VIOLATE: EVALUATING VIDEO-LANGUAGE MODELS ON FINE-GRAINED PHYSICAL IMPLAUSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Model-generated videos can look realistic while containing brief implausible events. Video-level evaluations cannot show which event occurred, when it occurred, or whether several events occur in the same video. We introduce VIOLATE, a benchmark for fine-grained physical implausibility detection. It contains 556 human-annotated synthetic videos with 1,552 events. Each annotation describes what goes wrong and when it happens. We also use 150 real videos to measure false detections. The benchmark supports two tasks. In event verification, a model decides whether a proposed implausible event is actually visible. In open-ended event detection, the model receives only the video and must find, describe, and localize every implausible event, or return no event when the video is plausible. Current models struggle with both tasks. Across CoT and non-CoT prompting, GPT-5.5 reaches the best off-the-shelf Matthews correlation coefficient of 0.084 for verification, while the best fine-tuned model reaches 0.087; both values show weak discrimination. For open-ended detection, Claude Opus 4.8 reaches 0.468 F1 for matching event descriptions and 0.336 F1 for their time intervals. We propose three-stage training with teacher distillation, reinforcement learning (RL), and self-distillation. These stages first learn from teacher-generated reasoning, then improve event recovery, and finally learn from the RL model's own predictions together with videos that contain no annotated implausible event. Its best variants reach 0.518 semantic F1 and 0.405 temporal F1. We will release VIOLATE to support evaluation and training for physical implausibility in model-generated videos.

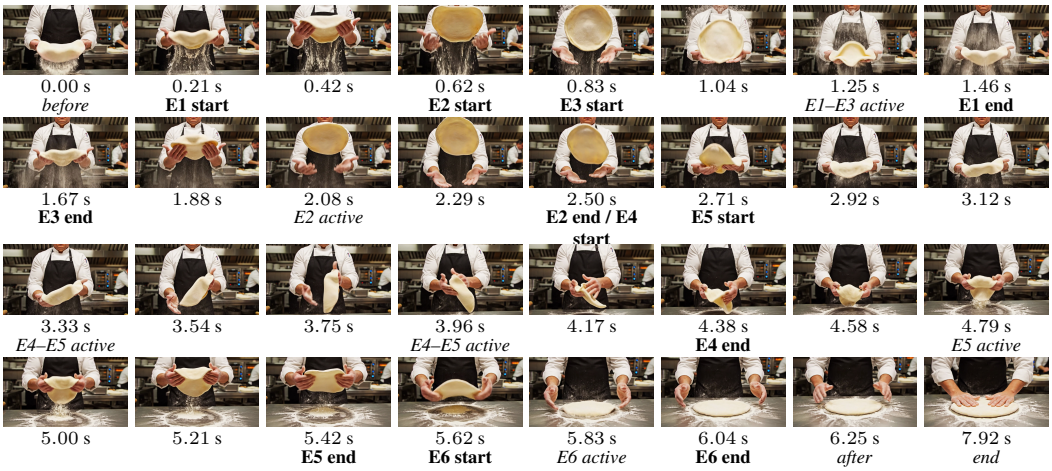
1 INTRODUCTION

Video generation models have improved rapidly, and many model-generated videos now look realistic at first glance. Existing benchmarks capture this progress through visual quality, temporal consistency, prompt alignment, and human preference (Liu et al., 2024; Huang et al., 2024; 2026; He et al., 2024). However, physical plausibility remains a distinct challenge. Even strong generators may cause an object to disappear, change identity, pass through a solid surface, or deform in an unnatural way (Bansal et al., 2025; Meng et al., 2025; Bansal et al., 2026; Zheng et al., 2025). Each implausible event may last only a few frames, and one video may contain many such events at the same time or at different times.

Current evaluation protocols do not capture this event-level structure. Most benchmarks assign one score or judgment to the entire video. This is useful for ranking generators, but it does not reveal which event occurred or when. Other benchmarks ask questions about a known physical rule or a predefined event (Bansal et al., 2025; Meng et al., 2025; Li et al., 2025; Motamed et al., 2025). These questions narrow the event that the model must examine. Recent work provides explanations, temporal annotations, and more detailed diagnostics (Wang et al., 2025; Zhang et al., 2026; Pothiraj et al., 2026). However, these benchmarks do not jointly test open-ended discovery, description, temporal localization, and the ability to return no event for a plausible video.

We study this problem as **fine-grained video implausibility detection**. Unlike a video-level judgment, the output is a variable-size set of localized events. The central challenge is to recover brief and subtle implausible events while returning no events for a plausible video.

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.



E1 (0.21–1.46 **s):** After the toss, the dough begins rotating but descends smoothly and lands on the same side without completing a visible flip.

E2 (0.62–2.50 **s):** The dough remains suspended unusually high and flat with little support, behaving like a rigid disk rather than flexible dough on a ballistic arc.

E3 (0.83–1.67 **s):** Large flour clouds burst around the dough without matching contact. Their amount and direction do not follow the dough's motion or gravity.

E4 (2.50–4.38 **s):** The dough abruptly changes from a round sheet to an elongated folded strip while the chef's hands do not drive the transformation.

E5 (2.71–5.42 **s):** The round dough folds and collapses into an irregular piece, then suddenly returns to a round disk.

E6 (5.62–6.04 **s):** On landing, the curved sheet instantly becomes a smooth, centered circular disk with little wrinkling or edge deformation.

Figure 1: A VIOLATE training example with six localized implausible events. We show 32 sampled frames in temporal order, including every annotated event boundary and the final frame. The intervals overlap, and one frame can mark both the end of one event and the start of another.

Figure 1 illustrates this event-level structure: one eight-second video contains six distinct, temporally overlapping implausible events that cannot be represented by a single video-level label.

We introduce VIOLATE, a human-annotated benchmark of synthetic videos, supplemented with real videos for measuring false detections. Annotators revise model proposals, add missed events, and adjudicate disagreements. Section 2 summarizes the data and quality controls, and Appendix B gives the full protocol and cost accounting.

We study two complementary tasks: **event verification**, which judges a supplied event description and interval, and **open-ended event detection**, which finds and localizes all events without candidate descriptions.

Current verifiers remain weak, so we train the detector directly against human event annotations. The best self-distilled variant reaches 0.518 semantic F1 and 0.405 temporal F1, improving over the strongest off-the-shelf detector.

In summary, our contributions are as follows:

- We introduce VIOLATE, with **556 human-annotated synthetic videos**, 1,552 localized event annotations, and 150 real evaluation videos.
- We provide a unified evaluation of **event verification**, **open-ended event detection**, and the ability to return no event on plausible videos.
- We show that current verifiers provide weak feedback and propose a **teacher distillation–RL–self-distillation** detector that improves event recovery while reducing false detections on plausible videos.

2 Task, Data, and Evaluation

2.1 Tasks

Scope: We use *implausible event* to mean a visible inconsistency in the depicted scene. Such an event can violate physical, causal, temporal, or spatial consistency. Examples include unsupported

motion, unexplained appearance or disappearance, object or body deformation, identity changes, and motion that contradicts contact. Camera motion, perspective, occlusion, editing, overlays, watermarks, and encoding artifacts do not count as implausible events.

Event verification: A model receives a video and one candidate event description. The candidate also states when the event begins and ends. The model returns `correct` when the described implausible event is visible during that interval. It returns `wrong` when the event is absent, inaccurately described, or physically ordinary. We write a candidate as (s, e, d), where s and e are the start and end times and d is the description.

Open-ended event detection: Here the model receives only the video. It must find every implausible event, describe it, and give its start and end times. The model is not told how many events occur or what they are. Formally, for video V, it predicts

$$\hat{Y}(V) = \{(\hat{s}_i, \hat{e}_i, \hat{d}_i)\}_{i=1}^{\hat{N}}, \tag{1}$$

where each output contains a description and two temporal boundaries. If the video contains no implausible event, the correct output is the empty set $Y = \emptyset$. Annotators record different implausible events separately, even when they happen at the same time. They record one persistent event only once.

2.2 Dataset and Annotation

VIOLATE contains 556 human-annotated synthetic videos, with 381 for training and 175 for evaluation. We source the prompts from the VideoPhy-2 hard subset (Bansal et al., 2026) and render the videos with Veo 3.1 Fast. Appendix A gives the prompt selection, split construction, generation configuration, and failed attempts.

Among the training videos, 344 contain at least one annotated implausible event and 37 contain none. The evaluation set includes 170 videos with annotated implausible events and 5 with none. The two splits contain 937 and 615 events, respectively, yielding 1,552 human-reviewed events in total. Videos with at least one annotated implausible event contain 3.02 events on average.

We use LLM-proposed events only as an annotation scaffold. For evaluation, two annotators independently review each complete video, and a third resolves their disagreements. Training annotations receive a separate quality-control pass. We also intersperse 25 real control videos during evaluation annotation.

As shown in Table 1, annotators substantially revise model proposals. On the synthetic evaluation videos, they reject 56.9% of GPT-5.4 proposals and 82.8% of Gemini 3.1 Pro proposals, and add 318 missed events. On the real controls, they accept no proposal unchanged. A targeted final audit found that the two retimed proposals and three human-added candidates could all be explained by ordinary physical behavior, so none is included as an implausible-event annotation. The small number of cases requiring this additional review also supports a low false-positive rate for the final annotations. Appendix B gives the pilot rounds, guidelines, adjudication protocol, cost, and training-annotation results.

Table 1: Human review of LLM-proposed events for the evaluation annotation batch. Percentages are computed over each model's proposals. Human-added events are reported once per split. As-is means accepted unchanged; retimed means kept with corrected temporal boundaries. On the real controls, a final audit found ordinary explanations for the two retimed and three human-added candidates, so none was retained as an implausible-event annotation.

