each of these violations breaks a different physical expectation, and one error magnitude cannot tell them apart. We therefore ask how distinct physical expectations can be learned and evaluated from plausible experience alone, with neither a paired clip nor violation labels.

Different violations surface at different moments and ask different questions, such as whether an object's disappearance and reappearance agree with occlusion (Aguiar & Baillargeon, 1999; Spelke, 1990), whether an interaction begins when two objects meet (Leslie & Keeble, 1987), and whether

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.

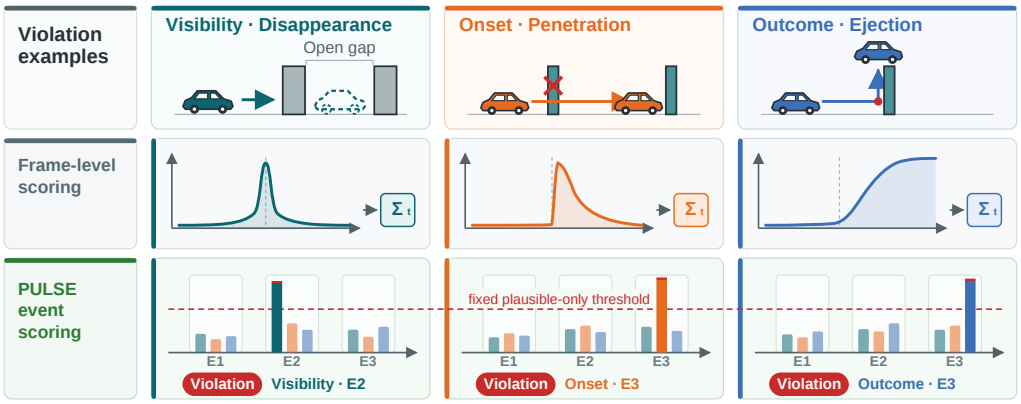


Figure 1: **From frame prediction errors to event-based decisions.** Accumulated frame errors have different scales across the three violations. PULSE instead evaluates three physical checks per event against normal ranges learned from plausible video. The largest deviation identifies the event and check behind the decision, without requiring a matched clip.

the state after contact is consistent (Kotovsky & Baillargeon, 1994; Hespos & Baillargeon, 2001; Baillargeon & Wang, 2002). Event structure can therefore separate these expectations and test each where its evidence lies (Figure 1, bottom row). This is the idea behind Physics Using Latent State-Events (PULSE), which organizes sequences of 3D bounding boxes into stable-motion intervals separated by event boundaries, following human event segmentation (Zacks et al., 2007). The three questions become three checks, Visibility, Onset, and Outcome, each trained on plausible videos alone to predict visibility or motion at its own event stage. Each check asks the same physical question of every video, and its errors on held-out plausible clips provide a normal range for judging new ones. A video is flagged when any check exceeds its range, and the largest relative deviation identifies the event, check, and objects behind the decision.

We evaluate on three benchmarks that test different aspects of this formulation. We build PhysEvent to provide controlled violations at different event stages together with ground-truth 3D boxes; it lets us examine which checks respond to each violation and which objects carry the strongest evidence, and vary state quality without changing the video. PhysicalConcepts and IntPhys extend the evaluation to other physical scenarios, where 3D boxes must first be recovered from video before the same parsing and checks apply. Together, these benchmarks test whether one decision rule can separate violations across scenes, violation types, and state-recovery conditions, with no matched clip to absorb these differences. PULSE achieves the best single-video detection and paired ranking among the evaluated continuous scorers (0.665 versus 0.519 balanced accuracy for the next best). Baselines with competitive paired ranking remain near chance on standalone decisions, and no choice of threshold closes that gap (Section D.3). Removing any check lowers standalone accuracy, and category-level responses show which expectation each violation challenges.

Our main contributions are:

- **A framework for learning and testing physical expectations.** PULSE organizes physical expectations around event stages, so that different violations can be recognized from a single video using plausible experience alone.
- **A benchmark for studying physical violations.** PhysEvent combines interventions at different event stages with ground-truth 3D boxes, making it possible to ask which expectation a violation breaks and how state quality affects its detection.
- **Evidence beyond paired ranking.** Across three benchmarks, strong paired ranking does not necessarily yield better standalone decisions, whereas event-aligned checks with plausible-only references do.

2 Related Work

Detecting physical violations in video Explicit physics models score individual videos against prescribed dynamics (Smith et al., 2019). Learned frame-level predictors instead score a clip by its prediction error (Mrowca et al., 2018; Bear et al., 2021; Garrido et al., 2025), but interpret that

error mainly within a matched plausible–implausible pair (Riochet et al., 2022; Piloto et al., 2022; Bordes et al., 2025). The same paired reference underlies likelihood-based evaluation (Yuan et al., 2026), while GeoPhys uses plausible and violated validation videos to select its readout (Internò et al., 2026). Language-based evaluators judge a clip directly (Liu et al., 2026), yet general-purpose video multimodal large language models (MLLMs) remain unreliable at judging physical plausibility (Jassim et al., 2024; Bordes et al., 2025; Ballout et al., 2025), and the stronger specialized ones buy their reliability with supervised training on labeled violations (Bansal et al., 2025; Motamed et al., 2025; Bansal et al., 2026). Scoring without a paired reference has been explored for other targets, such as geometric and photometric consistency in generated videos (Oh & Yun, 2026) and functional anomalies in industrial interactions (Li et al., 2025). PULSE keeps the plausible-only training of prediction-based methods but replaces the paired comparison with a per-check reference estimated from plausible videos, so a single clip receives a standalone decision without prescribed dynamics or violation labels.

Structured event representations A long line of work treats an activity as structured events rather than an undifferentiated frame sequence. Event Segmentation Theory holds that people parse continuous activity at perceived boundaries and update their event models there (Zacks et al., 2007; Kurby & Zacks, 2008). Structured video models such as temporal and causal And–Or graphs make this explicit, composing sub-events and grounding fluents like visibility whose changes are caused by interactions (Si et al., 2011; Pei et al., 2013; Qi et al., 2017; Xu et al., 2018). More recent work uses object trajectories and collision events for causal question answering (Yi et al., 2020) and physical state transitions for video-language understanding (Wu et al., 2026). Where these models parse events in order to recognize or explain them, PULSE parses them to test expectations, treating stable intervals and transitions as distinct sources of physical evidence and assigning Visibility to the intervals, Onset to the beginning of an interaction, and Outcome to its aftermath.

Benchmarks for physical violation detection Existing VoE benchmarks test physical expectations via controlled contrasts between plausible and implausible events (Riochet et al., 2022; Piloto et al., 2022; Jassim et al., 2024; Bordes et al., 2025; Weihs et al., 2022). PhysEvent follows this design, with matched plausible–implausible pairs that differ by an intervention at a specific event stage, so we can examine which checks respond to each violation and whether their responses match the expectation being challenged. Its ground-truth 3D boxes further let us hold a video fixed and degrade its object states on demand, separating the scorer's contribution from the quality of its input.

3 Physics Using Latent State-Events (PULSE)

PULSE parses object trajectories into events (Section 3.2), learns from plausible videos alone to predict visibility and state transitions at each event stage (Section 3.3), and judges a new video by how far each check's prediction error departs from its own plausible-validation reference (Section 3.4). The most deviant check scores the video and names the event, check, and objects behind the decision (Figure 2).

3.1 Overview and Notation

Input PULSE operates on object states rather than pixels. An upstream stage turns a video V into $\boldsymbol{X}(V) = \{\boldsymbol{x}_i^t\}_{i=1,t=1}^{N,T}$, where $\boldsymbol{x}_i^t$ holds object i's 3D box and velocity at time t. Frozen image features describing each object's appearance and its surroundings at event boundaries condition these states, and validity masks record which objects and prediction targets are observed. PhysEvent supplies ground-truth boxes natively; for IntPhys and PhysicalConcepts they are lifted from depth and masks (Section 4).

Three checks, four heads Parsing (Section 3.2) marks each event with Start, Middle, and End states, and PULSE asks one physical question at each stage: *Visibility* (should an object be visible?) over the stable interval, *Onset* (does the interaction begin as it should?) at the Start → Middle transition, and *Outcome* (is the state it leads to consistent?) at the Middle → End transition. Visibility is evaluated over two temporal supports, the stable interval and the short transition that follows it, implemented as two heads for the same prediction task; Onset and Outcome use one head each.

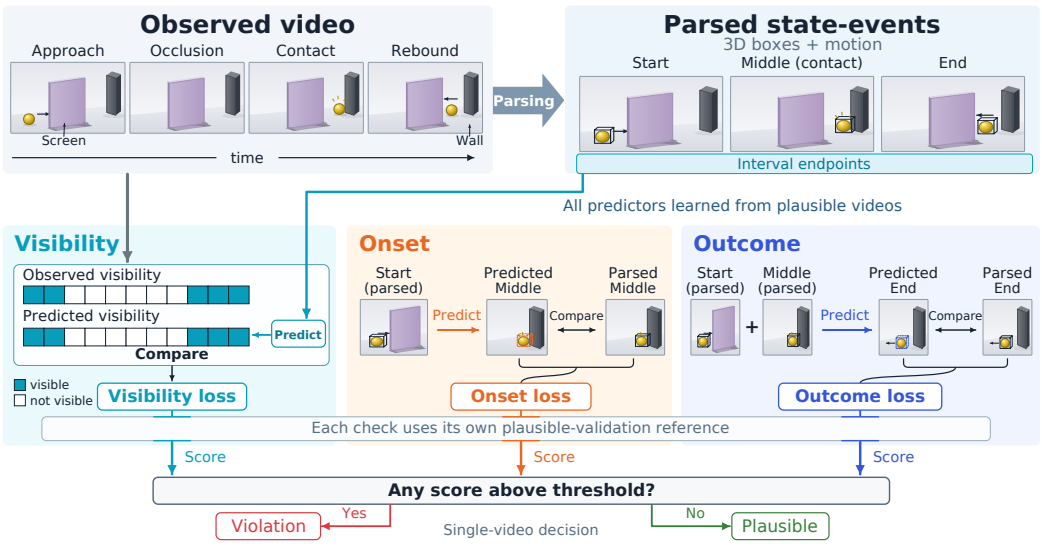


Figure 2: **Event parsing and stage-specific physical predictions.** A ball passes behind a screen, strikes a separate wall, and rebounds. Visibility compares its event-conditioned reconstruction with visibility in the video; Onset and Outcome compare predicted Middle and End states with parsed targets. Each check uses its own plausible-validation reference for the single-video decision.

The four heads $\mathcal{H} = \{\text{vis}^{\text{stable}}, \text{vis}^{\text{boundary}}, \text{ons}, \text{out}\}$ thus collapse into the three reported checks $\mathcal{C} = \{\text{vis}, \text{ons}, \text{out}\}$.

