VIDEO-XAI BENCH: A KNOWN-EVIDENCE BENCHMARK FOR SPATIOTEMPORAL EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video classifiers flag events for human review in public-safety and security monitoring. A classifier's decision is hard to trust unless a reviewer can check the evidence behind it. Explanation methods are meant to provide this evidence by highlighting the regions and moments of a video that drove the decision. Existing evaluations check if explanations look plausible or if removing the highlighted content lowers the model's confidence; they do not test if explanations point to the right place and moment. We introduce Video-XAI Bench, a benchmark for explanations of video classifiers. It tests explanation methods in two settings. The first uses real violence-recognition videos. The second adds a moving pattern to real videos at a known place and time, so we know exactly where and when the evidence is. On real videos, no method consistently performs best, as the best choice changes with the dataset, the model, and the score. With known evidence, scores based on the model's confidence do not reveal if an explanation finds the right moment, and some of the explanations that find it best partly reflect the input video rather than what the model learned. In a pilot with 18 participants, those who agreed that a video showed aggression often disagreed about which frames showed it, so human-marked evidence on real videos forms a window with soft edges. Our study concludes that video explanations should be tested against known evidence, not only against the model's confidence. Anon. code: https://anonymous.4open.science/r/video-xai-bench-iclr2027-A1BA/

1 INTRODUCTION

When a video classifier flags an event for a human reviewer, such as a security operator, the reviewer needs to know what triggered the flag. The trigger may be a safety-relevant incident, such as physical violence, or an incidental feature, such as camera motion or a background artifact. AI-based video surveillance is deployed in many countries (Feldstein, 2019), and its law-enforcement uses are regulated as high-risk (European Parliament and Council of the European Union, 2024). Post-hoc explanation methods can meet this need by providing evidence behind a decision. However, an explanation to the operator provides evidence of what the model actually used.

Explaining a video model is more challenging than explaining an image model. An image explanation shows *where* the evidence is, whereas a video explanation must also show *when*. The evidence may lie in motion or in the order of events, which no single frame contains. Yet, most video explanation methods are image methods extended along the time axis, such as class-activation maps, gradient attribution, and input masking (Selvaraju et al., 2017; Sundararajan et al., 2017; Petsiuk et al., 2018; Stergiou et al., 2019; Li et al., 2021), and so are their evaluations. Deletion and insertion (Petsiuk et al., 2018), the most common tests, remove or restore the most highlighted content and track change in the model's confidence; they measure how the model reacts to altered inputs, not if the highlighted region or moment is the evidence (Srinivas & Fleuret, 2021); and the altered inputs can distort the result (Hooker et al., 2019; Rong et al., 2022). Such metrics can also disagree with one another and rank the same methods differently (Tomsett et al., 2020). Other common checks have similar blind spots: temporal consistency rewards neighboring frames with similar maps and is satisfied by an explanation that is consistently wrong, and some explanations look plausible even after a model's parameters or labels are randomized (Adebayo et al., 2018). None of these checks asks if an explanation points to the right place *and* moment, and with no common benchmark, evaluation practices vary widely across studies (Nauta et al., 2023). What is missing is a way to test

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

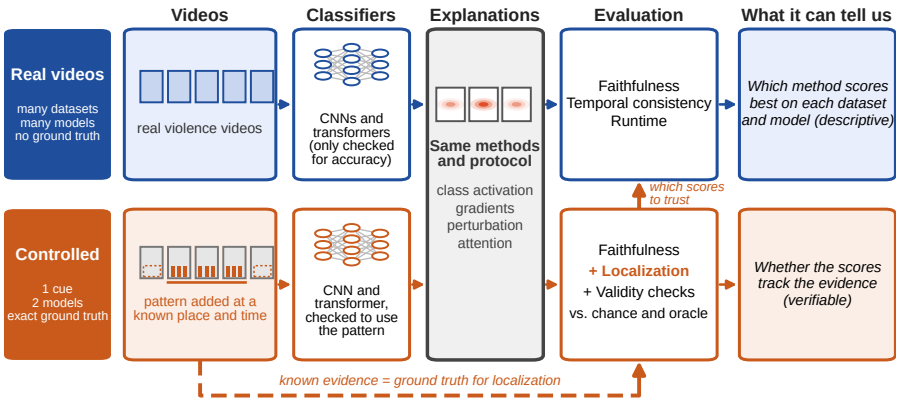


Figure 1: Video-XAI Bench: Real-video setting compares methods but cannot judge them; Controlled setting adds a cue at a known place & time so the same scores checked for known evidence.

video explanations against evidence whose location and timing are *known*, under one protocol that also tests if the explanations depend on the model.

We introduce Video-XAI Bench, a benchmark for evaluating explanations of video classifiers, instantiated on violence recognition. The benchmark pairs two complementary settings (Figure 1). (1) A *real-video setting* compares nine explanation methods from four families across five architectures and three datasets; because real videos carry no record of the evidence a model used, it can describe methods but not judge them. (2) A *controlled setting* asks if the scores track evidence we know is there. We add a synthetic moving pattern with a known location and time window to real videos, verify through task validation that the model relies on it, and score explanations against it. Scores are reported alongside chance and oracle references, with checks that each explanation depends on the model. Together, the two settings separate what can be observed on real data from what can be verified against known evidence. To the best of our knowledge, Video-XAI Bench is the first benchmark to score explanations of video classifiers against exact spatial and temporal evidence.

On real videos, no method consistently performs best. With known evidence, faithfulness scores, which track the model confidence do not indicate if an explanation finds the right moment: on R(2+1)D, better insertion goes with worse timing, a relation driven by a few methods, and on TimeSformer the two are unrelated. The explanations that localize best on R(2+1)D also depend only partly on the model.

In summary, our contributions are:

1. **Video-XAI Bench**, a two-setting benchmark that evaluates video explanations under one protocol and scores faithfulness, temporal consistency, localization, validity, and runtime separately.
2. **A controlled setting with exact spatial and temporal evidence**, matched controls, task-validity gates, chance and oracle baselines.
3. **Evidence that faithfulness, localization, and validity are distinct**: rankings shift across datasets, architectures, and metrics, and higher insertion does not mean better timing.
4. **A human-evidence pilot on real videos**: 18 participants who agreed that a clip showed aggression disagreed about which frames showed it, so human-marked evidence forms a soft window.

2 Related Work

Explaining video models. Video classifiers extend image networks along time, from inflated and factorized spatiotemporal convolutions (Carreira & Zisserman, 2017; Tran et al., 2018) and two-pathway and efficient designs (Feichtenhofer et al., 2019; Feichtenhofer, 2020) to space–time attention and masked pretraining (Bertasius et al., 2021; Tong et al., 2022); violence recognition, has small benchmarks from sports to surveillance footage (Bermejo Nievas et al., 2011; Cheng et al., 2021; Soliman et al., 2019). Their explanation methods largely come from images, extended to inputs: class-activation maps (Selvaraju et al., 2017; Chattopadhay et al., 2018; Jiang et al., 2021; Stergiou et al., 2019; Li et al., 2021), gradient attribution (Sundararajan et al., 2017; Smilkov et al.,

2017), masking (Petsiuk et al., 2018; Fong et al., 2019), and, for transformers, attention rollout and relevance propagation (Abnar & Zuidema, 2020; Chefer et al., 2021). Newer methods target efficiency (Wang & Liu, 2025) or concepts (Saha et al., 2024; Lee et al., 2025; Kowal et al., 2024), and their differing assumptions motivate a common evaluation.

Evaluating explanations. Explanations are usually evaluated with model-side proxies, such as deletion and insertion (Petsiuk et al., 2018) or retraining after feature removal (Hooker et al., 2019), often via toolkits and benchmarks (Hedström et al., 2023; Agarwal et al., 2022; Li et al., 2023b); these proxies measure how a model reacts to altered inputs, not if the highlighted content is the evidence (Srinivas & Fleuret, 2021). Benchmarks with known evidence address this for tabular and image data (Liu et al., 2021; Kim et al., 2022a; Arras et al., 2022; Hesse et al., 2023), and time-series and audio benchmarks score *where in time* methods place importance (Ismail et al., 2020; Crabbé & van der Schaar, 2021; Bolaños et al., 2025). For video, synthetic motion has been used for representation learning (Doughty et al., 2024), and models exploit static appearance cues (Li et al., 2023a), so a controlled video task needs controls that separate motion from appearance. Human studies if people can predict or verify a model's decision from (Hase & Bansal, 2020; Nguyen et al., 2021; Colin et al., 2022; Kim et al., 2022b); they measure usefulness to people, not if the highlighted content is the model's evidence, and none targets video. To our knowledge, no prior work scores explanations of video classifiers against exact spatial and temporal evidence while comparing methods across architectures and datasets under one protocol and checking that explanations depend on the model.

3 Video-XAI Bench

Video-XAI Bench has two goals: to compare explanation methods under a common protocol, and to check if the metrics used in that comparison can be trusted. It meets these goals with two settings (Figure 1). The real-video setting is deliberately broad (three datasets, five architectures, nine methods), but it is descriptive: real videos carry no record of the evidence a model used. The controlled setting is deliberately narrow (one cue, two architectures), but the evidence is exact. Each setting therefore covers what the other cannot. A pilot human study on five real videos (§ 5.4) asks what evidence looks like when it comes from people rather than from construction. § 4 describes the data, models, methods, and metrics.

Evaluation dimensions. The benchmark keeps five dimensions of explanation quality apart (definitions in § 4). *Faithfulness* asks if the model's confidence drops when the highlighted content is removed and recovers when it is restored. *Temporal consistency* asks if the map is stable across neighboring frames. *Localization* asks if the map covers the known evidence. *Validity* asks if the map depends on the learned model rather than on the input alone. *Runtime* is the time it takes to produce one map. Faithfulness is measured in both settings, temporal consistency and runtime in the real-video setting only, and localization and validity in the controlled setting only. Validity is necessary but not sufficient, since an explanation can depend on the model and still track spurious features. We keep the dimensions apart because they are easily conflated: a map can be stable, faithful by insertion, and even well localized, and still be invalid, for instance because it reproduces the input's own structure (Adebayo et al., 2018).

Controlled setting. In the controlled setting, a classifier sees a 32-frame clip and must decide if the clip contains a pattern that moves coherently, meaning, in the same direction by the same step on every frame. Explanations are then scored against the pattern's known place and time. *Before* scoring this, the setting must show if a model uses this evidence. So we required five properties: (P1) the task is learnable; (P2) the label depends on temporal *order*; (P3) it cannot be read from single frames; (P4) the evidence is exact in space and time; and (P5) matched controls separate order from appearance, energy, and source content. Prior work motivates controlled motion (Doughty et al., 2024) but not the cue or thresholds (Appx. Table 14).

The inserted pattern, which we call the *cue*, is a set of soft vertical stripes that drift sideways, called drifting grating, added to one 112×112 quadrant Q of a 32-frame Hockey Fight non-fight clip in frames 4–27 (Figure 2):

$$V_t'(x,y) = \mathrm{clip}\big(V_t(x,y) + \alpha\, w(x,y)\sin(2\pi x/\lambda - \phi_0 - k_t\,\omega),\ 0,\ 1\big), \tag{1}$$

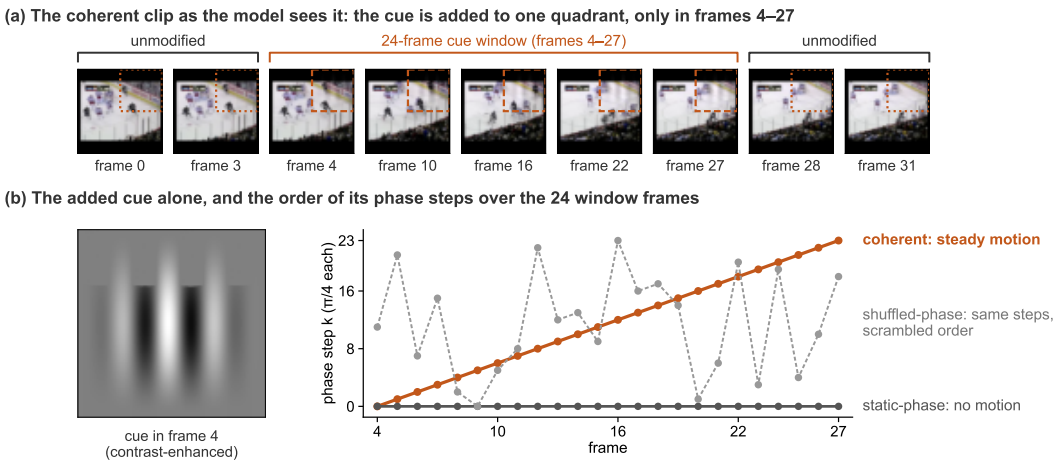


Figure 2: The cue in one test clip, as the models see it. (a) The cue occupies a randomly chosen quadrant (dashed box) in frames 4–27, barely visible since its amplitude matches the clip's own frame-to-frame change. (b) Left: the cue alone. Right: phase step per frame – coherent: one step ($\pi/4$) per frame; static: repeats it; shuffled: scrambles the order.

for $(x, y) \in Q$ and $t = 4, \ldots, 27$; all other pixels and frames are unchanged. Here w fades the stripes toward the edges of Q, $\lambda = 28$ px, $\omega = \pi/4$, ϕ_0 is a random offset, and α is three times the clip's own mean frame-to-frame change in Q, so the cue is as strong as the clip's own motion and is added to the video rather than pasted over it. Appx. C gives the exact, normalized form (Equation 5).