LLM	Total	As-is	Rejected	Retimed	Human-added
Synthetic (175 videos)					
GPT-5.4	511	32.9% (168)	56.9% (291)	10.2% (52)	318
Gemini 3.1 Pro	447	10.7% (48)	82.8% (370)	6.5% (29)	–
Real controls (25 videos)					
GPT-5.4	27	0.0% (0)	96.3% (26)	3.7% (1)	3
Gemini 3.1 Pro	36	0.0% (0)	97.2% (35)	2.8% (1)	–

The same data support two evaluation sets. Event verification uses 1,276 candidate descriptions from the 175 synthetic evaluation videos.

Open-ended detection evaluates complete videos rather than supplied candidates. It uses 325 videos: the 170 synthetic videos with annotated implausible events, the 5 synthetic videos with no annotated implausible event, and 150 real videos from ImplausiBench (Motamed et al., 2025). We use the first eight seconds of each real video and assign it an empty event list. Thus, this evaluation has 170 videos with annotated implausible events and 155 videos on which the model should return no event. The 25 real controls used during annotation are part of these 150 videos. Their candidate annotations are used only to audit annotation quality, not as event ground truth. Appendix E gives the full composition.

2.3 METRICS

Event verification: The positive class, IMP, contains candidate descriptions that correctly identify a visible implausible event. The negative class, ¬IMP, contains candidates that are absent, inaccurate, or physically ordinary. We report precision, recall, and F1 for each class. We also report accuracy, balanced accuracy, and Matthews correlation coefficient (MCC). MCC summarizes how well the verifier separates the two classes; a value near zero indicates weak discrimination.

Open-ended event detection: The detector uses free-form descriptions, so exact string matching is not appropriate. GPT-5.5 first decides which predicted and reference descriptions refer to the same implausible event. We then match predictions to references one-to-one. We compute semantic and temporal precision, recall, and F1 only on the 170 videos with annotated implausible events. Semantic metrics (sP/sR/sF1) measure whether the model found the correct implausible events, while temporal metrics (tP/tR/tF1) measure whether it assigned the correct intervals. On the separate set of 155 no-event videos, specificity is the fraction for which the model returns an empty list. We also report the average number of false events on these videos. GPT-5.5 only compares descriptions; it does not view the video or decide whether an event is physically valid. Appendix H gives the matching rules, metric definitions, one-to-one assignment, and judge prompt.

3 METHOD

3.1 OVERVIEW

We train one model to verify supplied event descriptions and a separate model to detect events directly from a video. We initially considered using the verifier to score detector outputs. However, Section 5.1 shows that the verifier is not reliable enough. Detector training therefore computes its reward by comparing predictions directly with the human event annotations.

3.2 EVENT VERIFIER

During training, a teacher model receives the video, the candidate event, and the human verdict. It writes a rationale that supports the verdict. A smaller student model receives the video and candidate but does not receive the human verdict. It learns to produce both the rationale and the verdict. At inference time, the student judges new candidates without any label. This is privileged distillation because the teacher receives information that the student never sees. Formally, for input $x = (V_{s:e}, d)$ and human verdict y, the teacher generates rationale r, and the student learns to predict (r, y) from x. We call this the chain-of-thought (CoT) condition. In the non-CoT condition, the model predicts only the verdict.

3.3 OPEN-ENDED EVENT DETECTOR

We train an autoregressive event-set predictor in three stages: teacher distillation, reference-based RL, and self-distillation with empty targets.

Stage I: Privileged teacher distillation. During training only, the teacher receives each video together with its human event list. It writes a scene summary, a timestamped account of ordinary events, and their causal relations. We combine this visible reasoning with the original human event

list, so the teacher cannot add, remove, or change the target events. When a video contains no annotated implausible event, we keep its target as `[]`, but the teacher still generates the scene summary, ordinary events, and causal relations. The student sees only the video and learns to produce the reasoning and event list. This stage teaches the output format, temporal grounding, and when to return no event. Appendix D gives the complete reasoning structure, teacher hint, and student prompt.

Stage II: Reference-based RL. Starting from the Stage I policy, we optimize the predicted event set directly. For prediction $\hat{Y}$ and human reference Y, the reward is

$$R_{\text{imp}}(\hat{Y}, Y) = \tfrac{1}{2}\big(\text{sF1}(\hat{Y}, Y) + \text{tF1}(\hat{Y}, Y)\big), \tag{2}$$

where sF1 and tF1 are the one-to-one semantic-matching and temporal-localization metrics defined in Section 2.3. Averaging them rewards both event recovery and temporal localization; Appendix H gives the complete judge, assignment, and metric definitions. We apply RL only to videos that contain annotated implausible events, so every rollout has at least one human reference event; an empty prediction receives zero reward.

In preliminary experiments, we also mixed real videos with empty references into RL. This did not reliably teach the policy to both detect implausible events and return `[]`. Depending on the proportion of the two sample types, the policy often degenerated toward reporting events on most videos or returning `[]` on most videos. This behavior resembles learning the event prior in the training mixture rather than learning to distinguish the two cases from visual evidence. We therefore use RL only to improve event recovery and address empty-list behavior separately in Stage III.

Stage III: Self-distillation with empty targets. Stage II never shows the model a video with no event. As a result, the RL model often predicts an implausible event even when the video is plausible. We first run the RL model on the training videos with annotated implausible events and retain its predicted lists $\hat{Y}_V^{\text{RL}}$. We then combine these predictions with plausible videos whose targets are empty:

$$\mathcal{D}_{\text{final}} = \{(V, \hat{Y}_V^{\text{RL}}) : V \in \mathcal{D}_{\text{impl}}\} \cup \{(V, \emptyset) : V \in \mathcal{D}_{\text{plaus}}\}. \tag{3}$$

Here $\mathcal{D}_{\text{impl}}$ contains videos with annotated implausible events and $\mathcal{D}_{\text{plaus}}$ contains videos with no annotated implausible event. We implement self-distillation from the RL checkpoint using the same token-level forward-KL objective as in Stage I. The RL model supplies its own targets for videos with annotated implausible events, while the empty examples teach it to return no event for a plausible video. This stage retains much of the event recovery learned through RL.

4 Experimental Setup

Shared setup: Across both tasks, we evaluate Claude Opus 4.8, GPT-5.5, GPT-5.4, GPT-5.4 mini, Gemini 3.1 Pro, and Gemma 4 models without task-specific training. Our trainable students are Gemma 4 E2B-it and E4B-it, with Claude Opus 4.8 or GPT-5.4 providing privileged training targets. All models receive timestamped JPEG frames capped at 40 frames, and evaluation uses greedy decoding. The task-specific frame selection, outputs, training, and evaluation protocols follow below.

4.1 Event Verification

Input and output: For each candidate description and interval, the verifier receives timestamped frames from the claimed window and two context frames on each side. The model outputs `correct` for a genuine, visible implausibility and `wrong` when the claim is absent, misdescribed, or physically plausible.

Training: The human training pool contains 1,432 claims. A video-level split provides 1,293 training claims and 139 validation claims. We select one checkpoint per run using validation MCC and average three runs per configuration. Appendix J gives the prompts, data composition, optimization, and checkpoint-selection details.

Evaluation: The main evaluation contains 1,276 candidate claims from the 175 synthetic evaluation videos. We report CoT results for the off-the-shelf and fine-tuned models in the main paper. This gives a matched comparison between the privileged-CoT students and the CoT teachers that produce their training rationales. Appendix J.5 reports the non-CoT and real-video verifier results.

4.2 Open-Ended Event Detection

Input and output: For detection, frames are sampled at $\sim$5 fps from the leading video window. In the CoT condition, the model first produces a visible JSON object with a scene summary, ordinary events, and causal relations, followed by the scored list of implausible events. This is a structured intermediate output, not hidden chain-of-thought. In the non-CoT condition, the model returns only the event list. We score only the final list in both conditions.

Training: We train the three stages described in Section 3 using forward-KL distillation, GRPO, and forward-KL self-distillation, respectively. Each stage uses held-out videos for checkpoint selection. Stages I and III use a 10% validation split, while Stage II uses a 320/24 train/validation split. Stages I and II select by validation tF1. Stage III selects by the harmonic mean of validation tF1 and specificity, measuring event localization on positive videos and empty-list behavior on empty-target videos. *RL-teacher distillation* isolates the value of RL initialization: it starts from the untuned E4B model and trains on exactly the same Stage III targets as the main E4B–Opus branch, including the Opus-branch RL predictions and forced empty targets. *Verifier-reward RL* tests whether a verifier can replace human references in Stage II. Starting from the Opus-distilled E4B checkpoint, it runs GRPO without reading the human event lists; GPT-5.5 labels each predicted event as `correct` or `wrong`. For a nonempty valid output with N predictions, C accepted predictions, total predicted duration D, and accepted duration D_C, its reward is $\frac{1}{2}(C^2/N + D_C^2/D)$. We also self-distill this verifier-reward policy using RL predictions for videos with annotated implausible events and forced empty targets for plausible videos. Appendices C and D give the data mixtures, optimization settings, reward handling, prompts, and ablations.

Evaluation: We evaluate on the 325-video set defined in Section 2. We compare the off-the-shelf models with the students after teacher distillation, RL, and self-distillation. All detector configurations in Table 4 are averaged over two runs. We use non-CoT as the primary off-the-shelf baseline because it tests direct event detection without requiring an auxiliary reasoning output. We discuss these results in the main paper and report the complete CoT/non-CoT comparison in Appendix Table 10. We also evaluate a filter→RL cascade. The matching teacher first classifies the whole video as plausible or implausible. A plausible decision returns `[]`; otherwise, the RL student produces the event list. Appendix F gives the complete cascade and scoring protocol.

5 Results

In result tables, green marks rank 1 and yellow marks rank 2 within a directly comparable group.

5.1 Event Verification

Table 2: Off-the-shelf verifiers with CoT on the synthetic verification slice. Imp denotes a correct candidate description and ¬Imp an incorrect one. Bal is balanced accuracy.