From states to a decision Let $\mathcal{D}^{+}_{\text{train}}$ and $\mathcal{D}^{+}_{\text{val}}$ hold only plausible videos. PULSE learns the checks on $\mathcal{D}^{+}_{\text{train}}$ and, at test time, maps $\boldsymbol{X}(V)$ to a continuous score $\mathcal{S}_{\text{final}}(V)$ together with the maximizing event e^*, check c^*, and object or object-pair index d^* that localize the evidence. A fixed threshold on the score, referenced to $\mathcal{D}^{+}_{\text{val}}$, turns it into a decision, and no step consults a paired clip. The check fixes *what* to predict, parsing fixes *where* to evaluate it, and the check's plausible reference measures *how far* the prediction error departs from its normal range.

3.2 State-Event Parsing

Parsing gives each check a well-defined interval to act on, separating questions about visibility, interaction onset, and subsequent motion. We fit each object's 3D trajectory with piecewise linear regression (Yang et al., 2016), mark changes in fitted velocity or direction as candidate event points, and merge nearby per-object changes into scene-level boundaries. Each resulting event e spans a stable interval and the transition that ends it, with states at times $t^{\text{s}}_e \leqslant t^{\text{m}}_e \leqslant t^{\text{e}}_e$. Start anchors the stable motion, Middle marks the observed onset of a motion change, and End records the state reached once the change is complete; adjacent events share a boundary, so $t^{\text{e}}_e = t^{\text{s}}_{e+1}$. The fitted trajectories also reconstruct object states during occlusion, so the intervals around these boundaries, and not only the boundary instants, can supply the evidence Visibility needs. Fitting, pruning, and merging parameters are in Section E.2.

In Figure 2, Start precedes the ball's passage behind the screen, Middle captures contact with the wall, and End records the rebound. The occlusion interval supplies evidence for Visibility, while contact and the subsequent motion supply evidence for Onset and Outcome.

3.3 Learning the Checks on Plausible Video

Visibility We predict an object's visibility from the states at the interval endpoints and compare the predicted sequence with its observed visibility in the video (Figure 2, left branch). For each event and interval role $r \in \{\text{stable}, \text{boundary}\}$, an edge feature $\boldsymbol{e}^{r}_{ij}$ encodes the camera-relative relation between objects i and j at the endpoints, together with the object features. The head selects relevant relations with a hard Gumbel-Softmax gate (Jang et al., 2017), aggregates them at object nodes, and jointly decodes a fixed-length sequence of visibility probabilities $\hat{v}^{r,t}_i$, where t indexes sampled

positions within the interval. The two roles have separate heads and separate plausible-validation references.

Onset and Outcome A relational graph network (Battaglia et al., 2016) with categorical decoders predicts discretized relative object-pair motion. Onset conditions on the Start state and predicts the Middle; Outcome conditions on Start and Middle and predicts the End. Both predictions are compared with the corresponding parsed states (Figure 2, middle and right branches). Relative states reduce dependence on absolute scene coordinates, and categorical rather than point predictions let a check keep several plausible outcomes open instead of committing to one.

Losses Let $\mathcal{A}^r_{\text{vis}}$ and $\mathcal{A}^\phi_{\text{dyn}}$ be the valid targets of visibility role r and transition check $\phi \in \{\text{ons}, \text{out}\}$, with the event index omitted for readability. With observed visibility $v_i^{r,t}$, predicted distribution $\hat{\boldsymbol{p}}^\phi_{ij,d}$, and target bin $y^\phi_{ij,d}$ for relative-state coordinate d,

$$\mathcal{L}^r_{\text{vis}} = \frac{1}{|\mathcal{A}^r_{\text{vis}}|} \sum_{(i,t)\in\mathcal{A}^r_{\text{vis}}} \text{BCE}(\hat{v}_i^{r,t}, v_i^{r,t}), \qquad \mathcal{L}^\phi_{\text{dyn}} = \frac{1}{|\mathcal{A}^\phi_{\text{dyn}}|} \sum_{(i,j,d)\in\mathcal{A}^\phi_{\text{dyn}}} \text{CE}(\hat{\boldsymbol{p}}^\phi_{ij,d}, y^\phi_{ij,d}). \tag{1}$$

We balance the four heads with fixed weights w_h inversely proportional to their losses under uniform predictions:

$$\mathcal{L}_{\text{train}} = \frac{1}{2|\mathcal{H}|} \sum_{h\in\mathcal{H}} w_h \, \mathcal{L}_h, \qquad w_h = \frac{|\mathcal{H}|/\log n_h}{\sum_{g\in\mathcal{H}} 1/\log n_g}, \tag{2}$$

where the class count n_h is 2 for Visibility and 1921 for Onset and Outcome. Validity masks exclude invalid objects, frames, and pairs; the architecture, the focal-loss factor, and optimization details are in Sections E.3 and E.4.

3.4 Scoring a Single Video

Per-head error A head may have many valid targets across the events of a video (*e.g.*, the sampled visibility outputs of every object). We take the maximum loss over these targets,

$$\ell_h(V) = \max_{a\in\mathcal{A}_h(V)} \ell_{h,a}(V), \qquad h \in \mathcal{H}, \tag{3}$$

so that a single localized failure survives instead of being averaged away by the many targets that look plausible.

Threshold calibration For each head h, we estimate the 10th and 90th percentiles (P10 and P90), $q_{h,0.10}$ and $q_{h,0.90}$, of ℓ_h on the plausible validation set $\mathcal{D}^+_{\text{val}}$, giving each dataset equal total weight. P90 serves as the head's threshold, and the error is expressed relative to the P10–P90 range:

$$S_h(V) = \frac{\ell_h(V) - q_{h,0.10}}{\max(q_{h,0.90} - q_{h,0.10}, \epsilon)}. \tag{4}$$

This maps P10 to zero and the threshold to one for every head, so heads with very different error scales are judged on one footing. P90 is a head-level reference rather than a bound on video-level false positives; Section D.3 reports the sensitivity to this choice.

Video decision The two visibility heads collapse into one check, $S_{\text{vis}} = \max(S^{\text{stable}}_{\text{vis}}, S^{\text{boundary}}_{\text{vis}})$, alongside S_{ons} and S_{out}. A video is scored by its most deviant check (Figure 2, bottom),

$$\mathcal{S}_{\text{final}}(V) = \max(S_{\text{vis}}(V),\, S_{\text{ons}}(V),\, S_{\text{out}}(V)), \tag{5}$$

and is flagged if any head exceeds its threshold:

$$\hat{y}(V) = \mathbb{1}[\exists h \in \mathcal{H} : S_h(V) > 1] = \mathbb{1}[\mathcal{S}_{\text{final}}(V) > 1]. \tag{6}$$

The maximizing check $c^* = \text{argmax}_{c\in\mathcal{C}} S_c(V)$, together with the event e^* and object or object-pair index d^* of the same maximizing target, reports *what* failed, *where*, and *on which object*. Section D.2 compares this maximum with softer aggregation rules.

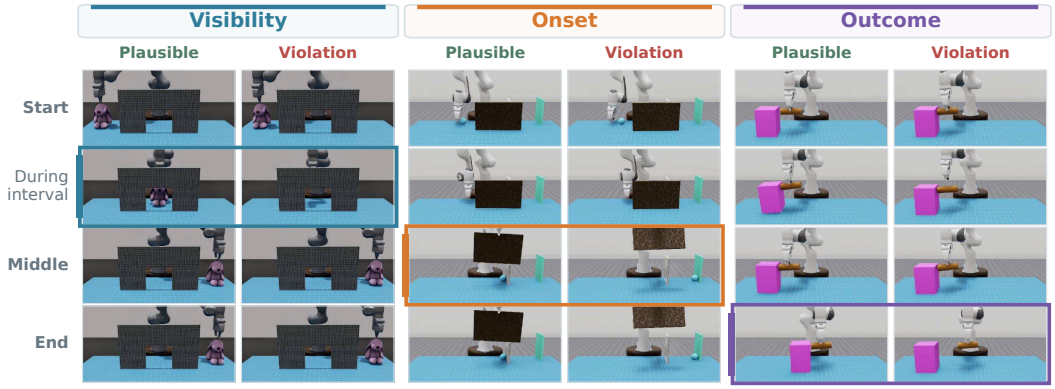


Figure 3: **Representative `PhysEvent` violations at three event stages.** Paired plausible and implausible clips differ only in object visibility (Visibility), the Start → Middle trigger (Onset), or the Middle → End state (Outcome). The stage targeted by each intervention is used only to evaluate check-level responses and never for training.

4 Benchmarks for Physical Violation Detection

We evaluate on `PhysEvent`, which we build, and on two public VoE benchmarks brought to the same object-state interface. `PhysEvent` is a synthetic benchmark of long, multi-event sequences built in Isaac Sim (Gao et al., 2026). Its interventions at different event stages let us examine which checks respond and which objects carry the strongest evidence, and each clip provides RGB, depth, instance masks, temporally aligned tracks, camera parameters, and native 3D boxes, so state quality can be varied while the video is held fixed.

Six categories across three event stages Six categories, each rendered as left/right variants under a fixed camera, target visibility during occlusion, interaction onset, or the state reached afterward (Figure 3). Continuity and Height alter object visibility relative to the occlusion geometry (Visibility); Collision and Gravity alter whether, where, or when an interaction begins (Onset); Attachment and Containment leave onset intact but change the state reached afterward (Outcome). Each intervention therefore targets a primary event stage, although its physical consequences can elevate several checks at once. We use this stage label only for analysis, asking whether a model trained with no violation-type supervision produces the expected check-level responses.

Training and test design `PhysEvent` contains 10,800 videos at 30 FPS (mean duration 66.5 s), including 7,200 plausible training and 2,400 plausible validation clips. Half of the training set consists of *event-non-occurrence* controls, in which the visibility change or interaction does not occur; these controls ground the learned expectations in object states rather than in a fixed event sequence. The 1,200 test videos form 600 matched plausible–implausible pairs, each differing only in the targeted intervention. Category statistics and split-integrity checks are in Section A.

Shared state interface PhysicalConcepts and IntPhys provide plausible training corpora separate from their evaluation videos, so the same checks can be learned and calibrated across different physical scenarios. For PhysicalConcepts we pair its released temporally consistent masks (Piloto et al., 2022) with monocular depth from Video Depth Anything (Chen et al., 2025); for IntPhys we use native depth and per-frame masks (Riochet et al., 2022), recovering temporally consistent identities with SAM2 where the released ones are insufficient (Ravi et al., 2025). In both, mask pixels are lifted with source-specific camera assumptions into axis-aligned 3D boxes that match the format of `PhysEvent`'s native boxes (Table A11; details in Section E.1).

