# M$^3$EC: A Multimodal Multidocument Benchmark for Event Extraction and Event Coreference Resolution

**Zaber Ibn Abdul Hakim, Najibul Haque Sarker, Aafiya Hussain, Hani Alomari, Alvi Md Ishmam, Ali Asgarov, Chia-Wei Tang, Chris Thomas**
Department of Computer Science, Virginia Tech
(zaberhakim666, najibulhaque, aafiyahussain, hani, alvi, aliasgarov, cwtang, christhomas)@vt.edu

**Dataset and code:** github.com/zaber666/m3ec-event-understanding

## Abstract

Understanding how the same real-world event is represented across text, video, and audio remains a fundamental challenge in information extraction. While prior datasets have explored events within individual modalities or limited modality pairs, a setting that combines all three modalities remains largely unexplored. We introduce M$^3$EC (**M**ulti**M**odal **M**ultidocument **E**vent & **C**oreference), a new benchmark for multimodal, multidocument event extraction and coreference resolution. It includes eventful videos, audio files, and news articles, grouped into clusters of related files. We also introduce a semi-automated annotation pipeline, AutoEventAnnotator, to generate automated annotations. Additionally, we construct a human-annotated test set for robust evaluation. To our knowledge, this is the first benchmark to support cluster-level multidocument event extraction and event coreference across text, video, and audio. Evaluation with strong multimodal LLMs shows that the majority of them fail to extract valid events, particularly in video and audio. We further show that model fine-tuned on the annotations generated by our automated pipeline achieve consistent improvements, including more than +30 F1 on video event extraction, +15 F1 on audio event extraction, and +40 CoNLL F1 on end-to-end coreference resolution, highlighting the usefulness of M$^3$EC for training and evaluation in multimodal event understanding.

## 1 Introduction

Real-world events are rarely described in a single source. A breaking news event may appear as a written article, a video clip showing the action, and an audio narration providing context. While each modality captures part of the event, none provides a complete description. Building systems that can extract events from such multimodal sources (**Event Extraction**) and determine whether different mentions refer to the same real-world occurrence (**Coreference Resolution**) remains a fundamental challenge.

While early progress in this domain was driven by text-only single-document datasets such as ACE 2005 (Walker et al., 2006), which covered 33 event types within a limited ontology, later benchmarks, including MAVEN (Wang et al., 2020), MAVEN-ERE (Wang et al., 2022), and ECB+ (Cybulska and Vossen, 2014), expanded event diversity and cross-document reasoning. Multimodal datasets such as M$^2$E$^2$ (Li et al., 2020) and VM$^2$E$^2$ (Chen et al., 2021) further demonstrate the value of visual grounding, but are limited to a single document or predefined article-image or article-video pairs with aligned information. More broadly, multimodal grounding research (Lin et al., 2023; Gu et al., 2024; Wang et al., 2025b) focuses on localizing visual or audio segments relevant to a textual query. However, grounding identifies where a described event occurs in a video or audio stream, while event coreference asks whether an event mentioned in one article is the same real-world event as one mentioned in another modality or document, requiring reasoning over participants, temporal consistency, and context.

Despite these advances, existing tasks fall short of modeling event understanding in real-world settings. As illustrated in Figure 1, correctly identifying event coreference often requires reasoning within and across documents and modalities. Prior formulations typically rely on some simplifying assumptions. Some assume **(1)** document isolation, where events are extracted and referenced within a single document (Wang et al., 2022). Others assume **(2)** modality isolation or pairing, where models operate on a single modality or predefined text–media pairs (Ayyubi et al., 2024; Li et al., 2020; Martinelli et al., 2025). A third line studies **(3)** grounding without coreference, localizing events



*Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 56843–56904
October 24-29, 2026 ©2026 Association for Computational Linguistics