				Imp			¬Imp		
Model	Acc	MCC	Bal	P	R	F1	P	R	F1
Claude Opus 4.8	.517	.052	.524	.499	.715	.588	.557	.333	.417
GPT-5.5	**.539**	**.070**	**.531**	.540	.298	.384	.539	.764	.632
GPT-5.4	.505	.005	.503	.485	.423	.452	.520	.582	.550
GPT-5.4 mini	.512	.001	.500	.483	.182	.264	.518	.818	**.635**
Gemini 3.1 Pro	.484	−.014	.494	.478	.771	**.590**	.505	.218	.304
Gemma 4 31B	.500	.010	.505	.486	.637	.551	.525	.372	.435

RQ1: Can current models verify candidate events reliably enough to supervise a detector? Table 2 shows that every evaluated off-the-shelf model has low MCC under CoT prompting. GPT-5.5 has the best MCC at only 0.070. The two teachers used for distillation reach 0.052 for Opus 4.8 and 0.005 for GPT-5.4. These values are close to zero, so the models only weakly separate correct event descriptions from incorrect ones. Some models obtain moderate accuracy or F1 for one class by predicting that class too often. Those scores therefore overstate their overall verification ability.

CoT does not consistently improve verification. Main Table 2 and Appendix Table 12 compare the same off-the-shelf models with and without CoT. Both prompts use the same frames, claim, and

Table 3: Gemma 4 base and CoT-SFT verifiers. A dash marks the model before fine-tuning. For SFT, the teacher receives the ground-truth `correct`/`wrong` verdict before generating each CoT rationale, but the student does not. SFT results are averages over three runs after per-run checkpoint selection by MCC. IMP denotes a correct candidate description and ¬IMP an incorrect one.

				IMP			¬IMP		
Student	Teacher	Acc	MCC	P	R	F1	P	R	F1
E4B	—	.484	−.058	.432	.226	.297	.501	.723	**.592**
E4B	Opus 4.8	**.541**	**.087**	.521	.596	.554	.567	.490	.524
E4B	GPT-5.4	.518	.044	.500	.627	**.556**	.545	.417	.472
E2B	—	.531	.056	.566	.119	.196	.527	.915	**.669**
E2B	Opus 4.8	**.545**	**.086**	.532	.465	.496	.554	.619	.585
E2B	GPT-5.4	.536	.078	.515	.618	**.562**	.564	.459	.505

verdict format; the CoT prompt requires a rationale, whereas the non-CoT prompt requires only the verdict. Switching from CoT to non-CoT makes GPT-5.4, Gemini 3.1 Pro, Gemma 4 31B, and GPT-5.4 mini predict `correct` more often. For GPT-5.4 mini, this raises recall on genuine implausible events from 0.182 to 0.498, but lowers recall on incorrect candidates from 0.818 to 0.505. The same switch raises GPT-5.5's MCC from 0.070 to 0.084, but lowers Claude Opus 4.8's MCC from 0.052 to -0.047. Thus, removing the rationale changes the models' class bias, but does not improve verification consistently. Appendix J.5.1 shows an evaluation example in which a CoT rationale incorrectly explains away a genuine implausible event.

Privileged distillation changes class-wise F1, but MCC remains low. Table 3 shows that the class-wise F1 changes depend on the teacher. Relative to Opus 4.8, the Opus-distilled students become better at rejecting incorrect candidates but worse at recognizing genuine implausible events. Their F1 on incorrect candidates rises from 0.417 to 0.524 and 0.585, while their F1 on genuine implausible events falls from 0.588 to 0.554 and 0.496. GPT-5.4-distilled students show the opposite pattern. Their F1 on genuine implausible events rises from 0.452 to 0.556 and 0.562, while their F1 on incorrect candidates falls from 0.550 to 0.472 and 0.505. MCC improves for every teacher–student pair. The Opus-distilled students reach 0.087 and 0.086, compared with 0.052 for their teacher. The GPT-5.4-distilled students reach 0.044 and 0.078, compared with 0.005 for their teacher. These gains are small in absolute terms. Privileged distillation changes the class-wise behavior, but it does not produce a strong verifier.

Implication for reward modeling. To train a detector, a reward model must reliably prefer a genuine event over an incorrect one. The low MCC values show that neither the teachers nor the students can do this consistently. A verifier-based reward could therefore reward an incorrect event or penalize a genuine one. We therefore compute the main RL reward by comparing detector outputs directly with the human annotations, as described in Section 3.3. These results motivate two choices evaluated next: using human references rather than verifier scores for the main RL reward, and treating event recovery and no-event recognition as separate training objectives.

5.2 OPEN-ENDED EVENT DETECTION

RQ2: How does training affect event recovery and no-event recognition? Without CoT, Claude Opus 4.8 provides the strongest event recovery, with 0.468 sF1, 0.336 tF1, and 0.716 specificity. We use this as the primary reference point for the trained detectors. Appendix Table 10 reports the complete CoT/non-CoT comparison, and Appendix G.2 gives paired examples.

Table 4 evaluates event recovery on 170 event-containing videos and empty-list behavior on a disjoint set of 155 no-event videos. These metrics therefore measure separate capabilities rather than a tradeoff on the same examples.

Stage I is conservative. Table 4 shows that teacher distillation reaches only 0.190–0.293 sF1 and 0.088–0.189 tF1, while retaining 0.897–0.929 specificity. This pattern is consistent with the limited positive supervision—344 videos with annotated implausible events—and the heavily empty-target training mixture, in which 1,889 of 2,233 targets are empty. Stage I therefore learns to avoid false events on no-event videos more reliably than it recovers events on event-containing videos. Because

Table 4: Detector results on 325 videos. Semantic metrics (sP/sR/sF1) and temporal metrics (tP/tR/tF1) use 170 videos with annotated implausible events; Spec. uses 155 no-event videos. fp/empty and fp/impl denote unmatched predictions per no-event and event-containing video, respectively. RL teacher dist. trains the untuned E4B model on Opus-branch RL targets.

Student	Teacher	Stage/init.	sP	sR	sF1	tP	tR	tF1	Spec.	fp/empty	fp/impl
E4B	Opus 4.8	Teacher dist.	.418	.208	.272	.360	.127	.184	.897	.26	1.05
		RL	.791	.386	**.518**	.417	.402	**.404**	.006	1.53	.36
		Self-dist.	.787	.347	.481	.432	.363	.389	.652	.58	.33
		filter→RL	.811	.258	.391	.441	.268	.330	.929	.11	.21
		RL teacher dist.	.867	.139	.233	.349	.118	.170	**.997**	**.01**	**.08**
		Verifier RL	.223	.440	.296	.018	.452	.034	.026	6.13	5.58
		Verifier self-dist.	.214	.234	.224	.051	.183	.080	.865	1.88	3.13
	GPT-5.4	Teacher dist.	.464	.217	.293	.387	.126	.189	**.923**	**.24**	.93
		RL	.869	.444	**.585**	.469	.463	**.459**	.000	1.42	.24
		Self-dist.	.825	.379	.518	.458	.376	.405	.745	.38	.28
		filter→RL	.871	.394	.541	.467	.403	.426	.742	.44	**.21**
E2B	Opus 4.8	Teacher dist.	.294	.143	.190	.231	.055	.088	**.929**	.25	1.23
		RL	.844	.438	**.576**	.427	.475	**.449**	.000	1.60	.29
		Self-dist.	.840	.362	.503	.419	.353	.379	.800	.37	.23
		filter→RL	.868	.296	.441	.439	.314	.366	**.929**	**.13**	**.16**

we do not perform a controlled data-scaling study, we do not attribute the low recovery solely to dataset size.

Reference-based RL recovers events but loses empty-list behavior. Across the three branches, RL raises sF1 from 0.190–0.293 to 0.518–0.585 and tF1 from 0.088–0.189 to 0.404–0.459. These are gains of 0.246–0.386 sF1 and 0.220–0.361 tF1. However, specificity drops by 0.891–0.929, reaching only 0.000–0.006: the RL models report an event on almost every no-event video. This is a failure to recognize the no-event regime, rather than a direct precision–recall tradeoff within the event-containing subset. For example, E4B–GPT-5.4 RL attains 0.869 semantic precision and only 0.24 unmatched predictions per event-containing video, yet produces 1.42 false events per no-event video.

Self-distillation restores specificity while retaining most RL gains. Relative to RL, self-distillation reduces sF1 by only 0.037–0.073 and tF1 by 0.015–0.070, while increasing specificity by 0.646–0.800. Because these changes are measured on disjoint subsets, they indicate that self-distillation adds the ability to return an empty list on no-event videos while preserving most of the event-recovery ability learned by RL. Among the self-distilled models, E4B–GPT-5.4 reaches the highest event scores, with 0.518 sF1, 0.405 tF1, and 0.745 specificity. Relative to the off-the-shelf Opus reference, it improves sF1 by 5.0 points and tF1 by 6.9 points while slightly improving specificity. Appendix I follows the same evaluation videos across the three stages and audits temporal localization.

RQ3: Which components are necessary, and can simpler alternatives replace them?

RL initialization matters. RL-teacher distillation trains the untuned E4B model on the same targets as the main E4B–Opus self-distillation branch. It reaches only 0.233 sF1 and 0.170 tF1, compared with 0.481 and 0.389 when training continues from the RL checkpoint. Its 0.997 specificity indicates that it primarily learns the empty targets. Its low 0.139 semantic recall on event-containing videos confirms that the RL-generated targets alone do not transfer the policy's event-recovery behavior (Appendix C.5).

Video-level filtering is a competitive but expensive alternative. For filter→RL, a frozen teacher classifies the complete video before event detection, returning `[]` for videos classified as plausible and retaining the complete RL event list otherwise. Filter→RL reaches 0.742–0.929 specificity. The GPT-5.4 cascade slightly exceeds its self-distilled counterpart in event F1 with similar specificity, but requires an additional teacher call per video. Appendix F gives the complete protocol.

Verifier rewards do not replace human references. Verifier-reward RL reaches only 0.296 sF1, 0.034 tF1, and 0.026 specificity. It therefore performs poorly in both regimes: on event-containing videos it also produces 5.58 unmatched predictions per video, while on no-event videos it pro-

duces 6.13 false events per video. Self-distillation raises specificity to 0.865, but sF1 remains 0.224 and tF1 0.080. Appendix C.6 shows that the policy can repeat broad claims that receive verifier reward despite poor localization. Together, these results show that, among the tested objectives, human-reference RL is substantially more effective for event recovery. They also show that RL initialization carries information not captured by its outputs alone, and empty-target self-distillation restores no-event recognition while retaining most of the reference-based RL model's event recovery, without requiring a teacher call at inference time. Video-level filtering can be competitive, but at substantially greater inference cost.

