EVENTTRIGGERBENCH: DISCOVERING PLAUSIBLE TRIGGER EVENTS FOR EXPLANATORY VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding videos requires more than recognizing what happens; it also requires identifying which earlier events make a later event plausible. We introduce EventTriggerBench, a benchmark for target-conditioned *plausible trigger attribution* in videos. Given an ordered sequence of localized events and a target event that may appear anywhere in the timeline, models must identify earlier events that provide plausible explanatory support for the target. EventTriggerBench is designed to test a key gap in current video understanding: models often capture temporal co-occurrence, but struggle to distinguish events that merely precede a target from those that meaningfully explain it. To study this gap, we provide auto-verified trigger labels and validate them with human judgments: when annotators independently select triggers without seeing the automatic labels, they agree with the automatic label in 77.7% of audited samples (substantial but imperfect agreement), and 88.3% of the automatic annotations are accepted in a separate audit. We evaluate a broad range of heuristic, sequential, graph-based, multimodal, causal, and LLM baselines. We further introduce EventGraph-Trigger, a multimodal event-graph reasoning framework for target-aware trigger attribution. Our experiments show that trigger attribution remains challenging for sequence, graph, video-based, and LLM-based baselines, especially when explanations involve non-adjacent events, multi-step event chains, or distinguishing direct triggers from incidental earlier events, and that EventGraph-Trigger remains the strongest among the representative models evaluated against independently collected human trigger labels. These findings suggest that target-conditioned trigger attribution, conditional on the existence of an observable plausible trigger, offers a useful diagnostic lens for moving beyond event recognition toward structured explanatory reasoning over video-grounded events.

1 INTRODUCTION

Videos are structured processes rather than independent collections of frames. When humans watch a video, they not only recognize visible events but also infer dependencies among them. A glass breaks after it is dropped; a person slips after water is spilled. Such reasoning goes beyond asking *what happened*; it also asks *why it happened* and *which earlier event plausibly explains the target*. This ability matters for accident analysis, anomaly explanation, robotics, and video agents.

Despite rapid progress in video-language modeling, existing benchmarks mainly evaluate recognition, retrieval, captioning, temporal grounding, or descriptive question answering (Carreira & Zisserman, 2017; Kay et al., 2017; Krishna et al., 2017; Lei et al., 2018). These tasks often allow models to succeed through local recognition, lexical matching, or temporal proximity, but rarely test whether a model can distinguish a meaningful trigger from an adjacent or correlated event. For example, in a cooking video, "smoke appears" may immediately precede "fire appears," but an earlier event such as "the stove was turned on" may provide a more informative explanation.

This challenge grows with the number of events. Candidate triggers may be non-adjacent, intermediate events may mediate the explanation, multiple earlier events may jointly contribute to an outcome, and distractors may be semantically plausible but irrelevant.

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.

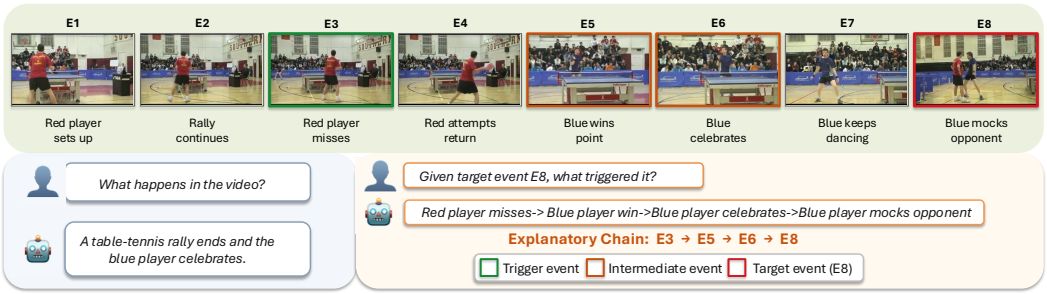

Figure 1: Illustration of the EventTriggerBench task. Top: The video is decomposed into a sequence of event clips, where the trigger event, intermediate events, and target event are highlighted in green, orange, and red, respectively. Bottom left: Conventional video understanding summarizes visible actions, such as the rally ending and the blue player celebrating. Bottom right: EventTriggerBench instead requires target-conditioned trigger attribution: given target event E_8, the model identifies the plausible trigger event E_3 and the explanatory chain $E_3 \rightarrow E_5 \rightarrow E_6 \rightarrow E_8$.

A growing body of work studies causal and explanatory reasoning in video. Explanatory and counterfactual VideoQA benchmarks (Xiao et al., 2021; Li et al., 2022; Wu et al., 2023) ask why- and what-if questions, CLEVRER (Yi et al., 2020) studies explanatory, predictive, and counterfactual reasoning in synthetic physical scenes, MECD (Chen et al., 2024) discovers causal relations among multiple video events, and CARVE (Le et al., 2025) studies causal abductive reasoning with trigger events and causal chains. EventTriggerBench does not claim to be the first benchmark for causal video reasoning in general. Existing settings, however, typically evaluate trigger attribution only indirectly through free-form answers, rely on synthetic environments, focus on final outcomes, or require recovering a full causal graph. In practical video analysis, the question is often target-conditioned: given an event of interest anywhere in the video, which earlier event best explains it?

We introduce EventTriggerBench, a benchmark for plausible trigger attribution in videos, which we formulate as the Video Trigger Attribution task. Given an ordered sequence of event clips and captions with a target event, the model must identify the earlier events annotated as plausible triggers. The task is closed-world: every target has at least one plausible trigger among its observed predecessors, so we evaluate attribution conditional on the existence of an observable plausible trigger rather than general trigger discovery. We also provide optional explanatory-chain annotations when the target is explained through intermediate events. Unlike final-outcome-only settings, EventTriggerBench allows target events at arbitrary temporal positions, enabling multiple samples per video and evaluating adjacent, non-adjacent, multi-hop, and multi-trigger attribution. Since captions carry most of the signal (Section 5.2), it is best understood as an event-level explanatory reasoning benchmark over video-grounded events.

Our benchmark construction follows a transparent event-centric pipeline. We obtain localized event clips and captions, select outcome-like or state-changing targets, use an LLM to propose trigger events and explanatory chains, and apply automated verification based on temporal constraints, counterfactual plausibility, chain coherence, and self-consistency confidence (Achiam et al., 2023; Wei et al., 2022; Wang et al., 2023). The resulting labels indicate plausible trigger-target relations rather than true causal ground truth, so we validate them with human judgments on 300 held-out test samples. Annotators who independently select triggers without seeing the automatic labels agree with them in 77.7% of samples, a substantial but imperfect agreement, and an annotation-visible audit accepts 88.3% of verified annotations (71.3% for raw proposals). Throughout, a *trigger* is a plausible explanatory trigger, and *causal* refers to explanatory and counterfactual plausibility rather than *causal identification*, which would require intervening on the data-generating process.

We also propose EventGraph-Trigger, a target-aware multimodal graph reasoning framework. It represents events with text, video, and temporal features, builds an event graph from temporal, semantic, visual, and learned relations (Kipf & Welling, 2017; Veličković et al., 2018; Rossi et al., 2020), scores each predecessor of the target as a candidate trigger, and optionally learns to recover the annotated explanatory path.

Our contributions are summarized as follows:

- We define Video Trigger Attribution, a closed-world, target-conditioned video reasoning task for identifying earlier events that plausibly explain a target event at any position in the timeline, without requiring recovery of a complete causal graph.
- We build EventTriggerBench with clips, captions, targets, auto-verified trigger labels, explanatory chains, and difficulty metadata, validated by a human audit and independent human trigger labels.
- We propose EventGraph-Trigger, a multimodal graph-based baseline for structured trigger attribution over multi-event sequences.
- We evaluate heuristic, sequence, graph, video, multimodal, causal-style, and LLM baselines against automatic and independent human labels, with ablations and difficulty, robustness, and insufficient-evidence analyses.

2 Related Work

Video understanding benchmarks. Action recognition benchmarks such as Kinetics (Carreira & Zisserman, 2017; Kay et al., 2017) evaluate whether models can recognize human actions from short clips, while video QA benchmarks such as TVQA (Lei et al., 2018) and STAR (Wu et al., 2021) test localized, compositional, and situated question answering. Dense-captioning datasets such as ActivityNet Captions (Krishna et al., 2017) provide temporally localized event descriptions. EgoSchema (Mangalam et al., 2023) and Video-MME (Fu et al., 2025) evaluate long videos. None of them explicitly evaluates target-conditioned trigger attribution.

Explanatory and counterfactual video QA. NExT-QA (Xiao et al., 2021) introduces causal and temporal questions about real-world videos, Causal-VidQA (Li et al., 2022) adds explanation, prediction, and counterfactual questions that require evidence and commonsense reasoning, and ACQUIRED (Wu et al., 2023) targets counterfactual questions in real-life videos. In these benchmarks, the explanandum is a natural-language question and trigger attribution is assessed only indirectly through the selected answer.

Physical and interaction-based causal reasoning. CLEVRER (Yi et al., 2020), CATER (Girdhar & Ramanan, 2020), and ComPhy (Chen et al., 2022) evaluate temporal, physical, and counterfactual reasoning in controlled synthetic environments, where the underlying dynamics are known. They lack the noisy event structure of real videos.