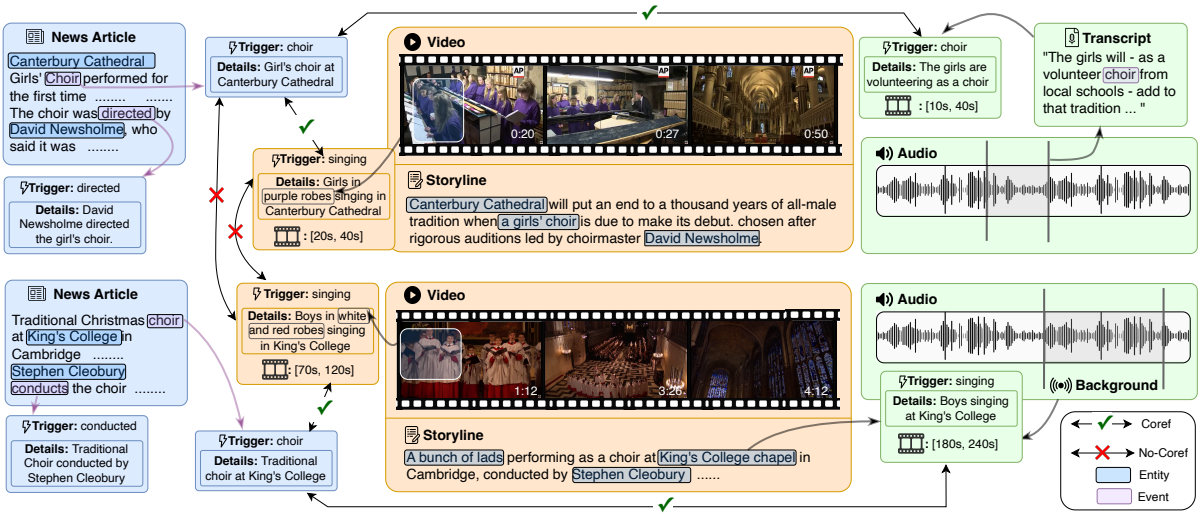



Figure 1: Example of multimodal, multidocument event extraction and coreference. The same triggers (e.g., choir, singing) appear across news articles, video, and audio, but refer to distinct events involving different entities and locations. Correctly resolving coreference requires reasoning across documents and modalities.

in media without linking them across documents or modalities (Zhang et al., 2024; Lin et al., 2023; Wang et al., 2025b). Finally, some work relies on **(4)** fixed or narrow event ontologies, limiting generalization to open-domain events (Liu et al., 2025; Wang et al., 2020, 2024; Sanders et al., 2024).

Moreover, no existing dataset supports event coreference across all modality combinations simultaneously, including *text–text*, *video–video*, *audio–audio*, and cross-modal pairs such as *text–video*, *text–audio*, and *video–audio*. Appendix F and H contain a detailed discussion of additional shortcomings of existing datasets. These limitations prevent joint event reasoning across heterogeneous modalities and documents, motivating a new benchmark that captures these interactions.

To support this task, we introduce M$^3$EC, a large-scale benchmark built from real-world news clusters containing text, video, and audio. It includes a large automatically annotated training set. It also contains a smaller human-annotated test set and an AI-generated test set, enabling systematic evaluation while supporting learning-based approaches. Event identification follows open-ended ACE-style trigger rules rather than a fixed ontology, encouraging generalization beyond predefined event categories. The dataset contains 3.61M event mentions and 53.24M coreference links across modalities and documents, with broad lexical event coverage (50.8% PropBank (Palmer et al., 2005), 86% NomBank (Meyers et al., 2004), 81.4% VerbNet (Kipper Schuler, 2005)), making it substantially larger than prior resources.

We construct the dataset using a semi-automated annotation pipeline, **AutoEventAnnotator**, which combines multimodal foundation models with multiple verification and filtering stages. The pipeline produces event and coreference annotations (**silver labels**) at scale, while enabling targeted human intervention to create a high-quality evaluation set (**gold labels**). Experiments show that even strong multimodal models struggle with consistent cross-modal and cross-document event extraction, underscoring the task's difficulty. Fine-tuning an open-weight foundation model with silver data yields substantial improvements, in many cases exceeding the performance of proprietary models.

In short, our contributions are as follows:

- We define multimodal, multidocument event extraction and coreference resolution as a unified task, requiring models to identify and link events across text, video, and audio in heterogeneous news clusters.
- We introduce M$^3$EC, a large-scale weakly labeled benchmark with a human-annotated test set to support the proposed task.
- We present AutoEventAnnotator, a semi-automated annotation pipeline for labeling multimodal, multidocument event extraction and coreference data at scale.
- We evaluate strong foundation models on the benchmark to characterize the difficulty of the

task, and fine-tune an open-weight foundation model on the silver labels to demonstrate the practical utility of the annotations produced by AutoEventAnnotator.