The phase step k_t controls how the stripes move, and this is what the classifier must detect. Coherent clips (label 1) use $k_t = t - 4$, so the stripes move right by 3.5 px per frame. Two controls (label 0) show the same stripes without coherent motion: the shuffled-phase control uses the same steps in a random order, so it differs from the coherent clip only in frame order. The static-phase control fixes $k_t = 0$, so the stripes do not move. Because any single frame shows only stationary stripes, the label exists only in the frame order, which is why we use a moving grating rather than a static pattern (Li et al., 2023a). A fourth, cue-absent variant serves audits only. We use 300 training, 75 validation, and 71 test sources, keep all variants of a source in one split, and treat the source as the unit of independence.

Near-perfect accuracy does not show that a model uses the intended cue. We, therefore, test each model with eight task-validity gates before any explanation is scored (Appx. Table 15). The main ones are: hiding the true region drops accuracy to the majority-class floor, while hiding another quadrant does not; every control type is classified correctly, a model trained on shuffled labels falls to the same floor; the label cannot be read from single frames under the tested probe; and nuisance features alone do not predict it. Two gates carry stated limits (see Appx. C.2).

The cue was found by a search in which shorter gratings and warps failed a pre-specified order-sensitivity gate ; the 24-frame grating passed (Appx. Table 16). No explanation method was used to choose it, and all pre-specified gates passed before any explanation on the final design was scored; one gate was formalized afterward. We do not claim that 24 frames is minimal or optimal, or that the program was preregistered (Appx. C.2).

4 Experimental Setup

Datasets & Models. On real videos we use Hockey Fight (Bermejo Nievas et al., 2011), RLVS (Soliman et al., 2019), and RWF-2000 (Cheng et al., 2021): professional ice-hockey footage, short real-life clips, and surveillance footage. We evaluate on each dataset's test partition after removing clips that duplicate, or for RWF-2000 share a source video with, a training or validation clip, which leaves 148, 299, and 95 clips; only 19 of the RWF-2000 clips are fights, so fight-class recall matters there (Appx. B, Table 4). We evaluate R(2+1)D-18 (Tran et al., 2018), SlowFast (Feichtenhofer et al., 2019), X3D-S (Feichtenhofer, 2020), TimeSformer (Bertasius et al., 2021), and VideoMAE

(Tong et al., 2022), one fully trained checkpoint per architecture and dataset (15 in total), fine-tuned as binary violence classifiers; their accuracy on the evaluation sample is 90.5–98.9% (Appx. Table 4). In the controlled setting we fine-tune R(2+1)D-18 and TimeSformer from Kinetics-400 with seeds 42, 123, and 7 plus one label-randomized checkpoint each (accuracies in Appx. C). All models are frozen during explanation and see $T = 32$ frames at 224×224 (16 for VideoMAE).

Explanation methods. We evaluate methods from 4 families (§ 2): class-activation maps (Grad-CAM++, LayerCAM), gradient attribution (Integrated Gradients, IG; SmoothGrad-IG), perturbation (Occlusion, RISE, Extremal Perturbation, EP), and transformer propagation (attention rollout (Abnar & Zuidema, 2020) and a gradient-weighted variant in the spirit of Chefer et al. (2021)). On real videos, IG, SmoothGrad-IG, Occlusion, and EP run on all five architectures and form the primary comparison; the CAM methods and RISE run on the convolutional networks, attention rollout on TimeSformer, and gradient-weighted rollout on VideoMAE. In the controlled setting, R(2+1)D runs seven methods and TimeSformer six, sharing IG, SmoothGrad-IG, Occlusion, RISE, and EP. Every method returns a saliency volume $S \in \mathbb{R}^{T \times H \times W}$ at input resolution, so every metric is defined on the same object. The explained class is the predicted class on real videos, even when the prediction is wrong, and the true class in the controlled setting, where only correctly classified coherent clips are explained. Each setting fixes one configuration per method before the runs. The configurations differ between settings, so scores are compared only within a setting. For example, IG and SmoothGrad-IG are smoothed and sharpened in the controlled setting but not on real videos (Appx. Tables 6 and 17).

Faithfulness. Deletion and insertion (Petsiuk et al., 2018) remove or restore the space–time positions with the highest signed saliency first (no absolute value is taken) and track the target-class probability; a faithful map makes the probability fall quickly under deletion and rise quickly under insertion. Their AUCs integrate this probability over eleven masking ratios from 0 to 1 on real videos and ten from 0.1 to 1 in the controlled setting, so AUCs are comparable within a setting only (lower deletion and higher insertion are better; Equation 3 in Appx. A). Since the scores are not normalized by the unmasked confidence, they also reflect model confidence: misclassified real clips average 0.59 deletion AUC and correct ones 0.72, so a low-confidence clip looks more faithful under deletion. We therefore compare methods only within a dataset–architecture cell.

Temporal consistency. Temporal consistency (TC) is the mean cosine similarity of consecutive saliency frames, each min–max normalized separately (formula in Appx. A). It measures stability, not correctness, and has no preferred direction: a map that repeats one spatial pattern in every frame scores 1 no matter when the evidence occurs.

Localization (controlled setting only). The cue's quadrant Q and frames 4–27 define a space–time support $G_{thw} \in \{0, 1\}$ with temporal window $W_t \in \{0, 1\}$: G_{thw} marks the cue's quadrant in frames 4–27 and W_t marks those frames. With positive saliency $A = \max(S, 0)$, spatiotemporal relevance-mass accuracy (ST-RMA) applies relevance-mass accuracy (Arras et al., 2022) to this support, and we split it into a temporal localization score (TLS) and spatial precision within the window (SPW):

$$\text{ST-RMA} = \frac{\sum_{t,h,w} A_{thw} G_{thw}}{\sum_{t,h,w} A_{thw}}, \quad \text{TLS} = \frac{\sum_{t,h,w} A_{thw} W_t}{\sum_{t,h,w} A_{thw}}, \quad \text{SPW} = \frac{\sum_{t,h,w} A_{thw} G_{thw}}{\sum_{t,h,w} A_{thw} W_t}, \tag{2}$$

TLS is the share of positive mass inside the true window and SPW the share of that in-window mass that falls on the true region; since $\text{ST-RMA} = \text{TLS} \times \text{SPW}$ per clip, they separate *when* from *where given the right time*. The Pointing Game scores if the maximum of the signed saliency S falls inside G. A map that ignores the support scores 0.75 on TLS (24 of 32 frames), 0.25 on SPW (one of four quadrants), and 0.1875, their product, on ST-RMA and the Pointing Game. Because chance on TLS is so high, we also report it chance-adjusted, with 0 at chance and 1 at the ceiling (Appx. A). Our main localization score is TLS, because timing is the question the controlled setting asks; ST-RMA also rewards location and can hide a timing failure.

Validity (controlled setting only). Following Adebayo et al. (2018), we randomize the network's weights block by block, from the output toward the input (cascading randomization), and compute

the Spearman correlation between each method's saliency and its trained-model saliency. A correlation that decays to about 0 indicates dependence on the learned model; one that plateaus does not. After full randomization, every method we mark as passing is at most 0.14 in absolute value and every method we mark as failing is at least 0.60, so no method falls near the boundary (Appx. C). In the label-randomization test, we explain a checkpoint trained on shuffled labels; localization should fall to chance.

Runtime and reference maps. Runtime is the wall-clock time to generate one explanation, reported for real videos on NVIDIA A100 GPUs (Appx. A). We also score model-free reference maps with the same code (random noise, a central Gaussian prior, and temporal, spatial, and full oracles; Appx. Table 20) and two model-free *input-energy* maps that test if a score merely tracks the pixel energy the cue adds.

Analysis. On real videos, following an analysis plan fixed before the runs, we compare the four common methods within each dataset–architecture cell, on the clips all four explained, with Holm-corrected paired Wilcoxon tests and effect sizes (Appx. B); nothing is pooled across datasets or architectures. Ranking transfer is the Kendall τ between the method rankings of two cells. Because we cannot verify that the evaluated clips were unseen during training, these results are exploratory. **In the controlled setting**, the source video is the unit of inference, because a source's variants and checkpoints are correlated. For each source and checkpoint, we rank the methods by a faithfulness score (insertion AUC, or negated deletion AUC) and by a localization score, and compute Kendall τ_b between the two rankings; we average over checkpoints and resample sources 10,000 times for 95% intervals. A negative τ means that on the same source the more faithful method localizes worse. Because this estimand depends on which methods are compared (the roster), we also report roster ablations and a paired cross-architecture contrast; analyses added after the protocol was frozen are marked as such in Appx. C.

5 Results

5.1 Comparing methods on real videos

On real videos, insertion and deletion favor different methods, and rankings transfer only partly across datasets and architectures. Among the four common methods (Table 1), Occlusion has the best insertion AUC in 10 of the 15 dataset–architecture cells and SmoothGrad-IG the best deletion AUC in 9, yet each is among the two worst methods on the other metric in 11 and 9 cells. Within a cell, the two metrics therefore tend to order the methods oppositely (mean Kendall $\tau = -0.31$).

Table 1: Real-video insertion AUC (↑) and deletion AUC (↓) for 4 methods run on every architecture, over the n clips all four explained; compare within a row only. Bold: best mean; *: beats every other method (paired Wilcoxon, Holm over 6 pairs, $p < 0.05$). All methods: Appx. Tables 9, 10, and 11.

			Insertion AUC ↑				Deletion AUC ↓			
Dataset	Architecture	n	EP	IG	Occ.	SG-IG	EP	IG	Occ.	SG-IG
Hockey Fight	R(2+1)D	148	.740	.703	**.906***	.689	.631	**.581**	.726	.583
	SlowFast	148	**.860**	.821	.857	.830	.651	.559	.752	**.556**
	X3D-S	148	.803	.706	**.897***	.722	.635	.638	.763	**.619**
	TimeSformer	148	**.948***	.856	.916	.871	.941	.691	.926	**.678**
	VideoMAE	144	**.948**	.926	.944	.935	.892	.766	.879	**.740**
RLVS	R(2+1)D	299	.894	.653	**.917***	.667	.740	.574	.590	**.568**
	SlowFast	299	.930	.868	**.963***	.886	.782	.606	.721	**.591***
	X3D-S	299	.871	.905	**.971***	.920	.725	.659	.836	**.636***
	TimeSformer	296	.951	.928	**.971***	.932	.856	.747	.747	**.737**
	VideoMAE	270	.810	.893	.900	**.904**	.846	.725	.890	**.697***
RWF-2000	R(2+1)D	95	.695	.866	.846	**.871***	**.485***	.734	.647	.741
	SlowFast	95	.569	.877	**.910***	.878	**.490***	.738	.799	.697
	X3D-S	95	.583	.937	**.976***	.946	**.505***	.778	.867	.777
	TimeSformer	92	.980	.909	**.986***	.915	.892	.887	**.883**	.884
	VideoMAE	95	.960	.968	**.978**	.971	.931	**.891**	.895	.896

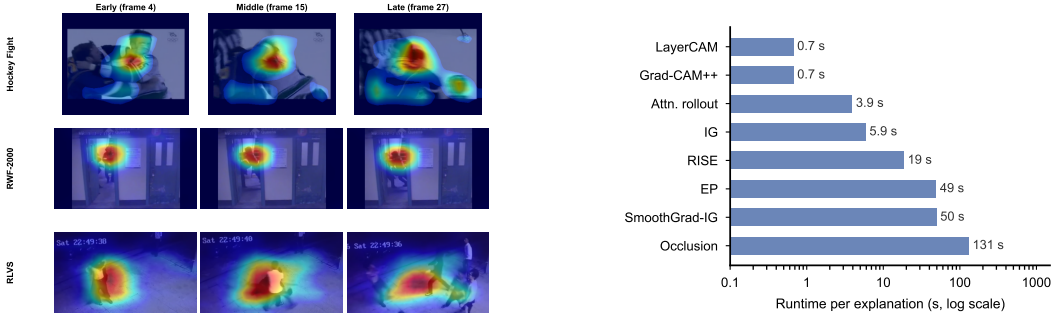
Only the insertion leader is reliable: Occlusion's insertion lead is statistically distinguishable from every other method (Holm-corrected) in 9 of its 10 cells, but SmoothGrad-IG's deletion lead only

in 3 of its 9, so deletion has no reliable leader on real videos. Even the insertion leads are small, because insertion is near its ceiling (per-clip effect sizes in Appx. Table 7).