5 Experiments

The three benchmarks jointly test whether learned physical expectations support standalone decisions across scenes, violation types, and levels of state quality, with no matched clip to absorb these differences.

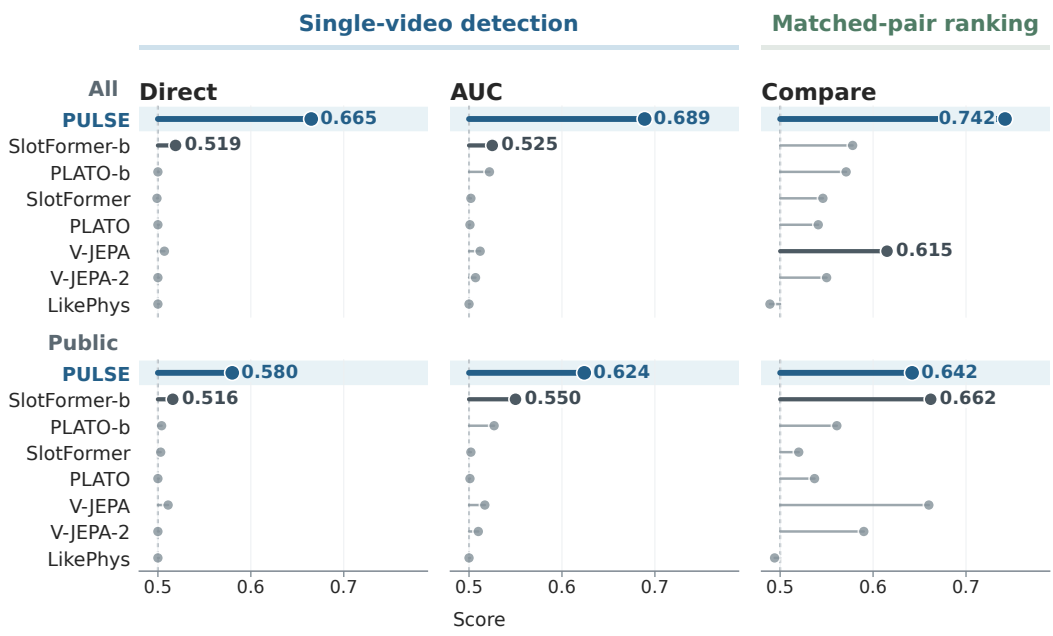


Figure 4: **Single-video detection and paired ranking.** Continuous scorers on All (`PhysEvent`, PhysicalConcepts, and IntPhys) and Public (PhysicalConcepts and IntPhys), with equal dataset weights. Trainable scorers show three-seed means; Table A2 lists all values with standard errors. Dashed lines mark chance.

Protocol We report three metrics. Direct is balanced accuracy for standalone decisions at a fixed threshold, and Compare is the fraction of matched pairs in which the implausible video scores higher, both macro-averaged over datasets; Mixed AUC, hereafter AUC, measures threshold-free separation on scores pooled across datasets with equal dataset weight. Only Compare uses the pairing. Every trainable scorer is trained, selected, and calibrated on plausible videos alone (Section B). We compare against the temporal predictors PLATO (Piloto et al., 2022) and SlotFormer (Wu et al., 2023), each in its original slot-based form and in a "-b" variant that receives the same 3D box states as `PULSE` together with whole-frame DINOv3 features; the frozen video representations V-JEPA (Bardes et al., 2024) and V-JEPA-2 (Assran et al., 2025); and the video-diffusion scorer LikePhys with Hunyuan T2V (Yuan et al., 2026). Our PLATO implementation closely matches the published results under the original evaluation protocol (Section F).

5.1 Single-Video Detection Beyond Paired Ranking

Trained jointly on plausible videos from the three benchmarks, `PULSE` leads the continuous scorers on all three metrics (Figure 4), with 0.665 Direct, 0.689 AUC, and 0.742 Compare against 0.519 and 0.525 for SlotFormer-b on Direct and AUC and 0.615 for V-JEPA on Compare. Direct and AUC are the demanding metrics here, since a fixed threshold and a pooled ranking both require one violation scale to hold across the three datasets.

The "-b" variants of PLATO and SlotFormer receive the same box states as `PULSE` yet remain near chance on Direct, so access to object geometry alone does not explain the standalone gain. Nor does the operating point. Across every plausible-validation percentile from P10 to P100, the baselines' Direct stays between 0.492 and 0.541 (Table A10), and replacing their temporal averaging with peak pooling raises AUC and Compare but leaves Direct near chance (Table A15).

On the Public split, the trained scorers are kept and recalibrated on plausible validation videos from PhysicalConcepts and IntPhys only. `PULSE` reaches 0.580 Direct and 0.624 AUC against 0.516 and 0.550 for SlotFormer-b, the next best on both, whereas SlotFormer-b and V-JEPA overtake it on Compare (0.662 and 0.660 versus 0.642) while trailing on Direct and AUC. Better matched-pair ranking therefore does not imply better single-video detection.

Because the `PhysEvent` test videos were not public at evaluation time, they also allow a contamination-free comparison with zero-shot MLLMs. GPT-5 and Gemini models (Singh et al., 2025; Team et al., 2023) (GPT-5.6 Luna and Gemini-3.8-Flash) receive sampled RGB frames with

the unchanged IntPhys 2 prompt (Bordes et al., 2025) and return Yes/No judgments, scored by Direct (Figure 5). Both stay near chance (0.553 and 0.568), whereas PULSE, whose task-specific training, selection, and calibration used plausible videos alone, reaches 0.821 on the same videos.

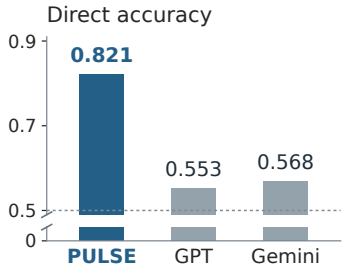


Figure 5: **Direct on `PhysEvent`.** Zero-shot MLLMs versus the all-dataset PULSE scorer on the 1,200 test videos.

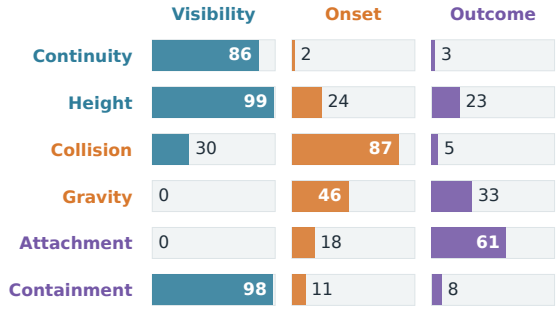
5.2 Event-Aligned Checks Organize the Decision

Removing any one of the three calibrated checks from the trained model lowers Direct (Figure 6a), most of all for Visibility. Removing Outcome lowers all three metrics, and removing Onset costs 1.5 percentage points of Direct with much smaller losses in AUC and Compare (per-dataset values in Table A5).

Onset's role is clearer in the per-category responses on PhysEvent (Figure 6b). Onset dominates for Collision and Gravity, Visibility for Continuity and Height, and Outcome for Attachment, so the dominant check matches the targeted stage in five of six categories without any violation-type supervision. Containment is the exception. Its intervention targets Outcome, yet Visibility crosses its threshold far more often (98.3% versus 8.0%), because the independent trajectory fits of Section 3.2 can miss how a hidden object moves with its container during occlusion, which distorts both its predicted visibility and the recovered inside–outside relation. Responses on the public benchmarks and the entity readouts behind Figure 6b are in Section C.

	Direct	AUC	Compare
Full	**0.665**	**0.689**	**0.742**
w/o Visibility	0.595 ↓ 7.0	0.628 ↓ 6.0	0.662 ↓ 8.0
w/o Onset	0.650 ↓ 1.5	0.688 ↓ 0.1	0.739 ↓ 0.3
w/o Outcome	0.630 ↓ 3.5	0.654 ↓ 3.4	0.711 ↓ 3.1

(a) Check ablation



(b) Per-category check responses

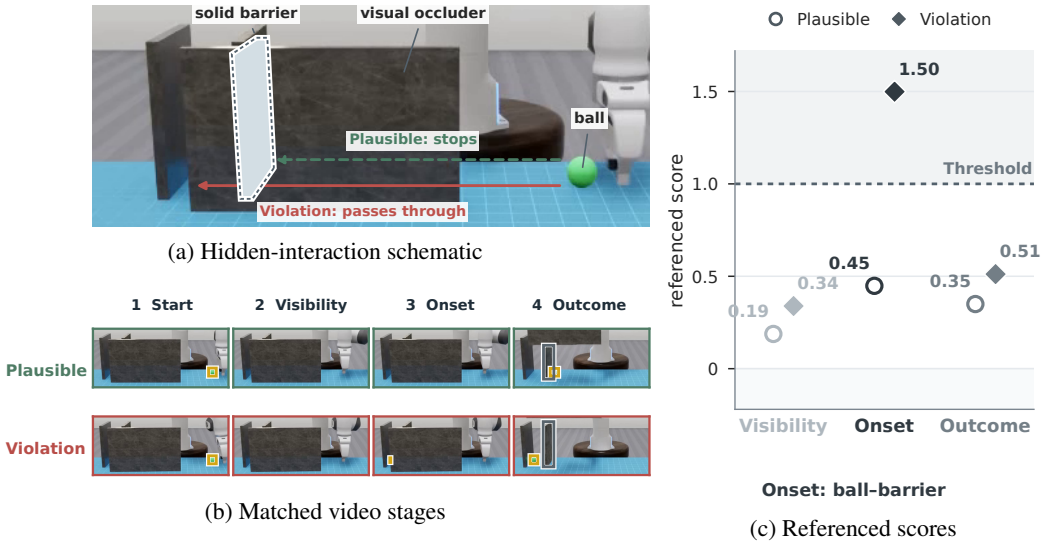
Figure 6: **Contributions and category responses of the event-aligned checks.** (a) Three-seed means for the full model and for each check removal; arrows show changes in percentage points. (b) Per-check detection rates (%) on the 100 violations per category; checks may co-activate. Label colors mark target stages; bold values mark dominant responses.

The collision pair in Figure 7 shows the readout at the object level. An occluder hides the ball's encounter with a solid barrier, so the ball is never seen passing through it; afterward it reappears at the barrier in the plausible video but beyond it in the violation. Visibility stays below threshold in both clips, as their occlusion is identical, and so does Outcome. Onset alone crosses the threshold, rising from 0.45 to 1.50, and its largest error falls on the ball–barrier pair (Figure 7b), the objects that should have collided, rather than on the occluder. The score that detects the violation thus also names the check and objects involved.