## 2 Related Works

Event extraction and event coreference resolution have been widely studied in NLP. Early benchmarks such as ACE 2005 (Walker et al., 2006) and TAC KBP (Mitamura et al., 2015) focused on identifying event mentions and resolving coreference within text using small ontologies and feature-based models. Later datasets such as MAVEN (Wang et al., 2020) expanded event coverage and enabled neural approaches, while ECB+ (Cybulska and Vossen, 2014) introduced cross-document event coreference in a text-only setting. Recent methods explore pipeline, joint, and generative formulations for event extraction and coreference (Huang and Peng, 2021; Lu and Ng, 2021; Wadden et al., 2019; Kriman and Ji, 2021), but remain largely limited to unimodal text.

More recent work has extended event understanding beyond text by incorporating visual and audio signals. Multimodal datasets such as $M^2E^2$ (Li et al., 2020), $VM^2E^2$ (Chen et al., 2021), VidSitu (Sadhu et al., 2021), VidEvent (Liang et al., 2025), and SpeechEE (Wang et al., 2024) study event extraction or grounding in video or speech, often in conjunction with paired text. However, these datasets typically operate at the level of single documents or predefined media–article pairs and provide limited support for event-level coreference, especially across modalities.

The NewsStories corpus (Tan et al., 2022) moves toward multimodal news understanding by clustering articles, images, and videos, but relies on weak alignment signals and lacks event-level or coreference annotations. REAMO (Zhang et al., 2024) proposes a unified multimodal information extraction framework across text, image, audio, and video, but its benchmark largely consists of single-instance examples, includes synthetic modality generation, and focuses mainly on grounding rather than cross-document or cross-modal event coreference. Other efforts such as Multi-HiEve (Ayyubi et al., 2024) explore multimodal event coreference between videos and paired text articles, but assume predefined pairs and do not support multidocument clusters or audio modalities.

In contrast to prior work, our benchmark targets multimodal, multidocument event extraction and event coreference resolution, requiring models to jointly reason over events across text, video, and audio in realistic news clusters.

## 3 Definitions and Task

**Events:** An *event* is a real-world occurrence (an action, a state change, or a happening) involving one or more participants. Each event has two core components: a *trigger* $\tau$, the word or phrase that directly expresses the occurrence, and *entities*, the participants (people, locations, times, etc.) that make the event unique. Within a document, an event is localized by $\ell$. For text, $\ell = (s, e)$ denotes the start and end token indices of the trigger span. For video and audio documents, $\ell = (t_s, t_e)$ denotes the temporal interval of the event. Each event also has an LLM-generated *phrase* $p$ that summarizes the event with globally salient entities from the full document context. This phrase absorbs the event entities and serves as the unit of comparison for coreference across modalities. Formally, an event is represented as a triple $e = (\ell, \tau, p)$.

**Coreference:** Two event mentions corefer if they describe the same real-world event. Coreference requires that the triggers are semantically similar and that the entities across the two mentions are mutually consistent. Coreference decisions are made by comparing the phrases $p$ of the two mentions, regardless of their source modalities, enabling uniform reasoning across all modality pairs without modality-specific comparison models. Coreference links may exist between events within the same document or across different documents, including across modalities.

**News Clusters:** A *news cluster* $\mathcal{C} = \{D_1, D_2, \ldots, D_N\}$ is a collection of text articles, videos, and audio documents that share one or more common real-world events. For example, a cluster about the Southern California wildfires would contain multiple articles, video clips, and audio files covering that event, each also containing document-specific events of its own. Organizing documents into clusters ensures that cross-document and cross-modal coreference relations are present by construction. An example of a news cluster, multimodal documents, events, and coreferences is portrayed in Figure 1.

**Task, Input, and Output:** The input to our task is a news cluster $\mathcal{C}$. The formal task is the extraction of event mentions from all the documents and all

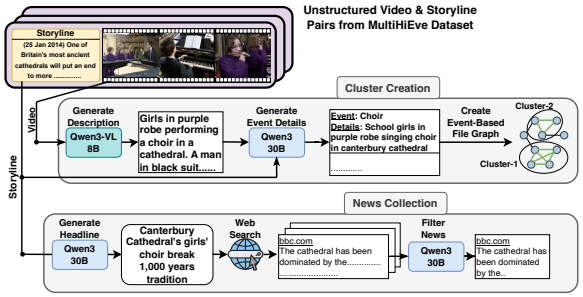



Figure 2: Overview of the pipeline for constructing multimodal news clusters from unstructured video and storyline pairs, combining event-based cluster formation with targeted news article retrieval.

the coreference links within the cluster. The output is **(1)** a set of event mentions $\{e_i\} = \{(\ell_i, \tau_i, p_i)\}$ extracted from all documents in the cluster, and **(2)** a binary coreference label for every event pair $(e_i, e_j)$ in the cluster.