6 Related Work

Physical reasoning: Controlled benchmarks isolate object permanence, collisions, materials, and physical prediction in synthetic environments (Riochet et al., 2018; Bordes et al., 2025; Yi et al., 2020; Bear et al., 2021; Zheng et al., 2024), while recent VLM benchmarks extend such reasoning to richer scenes and quantitative questions (Chow et al., 2025; Li et al., 2026). These settings typically specify objects, outcomes, or concepts; ours requires discovering and localizing an unknown number of brief inconsistencies.

Generated-video evaluation: General benchmarks score complete videos for visual quality, temporal consistency, prompt alignment, and human preference (Liu et al., 2024; Huang et al., 2024; 2026; He et al., 2024; 2025; Li et al., 2025; Zheng et al., 2025). Physics-focused benchmarks evaluate adherence to physical laws, trajectories, materials, scientific phenomena, and deliberately challenging prompts (Bansal et al., 2025; 2026; Meng et al., 2025; Guo et al., 2025; Motamed et al., 2026; Kang et al., 2025; Tragoudaras et al., 2026; Hu et al., 2026; Gu et al., 2026). Questions, grounding, and dependency graphs provide finer diagnosis (Lin et al., 2026; Pothiraj et al., 2026), but typically specify a question, principle, or prompt-level expectation; VIOLATE instead evaluates a variable-size event set without specifying the relevant law or event.

Implausibility and localization: Impossible Videos studies reasoning over impossible content (Bai et al., 2025); PhyDetEx evaluates detection and explanation, TRAVL studies video-language models as judges of physical implausibility, and Physion-Eval incorporates human reasoning and temporal localization (Wang et al., 2025; Motamed et al., 2025; Zhang et al., 2026). Complementary work studies gameplay anomalies and counterfactual implausibilities (Cao et al., 2026; Hao et al., 2025), while dense captioning and anomaly detection localize unusual events (Krishna et al., 2017; Sultani et al., 2018; Gao et al., 2026), but these events need not violate physics. VIOLATE instead evaluates open-ended, temporally localized discovery without specifying the event or physical rule in advance.

7 Conclusion

We introduced VIOLATE for finding, describing, and localizing physically implausible events in video while returning no event for a plausible video. Current models remain weak at both verification and open-ended detection. Teacher distillation gives only small verifier gains. For detection, RL improves event recovery but increases false detections, while self-distillation restores much of the ability to identify plausible videos. We hope VIOLATE supports video models that reason more reliably about physical consistency.

AI Use Statement

We used generative AI tools for writing assistance, including prose editing and LaTeX formatting, and for assistance with implementation, evaluation, and experiment-orchestration scripts. The research questions, method design, experimental protocol, interpretation, and conclusions are the authors' own. The authors reviewed and tested all AI-assisted code. The experimental pipeline produced all reported numbers, and the authors checked them against the evaluation outputs and logs. Generative models also form explicit parts of the method and evaluation. GPT-5.4 and Gemini 3.1 Pro generate candidate events for human review. Claude Opus 4.8 and GPT-5.4 generate privileged training targets. GPT-5.5 provides semantic alignments for evaluation and reference-based RL. We additionally

evaluate several off-the-shelf models as baselines and train the Gemma 4 student models reported in the paper. We take responsibility for the final text, claims, code, data, and experimental artifacts, including material produced with generative-AI assistance.

Ethics Statement

This work studies the detection and evaluation of physical implausibilities in AI-generated videos. Human annotators review generated videos and real control videos. The study evaluates the videos and model outputs rather than the annotators. We do not report personally identifying annotator information. We document the multi-stage guideline development, quality-control procedure, control videos, annotation effort, and cost in Appendix B. The benchmark may improve the auditing of generated video, but automated verifiers and detectors are not infallible. The low verifier MCC and observed errors on both event-containing and no-event videos show that systems can miss implausible events or incorrectly flag plausible content. Consequently, these tools should support rather than replace human review in safety-critical or high-stakes settings. Any release will include documentation of data provenance, annotation limitations, and applicable usage constraints. We take responsibility for the study design, released artifacts, and reported conclusions.

Reproducibility Statement

Section 2 defines the tasks, data splits, and metrics. Section 3 specifies the verifier and three-stage detector. Section 4 describes the experimental protocols. Appendix B documents annotation and quality control. Appendix C gives data mixtures and optimization settings. Appendix D provides the detector prompts. Appendix E specifies decoding and evaluation buckets. Appendix H specifies semantic matching, assignment, metrics, and the judge prompt. Appendix J provides verifier data, training details, prompts, and additional results. The source code, full prompts, pipeline configurations, evaluation scripts, and benchmark will be released upon publication.

References

Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2458–2483. PMLR, 2025.

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.

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.

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

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.

Meng Cao, Haoran Tang, Haoze Zhao, Mingfei Han, Ruyang Liu, Qiang Sun, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Order from chaos: Physical world understanding from glitchy gameplay videos. *arXiv preprint arXiv:2601.16471*, 2026.

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.

Shibo Gao, Peipei Yang, Yangyang Liu, Yi Chen, Han Zhu, Xu-Yao Zhang, and Linlin Huang. VAGU & GtS: LLM-based benchmark and framework for joint video anomaly grounding and understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4167–4175, 2026. doi: 10.1609/aaai.v40i6.42412.

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.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2VPhysBench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025.

Yutong Hao, Chen Chen, Ajmal Saeed Mian, Chang Xu, and Daochang Liu. Enhancing physical plausibility in video generation by reasoning the implausibility. *arXiv preprint arXiv:2509.24702*, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.127.

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. *arXiv preprint arXiv:2509.22799*, 2025.

Lanxiang Hu, Abhilash Shankarampeta, Yixin Huang, Zilin Dai, Haoyang Yu, Yujie Zhao, Haoqiang Kang, Daniel Zhao, Tajana Rosing, and Hao Zhang. Benchmarking scientific understanding and reasoning for video generation using VideoScience-Bench. In *European Conference on Computer Vision*, 2026.

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.

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

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.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017. doi: 10.1109/ICCV.2017.83.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1834.

Puyin Li, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-Fei, and Ehsan Adeli. QuantiPhy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33174–33184, 2026.

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. *arXiv preprint arXiv:2605.10806*, 2026.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024. doi: 10.1109/CVPR52733.2024.02090.

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.

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.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026. doi: 10.1109/WACV61042.2026.00099.

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.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018.

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

Zeqing Wang, Keze Wang, and Lei Zhang. Phydetex: Detecting and explaining the physical plausibility of t2v models. *arXiv preprint arXiv:2512.01843*, 2025.

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

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. *arXiv preprint arXiv:2603.19607*, 2026.

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

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B. Tenenbaum, and Chuang Gan. ContPhy: Continuum physical concept learning and reasoning from videos. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61526–61558. PMLR, 2024.

A Synthetic Video Sources and Generation

All released synthetic videos were generated with Veo 3.1 Fast (`veo-3.1-fast-generate-preview`) through Vertex AI. We generated eight-second videos at 1280×720, 24 fps, with a 16:9 aspect ratio and disabled audio. We selected Veo through informal inspection during dataset development because its outputs appeared more natural and showed fewer visible rendering artifacts, including unintended camera-angle changes. This choice was informed by the three annotation-training rounds, which included videos from ImplausiBench, Sora, and Veo, rather than by a controlled generator comparison.

VideoPhy-2 contains 3,940 captions for 197 actions (Bansal et al., 2026). Its official split contains 3,350 training captions and 590 test captions. The hard subset contains 60 actions for which CogVideoX-5B obtains zero joint semantic and physical adherence. It contains 1,020 training captions and 180 test captions. Both benchmark splits use these same 60 actions, and every action has at least one successful video in each split. No caption is shared between splits. For both splits, we send Veo the longer caption upsampled by Mistral-NeMo rather than the short original caption. Table 5 summarizes the generation runs.

Table 5: Synthetic-video generation. After deduplication, the hard training pool contains 1,013 prompts; we attempted generation for 393 prompts from the shuffled pool.

Split	Pool	Attempted	Failed	Retained	Prompt version
Train	1,013	393	12	381	Upsampled caption
Evaluation	180	180	5	175	Upsampled caption

B Human Annotation Guidelines and Results

Protocol development and cost accounting: The reported totals of approximately **1,500 hours of human work** and **$67K** cover pilot rounds, annotation, review, adjudication, and quality control. They are not limited to the 381 training and 175 synthetic evaluation videos in the released benchmark. Before annotating the final dataset, we conducted three annotator-training and protocol-development rounds using videos from ImplausiBench, Sora, and Veo.

In **round 1**, annotators independently viewed **20 videos** and produced implausibility annotations directly from the videos, without candidate events. Agreement was poor. Annotators often selected different events, used different levels of descriptive detail, and assigned different temporal boundaries. In **round 2**, conducted on **24 videos**, we introduced candidate implausible events as a shared annotation scaffold. Annotators reviewed these candidates and could accept, correct, or reject them, while still adding events missed by the candidates. This produced a more consistent decomposition of each video into events. In **round 3**, conducted on **15 videos**, we finalized the guidelines and aligned annotators on event identity, description specificity, temporal localization, duplicate handling, and the treatment of missed candidate events. The final dataset was annotated using the procedure established in this third round.

The reported cost also includes annotation of **25 real control videos** interspersed with the 175 synthetic evaluation videos, as well as the quality-control and adjudication passes described below. Thus, the time and cost totals should be interpreted as the resources required for the entire annotation process, including protocol development, annotator training, controls, and final review, rather than as the direct cost of annotating the 556 released synthetic videos alone.