5.3 State Quality Shapes Event-Level Scoring

Holding the PhysEvent videos fixed lets us vary the evidence available to the checks without changing the physical events. Errors in object states alter both the motion the checks see and the parsed boundaries that decide when each check is evaluated. Figure 8 compares native states, six levels of box perturbation, predicted boxes, and predicted depth, each with training and plausible-validation calibration matched to its state interface and all at seed 42 (numbers in Table A7; seed variation of the all-dataset model is small, Table A2).

For the predicted-box control, 2D localization is held fixed with prompts derived from ground-truth masks, and WildDet3D (Huang et al., 2026), fine-tuned on the PhysEvent training split, predicts 3D boxes from these prompts and camera intrinsics without ground-truth depth.



(a) Hidden-interaction schematic

(b) Matched video stages

(c) Referenced scores

Figure 7: **Score-aligned readout for a collision pair.** (a) The x-ray view exposes the hidden barrier and counterfactual paths. (b) Boxes mark the ball–barrier pair returned by the violation's Onset readout across all three seeds, with the same objects highlighted in the plausible clip. (c) Mean scores over three seeds; Onset alone exceeds the threshold in every run.

Performance varies non-monotonically across noise levels. Each condition repeats event parsing, training, and calibration, so these results characterize full-pipeline sensitivity. The predicted-box condition keeps most of the native-state performance and beats predicted depth on all three metrics; its AUC is 0.845 against 0.843 for native states and 0.682 for predicted depth. `PULSE` can therefore operate on reconstructed states, but the quality of 3D recovery bounds the reliability of its judgments.

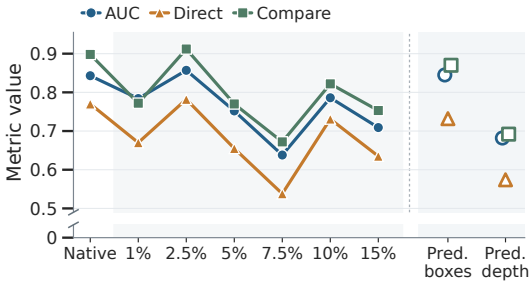


Figure 8: **State-quality controls.** Training and calibration use `PhysEvent` alone and match the evaluation state interface; all conditions use seed 42.

6 Conclusion

Judging one video in isolation requires a reference, but that reference need not be a matched clip. `PULSE` draws it from plausible experience, separating physical expectations by event stage and judging each against its own normal range. Across three benchmarks, this formulation gives the strongest overall single-video detection among the evaluated continuous scorers and leads in aggregate paired ranking. The same decision retains the event, check, and objects carrying the strongest evidence, so the judgment can be inspected. Event structure thus connects what a model learns to expect with the particular expectation an observation challenges.

Limitations and future work `PULSE` depends on consistent object identities and accurate 3D trajectories, which are hard to recover under camera motion and occlusion. Relation-aware parsing could improve hidden-state recovery in interactions such as containment, where independent trajectory fits do not model how objects move together. Zero-shot transfer beyond the three benchmark domains also lies beyond the current scorer. Advances in tracking and 3D reconstruction (Ravi et al., 2025; Chen et al., 2025; Huang et al., 2026) could let the same checks learn from diverse interactions in large video collections such as HowTo100M (Miech et al., 2019), and such broader experience is the natural path toward more general single-video detection. Finally, the Visibility–Onset–Outcome vocabulary covers stable intervals and discrete transitions. Deformation, sustained contact, friction, and changes whose evidence accumulates without a clear boundary call for richer event descriptions and further checks, which the event–check formulation can accommodate by giving each new expectation its own temporal support and plausible reference.

AI USE DISCLOSURE

Generative AI tools were used to improve the clarity and presentation of the manuscript, to assist with literature retrieval and reference checking, and to assist with software implementation. The research questions, hypotheses, methodology, and experimental design were developed by the authors without generative AI. All AI-suggested references were verified by the authors, all AI-assisted edits were reviewed and revised by the authors, and all AI-assisted code was reviewed by two members of the research team. The authors take full responsibility for the content of this work, including its text, claims, code, and results.

REPRODUCIBILITY STATEMENT

Sections 3 to 5 specify the `PULSE` formulation, the plausible-only calibration rule, and the evaluation protocol. The appendix documents `PhysEvent` construction and splits, external-dataset preprocessing, object-state construction, model architecture and optimization, baseline configurations, and complete per-dataset results. The submission includes an anonymous repository with the preprocessing, training, calibration, and evaluation code and the configurations used in this paper (`https://anonymous.4open.science/r/PULSE-6EFC/README.md`). `PhysEvent` will be released upon publication.

REFERENCES

Andréa Aguiar and Renée Baillargeon. 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive Psychology*, 39(2):116–157, 1999. 1

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 7, A10

Renée Baillargeon and Su-hua Wang. Event categorization in infancy. *Trends in Cognitive Sciences*, 6(2):85–93, 2002. 2

Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985. 1

Mohamad Ballout, Serwan Jassim, and Elia Bruni. Pixels to principles: Probing intuitive physics understanding in multimodal language models. *arXiv preprint arXiv:2507.16572*, 2025. 3

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 *Proceedings of International Conference on Learning Representations (ICLR)*, 2025. 3

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 *Proceedings of International Conference on Learning Representations (ICLR)*, 2026. 1, 3

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. 7, A10

Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 5

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences (PNAS)*, 110(45):18327–18332, 2013. 1

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021. 2

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. 1, 3, 8, A11

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 6, 9, A8

Sicong Gao, Maurice Pagnucco, Tomasz Bednarz, and Yang Song. Nvidia isaac sim: Enabling scalable, gpu-accelerated simulation for robotics. *arXiv preprint arXiv:2606.03551*, 2026. 6, A1

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025. 2

Susan J Hespos and Renée Baillargeon. Reasoning about containment events in very young infants. *Cognition*, 78(3):207–245, 2001. 2

Weikai Huang, Jieyu Zhang, Sijun Li, Taoyang Jia, Jiafei Duan, Yunqian Cheng, Jaemin Cho, Matthew Wallingford, Rustin Soraki, Chris Dongjoo Kim, Shuo Liu, Donovan Clay, Taira Anderson, Winson Han, Ali Farhadi, Bharath Hariharan, Zhongzheng Ren, and Ranjay Krishna. Wilddet3d: Scaling promptable 3d detection in the wild. *arXiv preprint arXiv:2604.08626*, 2026. 8, 9, A6

Christian Internò, Alexander Pondaven, Habon Issa, Fabio Pizzati, Francesco Pinto, Markus Olhofer, Ivan Laptev, Philip Torr, Eero P Simoncelli, Barbara Hammer, et al. Geophys: The geometry of physical plausibility. *arXiv preprint arXiv:2606.20707*, 2026. 3

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 4

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. Grasp: a novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. 3

Laura Kotovsky and Renée Baillargeon. Calibration-based reasoning about collision events in 11-month-old infants. *Cognition*, 51(2):107–129, 1994. 2

Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2):72–79, 2008. 3

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. 1

Alan M Leslie and Stephanie Keeble. Do six-month-old infants perceive causality? *Cognition*, 25 (3):265–288, 1987. 1

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. A10

Chang Liu, Yunfan Ye, Qingyang Zhou, Xichen Tan, Mengxuan Luo, Zhenyu Qiu, Wei Peng, and Zhiping Cai. Video-hoca: A diagnostic benchmark for physical anomaly reasoning in video-llms. *arXiv preprint arXiv:2602.19571*, 2026. 3

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. A10

Francesco Margoni, Luca Surian, and Renée Baillargeon. The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*, 131(3):716–748, 2024. 1

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 9

Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility. *arXiv preprint arXiv:2510.07550*, 2025. 1, 3

Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li F Fei-Fei, Josh Tenenbaum, and Daniel L Yamins. Flexible neural representation for physics prediction. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

Amy Needham and Renee Baillargeon. Intuitions about support in 4.5-month-old infants. *Cognition*, 47(2):121–148, 1993. 1

Yun Oh and Sukmin Yun. Reference-free assessment of physical consistency in world model-based video generation. *arXiv preprint arXiv:2606.22363*, 2026. 3

Mingtao Pei, Zhangzhang Si, Benjamin Z Yao, and Song-Chun Zhu. Learning and parsing video events with goal and intent prediction. *Computer Vision and Image Understanding (CVIU)*, 117 (10):1369–1383, 2013. 3

Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9): 1257–1267, 2022. 1, 3, 6, 7, A5, A8, A12

Siyuan Qi, Siyuan Huang, Ping Wei, and Song-Chun Zhu. Predicting human activities using stochastic grammar. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. 3

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2025. 6, 9, A8

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys 2019: A benchmark for visual intuitive physics understanding. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(9):5016–5025, 2022. 1, 3, 6, A5, A8

Zhangzhang Si, Mingtao Pei, Benjamin Yao, and Song-Chun Zhu. Unsupervised learning of event and-or grammar and semantics from video. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2011. 3

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. A8

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 7

Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

Elizabeth S Spelke. Principles of object perception. *Cognitive Science*, 14(1):29–56, 1990. 1

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. A9

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. A10

Luca Weihs, Amanda Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking progress to infant-level physical reasoning in ai. *Transactions on Machine Learning Research*, 2022. 3

Shang Wu, Haoran Lu, Songling Liu, Chenwei Xu, Lie Lu, Pranav Maneriker, Fan Du, Zhaoran Wang, and Han Liu. Apt: Atomic physical transitions for causal video-language understanding. *arXiv preprint arXiv:2606.18586*, 2026. 3

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023. 7

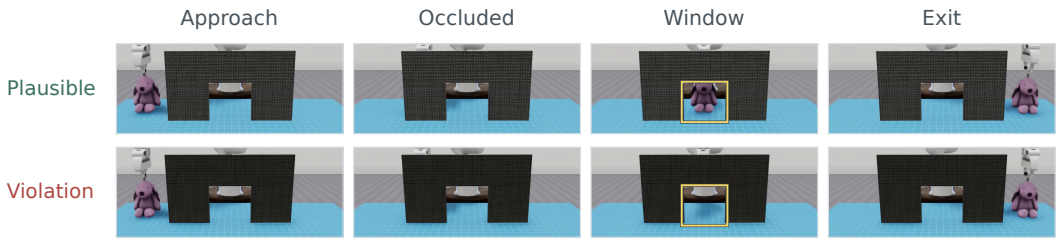
Yuanlu Xu, Lei Qin, Xiaobai Liu, Jianwen Xie, and Song-Chun Zhu. A causal and-or graph model for visibility fluent reasoning in tracking interacting objects. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

Lingjian Yang, Songsong Liu, Sophia Tsoka, and Lazaros G Papageorgiou. Mathematical programming for piecewise linear regression analysis. *Expert Systems with Applications*, 44:156–167, 2016. 4

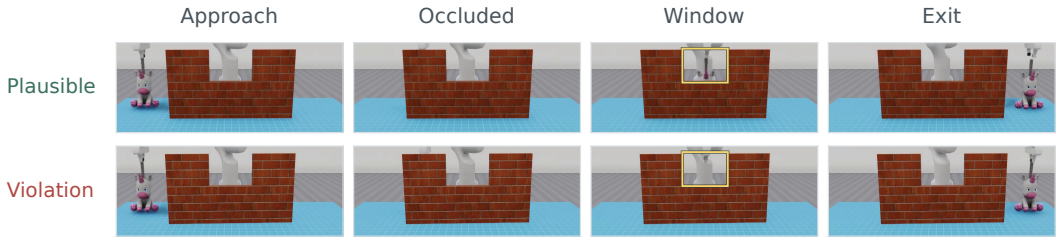
Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020. 3

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2026. 3, 7, A11, A12

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007. 2, 3



(a) **Continuity (Visibility).** The object fails to appear at the lower window but emerges on the far side.