## 4 Data Collection and Label Generation

In the whole data collection and label generation procedure, we use Qwen3-30B (Yang et al., 2025) as the LLM, Qwen3-VL-8B (Bai et al., 2025a) as the video LLM, and Qwen2-Audio-7B (Chu et al., 2024) as the audio LLM unless otherwise specified.

### 4.1 News Cluster Creation

We construct clusters around news videos for three reasons. First, videos can be curated to contain rich visual actions and usable entity information. Second, audio is extracted directly from the video track, ensuring natural alignment. Finally, given the video and audio modalities, retrieving related text articles via web search using video details as a query is feasible at scale, whereas the reverse direction, web-scale video search, is not. An overview of the process is shown in Figure 2.

We build on the MultiHiEve video dataset (Ayyubi et al., 2024), which pairs each video with an editor-written YouTube description ("storyline"). Storylines provide a critical advantage: while video alone yields only visual entities, storylines supply explicit named entities, such as actors, locations, and organizations, that are essential for distinguishing between otherwise visually similar events.

**(1)** For each video, we prompt an LLM to extract event triggers following ACE-style rules (Linguistic Data Consortium, 2005) and their associated entities. Each trigger and its entities are then converted into a concise descriptive phrase $p$ (e.g., *"police arrested a suspect in New York"*), forming a normalized event representation used in all downstream steps. **(2)** Of the original 400K videos, we retain approximately 330K by removing videos that does't contain observable real-world events. This removes non-eventive videos, including studio commentary, interviews without actions, and generic reporting segments. **(3)** To identify related videos, we match events across the dataset using a two-stage approach: dense embedding similarity (Chen et al., 2024) retrieves the top-10 candidate matches per event (producing $\sim$18M candidate pairs), which are then verified by an LLM comparing triggers, entities, and context. This design ensures scalability without sacrificing precision. **(4)** Verified matches form a weighted graph where each video forms a node and edge weights reflect the number of shared events. We retain only edges with weight greater than a 3 and apply the Bron–Kerbosch algorithm (Bron and Kerbosch, 1973) to find maximal cliques, each forming a video cluster. Clique-based clustering guarantees that every video in a cluster shares verified events with other videos. Audio tracks are extracted directly from the clustered videos; non-English or acoustically empty files are discarded.

Finally, for each video in the cluster, we generate a search query from the video storyline and description that is used in the Google Search API (Google for Developers, 2025) for relevant article retrieval. Retrieved articles are verified by an LLM for relevance to the cluster's events, yielding text documents grounded in the same real-world events present in the video and audio tracks. More details are provided in Appendix G.

### 4.2 AutoEventAnnotator: Semi-automated event and coreference label generation

Given a cluster of multimodal documents, AutoEventAnnotator generates event and coreference labels in two stages: per-modality event extraction, followed by cluster-level coreference annotation.

**Text Event Extraction:** **(1)** We prompt an LLM to extract candidate event triggers from each sentence using ACE-style rules (Linguistic Data Consortium, 2005), without restricting to a predefined event ontology. **(2)** We filter triggers using NomBank (Meyers et al., 2004), PropBank (Palmer et al., 2005), and VerbNet (Kipper Schuler, 2005) to remove false positive event triggers wrongly predicted by the LLM. **(3)** For each retained trigger, we prompt the LLM with the full article to extract globally salient entities (actors, locations, times),

since relevant entities often appear far from the trigger sentence. **(4)** We use the trigger and entities to form the phrase $p$ with another LLM call.

**Video Event Extraction:** **(1)** We prompt a video LLM to generate a dense caption of the full video. **(2)** We apply the same event trigger extraction and filtering stage to the dense video caption. **(3)** We rewrite each trigger into a natural-language grounding query with an LLM that can be used to retrieve relevant moments from the video. **(4)** We use CG-DETR (Moon et al., 2026) to retrieve candidate temporal segments $(t_s, t_e)$ for each query. **(5)** We re-examine each retrieved segment with a VLM and discard it if the trigger event is not visually present, correcting false positives from moment retrieval. **(6)** We prompt an LLM to read the video description alongside the storyline to extract entities and synthesize the phrase $p$, combining them with visual entities from the video. **(7)** We merge overlapping event spans (with IoU $\geq 0.9$) by unifying their trigger and phrase information.