LLM-assisted candidate generation: Motivated by the poor agreement observed when annotators generated events from scratch in round 1, the final protocol uses LLM-generated candidate

events as a common annotation scaffold. For the evaluation data, GPT-5.4 and Gemini 3.1 Pro independently propose candidate implausible events and their corresponding frame ranges. For the training data, only GPT-5.4 is used for candidate generation. These proposals are *not* treated as ground-truth annotations.

The LLM proposals may be incomplete or incorrect. A model may flag an event that is not actually implausible, provide an inaccurate frame range, identify the same underlying event as another proposal, or miss an implausibility altogether. Annotators therefore inspect the full video rather than restricting their review to the proposed candidates.

Human annotation procedure: Following the procedure aligned in round 3, annotators first inspect the full video and then review each LLM-proposed event, assigning one of three outcomes. An event is marked *as-is* when the proposed implausibility and its frame range are correct. If the implausibility is present but its temporal localization is incorrect, the annotator supplies an *updated frame range*. Otherwise, the proposal is *rejected*.

Annotators also search the full video for implausible events that are absent from the LLM proposals. When such an event is found, they add a new event with a human-written description and frame range. The LLM proposals therefore provide an initialization for annotation but do not constrain the final set of events.

Event descriptions and temporal localization: Annotators are instructed to describe each observable implausible event as specifically as possible. An annotation should state what becomes implausible and how the observed behavior violates physical or logical consistency. We favor localized, atomic descriptions over broad judgments about the overall realism of the video. The associated frame range should cover the temporal extent of the implausibility.

Duplicate resolution: Because multiple candidate proposals may refer to the same underlying event, annotators remove redundant proposals and retain a single annotation that best represents the event. This is particularly relevant for the evaluation data, where GPT-5.4 and Gemini 3.1 Pro generate candidates independently. Thus, each underlying implausible event appears only once in the final annotation.

Evaluation annotation and quality control: The synthetic evaluation set contains **175 videos**. Each video is independently reviewed by two human annotators. A third human reviewer resolves their disagreements and produces the final event annotations and temporal boundaries.

The complete evaluation annotation batch contains **200 videos**. It includes **175 synthetic evaluation videos** and **25 real control videos**. We intersperse the real videos with the synthetic videos during annotation and follow the same review procedure. They are used as quality-control examples and are not included in the final synthetic evaluation set.

Evaluation annotation results: Table 1 summarizes the review outcomes for candidate events proposed by each LLM. On the 175 synthetic videos, GPT-5.4 proposed 511 events, of which 32.9% were accepted as-is, 10.2% required updated frame ranges, and 56.9% were rejected. Gemini 3.1 Pro proposed 447 events, of which 10.7% were accepted as-is, 6.5% required updated frame ranges, and 82.8% were rejected. Annotators additionally identified **318 events** that were absent from the LLM candidate sets.

On the 25 real control videos, GPT-5.4 proposed 27 events and Gemini 3.1 Pro proposed 36. None of the 63 proposals were accepted as-is. During the initial review, 61 proposals were rejected, two were retained with corrected frame ranges, and annotators added three candidates. Because real controls should not be assumed flawless, we conducted a targeted final audit of these five cases. Each could be justified by ordinary physical behavior or the available visual evidence, so none was included as an implausible-event annotation. The small number of candidates requiring this additional review—five across 25 control videos, with no confirmed implausibility after adjudication—also supports a low false-positive rate for the final annotations. These control-video outputs are used only to audit the annotation process, and all 150 ImplausiBench real videos receive the empty target in evaluation.

Table 6: Human review of GPT-5.4-proposed events for the 381-video training split. Percentages are computed over proposals. As-is means accepted unchanged; retimed means kept with corrected temporal boundaries.

LLM	Total	As-is	Rejected	Retimed	Human-added
GPT-5.4	1087	46.8% (509)	45.5% (495)	7.6% (83)	345

Table 7: Stage I teacher-distillation mixture. Human event lists are kept identical to ground truth; plausible videos have an empty event target.

Subset	Videos	Events	Event target
Human implausible	344	937	unchanged GT list
Human plausible	37	0	[]
TRAVL real (first 8 s)	1,852	0	[]
Total	2,233	937	—

Training annotation results: For the 381-video training set, GPT-5.4 provides the candidate events. Each video receives an initial annotation followed by a separate quality-control pass from the same reviewer who adjudicates the evaluation annotations. GPT-5.4 proposes 1,087 events; annotators accept 509 (46.8%) unchanged, reject 495 (45.5%), and retain 83 (7.6%) with corrected boundaries. They also add 345 events absent from the proposals.

C Training and Implementation Details

C.1 Model Input and Output

We instantiate the student with Gemma 4 E2B-it or E4B-it. Each model receives the whole clip as JPEG frames sampled at $\sim$5 fps, capped at 40 frames using a leading window. Every frame is labeled with its timestamp. Train and evaluation use the same unhinted two-section detector prompt. Its first section contains an analysis JSON with a summary, observable events, and causal relations. Its second section is the scored JSON event array, with `[]` representing a plausible clip.

C.2 Stage I: Teacher Distillation

The human training batch contains 381 clips. It has 344 implausible clips and 37 plausible clips. The 344 event-containing training videos contain 937 annotated events. We add one leading 8-second window from each of 1,852 physically plausible TRAVL videos, producing 2,233 examples. The training examples are disjoint from the evaluation videos. Stage II uses the 923 annotations that have positive duration.

A frozen Claude Opus 4.8 or GPT-5.4 teacher produces the analysis target. Human event lists are supplied as hints and then copied unchanged into the event section. TRAVL and the 37 human-plausible clips receive plausible hints and empty event targets. The student never sees these hints. A video-level split with seed 42 and validation ratio 0.1 gives 2,010 training and 223 validation videos, with 1,889 empty targets. We use FSDP on 8 GPUs, batch size 16, 3 epochs, a 2×10^{-5} learning rate, cosine decay, and warmup ratio 0.03. The best checkpoint is selected by validation temporal F1.

C.3 Stage II: Reinforcement Learning

RL starts from the merged Stage I checkpoint and uses the 344 implausible clips. A video-level split gives 320 training clips and 24 validation clips. These clips contain 923 positive-duration events. We sample eight rollouts per prompt at $T{=}1$. GPT-5.5 constructs the same graded semantic edges and one-to-one assignment used for evaluation. With false-positive weight zero, the reward is

$$r = \frac{1}{2}\left(\mathrm{sF1} + \mathrm{tF1}\right). \tag{4}$$

Table 8: Stage III self-distillation mixture. Positive targets are greedy RL-policy event lists; plausible targets are forced to `[]`.

Subset	Videos	Events	Target source
Human implausible	344	RL-dependent	student, $T{=}0$
Human plausible	37	0	forced `[]`
TRAVL real (first 8 s)	1,852	0	forced `[]`
Total	2,233	RL-dependent	—

A well-formed empty array scores zero. An unparseable event section or an inverted interval receives -1. Zero-duration point events remain valid. We use GRPO without a KL penalty, cosine learning rate $5{\times}10^{-6}$, warmup ratio 0.15, train batch size 32, PPO mini-batch size 8, and 10 epochs. Selection uses validation temporal F1.

C.4 Stage III: Self-Distillation

The Stage II checkpoint greedily predicts ($T{=}0$) event lists for the same 344 implausible clips. These policy-generated lists replace ground truth as the positive supervised targets. We restore the 1,852 TRAVL windows and 37 human-plausible clips, forcing their event targets to `[]` while their analysis targets describe plausible content. The student input remains unhinted.

The Stage III mixture contains 2,233 videos, including 1,889 forced-empty examples. Training starts from the RL checkpoint. An ablation instead performs RL-teacher distillation: it fine-tunes the untuned base model on the same RL-generated targets. We reuse the Stage I video-level 90/10 split and optimization settings, with 3 epochs and batch size 16. We select the checkpoint by the harmonic mean

$$H_{\text{sel}} = \frac{2\,\text{tF1}\,\text{Specificity}}{\text{tF1} + \text{Specificity}}, \tag{5}$$

computed on the validation split. Here tF1 is computed on validation videos with event targets and specificity on validation videos with empty targets.

C.5 Stage III Initialization Ablation

The E4B–Opus self-distillation rows in Table 4 test whether the Stage III targets are sufficient without the RL initialization. Both conditions use the same target set. It contains greedy Section 2 predictions from the selected Opus 4.8 reference-based RL checkpoint on the 344 videos with annotated implausible events, forced `[]` targets on 1,852 TRAVL lead-eight-second windows, and forced `[]` targets on the 37 human empty videos. Training otherwise follows the Stage III configuration above.

The RL-initialized condition is the main E4B–Opus self-distillation result and averages its two runs. The RL-teacher-distillation condition starts from the untuned `gemma-4-E4B-it` model, uses the same RL-generated training targets, and is also averaged over two runs.

C.6 Verifier-Reward RL Baseline

This baseline tests whether an event verifier can replace human references in the Stage II reward. It uses Gemma 4 E4B initialized from the Opus 4.8 teacher-distilled policy. The GRPO configuration matches Stage II: 344 videos with annotated implausible events, a 320/24 train/validation split, eight rollouts at $T{=}1$, no KL penalty, cosine learning rate $5{\times}10^{-6}$, warmup ratio 0.15, batch size 32, PPO mini-batch size 8, and ten epochs. The two-section detector prompt is unhinted, and the Section 1 length penalty is disabled.

The reward never reads the human event list. A GPT-5.5 non-CoT verifier labels each predicted Section 2 event as `correct` or `wrong`. Failed or unparseable verifier calls count as `wrong`. Let N be the number of predicted events, C the number labeled `correct`, D the total predicted duration, and D_C the total duration of events labeled `correct`. For a nonempty valid prediction, the reward

components are

$$q_{\mathrm{s}} = \frac{C^2}{N}, \qquad q_{\mathrm{t}} = \frac{D_C^2}{D}, \qquad r = \tfrac{1}{2}(q_{\mathrm{s}} + q_{\mathrm{t}}). \tag{6}$$

These quantities are verifier-reward components, not evaluation F1 scores, and they are not bounded by one. A well-formed `[]` receives zero reward; malformed or invalid outputs receive -1. The false-positive weight is zero.