Rankings transfer only partly (Appx. Table 8). For insertion, the dataset matters more than the architecture: rankings of one architecture on two datasets agree at $\tau = 0.33$(1 = identical order), of two architectures on one dataset at 0.44–0.56; deletion rankings agree at 0.56 among the convolutional networks but only 0.19 between a convolutional network and a transformer. Temporal consistency ranks most stably (0.73) but largely tracks temporal resolution: attention rollout applies one spatial map to every frame and scores exactly 1 on 518 of 542 clips.

Runtime spans more than three orders of magnitude across architecture–method pairs (0.4 s to 701 s per explanation; Appx. Table 12) and is unrelated to insertion AUC (Spearman $\rho = 0.04$; Figure 3b). 99.0% of attempted explanations succeeded (Appx. B).

Grad-CAM++ is a cheap, strong insertion baseline on the convolutional networks: it has the highest insertion AUC of all methods in 3 of 9 cells and is within 0.04 of the best common method in every cell at a small fraction of its runtime (Figure 3b), though it is only mid-ranked on deletion (Appx. Tables 9 and 10). Its maps are coarse and nearly constant in time, which insertion rewards; in the controlled setting, it depends on the model yet localizes the cue at chance in time (chance-adjusted TLS $+0.07$; Table 2).



(a) (b)

Figure 3: Real videos. (a) Grad-CAM++ overlays from R(2+1)D on one video per dataset, at frames 4, 15, and 27 of the 32-frame clip. (b) Typical runtime per method: geometric mean of its median runtime across the architectures it runs on (log scale).

These are observations about one checkpoint per dataset and architecture, not a ranking of methods or a causal claim about architecture. The maps look plausible (Figure 3a), but real videos cannot say which ranking, or which map, is right, which is why we turn to known evidence.

5.2 Faithfulness versus localization of known evidence

On R(2+1)D, methods with better insertion AUC localize the evidence worse in time; on TimeSformer, the two are unrelated on methods both architectures share (Figure 4a–b). For R(2+1)D, within-source relation between insertion AUC and TLS is $\tau_b = -0.58$ (95% interval $[-0.59, -0.56]$, 71 test sources): on same video, the method with the higher insertion score tends to place less of its saliency in the cue window. Ranked by mean scores, the methods show the same negative sign in all six combinations of training seed and data split (Appx. Table 23). The disagreement concerns timing: insertion agrees positively with ST-RMA ($+0.38$), which also rewards location.

The divergence depends on which methods are evaluated (Figure 4c). It holds without IG and SmoothGrad-IG, which fail randomization, without Occlusion, and on the five methods shared with TimeSformer, and it vanishes only when all three are removed. It is therefore carried by the IG family and by Occlusion, whose timing is at chance by construction (it masks all frames at once).

TimeSformer shows no such divergence (Figure 4c): its full roster gives a weak positive relation that does not replicate across training seeds and data splits and disappears on the five shared methods ($+0.02$, 95% interval $[-0.07, +0.10]$). On the same sources and methods, the relation is 0.56 lower for R(2+1)D than for TimeSformer (paired contrast; intervals in Appx. Tables 21 and 22). On the shared methods, higher insertion thus indicates better timing on neither architecture; only the negative direction is specific to R(2+1)D.

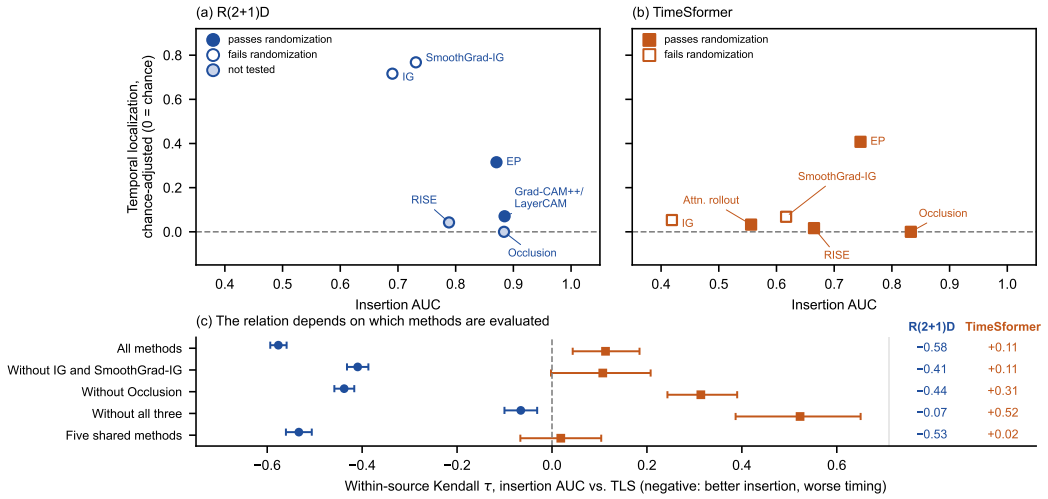


Figure 4: Faithfulness vs. timing in controlled setting (test split). (a, b) Mean insertion AUC and chance-adjusted TLS per method; hollow: fails parameter randomization, light: not tested (Table 2; Appx. Tables 18, 19). (c) Within-source Kendall τ_b between insertion AUC, TLS for each method set, with 95% source-resampling intervals.

A likely reason is that insertion cannot see timing: it is already at its ceiling (≈ 0.89) for a map that knows only the timing, only the location, or both (Appx. Table 20), so it rewards recovering the evidence anywhere in the clip; this is how Occlusion and the CAM methods reach near-ceiling insertion without selective timing. Deletion cannot tell a good map from a bad one either: random noise scores as well on deletion AUC as the full oracle (0.068 vs. 0.063). We do not show that the divergence transfers to natural footage or why the architectures differ.

Table 2: Validity. Rand.: Spearman correlation, saliency pre/post full randomization, on 20 validation clips (10 for the slowest TimeSformer methods); near 0: model-dependent, near 1: not. TLS (chance-adjusted, 0=chance) and Pointing (hit rate, chance 0.19), test split. Attention rollout becomes constant under enough randomization, scoring 0. "–": not run.

	R(2+1)D-18			TimeSformer		
Method	Rand.	TLS	Pointing	Rand.	TLS	Pointing
Grad-CAM++	−0.09 (pass)	+0.07	0.75		n/a	
LayerCAM	−0.14 (pass)	+0.07	0.74		n/a	
IG	0.75 (fail)	+0.72	0.09	0.91 (fail)	+0.05	0.00
SmoothGrad-IG	0.65 (fail)	+0.77	0.15	0.60 (fail)	+0.07	0.00
Occlusion	–	0.00	0.00	−0.08 (pass)	0.00	0.00
RISE	–	+0.04	0.62	0.05 (pass)	+0.02	0.53
Extremal Pert. (EP)	0.09 (pass)	+0.32	0.83	0.05 (pass)	+0.41	0.67
Attention rollout		n/a		0.00 (pass)	+0.03	0.19

5.3 Explanation validity

Methods differ sharply in if they depend on the learned model (Table 2): IG and SmoothGrad-IG fail on both architectures, keeping a correlation of 0.60–0.91 with their trained-model saliency after full randomization; every other method tested passes.

Passing does not guarantee localizing the cue: on R(2+1)D, Grad-CAM++ and LayerCAM pass but localize it at chance in time (chance-adjusted TLS $+0.07$), while the failing IG family has the highest raw TLS. Only EP passes randomization and localizes clearly above chance (TLS $+0.32$ and $+0.41$, Pointing Game ≥ 0.67) on both architectures. The IG family's high TLS is not a pixel-energy artifact, since the input-energy maps score at chance on timing (raw TLS 0.750–0.757), nor a fixed temporal bias, since on a checkpoint trained with shuffled labels its raw TLS falls to chance (IG $0.95 \rightarrow 0.75$, SmoothGrad-IG $0.98 \rightarrow 0.73$). Its maps thus mix input structure, which survives weight randomization, with timing learned from the labels; if IG's post-processing in this setting (§ 4) contributes to the input dependence is untested.

5.4 HUMAN-MARKED EVIDENCE ON REAL VIDEOS

The controlled setting supplies evidence by construction; on real videos, the only available reference is human judgment. We asked 18 participants to judge if five RWF-2000 clips (C01–C05) showed physical aggression and to mark the frames supporting their judgment; each clip was shown as 32 numbered frames over 16 seconds (Appx. D). The question is if people agree closely enough for their selections to serve as a temporal window like the cue's. All 18 saw aggression in C02 and C04, yet their selections overlapped only partly (mean IoU 0.541 and 0.739; Table 3): in C02 at least half selected frames 10–32, but at 75% the shared range narrowed to 21–32. C03 divided judgments about aggression itself (43.1% agreement despite its non-fight label), and no frame reached 50% in C01 or C03. Human-marked evidence on real videos is thus a window with soft edges, not the exact one the cue provides; scoring explanations against these windows is the next step.

Table 3: Judgment and evidence agreement. Agreement: % of pairs giving the same answer; IoU: nonempty selections only (C01: 0 pairs; C03: 3); 50%/75% sets: frames selected by at least that share; "none": no frame reached the threshold (one marked evidence in C01, three in C03).

		Judgment		Aggression evidence		
Clip	Label	Visible (Tot. 18)	Agreement %	Mean IoU	50% set	75% set
C01	non-fight	2	79.1	n/a	none	none
C02	fight	18	100.0	0.541	10–32	21–32
C03	non-fight	5	43.1	0.417	none	none
C04	fight	18	100.0	0.739	8–32	12–32
C05	fight	17	88.9	0.834	6–32	8–32

6 DISCUSSION

The results teach three lessons: (1) a ranking from one dataset, architecture, or metric should not be carried to another (§ 5.1), (2) faithfulness scores do not measure timing, and (3) insertion rewards recovering the evidence anywhere in the clip, so a map can reach near-ceiling insertion while localizing at chance in time, as Grad-CAM++ does on R(2+1)D; deletion has no reliable leader on real videos and cannot separate a random map from the oracle in the controlled setting. Neither score should therefore support a claim about when evidence occurs. Third, localization counts only if the explanation depends on the model: the IG family localizes best in time on R(2+1)D yet fails parameter randomization. On real videos, where evidence must come from people, it will be a window with soft edges, and localization scores there will need to allow for that.

7 LIMITATIONS

The controlled evaluation uses one synthetic cue on Hockey non-fight videos and two architectures, and its results are conditional on the method roster; task-validity and explanation-validity checks are necessary but not sufficient evidence that the cue is used. The real-video comparison is descriptive: one checkpoint per dataset and architecture, and test partitions whose independence from training cannot be fully verified. Perturbation metrics may introduce distribution shift, and we do not test if explanations help human auditors, so the findings describe this setting, not a leaderboard or a deployment test.