Multi-event causal discovery and abductive reasoning. MECD (Chen et al., 2024) builds on ActivityNet Captions and determines which premise events are causally related to the final event of a video, with GPT-4 annotations refined by five human annotators; an event-level causal diagram is then derived by treating events as results in turn. Visual abductive reasoning (Liang et al., 2022) infers plausible explanations of premise events. CARVE (Le et al., 2025) selects target events anywhere in a video's event graph and identifies their trigger events, with counterfactual labels from re-simulated synthetic physics videos and a smaller EpicKitchens-based set, and proposes CERN, a temporal-semantic event graph model. Table 1 compares these benchmarks directly. The novelty of EventTriggerBench lies in benchmark design rather than in a new task family. CARVE also allows targets at arbitrary positions, but its labels are derived from re-simulated physical interactions (or from a model on the smaller kitchen set), so it measures attribution under known physical dynamics. EventTriggerBench instead measures attribution among open-vocabulary events in real videos, where explanations often rest on goals, rules, or social conventions (e.g., a player celebrating because the opponent missed) that physical re-simulation cannot label, where semantically plausible distractors are common, and where the explanatory chain is evaluated in addition to the trigger. Unlike MECD, whose labels relate events only to the final event, EventTriggerBench labels intermediate targets and explanatory chains.

Event graphs and LLM-based reasoning. Scene graphs (Xu et al., 2017; Yang et al., 2023) and graph neural networks (Kipf & Welling, 2017; Veličković et al., 2018; Rossi et al., 2020) provide tools for reasoning over structured visual data. LLMs reason well with chain-of-thought, self-consistency, or tool use (Brown et al., 2020; Achiam et al., 2023; Wei et al., 2022; Wang et al., 2023; Yao et al., 2023).

Benchmark	Videos	Target specification	Targets per video	Explicit trigger selection	Multi-trigger	Explanatory chains	Label source
NExT-QA	Real	NL question	Multiple	No	–	No	Human
Causal-VidQA	Real	NL question	Multiple	No	–	No	Human
CLEVRER	Synthetic	NL question	Multiple	Partial	–	No	Simulator
MECD	Real	Final event	One	Yes	Yes	No	GPT-4 + human
CARVE	Synthetic + real	Any event	Multiple	Yes	Yes	No	Simulation + model-derived counterfactuals
EventTriggerBench	Real	Any event	Multiple	Yes	Yes	Yes	GPT-4 + verifier; human-validated

Table 1: Benchmark-level comparison with the most closely related video reasoning benchmarks.

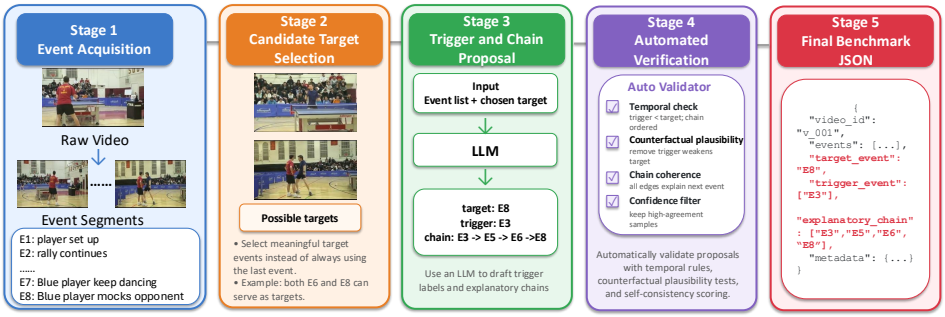

Figure 2: Overview of the EventTriggerBench construction pipeline: (1) event segmentation and captioning; (2) target selection, allowing intermediate targets such as E_6 and E_8; (3) LLM proposal of triggers and explanatory chains; (4) automated verification with temporal, counterfactual-plausibility, chain-coherence, and self-consistency checks; and (5) the final JSON sample (`target_event`, `trigger_event`, `explanatory_chain`, metadata). One video yields multiple target-conditioned samples.

3 Dataset Construction: EventTriggerBench

Our dataset pipeline consists of five stages (Figure 2): (1) event acquisition, (2) candidate target selection, (3) trigger and explanatory-chain proposal, (4) automated verification and quality control, and (5) benchmark assembly with difficulty metadata. We then validate the resulting labels with an independent human audit (Section 3.5). All prompts are given in Appendix F.

3.1 Event Acquisition

Each video is represented as an ordered set of events $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$. Each event contains a temporal interval, a caption, and optionally a clip path. If the source dataset already contains dense event captions, we directly use the provided temporal annotations after cleaning. For general videos without event annotations, we first detect shot boundaries with TransNetV2 (Souček & Lokoč, 2024) and merge adjacent shots into event segments, extract event clips, and generate event captions with Video-LLaVA-7B (Lin et al., 2024).

3.2 Candidate Target Selection

A central design choice is that target events are not restricted to the last event. We define a suitable target as an event whose occurrence can be meaningfully explained by earlier events. We select candidates using rule-based filters and LLM-assisted judgments. A target candidate must satisfy the following criteria:

- It is not the first event and has at least one predecessor.
- It expresses a state change, transition, completion, anomaly, or outcome-like event.

- It can naturally support a "why" question.
- It has at least one plausible earlier trigger event.

Examples include "the person falls" and "smoke appears". Since every target must have a plausible earlier trigger, the task is a choice among observed predecessors; no-trigger or insufficient-evidence targets are not included.

3.3 Trigger and Explanatory-Chain Proposal

For each target event, we ask an LLM (`gpt-4-turbo-2024-04-09`) to propose trigger event(s) and an optional explanatory chain using the ordered event captions. The prompt explicitly defines a trigger as an earlier event whose absence would likely prevent or substantially change the target. The LLM must output: (i) the main trigger, (ii) possible additional triggers, (iii) the explanatory chain from trigger to target, (iv) a short natural-language explanation, and (v) a confidence score. This output is a candidate annotation, not ground truth.

Label semantics. The main trigger is the first event of the proposed explanatory chain; later chain events are mediators, not triggers, and the counterfactual wording tests whether an event initiates this process, not whether it is the unique cause. Additional triggers are other consistently proposed initiating events, without distinguishing sufficient, jointly necessary, or alternative triggers. Since mediators can be plausible proximal explanations, we also report a chain-lenient accuracy (Appendix E).

3.4 Automated Verification

Given an LLM-generated proposal consisting of a target event, trigger event, and optional explanatory chain, the verifier evaluates whether the proposal satisfies temporal, counterfactual, and semantic constraints. The verifier uses the same model snapshot as the proposer but a separate prompt and independent inference runs. It applies four checks.

Temporal check. The trigger must occur before the target, and all events in an explanatory chain must be temporally ordered.

Counterfactual-plausibility check. A valid trigger should not only precede the target but also plausibly contribute to it. The verifier considers whether the target would still plausibly occur if the proposed trigger were removed, altered, or replaced with a neutral event. If removing the trigger does not weaken the explanation for the target, the proposal is treated as a temporal correlation rather than a plausible trigger. This assesses plausibility and is not an intervention on the data-generating process.

Chain coherence check. For samples with explanatory chains, each adjacent pair of events in the chain is evaluated for semantic continuity and explanatory plausibility, preventing chains that skip necessary intermediate steps or include unrelated events.

Confidence filter. For each candidate target, we query the proposal model $K = 5$ times with different decoding seeds and retain a sample only when the proposed trigger and chain are stable across generations, with an agreement-based confidence of at least $\tau = 0.7$. Samples with low agreement are marked as ambiguous and removed.

The verifier outputs *accept*, *revise*, or *discard* for each proposal, with rates of 64.8%, 21.5%, and 13.7%, respectively. Revised samples are corrected by removing unsupported edges, shortening invalid chains, or selecting a higher-confidence trigger; discarded samples are excluded.

3.5 Human Validation

We randomly sample 300 held-out test examples; three annotators view the clips and captions and judge target suitability, trigger plausibility, and chain coherence, and whether to accept, revise, or reject the annotation (Appendix B).

As shown in Table 2, 88.3% of the verified annotations are accepted, with substantial agreement on the accept/revise/reject decision (Fleiss' $\kappa = 0.81$; Fleiss, 1971). Automated verification raises human acceptance from 71.3% for raw LLM proposals to 88.3% (Table 3). Treating the human judgment on the raw proposals for the same 300 targets as the reference, the automated verifier identifies *retainable* proposals (those accepted as written or after revision) with 89.6% precision

Metric	Result
Target suitability	92.7%
Trigger correctness	87.3%
Chain correctness†	84.7%
Overall acceptance rate	88.3%
Fleiss' κ (accept/revise/reject)	0.81

Table 2: Human audit of 300 randomly sampled test annotations by three annotators. †Among samples with chain annotations.

Annotation stage	Human acceptance
Raw LLM proposals	71.3%
After automated verification	88.3%
Verifier precision (retainable)	89.6%

Table 3: Human acceptance before and after automated verification, and precision of the verifier against human judgments.

(definition and counts in Appendix B). Because annotators see the automatic annotation, this audit may be inflated by anchoring; Section 5.3 gives a stricter estimate. Shared priors between proposer and verifier cannot be fully eliminated. A verifier from another model family (Qwen2.5-72B) accepts, revises, and discards 78.3/13.1/8.6% of the test labels, and a fully independent Qwen2.5-72B pipeline (accept/revise/discard 61.8/25.3/12.9% of its own proposals) reproduces 74.6% of the labels and preserves the ranking of the six models evaluated on its labels (Appendix B).

4 EventGraph-Trigger Model

We propose EventGraph-Trigger, a target-aware multimodal graph reasoning framework for video trigger attribution. Instead of treating events as a flat sequence, EventGraph-Trigger represents the video as an event graph, where nodes are events and edges capture temporal, semantic, visual, or learned dependencies.

Trigger reasoning is relational: a trigger is determined not only by its description but also by how it connects to the target through intermediate steps. The framework contains four modules, illustrated in Figure 3: event representation, event graph construction, target-aware graph reasoning, and trigger prediction. Algorithm 1 in Appendix C summarizes the full procedure.