**Audio Event Extraction (Speech):** **(1)** We transcribe the audio with word-level timestamps using Whisper (Radford et al., 2023). **(2)** We apply the same ACE-style trigger extraction and filtering to the transcription. **(3)** We prompt an LLM to extract entities from the transcription and synthesize the phrase $p$. **(4)** Since a single trigger word is too short to ground temporally, we prompt the LLM to identify the minimal contiguous spoken phrase fully expressing the event; its word-level timestamps define the interval $(t_s, t_e)$.

**Audio Event Extraction (Background Sound):** **(1)** We segment the audio into 10-second windows and prompt Qwen2-Audio-7B to generate textual descriptions of sounds in each window. **(2)** We prompt an LLM to remove ambient, non-eventive sounds (wind, background music, etc.), retaining only causal events such as impacts, crowd reactions, etc. **(3)** We verify each retained sound against the corresponding video segment and storyline to remove hallucinated detections and normalize labels to canonical forms. **(4)** We merge temporally adjacent windows sharing the same normalized label into a single event interval $(t_s, t_e)$ to give the location of a sound event. **(5)** We prompt an LLM to synthesize the phrase $p$ from the sound label, video-clip description, and storyline.

**Coreference Link Annotation.** Regardless of source modality, each extracted event is represented by a textual phrase $p$ that encodes the trigger, globally salient entities, and broader event context. This phrase serves as a modality-agnostic event representation: constructing it from raw video, audio, or news article requires genuine multimodal understanding, but once formed, it enables uniform coreference reasoning across all modality pairs through text-based comparison. For each event pair $(e_i, e_j)$ in a cluster, we prompt an LLM to compare their phrases to determine whether both refer to the same real-world event, assigning a binary coreference label. This produces pairwise annotations across all six modality combinations. A detailed description of each of these steps is provided in Appendix J.

### 4.3 Test Set Annotation

We employ six annotators to label the test set. They are provided with annotation guidelines (Prompt V.1) covering event and coreference rules, along with annotated examples. We also hold discussion sessions to resolve ambiguities. Since no existing tool supports multimodal cluster annotation, we develop a custom web-based interface (Appendix U) that visualizes full event clusters with extracted events and coreference links, and allows annotators to add, delete, and modify events and coreferences. We adopt two complementary annotation settings: Human Annotation and AI-assisted annotation.

In the **human annotation set**, we provide the annotators with labels generated by AutoEventAnnotator to verify exhaustively. The annotation tool supports adding, deleting, and modifying both events and coreferences. Each annotator labels two different clusters, and each cluster is independently annotated by two annotators, yielding six densely annotated clusters. We compute inter-annotator agreement using the same metrics as our experiments (Section 5) and retain clusters with both event extraction and coreference agreement above 0.80, resulting in five final clusters.

In exhaustive human annotation, coreference link annotation is the most challenging component because the number of possible event pairs within a cluster can grow to several thousand. This limits the scale of fully human-annotated data. To increase coverage, we construct an **AI-assisted annotation set**, where annotators manually complete event annotations following the same procedure as above, and **Claude Opus 4.7** verifies all possible within-cluster event pairs to generate coreference labels. This process yields an additional 10 annotated clusters. To preserve the integrity of the two evaluation sets, we exclude these clusters from the training set.

| Dataset | #Clusters | #Docs | | | #Events | | | #Coref Links | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Text | Video | Audio | Text | Video | Audio | T–T | T–V | T–A | V–V | V–A | A–A |
| VM$^2$E$^2$ (Chen et al., 2021) | ✗ | 860 | 860 | ✗ | 4164 | 2702 | ✗ | ✗ | 852 | ✗ | ✗ | ✗ | ✗ |
| MultiHiEve (Ayyubi et al., 2024) | ✗ | 526 | 526 | ✗ | 2984 | 2529 | ✗ | ✗ | 4601 | ✗ | ✗ | ✗ | ✗ |
| VidEvent (Liang et al., 2025) | ✗ | ✗ | 1110 | ✗ | ✗ | 23989 | ✗ | ✗ | ✗ | ✗ | 17525 | ✗ | ✗ |
| NewsStories (Tan et al., 2022) | **350000** | **931679** | **333357** | **333357** | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Ours (M$^3$EC) | 20384 | 71051 | 55251 | 30216 | **2.98M** | **375K** | **259K** | **41.2M** | **1.67M** | **4.1M** | **480K** | **2.31M** | **3.48M** |
| M$^3$EC - Test (Human) | 5 | 14 | 30 | 23 | 532 | 180 | 171 | 2955 | 136 | 305 | 276 | 367 | 420 |
| M$^3$EC - Test (AI) | 10 | 22 | 33 | 23 | 823 | 219 | 204 | 1128 | 120 | 80 | 107 | 116 | 111 |