Self-distillation uses greedy predictions from the verifier-reward policy on the 344 videos with annotated implausible events and forced `[]` targets on 1,852 TRAVL windows and 37 human empty videos. Training uses three epochs, batch size 16, and learning rate 2×10^{-5}. Evaluation uses greedy decoding on the same 325 videos, one-to-one matching, and GPT-5.5 match judge as the other detector results.

The two verifier-reward rows in Table 4 are each averaged over two runs.

D Detector Prompts

D.1 Student Two-Section Detector Prompt (Training and Evaluation)

System prompt:

```
You are an expert at detecting physically implausible events in videos
(real or synthetic). Each frame you receive is labelled with its timestamp.

Answer in two sections, in this exact order.

SECTION 1 -- ANALYSIS. Output ONLY a JSON object, no code fences:

{
  summary: ...,
  events: [
    { time: 00:02-00:05, description: ... }
  ],
  causal_relations: [ ... ]
}

* summary: Briefly describe the overall video.
* events: List the main observable events in chronological order.
  Every event must include its approximate start and end time.
* causal_relations: Describe clear cause-and-effect relationships
  between events. Use an empty list if none are clear.

Do not include implausible_events in this object.

SECTION 2 -- IMPLAUSIBLE EVENTS. Output ONLY a JSON array, no code
fences, one object per physically implausible event in chronological
order:

[
  { time: 00:04-00:05, description: ... }
]

List ALL physically implausible events. For each, describe which object,
what happens, and why it breaks physics, and include the time when it
occurs. Return [] if the video is entirely plausible. Write nothing
after the JSON.

Write each time as MM:SS-MM:SS converted from the frame t= labels you
see (t=2.50s -> 00:02.50). Include fractional seconds when the frame
labels have them. Anchor every boundary to a provided frame timestamp.

Do NOT treat the following as implausible events:
  - Watermarks, logos, or on-screen overlays (including them shifting
    position or jumping around).
  - Any meta or non-physical aspects of the video (e.g., encoding
    artifacts that are not about scene physics, camera metadata, or
    post-production overlays).
Only report implausibilities in the actual physical scene: objects,
people, motion, causality, and spatial/temporal consistency of the
depicted world.
```

User prompt:

```
VIDEO ({N} frames in temporal order, each labelled with its timestamp
in seconds):
frame 1 (t={t1:.2f}s):
<image>
...

Analyze the video in two sections. SECTION 1 -- ANALYSIS: one JSON
object with summary, events (all observable events in chronological
order), and causal_relations (do not include implausible_events).
SECTION 2 -- IMPLAUSIBLE EVENTS: ONLY a JSON array of physically
implausible events (ALL of them), each with time (MM:SS-MM:SS
converted from the frame t= labels) and description. Return [] in
Section 2 if the video is entirely plausible. Write nothing after the
JSON.
```

During Stage II, the analysis prefix is unscored and only the event list is optimized. Its detection content matches the one-object teacher prompt below.

D.2 Teacher Prompt

System prompt: The teacher returns one JSON object with `summary`, `events`, `causal_relations`, and `implausible_events`. An empty prediction is represented as `"implausible_events": []`. The watermark, overlay, and metadata exclusions and the MM:SS timing rules are identical to the student system prompt.

Unhinted user prompt for evaluation:

```
VIDEO ({N} frames in temporal order, each labelled with its timestamp
in seconds):
frame 1 (t={t1:.2f}s):
<image>
...

Analyze the video and return one JSON object with summary, events
(all observable events in chronological order), causal_relations,
and implausible_events (ALL physically implausible events). Each
event object must have time (MM:SS-MM:SS converted from the frame
t= labels) and description. If the video is entirely plausible,
return implausible_events: []. Return only the JSON.
```

Ground-truth hint for the human-annotated training set: The system prompt is unchanged. The following text is appended after the frames. We keep Section 2 identical to ground truth, so the student target contains the annotated events rather than teacher-generated variations.

```
These physically implausible events are present (fixed, complete).
Put EXACTLY this list in implausible_events (same descriptions and
time spans; do not add, drop, merge, split, or retime). Still write
summary, events, and causal_relations from what you observe.
Do not mention that a list was provided.

- {MM:SS-MM:SS} {description}
...
```

Plausible hint for TRAVL and human-empty clips: Again, the system prompt is unchanged and the following text is appended after the frames:

```
This clip is likely physically plausible (often real footage) and may
have no physically implausible events. Unless you see a clear physics
violation in the depicted scene, return implausible_events: [].
Still fill in summary, events, and causal_relations from what
you observe.
```

We convert each retained teacher response into the two-section student target. The student never receives either hint.

D.3 Video-Level Filter Prompt

The frozen teacher in the filter→RL cascade receives the detector input frames followed by no event claim. The non-CoT video-level filter system prompt is:

```
You are an expert at judging whether a video contains physically
implausible events.
```

Table 9: Composition of the 325-video detector evaluation. Event and temporal metrics use only videos with annotated implausible events. Specificity uses only videos whose correct output is an empty list.

Bucket	Synthetic	Real	Total	GT events	Metrics
Annotated implausible events	170	0	170	615	sP/sR/sF1, tP/tR/tF1, fp/impl
No event target	5	150	155	0	specificity, fp/empty
All videos	175	150	325	615	—

```
CONTEXT: The video you will see was generated by a video-generation
model (a text-to-video diffusion / autoregressive system). Such models
routinely produce frames that violate everyday physics, causality,
anatomy, or object permanence -- but they also produce perfectly
plausible videos. Do not assume the footage is real, and do not assume
it is broken either; judge only what the frames actually show.

You will be given a sequence of frames sampled at a fixed rate from the
generated video, in temporal order. Each frame is preceded by a text
caption frame N (t=...s): giving its real-time timestamp in seconds.

Your job: decide whether the video as a whole is

* implausible -- at least one genuine physical implausibility is
  visible somewhere in the clip: objects defying gravity / momentum /
  rigidity, shapes morphing without cause, objects appearing or
  disappearing, limbs deforming unnaturally, motion that contradicts
  contact, liquid flowing uphill, teleportation, identity swaps,
  mid-air direction changes without external force, etc.; OR

* plausible -- nothing in the clip genuinely breaks physical,
  causal, temporal, or spatial consistency. Ordinary explanations
  count as plausible: camera motion, perspective, occlusion, motion
  blur, reflections, shadows, editing, overlays, compression
  artefacts, or an unusual but real physical effect.

Compare frames against each other rather than judging any single frame
in isolation: many violations are only visible as an inconsistency
across time.

Do NOT treat watermarks, on-screen overlays, encoding artefacts, or
camera metadata as implausibilities -- only the depicted physical scene
counts.

Do NOT explain or reason. Output ONLY your final verdict -- exactly one
word, either implausible or plausible -- between <answer> and
</answer> tags, with nothing before or after.

Example format:
<answer>implausible</answer>
```

E Evaluation Set and Decoding

E.1 Detector Inputs, Evaluation Buckets, and Selection

Section 4.2 gives the shared detector input, output, and decoding protocol. Table 9 gives the exact evaluation-set composition. The 25 real controls retain annotations for quality auditing, but all real videos have the empty event target `[]`.

Checkpoint selection uses only the held-out validation partition from the corresponding training stage. Stages I and III use a 10% validation split; Stage II uses 24 of its 344 videos for validation. Stages I and II select the checkpoint with the highest validation tF1. Stage III selects the checkpoint with the highest harmonic mean of validation tF1 and specificity. The 325-video evaluation set is used only for final reporting.

F Video-Level Filter→RL Cascade

Filter→RL places a frozen teacher video-level classifier before the RL event detector.

Table 10: Off-the-shelf detectors with and without CoT on the 325-video evaluation set, using GPT-5.5 as the match judge. Colors rank models separately within each prompting condition. The s-columns score event descriptions, the t-columns score time intervals, and Spec. measures correct empty lists on no-event videos.

Model	Prompt	sP	sR	sF1	tP	tR	tF1	Spec.	fp/empty	fp/impl
Claude Opus 4.8	CoT	.564	.347	**.430**	.383	.251	**.303**	.890	.16	.98
	No CoT	.520	.426	**.468**	.351	.321	**.336**	.716	.57	1.44
GPT-5.5	CoT	.802	.216	.341	.703	.118	.203	.961	.06	.19
	No CoT	.802	.242	.372	.600	.190	.289	.910	.10	.22
GPT-5.4	CoT	.719	.231	.349	.615	.113	.191	.955	.09	.33
	No CoT	.680	.284	.400	.555	.236	.331	.884	.18	.49
GPT-5.4 mini	CoT	.729	.082	.148	.433	.055	.098	.981	.02	.11
	No CoT	.750	.266	.393	.387	.278	.323	.465	.61	.32
Gemini 3.1 Pro	CoT	.626	.316	.420	.490	.169	.252	.200	1.12	.69
	No CoT	.630	.321	.425	.454	.202	.280	.168	1.21	.69
Gemma 4 31B	CoT	.696	.284	.403	.514	.155	.238	.639	.72	.45
	No CoT	.622	.295	.400	.437	.225	.297	.258	1.19	.65
Gemma 4 E4B	CoT	1.000	.002	.003	.000	.000	.000	**.987**	**.01**	**.00**
	No CoT	1.000	.002	.003	1.000	.001	.002	**.994**	**.01**	**.00**

Filter: The matching teacher—Claude Opus 4.8 for Opus branches and GPT-5.4 for GPT-5.4 branches—receives the same timestamped JPEG frames used for detector evaluation. The input uses approximately 5 fps and at most 40 frames from the leading window. No candidate event is attached. With the non-CoT video-level filter prompt, it returns only `<answer>implausible</answer>` or `<answer>plausible</answer>`. The former means that at least one genuine implausible event is visible somewhere in the clip. The latter means that none is visible. Watermarks, overlays, and encoding artifacts do not count. We use greedy decoding and one filter call per video, for 325 calls per teacher. An unparseable verdict closes the gate and is treated as `plausible`.