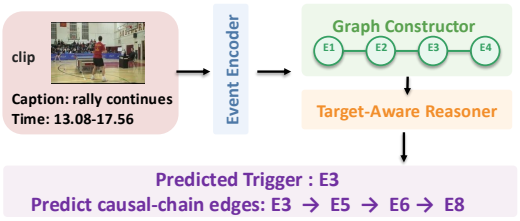

Figure 3: Architecture of EventGraph-Trigger. The model encodes each event using its clip, caption, and timestamp, constructs an event graph over the video timeline, and performs target-aware reasoning to identify the trigger event. An optional chain head predicts intermediate explanatory-chain edges, providing an interpretable path from the trigger to the target.

4.1 Event Representation

Each event e_i in the ordered sequence $\{e_1, \ldots, e_N\}$ has a caption, a clip, and a temporal interval. We encode them with a text encoder, a video encoder, and a temporal encoder, obtaining h_i^{text}, h_i^{video}, and h_i^{time}, and fuse them with a lightweight network:

$$h_i = \mathrm{MLP}\big([h_i^{text}; h_i^{video}; h_i^{time}]\big), \quad (1)$$

where $[\cdot;\cdot]$ denotes concatenation. Text-only and video-only variants drop h_i^{video} or h_i^{text}, respectively. The fused embedding h_i serves as the node feature for graph reasoning.

4.2 Event Graph Construction

We construct an event graph $G = (V, E)$ whose nodes are events. For each temporally valid event pair $i < j$, the graph edge score combines temporal, semantic, visual, and learned relations:

$$a_{ij} = \lambda_t\, a_{ij}^{temp} + \lambda_s\, a_{ij}^{sem} + \lambda_v\, a_{ij}^{vis} + \lambda_l\, a_{ij}^{learned}, \quad (2)$$

where a_{ij}^{temp} encodes temporal adjacency, a_{ij}^{sem} the semantic similarity of the event captions, a_{ij}^{vis} the visual similarity of the event clips, and $a_{ij}^{learned}$ a relation score inferred from the event representations and their temporal distance; $\lambda_t, \lambda_s, \lambda_v, \lambda_l$ weight the four relation types. A temporal mask removes

all future-to-past edges, so information flows only from earlier to later events, and for a target e_T only predecessors $\{e_i \mid i < T\}$ are candidate triggers.

4.3 Target-Aware Graph Reasoning

Trigger attribution is target-dependent: the same event may explain one target but be irrelevant to another. We therefore perform target-conditioned graph message passing: given the target representation h_T, each event aggregates information from its graph neighbors, weighted by the edge scores a_{ij} and modulated by the target query. The formal update rule is given in Equation 4 (Appendix C). Because edges point forward in time, a candidate's own representation encodes only its history; multi-hop evidence reaches its score through h_T, which aggregates messages along paths from the candidate through intermediate events to the target, and through the chain head, which predicts these paths explicitly.

4.4 Trigger Prediction and Training Objective

After target-aware graph reasoning, the model predicts the probability that each predecessor event triggers the specified target by comparing its updated representation with that of the target. Trigger prediction is supervised with the cross-entropy over candidate predecessors in single-trigger samples, or with the binary cross-entropy per candidate in multi-trigger samples, which allows multiple events to be selected; the number of annotated triggers is used only to select the training loss. At inference, every sample is scored by the same head without access to the trigger cardinality: candidates are ranked by their scores for Acc. and the ranking metrics, and the predicted trigger set for set-level metrics contains all candidates whose sigmoid score exceeds a validation-tuned threshold on temperature-calibrated scores (Appendix C). When explanatory-chain annotations are available, an auxiliary edge head is supervised with the binary cross-entropy over adjacent edges in the annotated explanatory chain. The overall objective is

$$L = L_{\text{trigger}} + \beta\, L_{\text{chain}}, \tag{3}$$

where β balances the two terms and $L_{\text{chain}} = 0$ for samples without chain annotations. The chain head also yields an interpretable event path from the trigger to the target.

5 Experiments

5.1 Setup and Dataset Statistics

Table 4 summarizes EventTriggerBench. The benchmark contains 1,240 videos, 8,936 event segments, and 3,482 target-conditioned trigger samples. Each video contains 7.2 events and produces 2.8 trigger samples on average, confirming that targets are not restricted to the final event. The average trigger distance is 2.4 events, the Previous Event heuristic is correct on only 31.6% of samples (Table 5), and 18.6% of samples contain multiple triggers; 2,146 samples (61.6%) carry explanatory-chain annotations. Dependencies thus span several events rather than long-form videos: 37.3% / 10.5% of samples have a trigger distance $\geq$3 / $\geq$5, with medians of 2 (distance) and 7 (events per video) (Table 12), so the benchmark tests non-adjacent, multi-event reasoning rather than long-video reasoning. A video-level split (70/10/20) yields 2,437 / 348 / 697 training, validation, and test samples. On multi-trigger samples, Acc. corresponds to Hit@1, while MAP, NDCG, and the set-level metrics in Section 5.2 use the full trigger set. Tasks, metrics, and baseline details are given in Appendix D.

Videos	1,240	Avg. events / video	7.2
Event segments	8,936	Avg. targets / video	2.8
Trigger samples	3,482	Avg. trigger distance	2.4
Multi-trigger	18.6%	Avg. chain length	3.1

Table 4: Dataset statistics for EventTriggerBench.

5.2 Main Results

Table 5 shows a clear progression from simple heuristics to structured multimodal reasoning. Temporal heuristics remain far below learned models, so temporal proximity alone is insufficient; sequence models improve by using event semantics, and graph-based models improve further.

Category	Model	Modality	Acc.	Top-3	MRR	MAP	NDCG
Heuristic	Random	–	18.4	-	-	-	-
	Previous Event	Time	31.6	-	-	-	-
	First Event	Time	22.8	-	-	-	-
	Temporal Nearest	Time	35.7	-	-	-	-
Sequence	BiLSTM (Graves & Schmidhuber, 2005)	Text	43.2	69.4	55.6	48.1	61.3
	Transformer Encoder (Vaswani et al., 2017)	Text	50.6	75.8	62.4	55.3	68.2
Graph	GCN (Kipf & Welling, 2017)	Text+Graph	54.1	78.3	65.7	58.9	70.6
	GAT (Veličković et al., 2018)	Text+Graph	57.8	80.9	68.6	61.7	73.5
	TGN (Rossi et al., 2020)	Text+Graph	59.3	82.1	70.2	63.4	75.1
Video-based	VideoMAE Feature + MLP (Tong et al., 2022)	Video	41.5	67.1	53.1	46.8	59.4
	TimeSformer (Bertasius et al., 2021)	Video	44.6	70.6	56.4	49.5	62.1
Multimodal	Matched Multimodal Transformer (no graph)	Video+Text+Time	62.1	83.9	73.5	66.9	78.2
	Relation-biased Transformer (no graph)	Video+Text+Time	65.3	86.1	76.2	69.8	80.9
	Qwen2-VL-7B (Wang et al., 2024)	Video+Text	56.0	79.3	67.1	60.4	72.2
	GPT-4o (OpenAI, 2024)	Video+Text	64.3	85.5	74.8	68.3	79.9
	Video-LLaVA-7B (Lin et al., 2024)	Video+Text	48.2	72.9	59.8	53.0	65.4
Causal	Event Masking (adapted from Chen et al., 2024)	Text	60.8	83.4	72.1	65.6	77.0
	VGCM (Chen et al., 2024)	Video+Text	64.6	85.9	75.3	69.0	80.1
	CERN (Le et al., 2025)	Video+Text+Graph	66.9	87.4	77.6	71.2	82.3
	Counterfactual Event Scoring (adapted from Yi et al., 2020)	Text+Video	63.4	85.1	74.6	68.2	79.3
LLM	GPT-4 Zero-shot (Achiam et al., 2023)	Text	57.0	79.2	67.5	61.1	72.8
	GPT-4 Few-shot (Achiam et al., 2023)	Text	61.7	83.1	71.8	65.4	76.6
	GPT-4 CoT (Wei et al., 2022)	Text	65.9	86.2	75.9	69.7	80.8
	ReAct (Yao et al., 2023)	Text	67.4	87.1	77.2	71.0	82.1
Ours	EventGraph-Trigger	Video+Text+Graph	72.7	90.4	82.3	76.5	86.7

Table 5: Main results on EventTriggerBench. We compare heuristic baselines, sequence models, graph neural networks, video-based models, causal masking methods, LLM prompting baselines, and our EventGraph-Trigger.

Model	Human-validated ($n = 300$)	Disagreement ($n = 67$)
Transformer Encoder	48.7	28.4
GAT	55.3	34.3
GPT-4 CoT	63.0	37.3
Matched MM Transformer	59.3	35.8
VGCM	61.7	34.3
CERN	64.0	38.8
EventGraph-Trigger	69.7	44.8

Table 6: Accuracy against human majority labels on the audited test samples and on the human–LLM disagreement subset.

Video-only models underperform text-based models, likely because captions already provide high-level semantic abstraction. Removing video features from EventGraph-Trigger costs 3.3 points and removing text 9.0 (Table 7), masking all captions reduces accuracy to 49.6, and generated instead of human-written captions on the same densely captioned videos cost 3.2 points (Table 15), so the benchmark relies substantially on caption semantics. EventGraph-Trigger is correct only with video features on 9.8% of test samples and only without them on 6.5% (Appendix E); EventTriggerBench therefore mainly evaluates reasoning over event-level descriptions grounded in clips, with video as complementary rather than primary evidence; on the 142 test samples where both text-only EventGraph-Trigger and GPT-4 CoT fail, the full model is correct on only 30.3%. The Matched Multimodal Transformer, which receives exactly the same inputs and chain supervision as EventGraph-Trigger and has a matched parameter count (14.2M vs. 13.8M) but no event graph, reaches 62.1 accuracy (vs. 72.7). This controls for inputs, supervision, and capacity but not for explicit pairwise relations: a Relation-biased Transformer that adds the edge scores of Equation 2 as attention biases reaches 65.3, so about 3 of the 10.6 points stem from the relation features and the rest from sparse, target-conditioned message passing. VLMs reach 48.2–64.3. The closest task-specific models, CERN and VGCM, reach 66.9 and 64.6; like GPT-4 CoT and ReAct, they lag behind EventGraph-Trigger, which is best on all metrics. Since Acc. needs only one hit on multi-trigger samples, we also evaluate predicted sets: EventGraph-Trigger reaches set-level precision/recall/F1 of 71.8/66.2/68.9 (CERN F1: 63.9) but an exact set match of only 41.5 (Table 17). Baseline details are in Appendix D.