Table 1: Comparison of M$^3$EC with existing multimodal event extraction and coreference datasets across modality coverage, scale, and coreference pairs. T, V, and A denote text, video, and audio, respectively. Missing modalities or unavailable statistics are marked with ✗. Comparison with more existing datasets is provided in Appendix E.

| Dataset | PropBank(%) | NomBank(%) | VerbNet(%) |
|---|---|---|---|
| MAVEN (Wang et al., 2020) | 0.48 | 1.19 | 0.41 |
| SpeechEE (Wang et al., 2024) | 0.85 | 1.32 | 1.65 |
| VidEvent (Liang et al., 2025) | 1.17 | 1.55 | 3.41 |
| GLEN (Li et al., 2023) | 12.49 | 24.42 | 14.46 |
| M$^3$EC | **50.77** | **85.98** | **81.41** |

Table 2: Coverage of lexical event banks across different datasets. Our dataset achieves the highest coverage.

| Modality | #Silver | #Removed | #Added | #Modified | #Gold |
|---|---|---|---|---|---|
| Text | 538 | 20 | 14 | 29 | 532 |
| Video | 215 | 46 | 11 | 11 | 180 |
| Audio | 185 | 15 | 1 | 18 | 171 |
| Total | 938 | 81 | 26 | 58 | 883 |

Table 3: Silver versus gold events on the human-annotated test split, with annotator correction breakdown.

## 5 Evaluation Metric

**Events:** We independently evaluate event extraction for each modality at the news-cluster level. A predicted event matches a gold event only if both location and semantic matches hold.

*Location match* is defined per modality. For text, the predicted trigger span must exactly match the gold span; for video and audio, the temporal Intersection-over-Union (IoU) between the predicted and gold intervals must exceed a threshold $t$. *Semantic match* is common across all modalities: an LLM judge determines whether the predicted event phrase and the gold event phrase describe the same real-world event. The evaluation of this semantic judge and robustness analysis is provided in Appendix N and O, respectively.

Based on these matches, we compute true positives (TP), false positives (FP), and false negatives (FN) per cluster to compute the F1 score. For video and audio, to account for temporal boundary uncertainty, we evaluate across multiple IoU thresholds $\{0.30, 0.35, \ldots, 0.60\}$ and report the mean F1.

**Coreference:** For event coreference resolution, we report CoNLL-F1 (Pradhan et al., 2012), a widely used metric in coreference analysis. CoNLL-F1 is defined as the average of the MUC, $B^3$, and $CEAF_e$ F1 scores, providing a balanced assessment of cluster quality. We additionally evaluate coreference at the link level by treating all within-cluster pairs as a binary classification task (coreferent / non-coreferent) and report macro F1. More details are provided in Appendix L.

## 6 Dataset Statistics and Analysis

*Does M$^3$EC overcome the limitations of prior multimodal event datasets?* **M$^3$EC is the only dataset to provide cross-document clusters, all three modalities, and all six coreference pair types with event and coreference labels simultaneously.** As shown in Table 1, prior datasets cover at most two modalities and a single coreference pair type where annotations were available, while M$^3$EC contains 20,384 clusters spanning 156K documents, 3.61M event mentions, and 53.24M coreference links across all six modality pair types. Because news articles are retrieved from heterogeneous web sources rather than being paired with specific videos, the dataset also avoids the predefined text–media pairing assumption of prior multimodal benchmarks (Ayyubi et al., 2024; Li et al., 2020).

**M$^3$EC exhibits broad lexical diversity over canonical event predicates.** Table 2 shows that M$^3$EC triggers overlap with 50.77% of PropBank (Palmer et al., 2005), 85.98% of NomBank (Meyers et al., 2004), and 81.41% of VerbNet (Kipper Schuler, 2005) event predicate inventories, compared with 12.49% / 24.42% / 14.46% for GLEN, the strongest prior dataset, and below 3.5% for existing multimodal benchmarks. This indicates substantially broader lexical coverage of canonical event predicates than prior multimodal event benchmarks.