8 CONCLUSION

We introduced Video-XAI Bench to ask if the scores used to evaluate video explanations track the evidence a model uses. On real videos, insertion and deletion favored different methods, deletion had no reliable leader, and rankings transferred only partly across datasets and architectures. In a controlled setting with known evidence, higher insertion indicated better timing on neither architecture's shared methods, and on R(2+1)D-18 the two diverged across all training seeds and data splits ($\tau = -0.58$, its size depends on the method roster), and the maps that localized best in time also failed parameter randomization. Video explanations should thus be judged on faithfulness, localization, and validity separately, against chance and oracle references, not by one score; more cues, architectures, and human-auditor studies come next.

AI USE STATEMENT

Generative AI tools were used to assist with parts of this work, including code, analysis, and writing. The authors reviewed and verified all AI-assisted content and take full responsibility for the paper.

ETHICS STATEMENT

The human study was reviewed by our institution's institutional review board and approved.This work studies explanation methods for video violence classifiers and may inherit biases, privacy concerns, and representational limitations from the underlying datasets. The controlled benchmark uses existing non-fight source videos and synthetic overlays; it does not establish that explanations are safe or reliable for operational surveillance decisions. The three datasets are publicly available research datasets, obtained from their public releases and used under their published terms; we do not redistribute any video, and RWF-2000 in particular restricts commercial use and redistribution. Figure 3 and Appx. Figure 5 show a few frames with explanation overlays for illustration only. Participants confirmed eligibility and consent, could skip clips or stop, and personally identifying information was excluded from analysis.

REPRODUCIBILITY STATEMENT

§ 4 and Appendices A–C specify the datasets, preprocessing, models, seeds, explanation configurations, metrics, statistical procedures, exclusions, failures, and the software, hardware, and runtime protocol, and state which analyses were added after the protocol was frozen. Tables and figures are generated by scripts from stored result files, and every real-video explanation is stored as a per-clip record with SHA-256 hashes of its checkpoint and source video. Link to anonymized repository with code, configurations, and result files is here: https://anonymous.4open.science/r/video-xai-bench-iclr2027-A1BA/ Appx. D documents the human-study instructions, response coding, agreement measures, and sensitivity checks.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of post hoc model explanations. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. doi: 10.1016/j.inffus.2021.11.008.

Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. In *International Conference on Computer Analysis of Images and Patterns*, volume 6855 of *Lecture Notes in Computer Science*, pp. 332–339, 2011. doi: 10.1007/978-3-642-23678-5_39.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824, 2021.

Cecilia Bolaños, Leonardo Pepino, Martin Meza, and Luciana Ferrer. Benchmarking time-localized explanations for audio classification models. In *Proceedings of Interspeech*, pp. 211–215, 2025. doi: 10.21437/Interspeech.2025-288.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 839–847, 2018. doi: 10.1109/WACV.2018.00097.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021. doi: 10.1109/CVPR46437.2021.00084.

Ming Cheng, Kunjing Cai, and Ming Li. RWF-2000: An open large scale video database for violence detection. In *International Conference on Pattern Recognition (ICPR 2020)*, pp. 4183–4190, 2021. doi: 10.1109/ICPR48806.2021.9412502.

Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in neural information processing systems*, 35:2832–2845, 2022.

Jonathan Crabbé and Mihaela van der Schaar. Explaining time series predictions with dynamic masks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2166–2177, 2021.

Hazel Doughty, Fida Mohammad Thoker, and Cees G. M. Snoek. LocoMotion: Learning motion-focused video-language representations. In *Proceedings of the Asian Conference on Computer Vision*, volume 15474 of *Lecture Notes in Computer Science*, pp. 3–24, 2024. doi: 10.1007/978-981-96-0908-6_1.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024.

Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 200–210, 2020. doi: 10.1109/CVPR42600.2020.00028.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6201–6210, 2019. doi: 10.1109/ICCV.2019.00630.

Steven Feldstein. The global expansion of AI surveillance. Working paper, Carnegie Endowment for International Peace, 2019.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2950–2958, 2019. doi: 10.1109/ICCV.2019.00304.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5540–5552, 2020.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3958–3968, 2023. doi: 10.1109/ICCV51070.2023.00368.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901. doi: 10.5169/seals-266450.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. LayerCAM: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. doi: 10.1109/TIP.2021.3089943.

Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Sanity simulations for saliency methods. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11173–11200, 2022a.

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, volume 13672 of *Lecture Notes in Computer Science*, pp. 280–298. Springer, 2022b. doi: 10.1007/978-3-031-19775-8_17.

Matthew Kowal, Achal Dave, Rares Ambrus, Adrien Gaidon, Konstantinos G. Derpanis, and Pavel Tokmakov. Understanding video transformers via universal concept discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10946–10956, 2024. doi: 10.1109/CVPR52733.2024.01041.

Jongseo Lee, Wooil Lee, Gyeong-Moon Park, Seong Tae Kim, and Jinwoo Choi. PCBEAR: Pose concept bottleneck for explainable action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2681–2690, 2025. doi: 10.1109/CVPRW67362.2025.00253.

Haoxin Li, Yuan Liu, Hanwang Zhang, and Boyang Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19854–19866. IEEE, 2023a. doi: 10.1109/ICCV51070.2023.01823.

Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. M4: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023b.

Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1119–1128, 2021. doi: 10.1109/WACV48630.2021.00116.

Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s):1–42, 2023. doi: 10.1145/3583558.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436, 2021.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18770–18795, 2022.

Avinab Saha, Shashank Gupta, Sravan Kumar Ankireddy, Karl Chahine, and Joydeep Ghosh. Exploring explainability in video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 8176–8181, 2024.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. arXiv:1706.03825, 2017.

Mohamed Mostafa Soliman, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. Violence recognition from videos using deep learning techniques. In *IEEE International Conference on Intelligent Computing and Information Systems*, pp. 80–85, 2019. doi: 10.1109/ICICIS46948.2019.9014714.

Suraj Srinivas and François Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. In *International Conference on Learning Representations*, 2021.

Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Remco C. Veltkamp, and Ronald Poppe. Saliency tubes: Visual explanations for spatio-temporal convolutions. In *IEEE International Conference on Image Processing*, pp. 1830–1834, 2019. doi: 10.1109/ICIP.2019.8803153.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6021–6029, 2020. doi: 10.1609/aaai.v34i04.6064.

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

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018. doi: 10.1109/CVPR.2018.00675.

Zerui Wang and Yan Liu. STAA: Spatio-temporal attention attribution for real-time interpreting transformer-based AI video models. *IEEE Access*, 13:101647–101661, 2025. doi: 10.1109/ACCESS.2025.3575440.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Appx. A gives protocol details shared by the two settings, Appx. B the real-video benchmark, and Appx. C the controlled setting.

A Shared Protocol Details

Explanation generation. Every method returns one saliency volume at input resolution, $T \times 224 \times 224$ with $T = 32$ (16 for VideoMAE on real videos), and every metric is computed on that volume. On real videos the target is the predicted class, including for misclassified clips, so that the explanation describes the decision the model made; in the controlled setting it is the true class, and only correctly classified coherent clips are explained (Appx. C). In the controlled setting, spatial and temporal evidence are evaluated on this one volume: the temporal marginal (mass per frame) against the true interval, the spatial marginal within the interval against the true region, and the full volume against the true space–time box.

Metric details. With input x, target class c, class probability p_c, per-video mean baseline b, and a mask M_k selecting the fraction ρ_k of space–time positions with the highest signed saliency, deletion and insertion evaluate

$$D_k = p_c\big((1 - M_k) \odot x + M_k \odot b\big), \qquad I_k = p_c\big(M_k \odot x + (1 - M_k) \odot b\big). \tag{3}$$

The masks M_k rank space–time positions by signed saliency, with no absolute-value or positive-part transformation, and break ties by stable order. Deletion and insertion AUC integrate D_k and I_k (Equation 3) by the trapezoidal rule. The controlled setting uses ten ratios, $\rho_k \in \{0.1, \ldots, 1.0\}$, omitting the zero-masking endpoint, so its integration interval is 0.9 long; the real-video runs use eleven, $\rho_k \in \{0, 0.1, \ldots, 1.0\}$, over an interval of length 1 whose first point is the unmasked input for deletion and the baseline for insertion. Absolute AUCs are therefore comparable within a setting but not across settings. *Temporal consistency* (TC) is the mean cosine similarity of consecutive saliency frames, each min–max normalized separately so that temporal variation is preserved,

$$\mathrm{TC} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\langle \hat{S}_t, \hat{S}_{t+1} \rangle}{\|\hat{S}_t\|_2 \|\hat{S}_{t+1}\|_2}, \tag{4}$$

where $\hat{S}_t$ is frame t of S, min–max normalized, and a spatially constant frame is set to zero. A map that repeats one spatial pattern in every frame scores exactly 1, so TC reflects a method's temporal resolution as well as its stability. In the controlled setting, a localization score s with chance level s_0 is adjusted as $s_{\mathrm{adj}} = (s - s_0)/(1 - s_0)$ for $s \geq s_0$ and $(s - s_0)/s_0$ otherwise, so it runs from -1 through 0 (chance) to 1 (ceiling); clips with zero total positive saliency mass are scored 0 on ST-RMA, TLS, and SPW rather than excluded.

Software, hardware, and runtime protocol. Real-video explanations and their metrics were computed on a university research-computing cluster with NVIDIA A100 GPUs (40 GB of GPU memory), one GPU per job, using PyTorch 2.7.1 with CUDA 11.8. Controlled-setting experiments ran on a local Windows 11 machine with an Intel Xeon W-2145 CPU at 3.70 GHz, 64 GB of system memory, and an NVIDIA GeForce RTX 5060 Ti GPU with 16 GB of memory, in a Conda environment with Python 3.9.25, PyTorch 2.8.0+cu128, Torchvision 0.23.0+cu128, PyTorchVideo 0.1.5, Captum 0.7.0, Transformers 4.36.2, NumPy 1.26.4, SciPy 1.13.1, and scikit-learn 1.6.1. Runtime is the wall-clock time to generate one explanation and is reported for real videos only; Table 12 gives the median over all explanations of each architecture–method pair.

B Real-Video Benchmark Details

Evaluation sample. The dataset registries contain 1,000 Hockey Fight (700/150/150), 2,000 RLVS (1,400/300/300), and 1,989 RWF-2000 (1,392/298/299) videos (train/validation/test). We evaluate the test partitions after removing clips whose bytes match a training or validation clip, which leaves 148 Hockey Fight, 299 RLVS, and 291 RWF-2000 clips. Several RWF-2000 clips can come from one source video, since their filenames differ only in a numeric suffix; 208 of the 291 share such a source name with a training or validation clip and are removed, and the 12 unique fight clips of the original release that belong to no partition are added. The RWF-2000 sample therefore has 95 clips (76 non-fight, 19 fight) from 87 source groups, only 11 of which contain fight clips, and its intervals resample source groups. Table 4 gives the sample and each checkpoint's accuracy on it, and Table 5 the validation accuracy logged during training. The archived launch scripts specify

pretrained initialization for R(2+1)D, SlowFast, and X3D-S, 20 training epochs for these three and TimeSformer, and at most 40 epochs for VideoMAE; we report these settings as specified, since the runs did not bind them immutably to the checkpoints. Two limits keep this evaluation descriptive and exploratory rather than held-out: the training runs did not record immutable split hashes, so independence from the training data cannot be fully verified, and 25 clips (14 Hockey Fight, 11 RLVS) appeared in earlier explanation experiments of this project. Short videos repeat the last frame, horizontal flips are used only in training, and there is no test-time augmentation.

Table 4: Real-video evaluation sample and checkpoint accuracy on it: accuracy in % with recall on the fight class in parentheses. Explanations target the predicted class, including misclassified clips.

Dataset	Clips	Non-fight/fight	R(2+1)D	SlowFast	X3D-S	TimeSformer	VideoMAE
Hockey Fight	148	73/75	97.3 (95)	97.3 (96)	93.2 (88)	93.9 (91)	95.3 (96)
RLVS	299	149/150	97.3 (96)	98.0 (97)	98.3 (99)	94.6 (99)	98.0 (100)
RWF-2000	95	76/19	97.9 (100)	96.8 (100)	98.9 (100)	92.6 (63)	90.5 (58)

Table 5: Validation accuracy (%) logged during training of each real-video checkpoint. Split hashes were not recorded, so values are reported as logged.