5.3 Evaluation Against Independent Human Labels

Since high scores could reflect alignment with the annotation pipeline, annotators also independently select the most plausible trigger(s) for the same 300 samples from clips and captions, without seeing any automatic output; the majority vote is the human reference. The automatic labels match the

human majority in 77.7% of samples, with Fleiss' $\kappa = 0.68$ over per-candidate selection decisions (Appendix B). Of the 67 disagreements, 29 involve a human trigger outside the automatic label and chain, 21 an intermediate event of the automatic chain (a more proximal explanation), and 17 partially overlapping multi-trigger sets. The instructions did not state the first-event convention of Section 3; counting mediators as matches raises agreement to 84.7%.

Table 6 evaluates models, without retraining, on these labels and on the 67-sample disagreement subset. EventGraph-Trigger remains the strongest model, outperforming GPT-4 CoT by 6.7 points; under a paired bootstrap that resamples videos rather than samples (Efron & Tibshirani, 1994), its advantages over GPT-4 CoT, CERN, ReAct, and the Matched Multimodal Transformer are all significant (Table 11). Its gap to the Matched Multimodal Transformer (10.4) is similar to that on automatic labels (10.6). GPT-4 CoT, which shares the annotation model family, drops from 65.9 on automatic labels to 63.0. On the disagreement subset, where reproducing the automatic label is an error, EventGraph-Trigger still performs best. Since all models are trained on automatic labels, this shows only partial transfer to human labels and does not rule out annotation-induced biases.

Model Variant	Acc.	Δ
Full EventGraph-Trigger	72.7	–
w/o video features	69.4	-3.3
w/o text features	63.7	-9.0
w/o temporal encoding	67.9	-4.8
w/o event graph	61.8	-10.9
w/o learned edges	68.4	-4.3
w/o target conditioning	59.5	-13.2
w/o chain supervision	69.0	-3.7
random graph	57.4	-15.3
reversed temporal order	55.8	-16.9
Single-relation graphs		
Temporal edges only	61.5	-11.2
Semantic edges only	64.1	-8.6
Visual edges only	58.8	-13.9
Learned edges only	66.7	-6.0

Table 7: Ablation study of EventGraph-Trigger. Target conditioning and event graph structure are the most critical components; the bottom block keeps a single relation type in the graph.

5.4 Ablation Studies

In Table 7, removing target conditioning or the event graph causes large drops, and random graphs or reversed temporal order degrade most, suggesting that event relations and temporal direction matter. Every single-relation graph underperforms the full hybrid graph (by at least 6.0 points), and performance is stable across the sparsification level k (Table 13). All models degrade with trigger distance, sequence length, and multiple triggers, while the gain of EventGraph-Trigger over the Matched Multimodal Transformer grows with trigger distance (8.1→13.9) and chain length (9.4→12.8; Table 14). In an insufficient-evidence stress test (target placed after another video's events), EventGraph-Trigger with an abstention threshold rejects 78.5% of these targets while keeping 70.9 accuracy on the original samples (Table 18). EventGraph-Trigger matches the annotated chain exactly in 48.6% of cases (36.3 without chain supervision; Table 19). Counting chain mediators as correct raises accuracy to 80.1 for EventGraph-Trigger and 74.2 for GPT-4 CoT (Table 16), so part of the remaining error reflects trigger granularity rather than wrong explanations.

6 Conclusion

We introduced EventTriggerBench, a closed-world benchmark for target-conditioned plausible trigger attribution in videos, in which targets may occur anywhere in the timeline and labels include trigger sets and explanatory chains. Its auto-verified labels agree with independent human labels in 77.7% of audited samples, a substantial but imperfect agreement, and re-annotation with a different model family preserves the ranking of the evaluated models. Across heuristic, sequence, graph, video, multimodal, causal-style, and LLM baselines, trigger attribution remains challenging, especially for non-adjacent triggers, multi-hop chains, and multi-trigger targets, where exact set match stays below 42%. EventGraph-Trigger, which combines multimodal event representations with target-conditioned graph reasoning, is the strongest model on both automatic and human labels, and its gain over parameter-matched Transformers without the graph, including one with the same pairwise relation scores, grows with trigger distance and chain length. The benchmark still relies on LLM-assisted labels, strong caption semantics, a first-event trigger convention, and a closed-world assumption (Appendix A); open-world targets without observable triggers, larger-scale human annotation, and longer event sequences are natural next steps.

AI USE STATEMENT

We used large language models (LLMs) in four ways. (1) Writing: LLMs were used to polish the language of the paper, including grammar and wording. (2) Dataset construction: LLMs are part of the benchmark pipeline and are used for target selection, trigger and explanatory-chain proposal, automated verification, and event captioning for videos without dense annotations (Section 3). LLMs also serve as baselines (Section 5.1). (3) Experimental design: LLMs helped design part of the ablation studies. (4) Ideation: we brainstormed with LLMs during the design of some ideas. The authors made all final decisions on the research ideas, experimental design, and analyses, verified all LLM-assisted content, and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

Section 3 specifies the segmentation and captioning models, LLM snapshots, self-consistency settings, and verification outcomes used to construct EventTriggerBench, Section 5.1 gives the data splits, and Appendix F provides all prompts verbatim. Appendix B describes the human validation protocol. The EventGraph-Trigger model is specified in Section 4 and Algorithm 1. We will release the benchmark annotations, split files, preprocessing scripts, evaluation code, training configurations, and model checkpoints upon publication; for source videos that cannot be redistributed, we will provide video identifiers, timestamps, and reconstruction scripts.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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. PMLR, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Joao 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. 6299–6308, 2017.

Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, and Weiyao Lin. MECD: Unlocking multi-event causal discovery in video reasoning. *Advances in Neural Information Processing Systems*, 37:92554–92580, 2024.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. ComPhy: Compositional physical reasoning of objects and events from videos. In *International Conference on Learning Representations*, 2022.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for compositional actions and temporal reasoning. In *International Conference on Learning Representations*, 2020.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

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

Thao Minh Le, Vuong Le, Kien Do, Sunil Gupta, Svetha Venkatesh, and Truyen Tran. Finding the trigger: Causal abductive reasoning on video events. *arXiv preprint arXiv:2501.09304*, 2025.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379, 2018.

Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21273–21282, 2022.

Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. Visual abductive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15565–15575, 2022.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020.

Tomáš Souček and Jakub Lokoč. TransNet V2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11218–11221, 2024. doi: 10.1145/3664647.3685517.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Chandra, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. ACQUIRED: A dataset for answering counterfactual questions in real-life videos. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11753–11770, 2023.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5419, 2017.

Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18675–18685, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

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

Appendix

A Limitations

EventTriggerBench has several limitations. First, its annotations are *auto-verified labels of plausible trigger-target relations* rather than true causal ground truth. The verification module checks temporal validity, counterfactual plausibility, chain coherence, and self-consistency, but these checks assess the plausibility of an explanation and do not constitute causal identification in the interventionist sense. Second, the proposer and verifier are based on the same LLM family. Although they use separate prompts and independent inference runs, and the verifier agrees well with human judgments (Section 3.5), shared language-model priors cannot be completely eliminated. Self-consistency shows that a proposal is stable under repeated sampling, not that it is correct, and acceptance by a verifier from the same model family is not independent validation. Our human audit and human-label evaluation (Section 5.3) cover 300 test samples, the training and validation labels are not independently audited, the audit may be affected by anchoring because annotators see the automatic annotation, and models are trained on the automatic labels, so annotation-induced biases cannot be ruled out. The cross-family re-annotation (Appendix B) changes only the test labels; the training labels still come from GPT-4, and an independently human-annotated training subset would be needed to test whether models learn annotation-specific preferences. Larger-scale independent human annotation would further strengthen the benchmark. Third, the benchmark depends on the quality of event segmentation and captions. Noisy boundaries or incomplete captions may affect target selection, trigger proposals, and explanatory-chain labels, even though event clips are retained as visual evidence. Because captions provide strong high-level semantics, text features currently contribute more than video features (Table 7); a more targeted evaluation of cases where captions are intentionally incomplete or noisy is an important direction for isolating the contribution of visual evidence. Fourth, every target has at least one plausible trigger by construction, so the main benchmark does not evaluate whether a model can recognize that no observable trigger exists or that the evidence is insufficient; our insufficient-evidence stress test (Table 18) uses synthetic negatives and is only a partial substitute for naturally occurring unexplained targets. Its event sequences are also relatively short (7.2 events per video on average), so it tests multi-event rather than long-form video reasoning. The proposal prompts also use causal wording (e.g., "true cause"), which may push proposals toward stronger explanations than the benchmark can support. Fifth, the first-event trigger convention (Section 3) is a design choice: annotators who do not follow it sometimes select more proximal mediators, multi-trigger labels may conflate individually sufficient triggers, jointly necessary triggers, and unresolved alternatives, and chain negatives may include valid but unannotated explanatory relations. The Relation-biased Transformer controls for explicit pairwise relations, but other architectural differences remain, so the gain of EventGraph-Trigger should not be attributed to message passing alone. A human-verified subset in which triggers cannot be determined from captions would be needed to test video understanding more directly. Finally, EventTriggerBench focuses on target-conditioned trigger attribution rather than full causal graph discovery. Future work can extend it with broader domains, larger-scale human auditing, richer multimodal annotations, and more explicit counterfactual validation.