Generate: For a `plausible` verdict, the cascade emits `[]` and does not run the student. For an `implausible` verdict, it runs the corresponding RL checkpoint from Table 4 greedily with the two-section detector prompt and retains the complete Section 2 event list. The filter does not remove individual events. Because both components are greedy, this is equivalent to generating on all videos and replacing the event list with `[]` on filter-negative videos. For each teacher, one set of 325 filter verdicts is shared across its E2B and E4B RL checkpoints.

Scoring: We score the gated lists with the same protocol as every other detector. We use GPT-5.5 semantic matching and one-to-one assignment on the 170 videos with annotated implausible events. We compute empty-set metrics on the 155 no-event videos. A false `plausible` verdict removes every prediction for a video with annotated implausible events and therefore reduces recall. A true `plausible` verdict emits the correct empty list on a no-event clip and improves specificity. A false `implausible` verdict retains the RL list. Its events count as empty-set false positives. The verbatim filter prompt appears in Appendix D.3.

G Additional Detector Results

G.1 Off-the-Shelf Detector Results

Table 10 compares CoT and non-CoT prompting under the same evaluation protocol. Rows for each model are adjacent to make the prompting tradeoff explicit. Here CoT means a visible JSON object containing a scene summary, ordinary events, and causal relations before the final list of implausible events. It is not hidden chain-of-thought. The non-CoT prompt asks only for the final event list. Both prompts exclude watermarks, overlays, and encoding artifacts.

G.2 Same-video CoT Comparisons

Video with no annotated implausible event: In a real ball-pit video, the correct output is an empty list. GPT-5.4 mini with CoT describes small ball movements as ordinary contact and settling,

then returns no event. Without CoT, it invents three levitating balls. The same pattern appears in a real roller-coaster video: CoT describes an ordinary gravity-driven drop, while non-CoT claims that the train tilts beyond what the rails allow. These examples show how CoT can raise specificity by encouraging an ordinary explanation, even when the same behavior hides real implausible events in synthetic videos.

H Detector Scoring Protocol

H.1 Semantic Matching

For every video with non-empty ground truth, we make one deterministic GPT-5.5 call with the complete ground-truth and predicted event lists. The judge outputs all semantically valid candidate edges rather than selecting a final assignment. Each edge contains a ground-truth index, a prediction index, and a grade $g_{ij} \in \{1, 2\}$. The judge assesses semantic correspondence rather than exact wording or timestamp agreement and does not choose the final matching.

The two grades distinguish semantic equivalence from differences in event granularity. Grade 2 denotes essentially the same physical implausibility. Grade 1 denotes a valid more-specific component or a broader description that clearly contains the annotated event. Both grades count as valid event detections. This design does not require the annotator and model to describe the same phenomenon at exactly the same level of detail. The assignment still prefers matches with equivalent detail when it has a choice.

Invalid indices and grades outside $\{1, 2\}$ are discarded. If the same pair is returned multiple times, the final returned grade is retained. If either the ground-truth list or the prediction list is empty, the semantic judge is not called.

H.2 One-to-One Assignment

Let G be the set of ground-truth events, P the set of predicted events, and $R \subseteq G \times P$ the set of semantic edges returned by the judge. For each $(i, j) \in R$, let $g_{ij} \in \{1, 2\}$ denote its semantic grade and I_{ij} its temporal intersection.

We assign each valid edge the weight

$$w_{ij} = c_{\text{scale}} + i_{\text{scale}} g_{ij} + I_{ij}, \qquad (i, j) \in R, \tag{7}$$

and assign weight zero to non-edges, where

$$i_{\text{scale}} = \sum I + 1, \tag{8}$$

and

$$c_{\text{scale}} = i_{\text{scale}} \left(\sum g + 1 \right). \tag{9}$$

We then solve for the maximum-weight one-to-one assignment using the Hungarian algorithm. By construction, the scaling implements the lexicographic objective

$$\text{\# matched pairs} \succ \text{total semantic grade} \succ \text{temporal intersection}. \tag{10}$$

The assignment first maximizes the number of matched events. It then prefers a grade-2 correspondence over a grade-1 correspondence. It uses temporal overlap only as the final tie-breaker.

H.3 Semantic, Temporal, and False-Positive Metrics

Let N_{GT} and N_{pred} be the numbers of ground-truth and predicted events on the 170 videos with annotated implausible events only. Here, $N_{\text{GT}} = 615$. Let TP be the number of assigned grade-1 or grade-2 pairs. Semantic precision, recall, and F1 are

$$\text{sP} = \frac{\text{TP}}{N_{\text{pred}}}, \qquad \text{sR} = \frac{\text{TP}}{N_{\text{GT}}}, \qquad \text{sF1} = \frac{2\,\text{sP}\,\text{sR}}{\text{sP} + \text{sR}}. \tag{11}$$

Each metric is set to zero when its denominator is zero.

For matched intervals $[s_i, e_i]$ and $[s_j, e_j]$, let

$$I_{ij} = \max\left(0, \min(e_i, e_j) - \max(s_i, s_j)\right). \tag{12}$$

We set $I_{ij} = 0$ when a span is missing or inverted. Duration-pooled temporal recall and precision are

$$\mathrm{tR} = \frac{\sum_{(i,j)\in M} I_{ij}}{\sum_{i\in G}(e_i - s_i)}, \qquad \mathrm{tP} = \frac{\sum_{j\in M} \mathrm{union}(\{I_{ij}\})}{\sum_{j\in P}(e_j - s_j)}, \tag{13}$$

with $\mathrm{tF1} = 2\,\mathrm{tP}\,\mathrm{tR}/(\mathrm{tP}+\mathrm{tR})$. The precision numerator measures the union of overlap within each matched prediction, preventing duplicated overlap from being counted twice. Durations are pooled rather than averaged per event, so long events receive more weight. Point events have zero duration and therefore do not contribute to temporal F1.

For videos whose correct output is empty, specificity is

$$\mathrm{Specificity} = \frac{|\{v : \hat{Y}_v = \emptyset\}|}{155}. \tag{14}$$

We also report

$$\mathrm{fp/empty} = \frac{1}{155}\sum_{v\in\mathcal{E}} |\hat{Y}_v|, \qquad \mathrm{fp/impl} = \frac{N_{\mathrm{pred}} - \mathrm{TP}}{170}, \tag{15}$$

where $\mathcal{E}$ is the 155-video no-event set. This set includes five synthetic videos and 150 real videos. No-event videos are never sent to the semantic judge. A detector that fires on every clip can obtain strong semantic F1 with zero specificity. A silent detector can obtain perfect specificity with zero recall on videos that contain annotated implausible events.

Verifier metrics evaluate a different task and are not comparable to detector sF1. For verification, `correct` is the positive class. Synthetic metrics use 1,276 labeled claims, rather than the detector-evaluation buckets. We report per-class precision, recall, and F1, together with

$$\mathrm{MCC} = \frac{\mathrm{TP}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{16}$$

setting MCC to zero when the denominator is zero.

H.4 Semantic Match-Judge Prompt

System prompt:

```
You are given two numbered lists from a SINGLE AI-generated video:
the GROUND-TRUTH physically-implausible events, and the PREDICTED
implausible events. Using both lists together, output every
(ground-truth, prediction) pair that describes the SAME implausibility.
Grade each pair: 2 if the prediction describes essentially the same
implausibility as the ground-truth event; 1 if it is a correct
MORE-SPECIFIC part of it, or a correct BROADER description that clearly
includes it. Ignore wording and exact timestamps, and do NOT output
pairs for unrelated events. A ground-truth event may relate to several
predictions and a prediction to several ground-truth events -- list every
such pair; the one-to-one assignment is handled downstream. Answer with
ONLY a JSON array like [{gt: 1, pred: 3, grade: 2}], using 1-based
numbers from the lists, or [] if nothing matches. Output nothing else.
```

User prompt:

```
GROUND-TRUTH events:

1. (t={start:.2f}-{end:.2f}s) {description}

...

PREDICTED events:

1. (t={start:.2f}-{end:.2f}s) {description}

...
```

Table 11: Same-video outputs from one representative E4B–Opus 4.8 run. A match is a one-to-one semantic match to a human event.

Video	Stage	Pred.	Match	Main behavior
Boy and ball	Teacher dist.	0	0	Returns `[]`
	RL	2	2	Recovers long events; one broad span
	Self-dist.	4	4	Separates more events; spans remain wide
Waterfall	Teacher dist.	0	0	Correct `[]`
	RL	2	0	Flags water flow and camera motion
	Self-dist.	0	0	Correct `[]`

I Qualitative Detector Analysis

I.1 Changes Across the Three Training Stages

Table 11 follows one synthetic video with six annotated implausible events and one real video with an empty target. In the synthetic video, a boy juggles a soccer ball near a brick wall. The implausible events include a missed kick that still moves the ball, a prolonged floating trajectory, the ball entering the wall, an implausible rebound, and continued spinning after the catch. Teacher distillation returns an empty list. RL recovers two trajectory implausibilities, but one prediction spans 2.00–7.92 seconds and covers several distinct events. Self-distillation produces four matched predictions and separates some of these events, although its intervals remain wider than the human annotations. The two misses are short: a 0.63-second contact implausibility and a 0.21-second wall intersection.

The real video shows a waterfall and has target `[]`. Teacher distillation correctly returns an empty list. RL instead reports inconsistent water flow and inconsistent camera motion as two implausible events. Self-distillation again returns an empty list.

I.2 Temporal Localization

Across every trained detector in Table 4, temporal F1 is lower than semantic F1. The models therefore recover event identity more reliably than precise temporal boundaries. The boy-and-ball example in Table 11 illustrates this pattern: the RL prediction finds the relevant trajectory implausibility but spans several distinct events.