Dataset	R(2+1)D	SlowFast	X3D-S	TimeSformer	VideoMAE
Hockey Fight	97.33	98.67	96.67	95.33	94.67
RLVS	98.00	98.33	99.00	97.67	98.67
RWF-2000	90.54	94.26	93.58	87.50	93.24

Configuration and primary analysis. All real-video explanations use one configuration locked before the runs (Table 6), and every explanation uses a seed derived from a hash of the protocol, checkpoint, clip, and method (base seed 42) with deterministic algorithms enabled; the four common methods have identical parameters on every architecture, whereas the CAM target layers and the EP implementation differ between convolutional networks and transformers. The analysis plan was also fixed before the runs. Within each dataset–architecture cell, the four common methods are compared on the clips that all four explained, by paired Wilcoxon signed-rank tests with Holm correction over the six pairs. Table 7 reports, for the two best methods, the mean per-clip advantage with a 95% bootstrap interval (2,000 resamples) and the share of clips on which the leader wins. With 92–299 clips per cell, small differences reach significance, so we interpret effect sizes rather than p-values; the Holm correction applies within each cell's six pairs, not across the 15 cells, so counts of significant leads are descriptive. Rank agreement (Table 8) is the mean Kendall τ between the four-method rankings of two cells. Nothing is pooled across datasets or architectures.

Table 6: Real-video explanation configuration, locked before the runs. Parameters of the common methods are identical across architectures; architecture-specific choices are stated.

Method	Configuration
Grad-CAM++, LayerCAM	Last convolutional block (R(2+1)D `layer4`, SlowFast slow pathway of block 4, X3D-S block 4); not applicable to the transformers. Grad-CAM++ uses the common closed-form weights from powers of first-order gradients, not exact higher-order derivatives.
IG	Black baseline, 50 steps, mean absolute channel aggregation; no smoothing, clipping, or sharpening.
SmoothGrad-IG	As IG; 10 noise samples, noise level 0.15.
Occlusion	32×32 patch, stride 16, mean baseline, sliding temporal window of 4 frames (temporally resolved, unlike the controlled task).
RISE	1000 masks, probability 0.5, $8 \times 8 \times 8$ grid, random shift, smoothing; convolutional networks only (empty saliency on both transformers).
EP	Preservation variant, 400 iterations, learning rate 0.1, L1 weight 1.0, total-variation weight 0.2, blur baseline $\sigma = 10$; separate convolutional and transformer implementations.
Attention rollout	TimeSformer only; entropy head fusion, residual connections; one spatial map weighted per frame.
Grad.-weighted attention rollout	VideoMAE only; mean head fusion, spatial $\sigma = 8$, temporal $\sigma = 1$ smoothing.

Full results, failures, and runtime. Tables 9, 10, and 11 give every method's insertion AUC, deletion AUC, and temporal consistency per cell, averaged over its successful explanations. Of 16,802 attempted explanations, 170 failed: empty RISE maps (SlowFast 90, X3D-S 32, R(2+1)D 9), empty VideoMAE Occlusion maps (33), and six cluster I/O errors for TimeSformer Occlusion; all other architecture–method cells have none. Failures are recorded, never imputed. Because they depend on the method and architecture (RISE fails on 85 of 299 RLVS clips for SlowFast), these means are conditional on a successful explanation and complement rather than replace the paired analysis, which uses only clips that every compared method explained. Table 12 gives the runtime of each method.

Table 7: Primary paired analysis of the four common methods: best versus second-best method by mean, the mean per-clip advantage of the best method with a 95% bootstrap interval (clips resampled; RWF-2000 resamples filename-derived source groups), the share of clips on which the best method beats the runner-up (ties count half), and the number of the six method pairs that differ after Holm correction.

		Insertion AUC				Deletion AUC			
Dataset	Arch.	Best vs. 2nd	Advantage [95%]	Share	Pairs	Best vs. 2nd	Advantage [95%]	Share	Pairs
Hockey Fight	R(2+1)D	Occ. vs. EP	0.166 [0.130, 0.201]	0.59	6/6	IG vs. SG-IG	0.002 [-0.005, 0.008]	0.52	5/6
	SlowFast	EP vs. Occ.	0.003 [-0.019, 0.025]	0.43	1/6	SG-IG vs. IG	0.003 [-0.003, 0.009]	0.49	5/6
	X3D-S	Occ. vs. EP	0.095 [0.072, 0.119]	0.78	5/6	SG-IG vs. EP	0.016 [-0.033, 0.066]	0.45	4/6
	TimeSformer	EP vs. Occ.	0.032 [0.015, 0.051]	0.56	6/6	SG-IG vs. IG	0.014 [-0.004, 0.034]	0.51	4/6
	VideoMAE	EP vs. Occ.	0.004 [-0.016, 0.024]	0.49	2/6	SG-IG vs. IG	0.026 [-0.016, 0.067]	0.49	4/6
RLVS	R(2+1)D	Occ. vs. EP	0.022 [0.012, 0.034]	0.42	6/6	SG-IG vs. IG	0.006 [0.000, 0.012]	0.54	3/6
	SlowFast	Occ. vs. EP	0.033 [0.023, 0.044]	0.60	6/6	SG-IG vs. IG	0.015 [0.005, 0.024]	0.60	5/6
	X3D-S	Occ. vs. SG-IG	0.051 [0.038, 0.066]	0.84	4/6	SG-IG vs. IG	0.022 [0.017, 0.028]	0.75	6/6
	TimeSformer	Occ. vs. EP	0.020 [0.012, 0.030]	0.93	6/6	SG-IG vs. Occ.	0.010 [-0.037, 0.053]	0.55	4/6
	VideoMAE	SG-IG vs. Occ.	0.004 [-0.008, 0.017]	0.47	5/6	SG-IG vs. IG	0.028 [0.021, 0.034]	0.64	6/6
RWF-2000	R(2+1)D	SG-IG vs. IG	0.005 [-0.000, 0.010]	0.67	6/6	EP vs. Occ.	0.162 [0.093, 0.228]	0.78	6/6
	SlowFast	Occ. vs. SG-IG	0.032 [0.008, 0.057]	0.79	5/6	EP vs. SG-IG	0.207 [0.120, 0.292]	0.74	6/6
	X3D-S	Occ. vs. SG-IG	0.030 [0.012, 0.056]	0.77	6/6	EP vs. SG-IG	0.272 [0.110, 0.429]	0.74	5/6
	TimeSformer	Occ. vs. EP	0.006 [0.001, 0.012]	0.55	6/6	Occ. vs. SG-IG	0.001 [-0.039, 0.045]	0.20	5/6
	VideoMAE	Occ. vs. SG-IG	0.007 [-0.010, 0.024]	0.58	1/6	IG vs. Occ.	0.005 [-0.000, 0.011]	0.53	0/6

Table 8: Mean Kendall τ between the rankings of the four common methods on real videos (1 = identical order, 0 = unrelated), over pairs of dataset–architecture cells; deletion is ranked with lower as better. Brackets: 95% intervals from resampling clips within each cell (1,000 draws). They reflect clip sampling only, not the choice of datasets and architectures, and because τ over four methods takes few values an interval can start at the point estimate. With one transformer pair per dataset, the transformer row is not estimable. Within a cell, insertion and (negated) deletion rank the methods at mean $\tau = -0.31$.

Rankings compared	Pairs	Insertion	Deletion	Temporal consistency
Same architecture, different datasets	15	0.33 [0.20, 0.42]	0.29 [0.11, 0.42]	0.73 [0.73, 0.78]
Same dataset, two convolutional networks	9	0.56 [0.48, 0.63]	0.56 [0.41, 0.70]	0.70 [0.70, 0.78]
Same dataset, two transformers	3	0.44 [0.00, 0.67]	0.56 [0.00, 0.78]	0.33 [0.33, 0.56]
Same dataset, convolutional vs. transformer	18	0.54 [0.31, 0.57]	0.19 [0.13, 0.48]	0.41 [0.41, 0.48]

Table 9: Insertion AUC (higher is better) for every real-video method, averaged over the valid explanations of each cell ("–": method not run on that architecture). LC: LayerCAM; AR: attention rollout; GAR: gradient-weighted attention rollout.

Dataset	Arch.	GC++	LC	IG	SG-IG	Occ.	RISE	EP	AR	GAR
Hockey Fight	R(2+1)D	.886	.884	.703	.689	.906	.650	.740	–	–
	SlowFast	.969	.969	.821	.830	.857	.709	.860	–	–
	X3D-S	.927	.926	.706	.722	.897	.844	.803	–	–
	TimeSformer	–	–	.856	.871	.916	–	.948	.928	–
	VideoMAE	–	–	.926	.936	.944	–	.948	–	.928
RLVS	R(2+1)D	.910	.907	.653	.667	.917	.873	.894	–	–
	SlowFast	.960	.958	.868	.886	.963	.908	.930	–	–
	X3D-S	.964	.960	.905	.920	.971	.899	.871	–	–
	TimeSformer	–	–	.926	.930	.971	–	.951	.879	–
	VideoMAE	–	–	.896	.906	.900	–	.804	–	.803
RWF-2000	R(2+1)D	.859	.856	.866	.871	.846	.744	.695	–	–
	SlowFast	.966	.964	.877	.878	.910	.600	.569	–	–
	X3D-S	.939	.951	.937	.946	.976	.901	.583	–	–
	TimeSformer	–	–	.910	.916	.986	–	.980	.934	–
	VideoMAE	–	–	.968	.971	.978	–	.960	–	.886

Table 10: Deletion AUC (lower is better) for every real-video method, averaged over the valid explanations of each cell ("–": method not run on that architecture). LC: LayerCAM; AR: attention rollout; GAR: gradient-weighted attention rollout.

Dataset	Arch.	GC++	LC	IG	SG-IG	Occ.	RISE	EP	AR	GAR
Hockey Fight	R(2+1)D	.589	.591	.581	.583	.726	.584	.631	–	–
	SlowFast	.762	.767	.559	.556	.752	.630	.651	–	–
	X3D-S	.643	.653	.638	.619	.763	.662	.635	–	–
	TimeSformer	–	–	.691	.678	.926	–	.941	.934	–
	VideoMAE	–	–	.764	.742	.879	–	.894	–	.760
RLVS	R(2+1)D	.633	.638	.574	.568	.590	.649	.740	–	–
	SlowFast	.647	.650	.606	.591	.721	.732	.782	–	–
	X3D-S	.776	.751	.659	.636	.836	.827	.725	–	–
	TimeSformer	–	–	.747	.737	.747	–	.857	.845	–
	VideoMAE	–	–	.735	.709	.890	–	.844	–	.805
RWF-2000	R(2+1)D	.649	.652	.734	.741	.647	.515	.485	–	–
	SlowFast	.747	.752	.738	.697	.799	.446	.490	–	–
	X3D-S	.761	.753	.778	.777	.867	.818	.505	–	–
	TimeSformer	–	–	.870	.868	.883	–	.881	.930	–
	VideoMAE	–	–	.891	.896	.895	–	.931	–	.775

Table 11: Temporal consistency (no better direction; it largely reflects each method's temporal resolution) for every real-video method, averaged over the valid explanations of each cell ("–": method not run on that architecture). LC: LayerCAM; AR: attention rollout; GAR: gradient-weighted attention rollout.

Dataset	Arch.	GC++	LC	IG	SG-IG	Occ.	RISE	EP	AR	GAR
Hockey Fight	R(2+1)D	1.000	1.000	.645	.701	.987	.994	.643	–	–
	SlowFast	.992	.992	.588	.593	.974	.959	.847	–	–
	X3D-S	.988	.986	.646	.663	.947	.994	.912	–	–
	TimeSformer	–	–	.694	.787	.962	–	.587	1.000	–
	VideoMAE	–	–	.592	.591	.846	–	.732	–	.942
RLVS	R(2+1)D	1.000	1.000	.609	.613	.974	.975	.691	–	–
	SlowFast	.992	.992	.542	.548	.968	.886	.749	–	–
	X3D-S	.978	.963	.597	.594	.932	.874	.836	–	–
	TimeSformer	–	–	.603	.626	.964	–	.485	1.000	–
	VideoMAE	–	–	.684	.684	.734	–	.654	–	.911
RWF-2000	R(2+1)D	1.000	1.000	.715	.742	.964	.995	.783	–	–
	SlowFast	.990	.990	.613	.620	.971	.991	.899	–	–
	X3D-S	.977	.952	.702	.691	.953	.848	.901	–	–
	TimeSformer	–	–	.765	.821	.970	–	.674	1.000	–
	VideoMAE	–	–	.717	.730	.936	–	.743	–	.920

Table 12: Median wall-clock time in seconds to generate one real-video explanation (over all clips and datasets), on one NVIDIA A100 40 GB GPU per task. VideoMAE takes 16 input frames, the other architectures 32.