B Human Validation Protocol

Sampling. We randomly sample 300 trigger samples from the held-out test set (697 samples). Each sample is judged independently by three annotators.

Audit of automatic annotations. For each sample, annotators view the event clips and captions, the target event, and the automatic trigger and explanatory chain, and answer four questions: (i) whether the selected target event is suitable for trigger reasoning; (ii) whether the annotated trigger plausibly explains the target; (iii) whether the explanatory chain is coherent; and (iv) whether the annotation should be accepted, revised, or rejected. A sample is counted as accepted when the majority of annotators accept it. To measure the effect of automated verification, annotators apply the same protocol to the raw LLM proposals for the same targets. Agreement on the accept/revise/reject decision is measured with Fleiss' κ (Fleiss, 1971). The verifier's precision is computed on the raw

proposals for the 300 audited targets, treating the human majority decision as the reference and mapping *accept* and *revise* to retained and *discard* to rejected, for both the verifier and the annotators. Trigger correctness and chain correctness are averaged over individual annotator judgments, whereas target suitability and overall acceptance are decided by majority vote per sample, which explains why acceptance can be slightly higher than trigger correctness. Table 8 gives the underlying counts.

Quantity	Numerator	Denominator	%
Target suitability (majority)	278	300 samples	92.7
Trigger correctness (per judgment)	786	900 judgments	87.3
Chain correctness (per judgment)	470	555 judgments (185 samples)	84.7
Overall acceptance, verified labels (majority)	265	300 samples	88.3
Overall acceptance, raw proposals (majority)	214	300 samples	71.3
Verifier precision (retainable)	232	259 retained proposals	89.6
Independent labels match automatic labels	233	300 samples	77.7
counting chain mediators as matches	254	300 samples	84.7

Table 8: Counts underlying the human validation results. The two 84.7% values are unrelated and have different denominators.

Cross-family verification. To partially decouple verification from the proposal model, we re-run the verification prompt of Appendix F on all 697 test labels with Qwen2.5-72B-Instruct at temperature 0. It accepts 78.3%, revises 13.1%, and discards 8.6% of the labels, thus retaining 91.4%; on the 300 audited samples, the 274 labels retained by both verifiers agree with the human majority in 81.8% of cases (224/274), compared with 34.6% (9/26) for labels rejected by the second verifier.

Independent human trigger labels. In a separate pass, annotators independently select the most plausible trigger event(s), without being told the first-event convention of Section 3, for each of the 300 samples after viewing the clips and captions. They are not shown the LLM proposal, verifier decision, confidence score, or explanatory chain. The majority vote forms the human reference label; for multi-trigger samples, a candidate is included in the reference set if at least two of the three annotators select it, and Fleiss' κ is computed over the per-candidate binary select/not-select decisions. The automatic label matches this reference in 77.7% of samples (set-level F1 of 71.2% on multi-trigger samples), leaving 67 samples in the human–LLM disagreement subset used in Section 5.3. Agreement on exact trigger selection is Fleiss' $\kappa = 0.68$. Significance is assessed with a paired bootstrap (Efron & Tibshirani, 1994) that resamples videos, so that targets from the same video are resampled together (Table 11).

Nature of human–LLM disagreements. We categorize the 67 disagreement samples by comparing the human reference set with the automatic trigger set and chain: in 29 (43.3%), the human trigger lies outside both the automatic trigger set and the automatic chain; in 21 (31.3%), annotators select an intermediate event of the automatic chain, i.e., a more proximal explanation of the same process; and in 17 (25.4%), the human and automatic trigger sets of a multi-trigger sample overlap only partially. Only the first category corresponds to a qualitatively different explanation; the second reflects ambiguity about the granularity of a trigger.

Cross-family re-annotation. To test whether our conclusions depend on the annotation model, we re-run the complete proposal, self-consistency, and verification pipeline of Section 3 with Qwen2.5-72B-Instruct on the 697 test targets, using the same prompts and settings. Its verifier accepts 61.8% (431), revises 25.3% (176), and discards 12.9% (90) of its own proposals, compared with 64.8/21.5/13.7% for GPT-4, leaving 607 of the 697 test targets with labels from both pipelines. On these 607 targets, the Qwen2.5 labels match the GPT-4 labels in 74.6% of cases (453/607). Of the 300 audited samples, 261 retain a Qwen2.5 label; on them, the Qwen2.5 labels match the human majority in 73.9% of cases (193/261), compared with 78.2% (204/261) for the GPT-4 labels. Table 9 evaluates models, trained on the GPT-4 training labels, against both label sets on the same 607 targets, so that all differences are paired. All models lose accuracy, GPT-4 CoT the most, and the ranking of the six evaluated models is identical under both label sets; we did not re-evaluate the remaining models, so we make no claim about their ranking.

Model	GPT-4 labels	Qwen2.5 labels	Δ
Transformer Encoder	51.1	46.3	-4.8
GAT	58.3	53.2	-5.1
Matched Multimodal Transformer	62.6	57.3	-5.3
CERN	67.4	61.8	-5.6
GPT-4 CoT	66.4	58.8	-7.6
EventGraph-Trigger	73.3	67.4	-5.9

Table 9: Accuracy against test labels produced by the original GPT-4 pipeline and by an independent Qwen2.5-72B proposer–verifier pipeline, on the 607 test targets labeled by both pipelines ($n = 607$; GPT-4-label accuracies therefore differ slightly from Table 5).

C EventGraph-Trigger Algorithm

Algorithm 1 summarizes the forward pass and training objective of EventGraph-Trigger (Equations 1–3), and Table 10 lists the implementation settings.

Edge terms. $a_{ij}^{temp} = \exp\big(-(j-i-1)/\sigma_t\big)$ decays with the number of intervening events; $a_{ij}^{sem} = \cos(h_i^{text}, h_j^{text})$ and $a_{ij}^{vis} = \cos(h_i^{video}, h_j^{video})$ are cosine similarities of caption and clip features; and $a_{ij}^{learned} = \sigma\big(\mathrm{MLP}([h_i; h_j; \phi(j-i)])\big)$, where ϕ embeds the temporal distance. The coefficients $\lambda_t, \lambda_s, \lambda_v, \lambda_l$ are learned, normalized with a softmax, and initialized uniformly. For sparsification, each event keeps its top-k incoming edges ($k = 4$) after masking future-to-past edges.

Target-conditioned message passing. At layer l, each event representation is updated as

$$g_i^{(l)} = \sigma\big(W_g[h_i^{(l)}; h_T^{(l)}]\big), \qquad h_i^{(l+1)} = \mathrm{ReLU}\Big(W_s h_i^{(l)} + g_i^{(l)} \odot \sum_{j \in \mathcal{N}(i)} \alpha_{ji} W_m h_j^{(l)}\Big), \tag{4}$$

where α_{ji} is the softmax of a_{ji} over the retained incoming edges $\mathcal{N}(i)$ and $\odot$ denotes element-wise multiplication. The trigger probability of a predecessor e_i is computed as $p(i \mid T) = \mathrm{softmax}_{i<T}\big(w^\top \tanh(W_z[h_i^{(L)}; h_T^{(L)}; h_i^{(L)} \odot h_T^{(L)}])\big)$, with a sigmoid in place of the softmax for multi-trigger samples.

Explanatory-chain decoding. The chain head outputs a probability $\hat{y}_{ij}$ for every forward pair $i < j$ within a sample. We decode the chain as the maximum-probability monotone path from the top-ranked trigger to the target, computed by dynamic programming over $\log \hat{y}_{ij}$ on the forward-only graph; since all edges point forward in time, the graph is acyclic and no cycles can arise. Only edges with $\hat{y}_{ij} > 0.5$ are allowed, and if no such path exists the chain defaults to the direct edge from trigger to target. The path length is therefore determined by the decoded path, and ties are broken in favor of the shorter path. During training, adjacent pairs along the annotated chain are positives and all other forward pairs within the sample are negatives. Because an annotated chain is a single LLM-proposed explanation, some of these negatives may be valid explanatory relations that were omitted; chain metrics therefore measure agreement with the annotated explanation rather than recovery of all plausible explanatory structures. Positive-unlabeled or soft-negative training for the chain head is a natural alternative that we leave to future work.

Calibration and ablation definitions. Because single-trigger samples are trained with a softmax and multi-trigger samples with per-candidate sigmoids over the same logits, we fit a single temperature T_c on the validation set by minimizing the binary cross-entropy of $\sigma(z_i/T_c)$ over all samples, and select the set-prediction threshold over $\{0.3, 0.4, \ldots, 0.7\}$ with T_c fixed (selected value: 0.5); the expected calibration error of the resulting scores is 0.041. *w/o target conditioning* removes the target gate from message passing and replaces h_T in the trigger scorer with the mean of all predecessor representations, so that candidate scores no longer depend on which event is the target.