(b) **Height (Visibility).** The tall object becomes fully hidden where its upper part should remain visible.

Figure A1: **`PhysEvent` intervention sequences.** Each pair progresses from left to right, with the plausible clip above its implausible counterpart. Columns align corresponding stages rather than equal elapsed times, and yellow boxes mark the windows or objects to compare.

Table A1: **Mean video duration by category (seconds).** Categories have equal numbers of videos in each split.

Continuity	Height	Collision	Gravity	Attachment	Containment
86	92	74	27	57	63

A PhysEvent Interventions and Splits

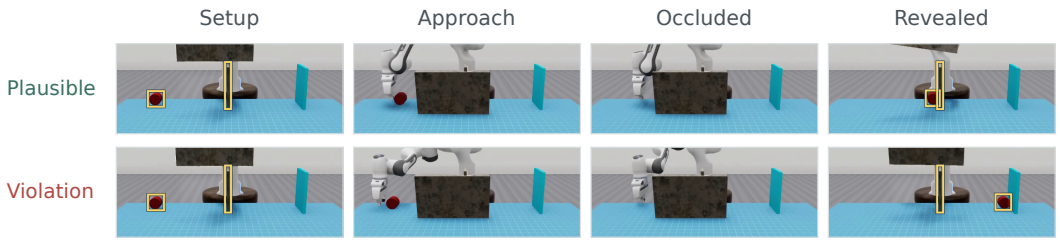
`PhysEvent` is generated in NVIDIA Isaac Sim (Gao et al., 2026), where a Franka arm sets up tabletop interactions. The simulator records RGB, depth, instance masks, object identities, camera parameters, and 3D boxes at 30 FPS. Object shape, color, size, and initial placement vary under a fixed camera, and each configuration is rendered as left and right variants. Figure A1 shows the six interventions as matched sequences. Matched clips share the camera, object geometry, and initial scene configuration, and each intervention leaves everything outside its target stage intact. Continuity and Height preserve motion direction and transit endpoints, Collision and Gravity preserve the motion before contact or support loss, and Attachment and Containment preserve the initial attachment or enclosure and the subsequent cylinder or cup path.

Splits. Initial scene configurations are disjoint across training, validation, and test. Each category contributes 1,200 plausible training videos, 400 plausible validation videos, and 100 test configurations, each rendered as one plausible and one implausible video. About half of the training clips are event-non-occurrence controls (Section 4). Table A1 gives the mean video duration by category.

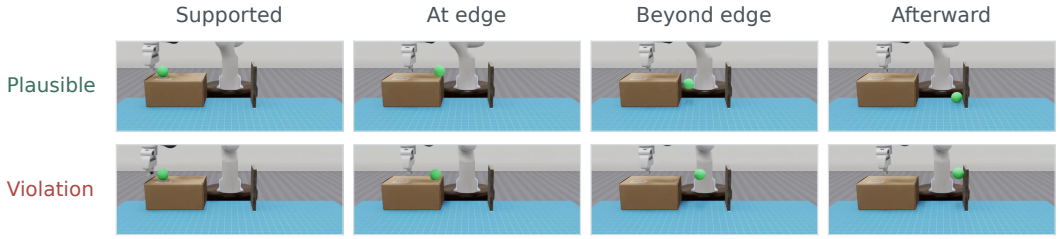
Data and code availability. We will release `PhysEvent` together with processed, ready-to-use versions of PhysicalConcepts and IntPhys, including compressed videos, the object-state representations used in our experiments, and split manifests. Redistributed material from the two public datasets keeps its original CC BY 4.0 license, with source attribution and documentation of our modifications.

B Evaluation Protocol and Complete Results

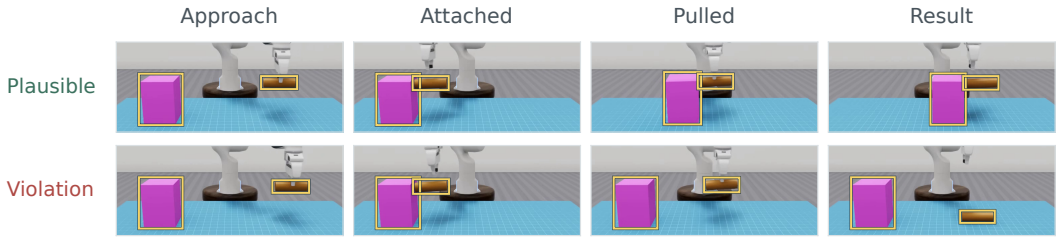
Evaluation set. The common evaluation set for continuous scorers contains 1,200 `PhysEvent` test videos, 5,000 PhysicalConcepts test videos (1,000 per physical property), and 360 IntPhys dev videos. Each dataset is balanced between plausible and implausible videos, giving 3,280 matched pairs in total.



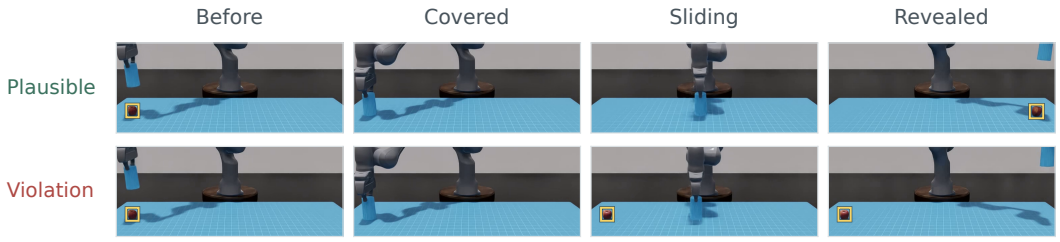
(c) **Collision (Onset).** The ball passes through the central solid barrier while hidden by the screen, instead of stopping at it.



(d) **Gravity (Onset).** The ball stays at support height after leaving the edge, rather than falling.



(e) **Attachment (Outcome).** The rigidly attached cube is left behind when the cylinder is pulled.



(f) **Containment (Outcome).** The cup slides along the table, but the enclosed object is left behind and exposed.

Figure A1: **`PhysEvent` intervention sequences (continued).**

Metrics and dataset weighting. Violation is the positive class, and chance is 0.5 for every metric. Direct is balanced accuracy, $\frac{1}{2}(\mathrm{TPR} + \mathrm{TNR})$, at each scorer's fixed threshold. Compare is the fraction of matched pairs in which the implausible video scores higher, with half credit for ties. Both are computed per dataset and macro-averaged. Mixed AUC, written AUC in all tables, instead pools the test scores of all datasets with per-video weight proportional to $1/|D_i^{\mathrm{test}}|$, so that each dataset carries equal total weight; it is not in general the mean of the per-dataset AUCs.

Thresholds and settings. Continuous baselines threshold their scalar video scores at the P90 of their plausible-validation scores; Physics Using Latent State-Events (`PULSE`) uses the per-head P90 rule in Equation (6). All-dataset results combine the three benchmarks. Public results restrict both calibration and evaluation to PhysicalConcepts and IntPhys, keeping the trained scorers and equal dataset weights throughout.

Complete results. Tables A2 and A3 give the aggregate and per-dataset values behind Figure 4, with standard errors for the trainable scorers. `PULSE` has the highest Direct on each of the three datasets, with SlotFormer-b closest on PhysicalConcepts (0.553 versus 0.594). On Compare,

Table A2: **Complete aggregate results for Figure 4.** Trainable scorers report mean ± standard error over three seeds; frozen models report point estimates. Each dataset has equal metric weight. For Public, calibration uses only the two included datasets.

Method	Direct	AUC	Compare
All datasets			
PULSE	0.665 ± 0.0081	0.689 ± 0.0046	0.742 ± 0.0037
PLATO-b	0.500 ± 0.0007	0.522 ± 0.0028	0.571 ± 0.0152
PLATO	0.500 ± 0.0003	0.501 ± 0.0000	0.541 ± 0.0091
SlotFormer-b	0.519 ± 0.0009	0.525 ± 0.0020	0.578 ± 0.0046
SlotFormer	0.499 ± 0.0004	0.502 ± 0.0000	0.546 ± 0.0047
V-JEPA	0.507	0.512	0.615
V-JEPA-2	0.500	0.507	0.550
LikePhys	0.500	0.500	0.489
Public			
PULSE	0.580 ± 0.0021	0.624 ± 0.0025	0.642 ± 0.0064
PLATO-b	0.504 ± 0.0017	0.527 ± 0.0012	0.561 ± 0.0075
PLATO	0.500 ± 0.0005	0.501 ± 0.0001	0.537 ± 0.0118
SlotFormer-b	0.516 ± 0.0013	0.550 ± 0.0015	0.662 ± 0.0015
SlotFormer	0.503 ± 0.0005	0.502 ± 0.0002	0.520 ± 0.0077
V-JEPA	0.511	0.517	0.660
V-JEPA-2	0.500	0.510	0.590
LikePhys	0.500	0.500	0.494

Table A3: **Per-dataset results for continuous scorers.** Trainable models report three-seed means; frozen models report point estimates. All-dataset references are used except in the final row, which recalibrates PULSE on the two Public datasets.