Method	R(2+1)D	SlowFast	X3D-S	TimeSformer	VideoMAE
Grad-CAM++	0.9	0.5	0.8	–	–
LayerCAM	0.9	0.4	0.8	–	–
Grad.-weighted attention rollout	–	–	–	–	1.1
Attention rollout	–	–	–	3.9	–
IG	5.7	2.6	2.8	25.8	6.7
RISE	30.4	15.1	13.9	–	–
EP	50.9	18.8	34.2	181.6	47.5
SmoothGrad-IG	49.3	22.1	20.6	228.0	58.4
Occlusion	147.0	67.9	64.2	700.6	85.1

C CONTROLLED-SETTING DETAILS

C.1 CONSTRUCTION AND TRAINING

Cue construction and audits. The cue is injected into 32-frame Hockey non-fight clips (224×224); a 112×112 quadrant is chosen uniformly per source. The cue is a signed, spatially tapered grating of vertical stripes that drifts horizontally, whose phase advances every frame, active in frames 4–27 (24 frames), leaving four unmodified frames on each side. For the k-th active frame ($k = 0, \ldots, 23$) the carrier is $r_k(x) = \sin(2\pi x/\lambda - \phi_0 - k\omega)$ with $\lambda = 28$ px (four cycles across a quadrant) and $\omega = \pi/4$ rad per frame (3.5 px of apparent displacement per frame, three phase revolutions over the window); the initial phase ϕ_0 is sampled per source. With $w(x, y)$ a two-dimensional Hann window over the quadrant and $\bar{r}_k = \sum w\, r_k / \sum w$, the unit-magnitude pattern is

$$g_k(x, y) = \frac{w(x, y)\,\big(r_k(x) - \bar{r}_k\big)}{\max_{x,y} \big|w(x, y)\,\big(r_k(x) - \bar{r}_k\big)\big|}, \tag{5}$$

which sums to zero over the quadrant, so the cue adds no net brightness, and the taper removes a hard rectangular edge. With δ_{base} the mean absolute adjacent-frame pixel difference inside the selected quadrant of the unmodified clip, the injection is

$$V_t'(x, y, c) = \text{clip}\big(V_t(x, y, c) + \alpha\, g_{t-4}(x, y),\, 0,\, 1\big), \qquad \alpha = 3\,\delta_{\text{base}}, \tag{6}$$

inside the quadrant for $t \in [4, 27]$ and $V_t' = V_t$ exactly elsewhere. The factor 3 and the hard clamp are project choices; some pixels saturate, so clipping and realized cue energy are audited per clip. Every source yields four matched variants: *coherent* (phases in order), *static-phase* (one phase repeated), *shuffled-phase* (the same phases in permuted order), and *cue-absent*. The classifier solves a binary *motion-coherence* task (coherent = 1; both controls = 0); the cue-absent variant serves construction audits only, so that an unmodified-versus-coherent comparison, which needs no temporal integration, is never read as evidence of order dependence. Each source contributes one positive and two negative examples (900/225/213 training/validation/test clips), and a class-balanced sampler corrects the imbalance; explanations use the coherent clip of each correctly classified source. We use 300 training, 75 validation, and 71 test sources; all variants of a source stay in one split, and the source video is the unit of independence throughout. Four official test sources with earlier controlled-setting materialization were excluded before the test build, and a 66-source subset additionally excludes five sources that appeared in earlier project artifacts (a positive-evidence scan, so we call the test split previously unexposed, not pristine). Audits verify zero cross-split overlap, exact out-of-window identity, mask reconstruction, quadrant balance, clipping, energy, hashes, and nuisance confounds.

Controlled-setting models. The controlled setting fine-tunes R(2+1)D-18 (Kinetics-400, full fine-tuning, 15 epochs, AdamW 10^{-4}, batch 4) and TimeSformer (Kinetics-400 base; AdamW 2×10^{-5} with warmup, batch 1 with accumulation 4–8, bf16, non-reentrant gradient checkpointing, which was required because the reentrant mode gave unstable gradients), with seeds 42, 123, and 7 plus one label-randomized checkpoint each. Model selection is the best validation accuracy. Validation accuracy is 100% for R(2+1)D and 99.6/97.3/99.6% for TimeSformer; the randomized TimeSformer never exceeds 63.1%. Table 13 summarizes the locked cue and training specification, and Table 14 the support for each design choice.

Table 13: Locked controlled-setting cue and R(2+1)D training specification.

Component	Setting
Clip geometry	$32 \times 224 \times 224$; one 112×112 quadrant
Cue support	frames 4–27 (24 active; 4-frame margins)
Grating	$\lambda = 28$ px; $\omega = \pi/4$ rad/frame (3.5 px/frame)
Boundary treatment	2D Hann taper
Amplitude	$\alpha = 3\delta_{\text{base}}$; hard clamp to $[0, 1]$
Classifier	Kinetics-initialized R(2+1)D-18; full fine-tuning
Optimization	15 epochs; AdamW; 10^{-4}; batch 4
Model selection	Best validation accuracy
Seeds	42, 123, 7

C.2 TASK-VALIDITY GATES AND CUE SEARCH

Task-validity gates. Table 15 lists the eight gates in full. Two carry limits: the single-frame gate (C5) shows non-decodability under the tested probe, not that the grating is invisible in a single

Table 14: Support for the controlled-setting design choices. **L**: motivated by literature (not necessarily this exact implementation); **M**: methodological or statistical reasoning; **P**: project-specific proposal evaluated empirically.

Choice	Type	Basis
Synthetic exact ground truth	L	Arras et al. (2022); Hesse et al. (2023); Nauta et al. (2023); does not prescribe this video cue
Parameter and label randomization checks	L	Adebayo et al. (2018)
Static-phase and shuffled-phase controls; single-frame gate	L, P	Static-bias findings (Li et al., 2023a); exact construction is ours
Hockey non-fight substrate	M, P	Cue must be the sole salient signal; homogeneous scenes
Four-quadrant geometry	P	Coarse CAM/mask resolution; simple chance anchors
Phase grating, $\lambda = 28$ px, $\omega = \pi/4$	P	Controlled-motion principle (Doughty et al., 2024); equation is ours
24-frame window; four-frame margin	P	Design search (Table 16); not shown minimal
Amplitude $\alpha = 3\delta_{\text{base}}$; Hann taper; hard clamp	P, M	Relative scaling and edge-artifact avoidance; clamp saturates some pixels
Gates C1–C8 and thresholds (e.g. 0.10 order effect)	P	Fixed before the corresponding outcomes, except C4
Source video as independent unit	M	Matched variants and checkpoints are correlated
Chance levels (TLS 0.75, ST-RMA 0.1875, SPW 0.25, Pointing 0.1875)	M	Analytic for this geometry; not universal constants

frame, and the window-specificity gate (C4) was formalized after two earlier operationalizations proved ill-posed for this geometry, so we present it as a documented post-hoc diagnostic.

Table 15: Task-validity gates (R(2+1)D results). For TimeSformer, C1, C4, and C8 also pass (true-support masking $99.6\% \rightarrow 41.8\%$ versus 92.4% for the wrong quadrant; at least 98.7% accuracy in every energy tercile); C2 and C3 follow from training, and C5–C7 test the dataset and are unchanged.

Gate	Tests	Result
C1	Targeted masking (P2, P4)	Masking the true support: accuracy $1.000 \rightarrow 0.667$ (majority floor); wrong quadrant: unchanged.
C2	Matched controls (P1, P5)	100% validation accuracy within every control type, all seeds.
C3	Label randomization	Validation collapses to the 0.667 majority floor (0% minority accuracy).
C4	Window specificity	Preserving the true window matters more than preserving as many frames elsewhere (floor reached at 16/24 versus 20/24 frames masked); a third operationalization, adopted post hoc.
C5	Single-frame leakage (P3)	Equivalence to chance under the tested probe and margin (two one-sided tests, $p < 0.001$).
C6	Source/nuisance confounds	Nuisance-only balanced accuracy 0.532 (threshold 0.60).
C7	Order sensitivity (P2)	Ordered minus shuffled effect 0.129, 95% interval $[0.107, 0.153]$ (threshold 0.10).
C8	Cue-energy strata	100% accuracy in low, middle, and high energy terciles.

Cue design search. Table 16 lists the candidates, in order, with the balanced-accuracy gain of ordered over shuffled presentation. A powered calibration of short gratings ($n \in \{2, 4, 8\}$ frames) was rejected: $n = 2$ had no order effect, $n = 4$ failed single-frame equivalence, and $n = 8$ gave a tightly estimated but insufficient effect (0.045). A rigid warp gave a null effect at full scale (100 sources), and a non-rigid shear warp, tested only at pilot scale (12 sources per amplitude, 2–24 px), never exceeded 0.083 and was non-monotonic in amplitude. We then extended the grating to 24 frames at a fixed phase step of $\pi/4$, so the coherent condition completes three phase revolutions instead of one; shuffling a longer low-frequency periodic signal decorrelates it more. The 24-frame build passed the gate.

Table 16: Cue candidates and the order-sensitivity gate: balanced-accuracy gain of ordered over shuffled presentation with a frozen probe; passing requires a 95% interval lower bound above 0.10, a threshold fixed before the calibration outcomes.

Candidate	Estimate [95% interval], sources	Outcome
Grating, 2/4/8 frames	-0.001 / failed 2D gate / 0.045 [0.028, 0.064], 352	rejected
Rigid warp (full-scale build)	0.000 [-0.015, 0.015], 100	rejected
Non-rigid shear warp, 2–24 px (pilots)	never above 0.083; 12 sources per amplitude	rejected
Grating, 24 frames	**0.129 [0.107, 0.153], 375**	**passed**

The 24-frame window is the largest the four-frame margins allow, and the earlier calibration protocol had prohibited falling back to the longest duration. The 24-frame build is a new parameterization with its own confirmatory gate, but we did not run a powered comparison of intermediate durations (20 and 24 frames both cleared the gate only at the point estimate in a smaller exploratory build). We therefore do *not* claim that 24 frames is minimal or optimal, nor that the program was preregistered: it is a sequential study of prospectively fixed gates, exploratory design search, post-hoc diagnostics, and corrective analyses.

C.3 Explanation configurations

Methods and compute. In the controlled setting, R(2+1)D uses Grad-CAM++, LayerCAM, IG, SmoothGrad-IG, Occlusion, RISE, and EP, and TimeSformer uses IG, SmoothGrad-IG, Occlusion, RISE, EP, and attention rollout; five methods are shared. Because we evaluate configured pipelines, IG is reported here as IG-Postprocessed (absolute channel aggregation, smoothing, percentile clipping, power sharpening), and SmoothGrad-IG reuses it; the real-video runs use IG without post-processing (Table 6). For compute, EP uses 150 rather than 400 iterations and SmoothGrad-IG 5 rather than 10 samples (a smoke test on R(2+1)D showed differences at noise level; this was not confirmed for TimeSformer EP), and TimeSformer RISE, EP, and SmoothGrad-IG are evaluated on at most 30 clips per checkpoint and split (registry order; seed-42 RISE ran on all clips and its first 30 rows are used). Table 17 lists the configurations.

Table 17: Locked controlled-setting explanation configurations (shared across architectures unless noted).

Method	Configuration
Grad-CAM++, LayerCAM	Final convolutional block (`model.layer4`); R(2+1)D only. Grad-CAM++ uses the common closed-form weights from powers of first-order gradients, not exact higher-order derivatives.
IG-Postprocessed	Black baseline; 50 steps; absolute-mean channel aggregation; spatial Gaussian $\sigma = 6.0$; temporal Gaussian $\sigma = 1.5$; 95th-percentile clipping; global normalization; power 3.5; no hard threshold.
SmoothGrad-IG	Same post-processing as IG; 5 noise samples (the real-video setting uses 10).
Occlusion	32×32 patch; stride 16; mean baseline; all frames masked together.
RISE	1,000 masks; probability 0.5; R(2+1)D: $8 \times 8 \times 8$ grid, batch 32, score centering, random shift, smoothing; TimeSformer: batch 4.
Extremal Perturbation	150 iterations (the real-video setting uses 400); TimeSformer: 16×16 patches.
Attention rollout	TimeSformer only; entropy head fusion, no discard, residual connections.