Setting	Value
Text encoder (checkpoint)	all-mpnet-base-v2 (Sentence-Transformers)
Video encoder (checkpoint)	VideoMAE-Base, Kinetics-400 fine-tuned; 16 frames per clip
Encoders frozen or fine-tuned	Frozen
Hidden dimension	512
Number of graph layers	3
Graph sparsification / thresholds	Top-k incoming edges, $k = 4$; $\sigma_t = 2$
Edge weights $\lambda_t, \lambda_s, \lambda_v, \lambda_l$	Learned, softmax-normalized, initialized to 0.25
Chain loss weight β	0.5
Optimizer / learning rate	AdamW / 1×10^{-4}, weight decay 0.01
Batch size / training epochs	32 / 30, early stopping on validation accuracy
Hyperparameter search space	lr $\in \{5\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$, layers $\in \{2,3,4\}$, $k \in \{2,4,8\}$, $\beta \in \{0.25, 0.5, 1\}$; 20 random configurations
Proposal decoding	temperature 0.7, top-p 0.95 for the $K = 5$ proposals; temperature 0 for verification
Hardware / training time	1× NVIDIA A100 80GB / about 2 hours of training

Table 10: Implementation settings of EventGraph-Trigger.

Algorithm 1 EventGraph-Trigger: forward pass and training objective for one sample

Require: Events $\{e_i\}_{i=1}^{N}$ with captions, clips, and temporal intervals; target index T; trigger label(s) Y; optional explanatory chain C

for $i = 1$ to N **do**
  encode caption, clip, and interval into $h_i^{text}, h_i^{video}, h_i^{time}$
  $h_i \leftarrow \mathrm{MLP}([h_i^{text}; h_i^{video}; h_i^{time}])$ ▷ Eq. 1
end for
for each temporally valid pair $i < j$ **do**
  $a_{ij} \leftarrow \lambda_t a_{ij}^{temp} + \lambda_s a_{ij}^{sem} + \lambda_v a_{ij}^{vis} + \lambda_l a_{ij}^{learned}$ ▷ Eq. 2
end for
apply the temporal mask to remove future-to-past edges
for each reasoning layer **do**
  update all event representations by target-conditioned message passing given h_T
end for
predict the trigger probability of each predecessor e_i, $i < T$
if $|Y| = 1$ **then** ▷ training only; $|Y|$ is never used at inference
  $L_{\text{trigger}} \leftarrow$ cross-entropy over candidate predecessors
else
  $L_{\text{trigger}} \leftarrow$ binary cross-entropy per candidate
end if
$L_{\text{chain}} \leftarrow$ binary cross-entropy over adjacent edges of C if C is available, else 0
return $L = L_{\text{trigger}} + \beta\, L_{\text{chain}}$ ▷ Eq. 3
Inference: rank candidates by score; predicted set $= \{i < T : \sigma(z_i) > \text{threshold}\}$; decode the chain (Appendix C)

D Experimental Setup

Tasks. We evaluate three tasks: trigger attribution, trigger ranking, and explanatory-chain prediction. Trigger attribution measures whether the top-ranked predicted event matches the annotated trigger. Trigger ranking evaluates the ranking quality over all predecessor events. Explanatory-chain prediction measures agreement with the annotated explanatory path, which is one plausible LLM-proposed explanation rather than a unique ground-truth path.

Metrics. We use Accuracy, Top-k Accuracy, Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), and NDCG for trigger ranking. For multi-trigger examples, a prediction is correct if it matches any annotated trigger for single-answer metrics, so Accuracy corresponds to Hit@1; MAP and NDCG evaluate all annotated triggers. Chain-lenient accuracy additionally counts a top-ranked intermediate event of the annotated chain as correct. For explanatory chains, we report Edge F1,

Path Exact Match, Path Recall, and chain-length accuracy (Table 19), as well as edge-level Precision, Recall, and F1 for chain-edge recovery (Table 21). For multi-trigger samples, we additionally report set-level Precision, Recall, F1, and exact match (Table 17).

Data split. The dataset is split by video rather than by sample to avoid leakage across targets from the same video, with 2,437 / 348 / 697 trigger samples in the training, validation, and test splits. Videos (868 / 124 / 248) are assigned to splits greedily so that the sample counts follow the 70/10/20 ratio, which is why the sample proportions are close to exact. The splits contain 18.6% / 18.7% / 18.7% multi-trigger samples (453 / 65 / 130) and 61.6% / 61.5% / 61.7% samples with chain annotations (1,502 / 214 / 430), and the test split follows the distributions in Table 12 to within 0.5 points. Unless otherwise noted, all results are reported on the test split.

Baselines. All methods, including LLM prompts, receive only the events up to and including the target. The two causal-style baselines are our adaptations to EventTriggerBench rather than methods proposed under these names in the cited works. *Event Masking* is adapted from the mask-based event prediction of MECD (Chen et al., 2024): a masked event prediction model trained on EventTriggerBench captions predicts the target event from its predecessors, each candidate is masked in turn, and candidates are ranked by the resulting drop in the target's predicted likelihood. *Counterfactual Event Scoring* is inspired by the counterfactual questions of CLEVRER (Yi et al., 2020): each candidate is removed from the event sequence, a video-text model scores the plausibility of the target given the remaining events, and candidates are ranked by the decrease in plausibility. The *Matched Multimodal Transformer* uses the same fused event embeddings as EventGraph-Trigger (Equation 1), including the target event, and encodes them with a 4-layer Transformer without any event graph; trigger scores, the chain head, and all losses (including chain supervision) are identical to EventGraph-Trigger, and its width is chosen to match the parameter count of EventGraph-Trigger (14.2M vs. 13.8M). The *Relation-biased Transformer* is identical to the Matched Multimodal Transformer but adds the edge score a_{ij} of Equation 2, with the same four relation terms and learned weights, as an additive bias to the attention logit from event i to event j, and applies the same temporal mask; it has 14.3M parameters. *CERN* (Le et al., 2025) and *VGCM* (Chen et al., 2024) are re-implemented from the official code and use the same frozen text and video features as EventGraph-Trigger. To support arbitrary target positions, each sample is truncated after the target, so that the target becomes the last event of the input sequence as in the original formulations, and both models score every predecessor with a trigger classification head trained with the trigger loss of Section 4. They do not receive chain supervision because their architectures have no chain head. All learned models, including EventGraph-Trigger, use the same hyperparameter-search budget of 20 configurations on the validation set. CERN and VGCM have 11.6M and 9.8M parameters and train in about 1.5 and 1.2 hours on the hardware in Table 10. *Video-LLaVA-7B* (Lin et al., 2024) is prompted zero-shot with 8 uniformly sampled frames per event clip, the ordered event captions, and the target event, and ranks candidates by the log-likelihood of each candidate event identifier. *ReAct* (Yao et al., 2023): the agent initially observes only the target event and the list of event identifiers with timestamps, and must retrieve information through text actions: `Inspect[event_id]` returns an event's caption and interval, `Lookup[entity]` returns the events whose captions mention an entity, and `Finish[answer]` returns the ranked triggers, with at most 8 steps. Unlike GPT-4 CoT, which receives the full event sequence in a single prompt, ReAct must actively gather evidence. All GPT-4 baselines use `gpt-4-turbo-2024-04-09` with temperature 0. *Qwen2-VL-7B* (Wang et al., 2024) and *GPT-4o* (OpenAI, 2024) are prompted zero-shot with 8 frames per event clip, the ordered captions, and the target event, and output a ranked list of candidate triggers.

E Additional Analyses

Event construction details. For the non-annotated videos, TransNetV2 boundaries with confidence above 0.5 are used; adjacent shots shorter than 2 s are merged with their neighbor, and segments shorter than 1 s or with failed decoding are removed. Video-LLaVA-7B captions are generated with greedy decoding from 8 uniformly sampled frames and cleaned by removing duplicated sentences and filler phrases.

Paired tests. Table 11 reports video-clustered paired bootstrap comparisons.

vs. EventGraph-Trigger	Set	Δ	95% CI
GPT-4 CoT	Human	6.7	[2.4, 11.0]
CERN	Human	5.7	[1.5, 10.0]
Matched MM Trans.	Human	10.4	[6.0, 14.7]
ReAct	Auto	5.3	[3.0, 7.7]
CERN	Auto	5.8	[3.4, 8.1]
Matched MM Trans.	Auto	10.6	[8.0, 13.1]
Relation-biased Trans.	Auto	7.4	[4.9, 9.9]

Table 11: Video-clustered paired bootstrap (10,000 resamples) on the human-labeled (Human) and full automatic (Auto) test sets.

Dataset distributions. Table 12 reports the distributions of sequence length, trigger distance, explanatory-chain length, and video duration.

Statistic	Bin	% of samples
Events per video	3–5	30.0
	6–8	43.9
	9+	26.1
Trigger distance (events)	1	33.8
	2	28.9
	3–4	26.8
	$\geq$5	10.5
Explanatory-chain length	2	27.4
	3	41.8
	$\geq$4	30.8
Video duration	$<$60 s	18.7
	60–180 s	58.4
	$>$180 s	22.9

Table 12: Distributions of EventTriggerBench statistics. Mean video duration: 124.6 s; median trigger distance: 2 events; median sequence length: 7 events.

Sensitivity to graph sparsification. Table 13 varies the number k of retained incoming edges per event.

k	2	3	4	6	8	Full
Acc.	70.4	71.9	72.7	72.5	71.7	70.9
NDCG	85.1	86.1	86.7	86.5	86.0	85.4

Table 13: Sensitivity of EventGraph-Trigger to the sparsification level k ("Full" keeps all predecessor edges).

Difficulty analysis. Table 14 breaks down accuracy by trigger distance, sequence length, explanatory-chain length, and trigger multiplicity; the last column gives the gain of EventGraph-Trigger over the Matched Multimodal Transformer (MM Trans.).