	PhysEvent			PhysicalConcepts			IntPhys		
Method	Direct	AUC	Compare	Direct	AUC	Compare	Direct	AUC	Compare
PULSE	0.821	0.893	0.921	0.594	0.633	0.679	0.580	0.629	0.626
PLATO-b	0.500	0.555	0.591	0.500	0.535	0.557	0.499	0.527	0.565
PLATO	0.498	0.504	0.551	0.500	0.503	0.551	0.503	0.500	0.522
SlotFormer-b	0.500	0.482	0.408	0.553	0.588	0.727	0.504	0.534	0.598
SlotFormer	0.499	0.505	0.598	0.501	0.507	0.530	0.498	0.500	0.509
V-JEPA	0.505	0.503	0.525	0.505	0.505	0.543	0.511	0.533	0.778
V-JEPA-2	0.500	0.494	0.468	0.500	0.514	0.553	0.500	0.511	0.628
LikePhys	0.498	0.497	0.478	0.500	0.500	0.499	0.500	0.500	0.489
PULSE (Public)	–	–	–	0.580	0.625	0.676	0.580	0.633	0.608

Table A4: **Operating characteristics of the all-dataset detector.** Three-seed means; every row uses the fixed all-dataset per-head references. TPR and TNR are violation and plausible-video recall.

Dataset	Direct	TPR	TNR
PhysEvent	0.821	84.4%	79.9%
PhysicalConcepts	0.594	47.6%	71.2%
IntPhys	0.580	90.9%	25.0%
All	0.665	74.3%	58.7%

SlotFormer-b leads on PhysicalConcepts (0.727 versus 0.679) and V-JEPA on IntPhys (0.778 versus 0.626), which is the source of the Compare reversal in the Public setting (Section 5).

Operating characteristics. Table A4 separates violation recall (TPR) from plausible-video recall (TNR) under the all-dataset decision rule of PULSE. Their macro-averages give Direct 0.665, with FPR 41.3% and FNR 25.7%. The low plausible-video recall on IntPhys (25.0%) is the cost of

Table A5: **Per-dataset Compare under check removal.** Entries are three-seed means; each ablation disables one calibrated check at evaluation time. Aggregate effects are in Figure 6a.

Setting	PhysEvent	PhysicalConcepts	IntPhys
Full model	0.921	0.679	0.626
w/o Visibility	0.804	0.602	0.580
w/o Onset	0.870	0.702	0.646
w/o Outcome	0.926	0.591	0.618

Table A6: **Check responses by property on the public benchmarks.** Percentages of implausible videos exceeding each check's threshold, using the fixed all-dataset references. Checks may co-activate, so rows need not sum to 100%.

Dataset	Property	n	Visibility	Onset	Outcome
IntPhys	Object permanence (O1)	60	88.3%	50.0%	41.7%
IntPhys	Shape constancy (O2)	60	75.0%	71.7%	36.7%
IntPhys	Spatio-temporal continuity (O3)	60	56.7%	55.0%	60.0%
PhysicalConcepts	Continuity	500	41.6%	5.6%	1.0%
PhysicalConcepts	Directional inertia	500	0.8%	0.0%	7.6%
PhysicalConcepts	Object persistence	500	57.0%	6.4%	9.6%
PhysicalConcepts	Solidity	500	32.0%	3.6%	5.2%
PhysicalConcepts	Unchangeableness	500	35.4%	9.4%	84.2%

thresholds shared across the three datasets; Section D.3 shows how raising the percentile trades it against violation recall.

C Check Contributions and Evidence Readouts

Check ablations. Table A5 gives per-dataset Compare after removing each check from the same calibrated model; the aggregate effect on all three metrics is in Figure 6a. Removing Visibility lowers Compare on all three datasets. Removing Outcome leaves PhysEvent almost unchanged but costs PhysicalConcepts nearly nine points, where Unchangeableness relies on Outcome (Table A6). Removing Onset lowers Compare on PhysEvent, whose Collision and Gravity interventions alter interaction onset directly (Figure 6b), yet raises it slightly on both public benchmarks, whose properties do not map one-to-one onto event stages; in aggregate it costs little Compare but 1.5 points of Direct.

Responses on plausible videos. Figure A2 complements the violation responses in Figure 6b with false-positive rates on their matched plausible counterparts, using the same trained scorers and thresholds.

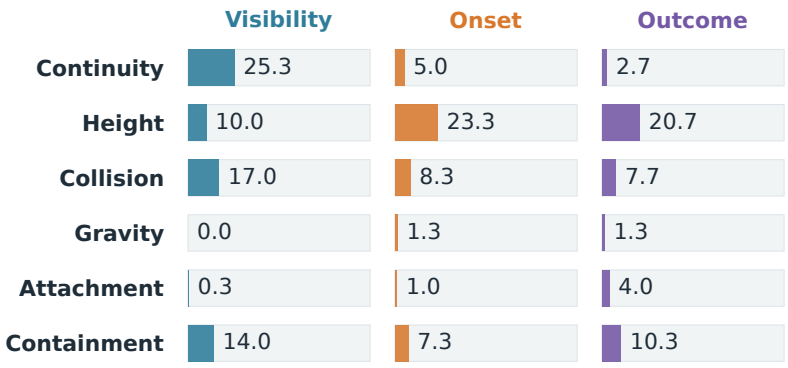


Figure A2: **Check responses on plausible PhysEvent videos.** Per-check false-positive rates (%) on the 100 matched plausible test videos per category, averaged over three seeds with the all-dataset thresholds unchanged. Checks may co-activate; gray tracks indicate 100%.

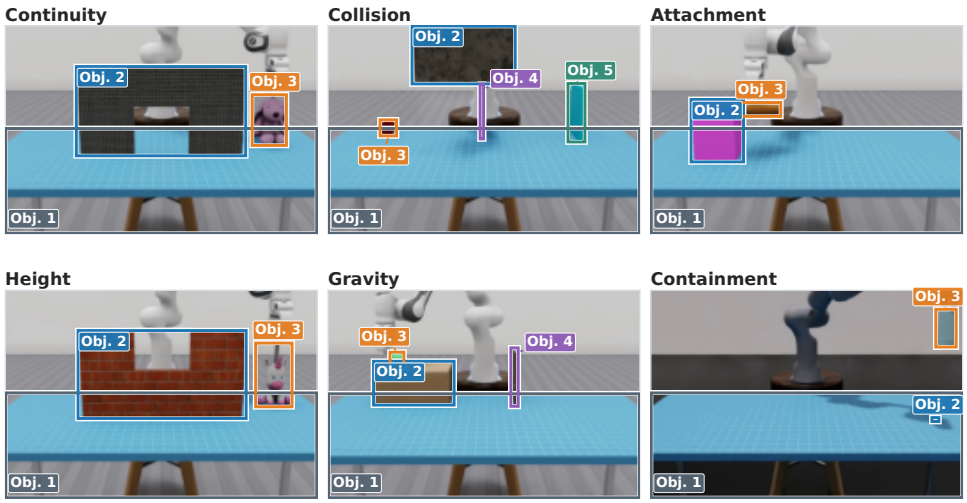


Figure A3: **Objects used in the entity readouts.** Boxes from native instance masks identify object IDs; colors match Figure A4. The robot is excluded from the scored object set.

	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 5
Continuity	0.5%	1.3%	98.2%	—	—
Height	4.3%	3.2%	92.5%	—	—
Collision	4.7%	3.2%	44.3%	40.0%	7.8%
Gravity	7.5%	12.2%	49.7%	30.7%	—
Attachment	21.7%	28.5%	49.8%	—	—
Containment	1.0%	96.8%	2.2%	—	—

Figure A4: **Entity readout frequencies across `PhysEvent` categories.** Each maximum-loss Visibility target contributes unit mass to its object; each relational target divides unit mass between its pair endpoints. Values are averaged over three seeds on the same 100 violations per category; dashes mark absent objects.

Public-benchmark property responses. The five PhysicalConcepts properties (Piloto et al., 2022) and three IntPhys blocks (Riochet et al., 2022) allow the same check-response analysis as Figure 6b. Table A6 uses the all-dataset references without property-specific calibration. On PhysicalConcepts, Visibility responds most to Continuity and Object persistence and Outcome to Unchangeableness, whereas Directional inertia (0.8%/0.0%/7.6% for Visibility/Onset/Outcome) and Solidity (32.0%/3.6%/5.2%) are detected more weakly. Both properties concern motion that changes or continues without a visible interaction, and the short PhysicalConcepts clips leave little stable motion on either side of a boundary for Onset and Outcome to compare. On IntPhys, Visibility responds most often to object permanence and shape constancy, while spatio-temporal continuity elicits similar rates across the three checks.

Entity readout. The maximizing target retains its event and entity indices, one object for Visibility or the object pair of the maximizing coordinate for Onset and Outcome. Figure A3 connects these indices to the `PhysEvent` scenes, and Figure A4 summarizes their frequencies across violations. Readouts concentrate on the moving object in Continuity and Height (Obj. 3; 98.2% and 92.5%) and on the ball and central barrier in Collision (Obj. 3/4). Gravity retains the ball but attributes more to

Table A7: **State-quality controls on fixed `PhysEvent` videos.** Each row uses `PhysEvent`-only training matched to its state interface and its own plausible-validation calibration at seed 42. IoU is the median IoU between perturbed and native boxes.

Condition	IoU	Direct	AUC	Compare
Native states	1.000	0.770	0.843	0.898
1% box noise	0.955	0.670	0.784	0.772
2.5% box noise	0.891	0.782	0.857	0.912
5% box noise	0.797	0.655	0.752	0.770
7.5% box noise	0.714	0.538	0.638	0.672
10% box noise	0.642	0.731	0.786	0.822
15% box noise	0.520	0.635	0.709	0.753
Predicted boxes (WildDet3D)	–	0.732	0.845	0.870
Predicted depth	–	0.574	0.682	0.692

Table A8: **Discretization and event-window sensitivity.** Each variant retrains and recalibrates the scorer at seed 42 after changing one setting; the default uses 1921 bins and a $1/3$-second merging window.

Factor	Value	Direct	AUC	Compare
Default	1921 bins, $1/3$ second	0.680	0.696	0.735
Discretization	1025 bins	0.671	0.700	0.741
Discretization	3841 bins	0.662	0.691	0.733
Event window	$2/3$ second	0.680	0.690	0.761
Event window	$1/6$ second	0.673	0.687	0.747

the barrier than to the support box, and Attachment retains the cylinder while splitting the attribution of its partner between the block and the table. Containment concentrates on the enclosed ball (Obj. 2; 96.8%) rather than on the ball–cup pair, consistent with its dominant Visibility response. Because `PhysEvent` supplies native identities, these indices have a consistent meaning across videos.

D State Quality and Scoring Sensitivity

D.1 State-Quality Controls

Table A7 gives the numbers behind Figure 8. All conditions train on `PhysEvent` alone and calibrate on its plausible validation split at seed 42, both on the same state interface as the evaluation; states, parsed events, and visual event context are recomputed for each interface. The non-monotonic results characterize sensitivity of this full pipeline.

Box perturbations. Independent zero-mean Gaussian noise is added to the six faces of each visible box, with standard deviation equal to the stated percentage of the box extent along each axis; Table A7 lists the resulting median IoU with the native boxes.