Inputs use (C, T, H, W) layout.

C.4 Analysis protocol and exclusions

Validity and analysis protocol. *Explanation validity.* Following Adebayo et al. (2018), in the cascading randomization test we re-initialize successive parts of the network from the output toward the input (R(2+1)D: classifier, then `layer4` through `layer1`, then the stem; TimeSformer: classifier, last 3, last 6, all 12 blocks, then patch embeddings), hold the explained class fixed, and compute the Spearman correlation between each method's saliency and its own saliency under the trained model, on 20 validation clips (10 clips and three stages for the slowest TimeSformer methods; Occlusion and RISE were not run for R(2+1)D). A correlation that decays to that of a random map (≈ 0) indicates sensitivity to learned parameters; one that plateaus above it does not (after full randomization, methods marked as passing are at most 0.14 in absolute value and those marked as failing at least 0.60); a constant map is scored 0. In the label-randomization test we explain a checkpoint trained on shuffled labels with the true label as target (that checkpoint predicts the majority class on every cue-present clip, so the correct-only filter would leave nothing) and expect localization to fall to chance. *Controlled-setting analysis.* The source video is the unit of inference; the three checkpoint evaluations of a source are correlated and never treated as independent videos. Kendall's tau between rankings of only 5–7 methods has almost no classical power, so the main analysis works *within* sources, with the estimator defined in § 4. This estimand is conditional on the method roster, so we probe its dependence on the roster by dropping the methods that fail randomization, by restricting to the five shared methods, and by a paired cross-architecture contrast on sources evaluated for both. We complement it with rank-level Kendall correlations in each of the six seed-by-split cells (Appx. C.6), Holm correction within pairwise families, and the 66-source subset. The per-source analysis, the input-energy maps, and a TimeSformer bf16 rerun were added after the protocol was frozen; we report the per-source analysis as the main estimate because rank-level tests over 5–7 methods have almost no power, and for R(2+1)D it agrees in sign with the rank-level results in every cell (Appx. C.6).

Explained clips. Only coherent cue-present clips that the model classifies correctly are explained: all 71 test sources for every R(2+1)D checkpoint, and 68, 56, and 60 for the three TimeSformer checkpoints, whose fp32 evaluation is stricter near the decision boundary than the bf16 evaluation used in training.

C.5 ADDITIONAL RESULTS

Method-level results. Table 18 gives all localization metrics for each R(2+1)D method, with chance-adjusted TLS in parentheses; TimeSformer is in Table 19. Table 22 gives the within-source agreement values, Table 20 the non-learned reference maps, and Figure 5 a qualitative example.

Table 18: R(2+1)D-18, held-out test split: per-source means over three checkpoints (n = sources; Pointing = mean hit rate over checkpoints). ST-RMA: spatiotemporal relevance-mass accuracy; TLS: temporal localization score; SPW: spatial precision within the window; Pointing: Pointing Game hit rate. Chance: ST-RMA 0.1875, TLS 0.75, SPW 0.25, Pointing 0.1875; parentheses give the chance-adjusted TLS (0 = chance, 1 = ceiling). ‡ fails the model-parameter randomization check.

Method	n	ST-RMA	TLS (adj.)	SPW	Pointing	Del. AUC ↓	Ins. AUC ↑
Grad-CAM++	71	0.551	0.768 (+0.07)	0.702	0.75	0.425	0.885
LayerCAM	71	0.548	0.767 (+0.07)	0.699	0.74	0.425	0.885
IG‡	71	0.408	0.929 (+0.72)	0.439	0.09	0.498	0.690
SmoothGrad-IG‡	71	0.272	0.942 (+0.77)	0.288	0.15	0.538	0.731
Occlusion	71	0.597	0.750 (-0.00)	0.796	0.00	0.420	0.884
RISE	71	0.206	0.761 (+0.04)	0.270	0.62	0.432	0.788
Extremal Pert. (EP)	71	0.430	0.829 (+0.32)	0.515	0.83	0.428	0.871

Table 19: TimeSformer, held-out test split: per-source means over checkpoints. n varies because RISE, EP and SmoothGrad-IG were capped at 30 clips per seed and correctly classified clips differ by seed. ST-RMA: spatiotemporal relevance-mass accuracy; TLS: temporal localization score; SPW: spatial precision within the window; Pointing: Pointing Game hit rate. Chance: ST-RMA 0.1875, TLS 0.75, SPW 0.25, Pointing 0.1875; parentheses give the chance-adjusted TLS (0 = chance, 1 = ceiling). ‡ fails the model-parameter randomization check.

Method	n	ST-RMA	TLS (adj.)	SPW	Pointing	Del. AUC ↓	Ins. AUC ↑
IG‡	68	0.227	0.763 (+0.05)	0.298	0.00	0.368	0.419
SmoothGrad-IG‡	27	0.236	0.767 (+0.07)	0.307	0.00	0.485	0.617
Occlusion	68	0.599	0.750 (+0.00)	0.799	0.00	0.251	0.833
RISE	30	0.206	0.754 (+0.02)	0.274	0.53	0.517	0.665
Extremal Pert. (EP)	27	0.457	0.852 (+0.41)	0.533	0.67	0.553	0.746
Attn. Rollout	68	0.228	0.758 (+0.03)	0.302	0.19	0.392	0.556

Table 20: Non-learned reference maps (R(2+1)D seed-42 checkpoint, validation split, 75 clips, scored by the same code as the explanation methods). The full oracle attains exactly 1 on every mass metric, confirming the metric pipeline. Insertion AUC is already at its ceiling (about 0.89) for a map that knows only the location and is uniform over time.

Reference map	ST-RMA	TLS	SPW	Pointing	Del. AUC	Ins. AUC
Random (IID noise)	0.188	0.750	0.250	0.20	0.068	0.114
Center prior	0.188	0.750	0.250	0.00	0.674	0.380
Temporal oracle	0.250	1.000	0.250	0.24	0.055	0.889
Spatial oracle	0.750	0.750	1.000	0.00	0.062	0.894
Full oracle	1.000	1.000	1.000	1.00	0.063	0.893

Additional faithfulness–localization checks. Table 21 states the architecture difference as a paired contrast on sources and methods common to both architectures (-0.56, $[-0.65, -0.46]$; validation -0.52); it stays negative with only Occlusion, RISE, and EP. For R(2+1)D the insertion–TLS relation is -0.57 on validation (75 sources) and negative for each checkpoint (test -0.52, -0.52, -0.69); the deletion–TLS relation has the same sign (-0.49; -0.18 without IG and SmoothGrad-IG) but carries less weight. At the rank level the pattern holds in all six seed-by-split cells (insertion vs. TLS -0.52 to -0.81; Appx. C.6), and average ranks shift by less than 0.1 in the 66-source subset.

For TimeSformer the full roster gives $+0.11$ ($[+0.04, +0.18]$, 27 sources), a roster effect, since the five shared methods give zero; the deletion–TLS divergence is present on shared methods (-0.25) but weaker, so we restrict the architecture claim to insertion versus TLS. The TimeSformer cells have 21–30 sources, one checkpoint (seed 123) negative on test, capped RISE, EP, and SmoothGrad-IG, and no robust rank-level comparison; a bf16 rerun of the classification gate, which recovers 33 of 438 clips, leaves the conclusions unchanged (full roster $+0.10$, $[+0.01, +0.18]$).

Occlusion ablation. In the controlled setting Occlusion masks all frames of a patch together, so its temporal localization is at chance by construction. Removing it gives an insertion–TLS relation of -0.44 ($[-0.46, -0.42]$) for R(2+1)D and $+0.31$ ($[+0.24, +0.39]$) for TimeSformer; removing both

Table 21: Paired cross-architecture contrast: mean within-source τ_b for R(2+1)D minus TimeSformer on sources evaluated for both (95% source-resampling intervals). Negative values: stronger behavioral–localization divergence for R(2+1)D.

Split	Roster	n	Ins. vs TLS	Del. vs TLS
val	Full rosters	30	-0.68 [-0.76, -0.60]	-0.49 [-0.56, -0.41]
val	Shared 5 methods	30	-0.52 [-0.61, -0.42]	-0.41 [-0.51, -0.30]
val	Shared 3 (Occl., RISE, EP)	30	-0.34 [-0.49, -0.20]	+0.00 [-0.18, +0.19]
test	Full rosters	27	-0.69 [-0.77, -0.62]	-0.41 [-0.49, -0.33]
test	Shared 5 methods	27	-0.56 [-0.65, -0.46]	-0.32 [-0.42, -0.22]
test	Shared 3 (Occl., RISE, EP)	27	-0.21 [-0.35, -0.07]	+0.30 [+0.16, +0.44]

Table 22: Per-source agreement between behavioral and localization method rankings (mean within-source Kendall τ_b over the methods, averaged over the three checkpoints per source; 95% source-resampling intervals, 10,000 draws). Negative values mean methods that score well on the behavioral metric localize worse. "Excl. failed checks" drops IG and SmoothGrad-IG; "shared 5" uses the five methods run on both architectures. Inference is over sources, conditional on the method roster. Locked fp32 evaluation.

Arch.	Split	Roster (k)	n src.	Ins. AUC vs TLS	Del. AUC vs TLS	Ins. AUC vs ST-RMA
R(2+1)D	val	full roster (7)	75	-0.57 [-0.59, -0.56]	-0.49 [-0.51, -0.47]	+0.38 [+0.33, +0.43]
R(2+1)D	val	excl. failed checks (5)	75	-0.40 [-0.42, -0.37]	-0.19 [-0.22, -0.16]	+0.44 [+0.39, +0.50]
R(2+1)D	val	shared 5 (5)	75	-0.52 [-0.55, -0.50]	-0.54 [-0.57, -0.51]	+0.40 [+0.35, +0.45]
R(2+1)D	test	full roster (7)	71	-0.58 [-0.59, -0.56]	-0.49 [-0.51, -0.46]	+0.38 [+0.33, +0.42]
R(2+1)D	test	excl. failed checks (5)	71	-0.41 [-0.43, -0.39]	-0.18 [-0.21, -0.15]	+0.45 [+0.40, +0.50]
R(2+1)D	test	shared 5 (5)	71	-0.53 [-0.56, -0.51]	-0.56 [-0.59, -0.53]	+0.39 [+0.33, +0.44]
TimeSformer	val	full roster (6)	30	+0.12 [+0.04, +0.20]	-0.01 [-0.08, +0.06]	+0.30 [+0.24, +0.37]
TimeSformer	val	excl. failed checks (4)	30	+0.14 [+0.06, +0.23]	-0.00 [-0.09, +0.08]	+0.47 [+0.38, +0.55]
TimeSformer	val	shared 5 (5)	30	-0.00 [-0.10, +0.09]	-0.14 [-0.23, -0.05]	+0.33 [+0.26, +0.40]
TimeSformer	test	full roster (6)	27	+0.11 [+0.04, +0.18]	-0.10 [-0.17, -0.04]	+0.30 [+0.22, +0.39]
TimeSformer	test	excl. failed checks (4)	27	+0.11 [-0.00, +0.21]	-0.15 [-0.21, -0.09]	+0.42 [+0.30, +0.54]
TimeSformer	test	shared 5 (5)	27	+0.02 [-0.07, +0.10]	-0.25 [-0.33, -0.17]	+0.33 [+0.22, +0.43]

Occlusion and the methods that fail randomization gives -0.07 ($[-0.10, -0.03]$; four methods) for R(2+1)D and $+0.52$ ($[+0.39, +0.65]$; three methods) for TimeSformer. On the four shared methods other than Occlusion (IG, SmoothGrad-IG, RISE, EP) the relations are -0.29 and $+0.27$, with a paired contrast of -0.57 ($[-0.70, -0.45]$; validation -0.45). These analyses were added after the protocol was frozen and involve few methods, so we treat them as sensitivity analyses.