Contribution of visual evidence. Comparing the full EventGraph-Trigger with its w/o-video variant on the 697 test samples, 68 samples (9.8%) are answered correctly only with video features and 45 (6.5%) only without them, which accounts for the net difference of 23 samples (3.3 points) in Table 7. The video-dependent samples are dominated by targets whose captions omit object state or contact (e.g., whether a ball touched the net), consistent with the edge-recovery results in Table 21. On the 142 test samples where both EventGraph-Trigger w/o video and GPT-4 CoT fail, the full EventGraph-Trigger is correct on 30.3%, GPT-4o on 23.9%, and the Matched Multimodal Transformer on 21.1%. Because this subset is defined by model failures, it is a diagnostic rather than

Setting	Transformer	GAT	LLM-CoT	MM Trans.	CERN	Ours	Gain vs. MM
Short distance	58.1	64.0	71.6	68.6	72.9	**76.7**	8.1
Medium distance	53.7	61.2	68.2	64.7	70.6	**74.1**	9.4
Long distance	41.5	49.6	58.8	54.2	58.5	**68.1**	13.9
3–5 events	56.5	62.7	69.9	67.5	71.3	**76.1**	8.6
6–8 events	51.3	58.5	67.3	61.8	66.7	**73.5**	11.7
9+ events	42.9	51.1	58.8	56.6	62.1	**67.6**	11.0
Chain length 2	55.1	61.9	69.5	66.9	71.2	**76.3**	9.4
Chain length 3	50.6	57.8	66.1	61.7	66.7	**73.3**	11.6
Chain length $\geq$4	44.7	50.8	59.8	56.1	59.8	**68.9**	12.8
Single trigger	52.0	59.1	67.4	63.7	68.4	**73.7**	10.0
Multi-trigger	44.6	52.3	59.2	55.4	60.0	**68.5**	13.1

Table 14: Difficulty analysis on the test split by trigger distance (short: 1, medium: 2, long: $\geq$3 events; 236/201/260 samples), sequence length (209/306/182 samples), explanatory-chain length (118/180/132 of the 430 chain-annotated samples), and trigger multiplicity (567/130 samples). All values are single-run accuracies computed from integer counts; within the distance, sequence-length, and multiplicity blocks, the numbers of correct predictions sum exactly to those underlying Table 5.

a benchmark split. Visual evidence therefore changes the attribution for a minority of samples, and most of the benchmark signal is carried by caption semantics.

Caption ablations. Table 15 tests how much the benchmark relies on captions. *Masked* replaces all captions with an empty string, so that multimodal models rely on video and timestamps only; *Incomplete* randomly deletes 50% of the content words in every caption. To separate the effect of caption source from that of the video source, *Human capt.* and *Generated capt.* evaluate the same test videos that come with human-written dense captions, using the original captions and captions generated by Video-LLaVA-7B, respectively. Text-only models cannot be evaluated in the masked setting.

Model	Full	Masked	Incomplete	Human capt.	Generated capt.
GPT-4 CoT (text)	65.9	–	51.2	67.8	63.4
GPT-4o (video+text)	64.3	43.5	55.1	65.9	62.7
Matched Multimodal Transformer	62.1	42.8	52.5	63.8	60.1
EventGraph-Trigger	72.7	49.6	61.7	74.1	70.9

Table 15: Accuracy under caption ablations.

Semantics of multi-trigger labels. A sample is marked as multi-trigger when the K proposals select different triggers that together form a consistent set (Appendix F). Such sets may contain individually sufficient triggers, jointly necessary triggers, or plausible alternatives that the annotation process cannot distinguish, and the current labels do not separate these cases. Because single-answer metrics count a top-ranked prediction as correct if it matches any annotated trigger, we also report set-level metrics below.

Trigger granularity. Under the first-event convention (Section 3), predicting a mediator of the annotated chain counts as an error, although the mediator may be a plausible proximal explanation. Table 16 additionally reports chain-lenient accuracy, which also accepts intermediate chain events; samples without chain annotations are scored as in Table 5.

Multi-trigger evaluation. Table 17 evaluates the complete predicted trigger set on multi-trigger samples, where each model predicts a set of triggers. Set-level F1 of EventGraph-Trigger varies by at most 1.4 points across thresholds in $\{0.4, 0.5, 0.6\}$ and is similar for samples with 2 and with 3 or more annotated triggers (69.4 vs. 67.8).

Model	Acc.	Chain-lenient Acc.
Transformer Encoder	50.6	58.8
GAT	57.8	65.7
Matched Multimodal Transformer	62.1	70.4
CERN	66.9	74.7
GPT-4 CoT	65.9	74.2
EventGraph-Trigger	72.7	80.1

Table 16: Standard and chain-lenient accuracy on the test set.

Model	Precision	Recall	F1	Set exact match	MAP
Transformer Encoder	55.3	47.1	50.9	22.3	52.8
GAT	60.2	52.7	56.2	27.7	58.6
GPT-4 CoT	66.4	58.9	62.4	33.8	66.0
Matched Multimodal Transformer	63.1	56.0	59.3	30.0	62.7
CERN	67.9	60.3	63.9	34.6	67.8
EventGraph-Trigger	71.8	66.2	68.9	41.5	72.1

Table 17: Set-level evaluation on the 130 multi-trigger test samples (18.7% of the test set; 18.6% benchmark-wide).

Insufficient-evidence stress test. Because every target has a plausible trigger by construction, we build two sets of synthetic negatives from the test split. *Transplanted*: the predecessors of each target are replaced by the first $T-1$ events of a different test video, so that no observed event is likely to explain the target. *Trigger removed*: all annotated triggers and chain events are deleted from the predecessors; because unannotated plausible triggers may remain, this set is noisier. Scored models abstain when the maximum calibrated trigger score falls below a threshold chosen on analogous validation negatives such that at most 5% of the original validation samples are rejected; GPT-4 CoT receives an explicit "none of the above" option. Table 18 reports the fraction of negatives on which each model abstains, and accuracy on the original test samples with abstention counted as an error.

Model	Abstention (transplanted)	Abstention (trigger removed)	Acc. on original (with abstention)
GPT-4 CoT (+ none option)	71.3	38.6	64.8
Matched Multimodal Transformer	62.4	29.7	60.3
CERN	66.7	31.4	65.0
EventGraph-Trigger	78.5	41.2	70.9

Table 18: Insufficient-evidence stress test with synthetic negatives.

Explanatory-chain prediction. Table 19 compares chain prediction on samples with explanatory-chain annotations.

Robustness. We evaluate robustness under controlled event-level perturbations: distractor injection, caption paraphrase, event omission, order perturbation, and visual-text mismatch. Distractor injection inserts events from other test videos at random positions that are not part of the annotated trigger or chain; caption paraphrase rewrites every caption with an LLM while preserving its meaning; and visual-text mismatch replaces the clip of one randomly chosen event that is neither the target, a trigger, nor a chain event with a clip from another video. Because the target, triggers, and chain events are left unchanged in these three settings, we treat them as preserving the annotated relation. Event omission removes a random non-target event, and order perturbation swaps two adjacent predecessor events. Each perturbation is applied once per test example with a fixed random draw. Table 20 shows that caption paraphrasing has only a small effect, while distractors, event omission, order perturbation, and visual-text mismatch cause larger drops because they disrupt the explanatory structure or multimodal consistency. Caption paraphrase, distractor injection, and visual-text mismatch preserve the annotated trigger relation, and EventGraph-Trigger performs best under all three. Event omission and order

Model	Edge F1	Path Exact Match	Path Recall	Chain-length Acc.
GAT	65.4	29.5	61.2	55.8
GPT-4 CoT	70.8	35.1	67.5	64.0
Matched Multimodal Transformer	68.9	32.8	65.3	58.6
EventGraph-Trigger w/o chain supervision	71.3	36.3	68.0	60.5
EventGraph-Trigger	82.6	48.6	79.4	71.6

Table 19: Explanatory-chain prediction on the 430 test samples with chain annotations. All metrics measure agreement with the annotated explanatory path, not recovery of the underlying causal structure.

perturbation may alter the underlying explanation, so evaluating them against the original label cannot separate model brittleness from a genuine change in the task; we report them for completeness only and do not use them as evidence of robustness.

Perturbation	Transformer	GAT	LLM-CoT	Ours
Original	50.6	57.8	65.9	**72.7**
Distractor injection	43.2	50.1	58.7	**68.6**
Caption paraphrase	48.9	55.5	63.8	**71.4**
Event omission	40.7	47.8	55.2	**65.9**
Order perturbation	38.5	45.2	53.5	**64.7**
Visual-text mismatch	41.9	49.4	57.1	**66.3**

Table 20: Robustness under controlled event-level perturbations.

Condition	Precision	Recall	F1
Text-only	79.6	75.9	77.7
Video-only	76.8	72.5	74.6
Video+Text	**84.1**	**81.2**	**82.6**
With distractors	78.9	74.6	76.7
Long event sequences	75.3	71.8	73.5

Table 21: Explanatory-chain edge recovery. Positive edges are adjacent event pairs along verified explanatory chains, not full event-graph ground truth.

Explanatory-chain edge recovery. Although EventTriggerBench does not provide complete event-graph annotations for every video, samples with explanatory-chain labels provide partial edge-level supervision. We therefore evaluate whether the constructed event graph can recover verified chain edges. For each sample with a chain annotation, we treat adjacent event pairs along the verified chain as positive edges; for example, a chain $E_3 \rightarrow E_5 \rightarrow E_6 \rightarrow E_8$ yields positive edges (E_3, E_5), (E_5, E_6), and (E_6, E_8). Other predecessor-to-successor event pairs within the same sample are treated as candidate negatives. Table 21 reports edge-level Precision, Recall, and F1. Video+Text edge recovery outperforms both Text-only and Video-only settings, showing that visual evidence provides complementary cues beyond event captions, especially when captions fail to capture object continuity, motion, or interaction changes. Performance decreases with distractor events and long event sequences. This evaluation only measures whether predicted event relations recover the annotated chain edges; it does not assume that EventTriggerBench provides full graph-level supervision.