J Verifier Training, Prompts, and Additional Results

J.1 Data and visual input

Verifier supervision comes from 1,432 human verdicts on candidate events from 376 training videos. The remaining five training videos contain no candidate claims and therefore do not contribute verifier examples. Genuine implausible events form the IMP class and use the target `correct`. Candidates that are not genuine implausibilities form the ¬IMP class and use the target `wrong`. A video-level 90/10 split with seed 42 yields 1,293 training claims (852 correct, 441 wrong) and 139 validation claims (85 correct, 54 wrong). We use the validation claims for checkpoint selection and do not rebalance the classes. The synthetic evaluation uses the 175 synthetic evaluation videos and contains 1,276 candidate claims. One video has no candidate claim. The real evaluation contains 66 candidate claims from the 25 annotated controls and 424 fabricated claims from the other 125 videos. All 490 real-video claims are labeled `wrong`, so we report ¬IMP recall and F1 on this slice.

Frames are pre-decoded at native fps divided by five. For each claim, the model receives frames within the annotated interval, two cached context frames on each side, and at most 40 frames total. Each frame is prefixed by its index and timestamp. Claude teacher inputs are resized to a maximum long side of 1,024 pixels. GPT-5.4 and student inputs retain the cache resolution.

J.2 Privileged teacher targets and optimization

For CoT training, GPT-5.4 and Claude Opus 4.8 independently generate rationales for all 1,432 claims. We use privileged distillation. Before generating each CoT rationale, the teacher receives

the human `correct`/`wrong` verdict. We instruct it to justify it without revealing that a label was supplied. Correct rationales must identify the visual evidence and violated rule. Wrong rationales must explain whether the event is absent, mis-described, or physically ordinary. Outputs are retained only when both tags parse and the answer matches the human verdict. Each CoT run distills a single teacher.

Students are Gemma 4 E4B-it and E2B-it. Training uses FSDP in bfloat16 with gradient checkpointing and SDPA, AdamW with learning rate 2×10^{-5}, cosine decay, warmup ratio 0.03, minimum learning-rate ratio 0.1, weight decay 0.1, betas $(0.9, 0.95)$, gradient clipping at 1.0, and a 32,768-token context. CoT runs train for three epochs with global batch 32 on two GPUs. We run seeds 101, 102, and 103.

Evaluation uses greedy decoding with uncapped generation. We parse the last answer span and count unparseable outputs as errors. For each seed, we select by MCC on the 139-claim validation split. Reported results are averages over the three runs.

J.3 Verifier prompts

Frames replace each `<image>` placeholder below. The claimed time window and natural-language description are filled separately for each event. The CoT and non-CoT conditions use the same visual evidence and decision rule; they differ only in the output instruction.

J.3.1 Shared system prompt

```
You are an expert at verifying claims about physically implausible events in videos.

CONTEXT: The video you will see was generated by a video-generation model (a text-to-video
     diffusion / autoregressive system). Such models routinely produce frames that violate
     everyday physics, causality, anatomy, or object permanence -- but they also produce
     perfectly plausible videos. Do not assume the footage is real, and do not assume it is
     broken either; judge only what the frames actually show.

You will be given:
  * A sequence of frames sampled at a fixed rate from the generated video, in temporal order
     . Each frame is preceded by a text caption frame N (t=...s): giving its real-time
     timestamp in seconds.
  * A CLAIM that one specific implausible event occurs in the video, along with its
     predicted (start_sec, end_sec) window.

Your job: decide whether the claim is

  * correct -- the described implausible event IS actually visible in the video at (or close
      to) the predicted window AND it genuinely violates physical, causal, temporal, or
     spatial consistency (objects defying gravity / momentum / rigidity, shapes morphing
     without cause, objects appearing or disappearing, limbs deforming unnaturally, motion
     that contradicts contact, liquid flowing uphill, teleportation, identity swaps, mid-air
      direction changes without external force, etc.); OR

  * wrong -- the described event is NOT actually visible in the video, is mis-described,
     points at something that is in fact consistent with normal physics, or the video is
     simply plausible. In any of these cases the implausibility claim does not hold.

Do NOT treat watermarks, on-screen overlays, encoding artefacts, or camera metadata as
     implausibilities -- only the depicted physical scene counts.
```

J.3.2 Non-CoT output instruction

For the non-CoT experiments, we append the following instruction:

```
Do NOT explain or reason. Output ONLY your final verdict -- exactly one word, either correct
      or wrong -- between <answer> and </answer> tags, with nothing before or after.

Example format:
<answer>correct</answer>
```

J.3.3 CoT output instruction

For the CoT experiments, the output instruction is instead:

```
Reason step by step inside <think>...</think>, then output exactly one word -- either
     correct or wrong -- between <answer> and </answer> tags.
```

Table 12: Off-the-shelf verifiers without CoT on the same synthetic slice. Bal is balanced accuracy and mF1 is macro-F1.

					Imp			¬Imp		
Model	Acc	MCC	Bal	mF1	P	R	F1	P	R	F1
Claude Opus 4.8	.472	−.047	.478	.458	.466	.657	.545	.484	.300	.370
GPT-5.5	.545	**.084**	.539	**.529**	.542	.371	.440	.547	.708	.617
GPT-5.4	.450	−.095	.453	.447	.442	.541	.487	.461	.365	.407
GPT-5.4 mini	.502	.003	.501	.501	.483	.498	.490	.519	.505	.512
Gemini 3.1 Pro	.474	−.037	.485	.425	.473	.795	.593	.479	.175	.257
Gemma 4 31B	.464	−.066	.477	.396	.468	.829	**.599**	.439	.124	.193
Gemma 4 E4B	.520	.025	.511	.486	.503	.273	.354	.525	.749	.618
Gemma 4 E2B	.511	−.016	.496	.393	.455	.073	.126	.516	.918	**.661**

```
Example format:
<think>
...your reasoning, citing frame captions like at frame 7 (t=1.46s)......
</think>
<answer>correct</answer>
```

J.3.4 User message

Both conditions use the same user-message template:

```
VIDEO (K frames in temporal order, each labelled with its timestamp in seconds):
frame 1 (t=0.00s):
<image>
frame 2 (t=0.21s):
<image>
...
frame K (t=T.TTs):
<image>

CLAIMED IMPLAUSIBLE EVENT (predicted window: start_sec=S.SS, end_sec=E.EE):
<description>

<FOOTER>
```

For the CoT condition, `<FOOTER>` is:

```
Is this implausibility claim correct or wrong? Respond in the <think>...</think><answer>
     correct|wrong</answer> format.
```

For the non-CoT condition, `<FOOTER>` is:

```
Is this implausibility claim correct or wrong? Do NOT explain. Respond with ONLY your
     verdict in the <answer>correct|wrong</answer> format.
```

J.4 Privileged teacher prompt

The CoT teacher receives the same frames and claim, followed by a private block:

```
TRAINING DATA NOTE (you only, the student will not see this): the
ground-truth verdict for this claim is correct|wrong.
```

The teacher system message explains that it is producing training data, requests a four-to-eight-sentence visual rationale, and forbids mentioning the supplied label (e.g., "ground truth" or "the label is"). Its completion must end with `<answer>correct|wrong</answer>`. The student receives neither this private note nor any other verdict hint.

J.5 Off-the-shelf baselines and full SFT metrics

Main Table 2 reports the CoT results; Table 12 gives the corresponding non-CoT results on the same 1,276-claim synthetic evaluation slice.

J.5.1 Qualitative CoT comparison

This example uses a synthetic evaluation clip distinct from the training example in Figure 1; the candidate claim and time window are identical across prompting conditions. The human verdict is

Table 13: Real-video slice: 490 claims from 150 real videos, all ground truth ¬IMP. The 25 annotated controls contribute 66 candidate claims, and the other 125 videos contribute 424 fabricated claims. All models use non-CoT prompting.

Model	Prompt	R ¬IMP	F1 ¬IMP
GPT-5.5	non-CoT	**.941**	**.970**
Claude Opus 4.8	non-CoT	.712	.832
GPT-5.4 mini	non-CoT	.710	.831
GPT-5.4	non-CoT	.573	.729
Gemini 3.1 Pro	non-CoT	.312	.476
Gemma 4 31B	non-CoT	.159	.275

Table 14: CoT results averaged over the same three SFT runs on the all-¬IMP real-video slice. Because every claim is negative, we report ¬IMP recall and F1.

Student	Prompt	Teacher	R ¬IMP	F1 ¬IMP
E4B	CoT	Opus 4.8	.463	.633
E4B	CoT	GPT-5.4	.463	.633
E2B	CoT	Opus 4.8	**.556**	**.715**
E2B	CoT	GPT-5.4	.459	.629

IMP (`correct`): the dough remains airborne for an implausibly long time. GPT-5.4 mini with CoT explains the motion as an ordinary toss and returns `wrong`. Without CoT, the same model returns `correct`. This example illustrates the operating-point shift between Main Table 2 and Appendix Table 12; it does not by itself establish that non-CoT is better.

```
CLIP: Pizza-dough evaluation video
WINDOW: start_sec=0.83, end_sec=3.12
HUMAN VERDICT: correct (Imp)
CLAIM:
The pizza dough rises and then hovers unrealistically high in the air
for an extended time, with very slow descent inconsistent with gravity.
```

GPT-5.4 mini with CoT:

```
<think>
The dough is clearly being tossed upward by the chef, so its initial
rise is physically plausible. From frame 3 (t=0.83s) through frame 14
(t=3.12s), it remains airborne near the chef's hands, but this is
consistent with a pizza-tossing motion in a short clip: the dough is
large, light, and still in the upward/near-apex phase for much of the
sequence. It does descend gradually by frames 14-16, so there is no
obvious teleportation or gravity-defying hover. The claim overstates
the implausibility.
</think>
<answer>wrong</answer>
```

GPT-5.4 mini without CoT:

```
<answer>correct</answer>
```

Because the real slice contains no positive examples, we report only negative-class recall and F1. Table 13 gives the complete non-CoT comparison. GPT-5.5 rejects 94.1% of the 490 real-video claims.

Main Table 3 gives the CoT student metrics. The Opus-distilled E4B and E2B students reach mean MCC values of 0.087 and 0.086, respectively. On the real-video slice, Table 14 shows that the Opus-distilled E2B student has the highest ¬IMP recall at 0.556.