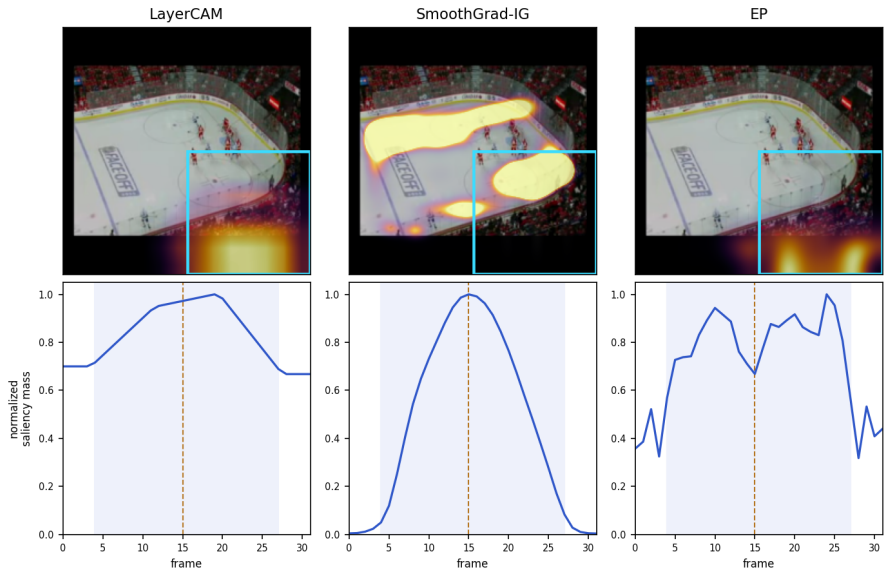


Figure 5: Qualitative example on one held-out test clip (R(2+1)D-18, seed 42). Top: saliency at the ground-truth window's midpoint frame with the true quadrant boxed; bottom: per-frame saliency mass with the true window (frames 4–27) shaded. LayerCAM concentrates spatial mass in the box but has a broad temporal profile that is only weakly window-specific; SmoothGrad-IG has a sharply window-centered temporal profile but places much of its spatial mass outside the box; EP is intermediate. One clip is shown for illustration only; quantitative results are in § 5.2.

C.6 RANK-LEVEL REPLICATION

Table 23 gives the rank-level agreement in each seed-by-split cell.

Table 23: Rank-level Kendall τ between method rankings (mean over sources) per seed and split, seven methods for R(2+1)D and six for TimeSformer. R(2+1)D: 75 (val)/71 (test) sources; TimeSformer: 21–30.

Arch.	Comparison	s42 val	s42 test	s123 val	s123 test	s7 val	s7 test
R(2+1)D	Ins. vs TLS	$-.62$	$-.62$	$-.62$	$-.52$	$-.71$	$-.81$
R(2+1)D	Del. vs TLS	$-.33$	$-.33$	$-.71$	$-.71$	$-.90$	$-.90$
TimeSformer	Ins. vs TLS	$+.20$	$+.20$	$-.07$	$-.47$	$+.47$	$+.47$
TimeSformer	Del. vs TLS	$-.33$	$-.33$	$-.60$	$-.47$	$-.07$	$-.07$

Both R(2+1)D comparisons have 95% intervals on one side of zero in all six cells; no TimeSformer comparison does. Rank-level Kendall p-values for R(2+1)D are not significant after Holm correction, since the test has almost no power with seven items.

D HUMAN JUDGMENTS AND VIDEO EVIDENCE

D.1 STUDY DESIGN AND METHODOLOGY

We conducted an asynchronous online study to examine how people judge physical aggression in short videos and identify the frames supporting their judgments. Participants completed the study independently through Google Forms. We recruited 18 participants through campus and community outreach, including flyers at a university campus and invitations shared through university groups and personal networks. Participation was voluntary and unpaid, with one submission allowed per person. Participants confirmed that they were at least 18 and could understand English instructions and view videos. They then read the consent information and agreed to participate. All personally identifying information was excluded from the analysis.

The study used five clips from RWF-2000, a dataset of surveillance videos used for violence detection (Cheng et al., 2021). We refer to them as C01–C05, in presentation order. C01 and C03 have non-fight labels in the dataset, while the other three have fight labels. Each study video presented 32 frames, or still images from the clip, in sequence over 16 seconds. Playback at two frames per second gave viewers more time to inspect each frame. A frame number above the footage and a timeline below it showed which frame was on screen, allowing participants to identify evidence by frame number rather than by time (Figure 6).

The clips show three people in an elevator (C01), a partly obscured interaction in a hallway (C02), a person standing on a table in a crowded cafeteria (C03), pushing and striking between people in a locker aisle (C04), and a physical confrontation in a barbershop (C05).

Each clip had separate questions for judgment and evidence. Participants first judged if physical aggression was visible, absent, or unclear. If they could see contact but were unsure how to interpret it, they could choose "Cannot tell." The evidence questions then asked them to mark frames showing aggression and, separately, frames supporting a non-violent interpretation. A five-level confidence rating and an optional comment followed each clip. Participants could replay videos, take breaks, or stop.

Before viewing these clips, participants received instructions and a separate practice example clip. The instructions defined physical aggression as visible hitting, kicking, aggressive pushing, striking with an object, or a clear attempt to physically harm someone. Participants were asked to mark the action and the motion needed to recognize it, rather than people gathering beforehand. Arguments, crowds, or fast movement alone were not enough to establish aggression. The example made this distinction concrete. People gathered near a doorway before a fight began, and the accompanying explanation identified frames 20 through 32 as showing the initial striking motion and the continuing fight. The example acknowledged that reasonable observers might choose slightly different boundaries.

Two practice questions then checked how to record evidence. The first asked participants to enter the demonstrated range as `20-32`, meaning all frames from 20 through 32. For separate ranges, the instructions explained that a semicolon kept them apart without including the gap. The second question asked what to enter if there was no aggression evidence. The answer was `NONE`, meaning



Figure 6: Example of the numbered video presentation used in the human study, showing frame 6 of 32 from clip C03. The frame counter and timeline identify the current frame so participants can report evidence using frame numbers.

no frames showed that type of evidence. Participants could instead enter `UNSURE` if they could not decide or locate evidence, or `SKIP` if they skipped a clip or it failed to load. For non-violent evidence, `WHOLE` meant that their judgment depended on the entire sequence and was encoded as `1-32`. We also retained the `WHOLE` label to report how often participants chose it. In the analysis, `NONE` counted as selecting no frames, while `UNSURE` and `SKIP` were excluded from frame-level calculations.

The instructions defined "physical aggression," but the first two participants saw judgment options referring to "violence," a term we had not separately defined. Following feedback, we replaced it with "physical aggression" so the options referred directly to the definition participants had been given. We retained the two earlier responses under the corresponding categories and repeated the analysis without them to check if their inclusion changed the findings. This check assessed the effect on our results, without assuming that participants understood the two terms identically.

For the overall judgments, we summarized how often participants saw aggression, saw none, or could not tell, and reported 95% Wilson confidence intervals (CIs) for the percentage seeing aggression (Wilson, 1927). We calculated this percentage out of all judgments, including "Cannot tell" answers. Event agreement measured how often two participants gave the same answer, including when both chose "Cannot tell."

For aggression evidence, we counted the percentage selecting each frame among respondents who provided a range or explicitly reported no evidence. Frames selected by at least half of these respondents formed the 50% consensus set. Each frame had to meet this threshold on its own. We also examined a stricter 75% threshold.

For evidence agreement, we compared frame selections among participants who marked aggression using mean pairwise intersection-over-union (IoU), or Jaccard similarity (Jaccard, 1901). For each pair, this is the number of frames both selected divided by the number either selected. Longer selections can overlap more simply because they cover more of the clip. We therefore also reported how many frames participants selected. Each participant contributed equally, regardless of confidence.

D.2 Results and analysis

All 18 participants completed the practice checks and judged all five clips, giving us 90 judgments. The evidence questions received 3 `UNSURE` answers for aggression evidence and 4 `SKIP` answers for non-violent evidence. These participants still answered the overall judgment question, and their answers appear in Table 24. We excluded `UNSURE` and `SKIP` from the frame calculations but kept their other evidence answers.

Clip	Label	Visible	Absent	Unclear	Visible % [95% CI]	Agreement %
C01	non-fight	2	16	0	11.1 [3.1, 32.8]	79.1
C02	fight	18	0	0	100.0 [82.4, 100.0]	100.0
C03	non-fight	5	11	2	27.8 [12.5, 50.9]	43.1
C04	fight	18	0	0	100.0 [82.4, 100.0]	100.0
C05	fight	17	0	1	94.4 [74.2, 99.0]	88.9

Table 24: Overall judgments for each clip. "Visible" and "Absent" count those who did or did not see physical aggression, and "Unclear" counts those who chose "Cannot tell." "Visible %" expresses the visible count out of all 18 judgments, with a 95% Wilson confidence interval in brackets. "Agreement %" shows how often two participants gave the same answer. "Label" is the dataset category.

All 18 participants saw aggression in C02 and C04, and 17 saw it in C05 (Table 24). Most saw none in the elevator scene, C01. The cafeteria scene, C03, was less clear despite its non-fight label. Here, 11 participants saw no aggression, 5 saw aggression, and 2 could not tell. Participants thus disagreed about whether aggression was present at all.

Even when everyone saw aggression, they did not necessarily select the same frames. Selections overlapped less in C02 than in C04 (mean IoU 0.541 versus 0.739). In C02, some participants began marking evidence as early as frame 2, while others began at frame 23. C05 had the greatest overlap among the fight clips, although participants also tended to select longer ranges (Table 25). Agreement on the overall judgment therefore did not mean agreement on where to mark the evidence.

Clip	Included n	Marked n	Mean IoU	Median length	50% set	75% set
C01	18	1	n/a	15	none	none
C02	18	18	0.541	19.5	10-32	21-32
C03	16	3	0.417	5	none	none
C04	18	18	0.739	24	8-32	12-32
C05	17	17	0.834	27	6-32	8-32

Table 25: Where participants found aggression evidence. "Included" counts respondents who entered frame ranges or `NONE`; `UNSURE` and `SKIP` are excluded. "Marked" counts only those who selected at least one frame. IoU compares pairs who both marked evidence, from 0 (no overlap) to 1 (identical selections). Median length is the median number of selected frames among those who marked evidence; n denotes the number of participants. The final columns list frames selected by at least 50% or 75% of included respondents. C01 has no IoU because only one person marked evidence. C03's IoU uses just three pairs.

At the 50% threshold, C02's shared evidence now forms one continuous range, frames 10–32. C04 and C05 also each had one continuous range. In C01 and C03, some participants selected evidence frames, but no individual frame was selected by at least half of the included respondents. For C03, 3 participants marked evidence and 2 entered `UNSURE` for the evidence frames. An empty set does not mean everyone saw no evidence.

For non-violent judgments, participants often relied on watching the entire sequence. They chose `WHOLE` 12 times for C01 and 8 times for C03, but never for the fight clips. These answers cover frames 1–32 and indicate that participants based their judgment on the full sequence.

We also checked if two decisions about including responses affected the shared evidence frames. One participant reported aggression in C01 and C03 but entered `NONE` for its evidence, and another participant reported aggression in C03 while entering `NONE`. Removing these three evidence answers left every clip's consensus sets unchanged at both the 50% and 75% thresholds. We obtained the same sets when we separately excluded the first two participants, who had seen judgment options

using "violence" before we changed them to "physical aggression." Thus, with all 18 responses included, neither sensitivity check changed the reported consensus sets at either threshold.

D.3 Limitations and future work

The study shows how agreement about whether aggression is present can coexist with different choices of evidence frames. With 18 participants recruited through campus and personal networks and only five clips, these patterns may not extend to other viewers or videos. The confidence intervals reflect uncertainty in responses to each clip under an assumption of independent participants; they do not account for how participants or clips were selected.

Slower playback gave participants more time to inspect the footage. It could not make actions hidden by the camera angle or other people visible. It may also have changed how participants interpreted motion. A larger study could include more varied scenes and compare playback speeds to assess if these choices affect judgments and frame selections. Comparing human selections with model explanations is another next step. Agreement would show that people and the model explanation point to similar frames, but would not by itself establish that the explanation faithfully reflects the model's decision. A follow-up intervention could ask if masking the frames people agreed showed aggression reduces the model's probability of aggression more than masking the same number of randomly selected frames (Petsiuk et al., 2018). This would test whether the model relies on the frames participants agreed were evidence of aggression, although masking may also alter the video's temporal structure.