Predicted boxes. WildDet3D (Huang et al., 2026), fine-tuned on the `PhysEvent` training split, predicts 3D cuboids from 2D box prompts derived from ground-truth masks and the known camera intrinsics, without ground-truth depth. The fixed prompts hold 2D localization constant, so the condition varies 3D recovery alone.

Predicted depth. Masks and tracks are held fixed. Under the fixed camera, Video Depth Anything estimates are calibrated frame by frame on visible table pixels against a fixed depth reference for the static table (Obj. 1), built from native depth in the training split.

D.2 Discretization and Check Aggregation

Table A8 varies the state discretization or the event-merging window at seed 42, retraining and recalibrating each variant under the all-dataset protocol. Bin-count changes keep the class-count weighting of Equation (2); window changes keep the per-object trajectory fits and recompute the

Table A9: **Final check aggregation.** All rows use the same seed-42 check scores and a decision threshold of 1, with Visibility first taking the maximum of its two heads. Softmax temperature is 1 and log-sum-exp uses log-mean-exp. Bold marks the best value per column.

Aggregation rule	Direct	AUC	Compare
Max check score	**0.680**	**0.696**	0.735
Top-2 check average	0.630	0.680	0.736
Softmax-weighted check score	0.622	0.685	0.748
Mean check score	0.567	0.657	**0.753**
Log-sum-exp check score	0.601	0.674	0.744

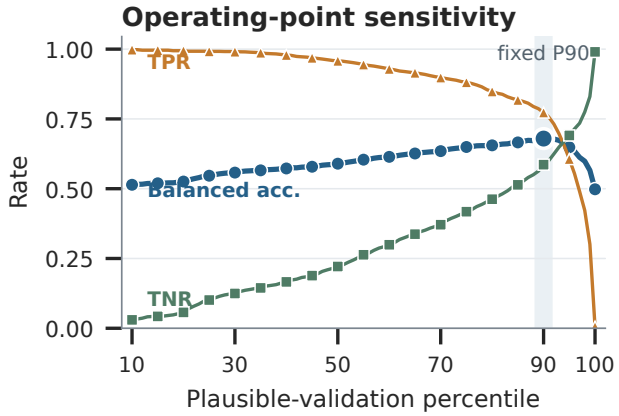


Figure A5: **Sensitivity to the upper plausible-validation percentile.** Direct, violation recall (TPR), and plausible-video recall (TNR) for the seed-42 model.

merged events and their visual context. Coarser and finer discretizations, and halving or doubling the window, move all three metrics by under three points.

Table A9 compares rules for combining the three check scores into the video score of Section 3.4. Maximum aggregation gives the highest AUC, supporting its use for separating plausible and implausible videos across scenes. Mean aggregation instead gives the highest Compare, illustrating that better within-pair ranking need not imply better cross-video separation.

D.3 Threshold Sensitivity

Each head is referenced to its own plausible-error distribution (Equation (4)), so every physical prediction task uses the same percentile criterion regardless of the shape of its error tail, and a video is flagged when any of the four heads crosses its boundary. P90 is therefore a head-level reference rather than a bound on video-level false positives. With four heads at 10% each, the video-level rate reaches $1 - 0.9^4 \approx 34\%$ even for independent heads; the realized rates are in Table A4.

Figure A5 varies the upper plausible-validation percentile from P10 to P100 for the seed-42 model. Raising it trades violation recall for plausible-video recall.

The six baselines in Table A10 are evaluated at every integer percentile from P10 to P100, of which the table shows the deciles. Across the full sweep their Direct stays between 0.492 and 0.541 despite large swings in violation and plausible-video recall, so no operating point closes the gap to PULSE. Results at the main operating point, with three-seed estimates for the adapters, are in Table A2.

E Method and Training Details

E.1 Object-State Inputs and Visual Context

All datasets use the same object-state interface; Table A11 distinguishes native annotations from estimated components.

Table A10: **Threshold sensitivity of continuous baselines.** Columns give plausible-validation percentiles. Calibration and metrics weight the three datasets equally; adapters use seed 42 and the JEPA models the main-comparison protocol. Video scores stay fixed throughout.

Method	P10	P20	P30	P40	P50	P60	P70	P80	P90	P100
Direct										
PLATO-b	0.505	0.518	0.535	0.535	0.539	0.509	0.513	0.502	0.501	0.501
PLATO	0.500	0.500	0.500	0.500	0.499	0.501	0.501	0.502	0.501	0.500
SlotFormer-b	0.492	0.501	0.508	0.528	0.538	0.540	0.524	0.521	0.517	0.501
SlotFormer	0.501	0.500	0.500	0.500	0.499	0.503	0.501	0.502	0.498	0.500
V-JEPA	0.522	0.504	0.499	0.500	0.500	0.505	0.511	0.508	0.507	0.500
V-JEPA-2	0.503	0.502	0.506	0.505	0.507	0.504	0.501	0.502	0.500	0.500
Violation recall (TPR, %)										
PLATO-b	93.8	88.3	82.2	64.4	60.7	51.0	46.2	40.0	33.2	0.4
PLATO	81.3	78.2	77.1	71.9	61.9	49.7	35.7	24.7	10.3	0.0
SlotFormer-b	95.1	87.4	75.4	64.8	64.5	62.9	58.3	51.2	37.0	0.2
SlotFormer	81.3	78.4	76.6	73.8	64.6	52.1	39.5	27.7	12.6	0.0
V-JEPA	93.2	81.6	66.6	58.6	54.1	46.4	35.1	21.7	12.3	0.0
V-JEPA-2	92.1	86.7	81.6	71.8	58.8	45.8	11.2	1.5	0.1	0.0
Plausible-video recall (TNR, %)										
PLATO-b	7.2	15.2	24.7	42.5	47.0	50.9	56.5	60.5	66.9	99.8
PLATO	18.8	21.8	22.9	28.1	38.0	50.5	64.6	75.7	89.8	100.0
SlotFormer-b	3.4	12.8	26.1	40.8	43.1	45.1	46.4	53.0	66.4	100.0
SlotFormer	18.9	21.6	23.5	26.2	35.3	48.5	60.7	72.8	87.1	100.0
V-JEPA	11.2	19.2	33.3	41.5	45.9	54.5	67.1	79.8	89.1	100.0
V-JEPA-2	8.5	13.8	19.6	29.2	42.6	55.0	89.0	98.9	99.9	100.0

Table A11: **Sources of the object-state inputs.** External boxes are lifted from depth and masks; `PhysEvent` supplies native 3D boxes.

Dataset	Depth	Mask/track	3D state
`PhysEvent`	GT	GT	native
PhysicalConcepts	estimated	GT	lifted
IntPhys	GT	GT/SAM2	lifted

Depth and identity. For IntPhys we use native depth and framewise instance masks (Riochet et al., 2022), propagating identities with SAM2 when the released mask identifiers are not temporally consistent (Ravi et al., 2025). For PhysicalConcepts the released masks already provide tracks (Piloto et al., 2022), and depth is estimated with the Video Depth Anything ViT-L checkpoint `video_depth_anything_vitl.pth` (Chen et al., 2025).

Metric scale and lifting. For PhysicalConcepts, the estimated depth is calibrated once per video against ground-plane geometry. Reference depths at released ground-mask pixels come from intersecting camera rays with the generator's ground plane ($y = 0\,\mathrm{m}$) under the released camera pose and an assumed 30° horizontal field of view, so no ground-truth depth is used. IntPhys uses the 90° horizontal field of view given in its training metadata, and when extrinsics are unavailable the camera pitch is estimated by fitting a plane to the native floor depth. Back-projected object pixels are transformed with the camera pose, and their axis-aligned bounds define the 3D boxes.

Frozen visual context. DINOv3 ViT-L/16 (Siméoni et al., 2025) supplies globally pooled 1024-dimensional image features from the checkpoint `facebook/dinov3-vitl16-pretrain-lvd1689m`. An object's fixed descriptor sums the features of masked, centered views over its visible frames, normalized by total foreground-pixel support. Event context uses object-centered views at Start and Middle that retain the surrounding objects; for a hidden object, the nearest earlier frame with a visible mask is used.

Table A12: **Inputs and targets of the four prediction heads.** All state inputs are interval endpoints; + denotes concatenation. Visibility outputs are jointly decoded sequences at sampled interval positions. Object descriptors and event context are frozen 1024-dimensional image features.

Head	Pair-state input	Visual conditioning	Prediction target
Stable Visibility	Start + Middle (2×24)	Object descriptor (1024)	64 binary visibility targets per object
Boundary Visibility	Middle + End (2×24)	Object descriptor (1024)	4 binary visibility targets per object
Onset	Start (24)	Object descriptor + Start context (2048)	Middle position relation (12)
Outcome	Start + Middle (2×24)	Object descriptor + Middle context (2048)	End position and velocity relation (24)

E.2 Event Parsing

The parser fits each object's trajectory with a piecewise linear function,

$$x(t) = x(0) + v(0)t + \sum_{i=1}^{n} \Delta v_i \max(0, t - t_i), \tag{A1}$$

where t_i are candidate event times and Δv_i velocity changes. The initial temporal grid is uniform with at most 50 candidate points, fewer for short clips. L-BFGS jointly optimizes the ordered event times and the trajectory parameters to minimize the squared error over visible frames and all six box coordinates. Greedy pruning then removes the candidate whose removal increases this per-object error least, and stops when the smallest increase exceeds 10 for `PhysEvent` and IntPhys or 5 for PhysicalConcepts. A $1/3$-second window merges the remaining per-object changes into scene-level boundaries.

For each merged cluster of motion changes, Middle and End bracket its earliest and latest times, rounded outward to source frames, and Start is the preceding cluster's End or the first frame for the initial event. A terminal triplet covers the remaining stable interval with Middle and End at the last frame. Checks whose associated interval has zero duration are masked out.

The fitted segments also reconstruct states during occlusion under the same piecewise-linear assumption. These parsed states provide the inputs and targets of every check. Visibility predicts visibility at sampled times within an interval from its endpoints; Onset and Outcome predict the Middle and End states.

E.3 State Representation and Check Architecture

Each object's state comprises a six-coordinate position box and its six coordinate velocities. Onset and Outcome use world-coordinate boxes, $[x_{\min}, y_{\min}, z_{\min}, x_{\max}, y_{\max}, z_{\max}]$; Visibility uses camera-projected endpoint states with (u, v, z) in place of (x, y, z), where the image coordinates and their velocities are scaled by $10/W$ or $10/H$ for image width W and height H and camera-axis depth stays metric. Writing a position or velocity six-vector as lower and upper three-vectors $\boldsymbol{m}, \boldsymbol{M}$, its ordered-pair representation is

$$[\boldsymbol{m}_i - \boldsymbol{m}_j,\ \boldsymbol{m}_i - \boldsymbol{M}_j,\ \boldsymbol{M}_i - \boldsymbol{m}_j,\ \boldsymbol{M}_i - \boldsymbol{M}_j],$$

and concatenating the 12 velocity and 12 position components gives the 24-dimensional pair state. The model accepts up to 12 objects and masks invalid slots and targets.