Error typology. We categorize incorrect predictions of Transformer, GAT, GPT-4 CoT, and EventGraph-Trigger into five types: target misunderstanding, temporal shortcut, distractor overfitting, causal-chain (explanatory-chain) break, and visual evidence failure. Figure 4 reports absolute counts. Counts are computed from the test predictions of the same model checkpoint (or, for GPT-4 CoT, the same run) that is evaluated in Table 5. Each error is checked against five criteria: *target misunderstanding*, the prediction explains a different event than the target; *temporal shortcut*, the prediction is the immediately preceding or temporally nearest event without explanatory support;

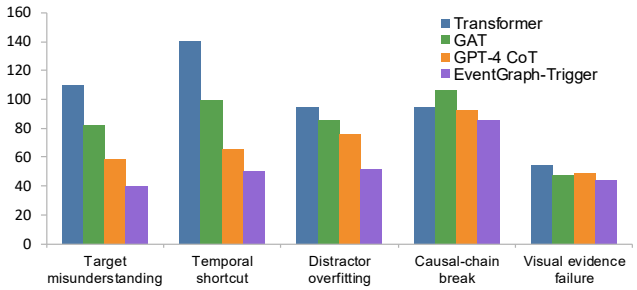

Figure 4: Error typology by absolute count of category labels. Categories are not mutually exclusive, so each model's bars sum to more than its number of errors. We compare Transformer, GAT, GPT-4 CoT, and EventGraph-Trigger across five failure categories: target misunderstanding, temporal shortcut, distractor overfitting, causal-chain break, and visual evidence failure.

distractor overfitting, the prediction is semantically related to the target but not part of its explanation; *explanatory-chain break*, the prediction is a chain event or connects to the target only through a missing link; and *visual evidence failure*, the correct attribution depends on visual information absent from the captions. The taxonomy is multi-label, and an error receives every category whose criterion it meets (e.g., a temporal shortcut that also breaks the chain). This yields 458 labels for 344 errors (Transformer), 390 for 294 (GAT), 317 for 238 (GPT-4 CoT), and 253 for 190 (EventGraph-Trigger), so the bars of each model sum to more than its number of errors. EventGraph-Trigger produces fewer errors than all baselines in every category. Transformer is most affected by temporal shortcuts, suggesting that sequence-only models often select nearby events without explanatory justification. GAT reduces some shortcut errors through event structure, while GPT-4 CoT remains vulnerable to semantically plausible distractors. The remaining errors of EventGraph-Trigger are mostly chain breaks, indicating that multi-hop reasoning over non-adjacent events remains the hardest case.

F Prompt Templates

We reproduce the prompts used in the construction pipeline verbatim. Some prompts and their intermediate output fields retain the word "causal" (e.g., `causal_chain`, `causal_consistency`); these correspond to the explanatory chain and the counterfactual-plausibility check, respectively. In the released benchmark JSON (Appendix F.5), the chain is stored in the `explanatory_chain` field.

F.1 Target Selection Prompt

Instruction. Given an ordered list of video events, identify which events are suitable target events for trigger reasoning. A suitable target event should be meaningful, outcome-like, state-changing, or naturally answerable by a "why did this happen?" question. Do not select events that are purely descriptive, static, repetitive, or have no plausible earlier cause.

Selection Criteria. An event should be selected as a target if it satisfies most of the following conditions:

- It describes an outcome, consequence, state change, completion, failure, or abnormal event.
- It can plausibly be explained by one or more earlier events in the same video.
- It is not simply a background description or a static scene observation.
- It is not the first event in the video, since no earlier trigger is available.
- It can be naturally converted into a causal question, such as "Why did this event happen?"

Prompt.

> You are given an ordered sequence of events from a video. Each event has an event ID, timestamp, and short description.
>
> Your task is to identify which events are suitable **target events** for trigger reasoning. A target event should be an event whose occurrence can be meaningfully explained

by earlier events. It should usually describe an outcome, state change, failure, success, reaction, or consequence.

Do not select events that are only descriptive, repetitive, or unlikely to have a meaningful earlier cause.

For each selected target event, provide:

1. the target event ID;
2. a short reason why it is suitable;
3. whether it is a local target, long-range target, or final-outcome target.

Events:

{EVENT_LIST}

Output format:

```
{
  "selected_targets": [
    {
      "target_event": "E5",
      "reason": "This event is an outcome caused by earlier actions.",
      "target_type": "final-outcome"
    }
  ],
  "rejected_events": [
    {
      "event": "E1",
      "reason": "This is an initial setup event with no earlier cause."
    }
  ]
}
```

F.2 Trigger Proposal Prompt

Instruction. Given an ordered list of video events and a selected target event, identify the earlier event or events that causally trigger the target. A trigger is not merely an event that happens before the target; it should be an event whose absence would likely prevent, weaken, or significantly change the target event.

Trigger Definition. We define a trigger event as an earlier event that satisfies the following criteria:

- It occurs before the target event.
- It causally contributes to the target event.
- Removing or changing it would likely change whether, when, or how the target occurs.
- It is more explanatory than a merely adjacent or semantically similar event.

Prompt.

You are given an ordered sequence of video events and a target event. Your task is to identify the earlier event or events that triggered the target event.

A trigger event is an event whose absence would likely prevent or significantly change the target event. Do not simply choose the event immediately before the target unless it is truly causal. If the target is caused through intermediate events, provide the full causal chain.

For each target, return:

1. the main trigger event;
2. any additional trigger events, if applicable;
3. the causal chain from trigger to target;
4. a short explanation;
5. a confidence score from 0 to 1.

Events:

{EVENT_LIST}

Target event: {TARGET_EVENT}
Output format:

```
{
  "target_event": "E8",
  "main_trigger": "E3",
  "additional_triggers": [],
  "causal_chain": ["E3", "E5", "E6", "E8"],
  "explanation": "E3 causes the red player to miss the ball. This leads to the blue
    player winning the point, celebrating, and then mocking the opponent.",
  "confidence": 0.87
}
```

F.3 Automated Verification Prompt

Instruction. After the trigger proposal step, we use a separate verification prompt to check whether the proposed target, trigger, and causal chain are valid. The verifier is asked to judge temporal validity, causal consistency, chain coherence, and confidence.

Prompt.

You are a verifier for a video trigger reasoning dataset. You are given an ordered list of events, a target event, a proposed trigger event, and a proposed causal chain.

Your task is to determine whether the proposed annotation is valid.

Check the following:

1. **Temporal validity**: Does the trigger occur before the target? Are all events in the causal chain temporally ordered?
2. **Causal consistency**: Does the trigger plausibly cause or contribute to the target, rather than merely occur before it?
3. **Chain coherence**: Does each event in the chain plausibly lead to the next event?
4. **Counterfactual plausibility**: If the trigger were removed, would the target become less likely or change significantly?
5. **Ambiguity**: Are there multiple plausible triggers, or is the proposed trigger uncertain?

Return one of three decisions:

- `accept`: the annotation is valid;
- `revise`: the annotation is mostly valid but needs correction;
- `discard`: the annotation is invalid or too ambiguous.

Events:
{EVENT_LIST}
Target event: {TARGET_EVENT}
Proposed trigger: {TRIGGER_EVENT}
Proposed causal chain: {CAUSAL_CHAIN}
Output format:

```
{
  "decision": "accept",
  "temporal_validity": true,
  "causal_consistency": true,
  "chain_coherence": true,
  "counterfactual_plausibility": true,
  "ambiguity": "low",
  "revised_trigger": "E3",
  "revised_chain": ["E3", "E5", "E6", "E8"],
  "explanation": "The trigger occurs before the target and plausibly initiates the
    causal chain.",
  "confidence": 0.91
```

```
}
```

F.4 Self-Consistency Prompting

To reduce noise from a single LLM generation, we query the proposal model multiple times using paraphrased prompts or different decoding seeds. A sample is retained only if the predicted trigger and causal chain are stable across generations or pass the verifier with high confidence.

Prompt Variant.

> Given the following video event sequence, answer the causal question carefully. Identify the event that best explains why the target event happened. Consider whether each earlier event is a true cause, an intermediate step, a distractor, or merely temporally adjacent.
>
> **Events:**
>
> {EVENT_LIST}
>
> **Question:** Why did {TARGET_EVENT} happen? Which earlier event triggered it?
>
> Return the answer as:

```
{
  "trigger": "E3",
  "causal_chain": ["E3", "E5", "E6", "E8"],
  "reasoning": "E3 changes the rally outcome, which leads to the blue player winning
    and then mocking the opponent.",
  "confidence": 0.88
}
```

Aggregation Rule. For each target event, we collect K independent proposals. We accept a trigger annotation if:

- the same trigger is selected by a majority of proposals;
- the average verifier confidence exceeds a threshold τ;
- the causal chain satisfies temporal and semantic consistency checks.

If the proposals disagree but contain a consistent set of multiple triggers, the sample is marked as multi-trigger. If the disagreement is severe, the sample is discarded.

F.5 Final JSON Schema

Each verified sample is stored in the following format:

```
{
  "sample_id": "v_001_t_E8",
  "video_id": "v_001",
  "events": [
    {
      "event_id": "E1",
      "start": 0.0,
      "end": 4.2,
      "caption": "The red player sets up for the rally.",
      "clip_path": "clips/v_001/E1.mp4"
    },
    {
      "event_id": "E2",
      "start": 4.2,
      "end": 8.9,
      "caption": "The rally continues.",
      "clip_path": "clips/v_001/E2.mp4"
    }
```

```
  ],
  "target_event": "E8",
  "trigger_event": ["E3"],
  "explanatory_chain": ["E3", "E5", "E6", "E8"],
  "metadata": {
    "target_type": "reaction",
    "trigger_distance": 5,
    "chain_length": 4,
    "multi_trigger": false,
    "verification_decision": "accept",
    "verification_confidence": 0.91
  }
}
```