Discretization. Each continuous component is discretized into 1921 bins, symmetric about zero with widths growing exponentially toward the tails. Bin indices $q \in \{0, \ldots, 1920\}$ are rescaled to $2q/1920 - 1$ before Fourier encoding (Tancik et al., 2020) and projection to 256-dimensional features; training adds zero-mean Gaussian noise with standard deviation 0.01 to the rescaled indices.

Heads. Table A12 lists the inputs, visual conditioning, and targets of the four heads. Visibility targets summarize the observed binary sequence over uniform interval bins; a hard Gumbel-Softmax gate selects relational evidence at object nodes, and a transposed-convolution decoder produces the

Table A13: **Shared training configuration.** Datasets are sampled equally.

Parameter	Value	Parameter	Value
Optimizer	AdamW (Loshchilov & Hutter, 2019)	Weight / EMA decay	0.01 / 0.999
Learning rate	10^{-4}	Warmup	1000 steps
LR decay	Cosine to 10^{-5}	Batch / dataset	64
Epochs	600	Train / val. batches	500 / 100 per epoch
Gumbel temperature	1.8 to 0.4	Annealing interval	10k–220k steps
Gradient clipping	1.0	Seeds	42 / 0 / 2026

Table A14: **Temporal inputs for the JEPA evaluators.** Frame step is measured in decoded frames; context lengths count input frames, including the prepended frame for PhysicalConcepts.

Dataset	Window length	Frame step	Context lengths
`PhysEvent`	16	15	2 / 4 / 6 / 8 / 10
PhysicalConcepts	16	1	2 / 4 / 6 / 8 / 10
IntPhys	16	2	2 / 4 / 6 / 8 / 10

visibility sequence. Onset and Outcome use six relational GNN layers with sum/max aggregation and six-layer, eight-head Transformer decoders (Vaswani et al., 2017) that jointly predict the target coordinates.

E.4 Training and Calibration

Plausible-only learning refers to the task-specific training, model selection, and calibration of the `PULSE` scorer; the pretrained perception and feature encoders supply frozen inputs, and no implausible video or violation annotation enters any of these steps. Training samples the three datasets equally. The losses in Equation (2) use class counts $(2, 2, 1921, 1921)$ for stable Visibility, boundary Visibility, Onset, and Outcome, and the focal-loss weighting with $\alpha = 0.5$ and $\gamma = 0$ (Lin et al., 2017), which reduces to cross-entropy scaled by $1/2$, supplies the common factor. Evaluation uses the EMA checkpoint with the best plausible-validation loss. Table A13 lists the shared configuration.

Validation data. Model selection and calibration use the same fixed plausible validation splits of 2,400 `PhysEvent`, 5,000 PhysicalConcepts, and 3,000 IntPhys videos. PhysicalConcepts uses held-out freeform videos, separate from its violation probes; IntPhys reserves plausible training videos for validation. These splits are disjoint from training and test, with equal total calibration weight per dataset.

F Baseline Protocols and Diagnostics

F.1 Scoring Protocols

PLATO and SlotFormer. The four entries are reimplementations that follow the original papers within a shared evaluation pipeline. Unsuffixed adapters predict object-slot sequences; the -b adapters replace object-slot inputs with the same 3D box states supplied to `PULSE`, while retaining whole-frame DINOv3 features for temporal prediction. Each is trained for 300,000 steps at seeds 42/0/2026, with the 10,400 plausible validation videos used for per-seed score referencing and threshold selection. Slot-input adapters decode predicted slots to RGB with a frozen decoder and predicted hard masks, and score a video by pixel SSE on 128×128 RGB frames in $[0, 1]$ averaged over predicted frame occurrences; box-input adapters use six-coordinate SSE averaged over real objects and predicted frame occurrences. Overlapping predictions contribute separately, padding is excluded, and validation and test share this readout.

V-JEPA and V-JEPA-2. We use the frozen ViT-H checkpoints `vith16.pth.tar` (V-JEPA, non-RoPE; Bardes et al., 2024) and `vith.pt` (V-JEPA-2; Assran et al., 2025) with center-square inputs at resolutions 224 and 256, respectively. Table A14 gives the sampling configurations, shared

Table A15: **Baseline error pooling.** Mean ± standard error over three seeds. All variants use the same predictions and plausible-only calibration.

Model	Pooling	Direct	AUC	Compare
All datasets				
PLATO-b	Mean	0.500 ± 0.0007	0.522 ± 0.0028	0.571 ± 0.0152
	Temporal max	0.509 ± 0.0005	0.537 ± 0.0016	0.616 ± 0.0070
	Object–time max	0.508 ± 0.0013	0.544 ± 0.0015	0.648 ± 0.0073
PLATO	Mean	0.500 ± 0.0003	0.501 ± 0.0000	0.541 ± 0.0090
	Temporal max	0.498 ± 0.0003	0.505 ± 0.0001	0.552 ± 0.0130
SlotFormer-b	Mean	0.519 ± 0.0009	0.525 ± 0.0020	0.578 ± 0.0046
	Temporal max	0.513 ± 0.0012	0.545 ± 0.0019	0.615 ± 0.0074
	Object–time max	0.509 ± 0.0003	0.555 ± 0.0020	0.610 ± 0.0071
SlotFormer	Mean	0.499 ± 0.0004	0.502 ± 0.0000	0.546 ± 0.0047
	Temporal max	0.498 ± 0.0004	0.505 ± 0.0003	0.558 ± 0.0065
Public				
PLATO-b	Mean	0.504 ± 0.0017	0.527 ± 0.0012	0.561 ± 0.0075
	Temporal max	0.510 ± 0.0009	0.546 ± 0.0002	0.586 ± 0.0095
	Object–time max	0.504 ± 0.0008	0.547 ± 0.0004	0.612 ± 0.0052
PLATO	Mean	0.500 ± 0.0005	0.501 ± 0.0001	0.537 ± 0.0117
	Temporal max	0.500 ± 0.0016	0.502 ± 0.0002	0.539 ± 0.0131
SlotFormer-b	Mean	0.516 ± 0.0013	0.550 ± 0.0015	0.662 ± 0.0015
	Temporal max	0.511 ± 0.0019	0.555 ± 0.0009	0.643 ± 0.0057
	Object–time max	0.505 ± 0.0028	0.557 ± 0.0012	0.627 ± 0.0075
SlotFormer	Mean	0.503 ± 0.0005	0.502 ± 0.0002	0.520 ± 0.0077
	Temporal max	0.500 ± 0.0005	0.502 ± 0.0001	0.546 ± 0.0081

by validation and test; windows advance by two sampled frames over each video's own length, and PhysicalConcepts retains all 15 frames and prepends the first frame to form one 16-frame window, with the duplicate in the context rather than the targets. A video's score averages, over valid windows, the minimum prediction loss across the prescribed context lengths. The same label-blind rule is used for all properties.

LikePhys with Hunyuan T2V. The released LikePhys scorer (Yuan et al., 2026) uses frozen Hunyuan T2V at 512×320 resolution. We sample 61 frames uniformly over each `PhysEvent` and IntPhys video, whereas PhysicalConcepts retains all 15 frames and prepends two copies of the first frame. All categories and splits use the same neutral scene-description prompt, guidance scale 6, and seed 42 reset for each video. The score averages the noise-prediction MSE over ten uniformly indexed noise timesteps and enters the shared protocol of Section B.

Binary MLLM references. GPT-5.6 Luna and Gemini-3.8-Flash are evaluated only on the 1,200 `PhysEvent` test videos, through the API IDs `gpt-5.6-luna` and `gemini-3.8-flash` with the unchanged IntPhys 2 `prompt1` (Bordes et al., 2025), temperature 0, and high reasoning effort. These evaluations were completed on September 7 and September 4, 2026, respectively. Frames are sampled at stride 60 and capped at 50 to respect the IntPhys 2 frame limit. Each video has one retained response per model, all of which yield a valid Yes/No decision with Yes denoting plausibility, and Direct is computed from these binary decisions.

F.2 Additional Diagnostics

Baseline error pooling. To test whether temporal averaging hides localized errors, we recompute each adapter's video score as a maximum instead of a mean over the same predictions (Table A15). Temporal max takes the largest per-frame SSE, averaged over objects for box-input adapters or summed over RGB pixels for slot-input adapters; box-input adapters also support an object–time

Table A16: **Label-aware JEPA context diagnostic.** Context is selected by Compare within each test property; thresholds use plausible validation scores under that context mixture.

Model	Dataset	Direct	AUC	Compare
V-JEPA	IntPhys	0.497	0.536	0.817
V-JEPA	PhysicalConcepts	0.519	0.527	0.637
V-JEPA	PhysEvent	0.507	0.505	0.572
V-JEPA-2	IntPhys	0.500	0.512	0.656
V-JEPA-2	PhysicalConcepts	0.501	0.514	0.574
V-JEPA-2	PhysEvent	0.500	0.502	0.535

max, the largest six-coordinate SSE over all object–frame targets. Checkpoints and temporal sampling are unchanged, and each reduction uses its own plausible-validation calibration under the dataset-balanced protocol of Section B. Peak pooling raises all-dataset AUC and Compare, but Direct stays between 0.498 and 0.513 across the peak-pooled rows, so temporal averaging alone does not account for the gap to PULSE (0.665 Direct).

PLATO tuple evaluation. On PhysicalConcepts, the PLATO slot-input adapter is also scored with the original four-video rule, under which the summed scores of two implausible clips must exceed those of their plausible counterparts (Piloto et al., 2022). Across 1,250 complete tuples its three-seed property-macro accuracy is $74.32\% \pm 2.53\%$, close to the 73.42% property macro-average of the original paper.

LikePhys PPE evaluation. Applying the released PPE aggregation to our IntPhys scores over the 90 four-video scenes with equal property weight gives 39.4%, against the published 30.5% (lower is better; Yuan et al., 2026).

JEPA context selection. Table A16 keeps the main JEPA checkpoints and temporal sampling but selects, for each test property, the context length with the highest Compare; this label-aware diagnostic differs from the label-blind main comparison. Dataset-balanced AUC is 0.518 for V-JEPA and 0.508 for V-JEPA-2, and Direct 0.508 and 0.500. Even with this selection, V-JEPA's 0.817 Compare on IntPhys coexists with 0.536 AUC and 0.497 Direct, preserving the distinction between paired ranking and standalone detection.