*How closely do AutoEventAnnotator silver la-*

| Video Pipeline | | | | Audio (Background) Pipeline | | | |
|---|---|---|---|---|---|---|---|
| **Stage** | **P** | **R** | **F1** | **Stage** | **P** | **R** | **F1** |
| CG-DETR | 62.15 | 85.67 | 72.03 | Qwen2-Audio | 45.20 | 85.73 | 59.19 |
| + Step (5) | 67.32 | 84.40 | 74.89 | + Step (2) | 68.43 | 81.21 | 74.27 |
| + Step (7) | 73.49 | 87.78 | **80.00** | + Step (3) | 84.50 | 80.30 | **82.35** |

Table 4: Contribution of different filtering stages in AutoEventAnnotator. P, R denote Precision and Recall.

*bels approximate the human-annotated gold set?*

**Silver event labels achieve an overall retention rate of 91.4%, with annotators adding only 26 events, a 2.9% miss rate, indicating high precision and strong recall.** Table 3 reports per-modality retention rates of 96.3% (text), 91.9% (audio), and 78.6% (video), where 46 of 215 video silver events were removed as hallucinations. Miss rates similarly vary by modality: annotators added 14 text events (2.6%), 11 video events (6.1%), and just 1 audio event (0.6%), confirming that video is the primary source of both false positives and false negatives. For coreference, only 4459 of 7277 silver links survive verification, reflecting the LLM's tendency to over-link during pairwise comparison (breakdown provided in Appendix K). Because verification of silver labels the gold set to the AutoEventAnnotator's output, we additionally ran a blind, from-scratch annotation study. Our released gold agrees with independent annotation at 96.26 F1 for events and 84.11 F1 for coreference, indicating the anchoring effect is small (Appendix R).

*How important are the intermediate verification stages?* **Verification stages contribute most where upstream models are noisiest.** Table 4 shows that Step (5) (Segment Verification) and Step (7) (Interval Merging) together improve AutoEventAnnotator's video event extraction F1 from 72.03 to 80.00 (+7.97). Likewise, Step (2) (Ambient Noise Filtering) and Step (3) (Visual Cross-Verification) improve the background sound pipeline's F1 from 59.19 to 82.35 (+23.16), nearly doubling precision (45.20 to 84.50). Text and speech pipelines, which start from cleaner upstream signals, do not require comparable filtering. This directly explains the modality-specific correction rates observed earlier: targeted filtering matters most where naive extraction hallucinates most.

## 7 Experiments and Results

### 7.1 Experiment Setup

We compare different systems for both event extraction and coreference resolution.

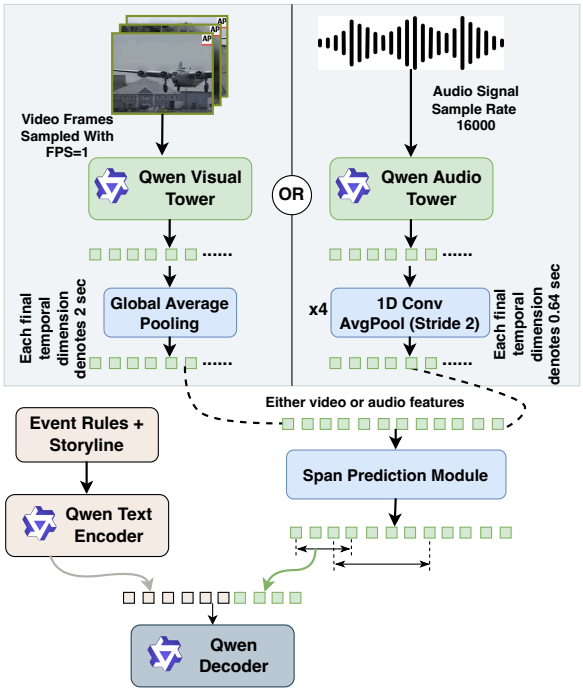



Figure 3: SpanQwen: Qwen-Omni's pipeline enhanced with a span prediction network (Lin et al., 2023), which helps locate the eventful video/audio segments. The features after Qwen's visual/audio tower are transformed into temporally meaningful features, which are passed onto the span prediction module.

**Traditional event extractors.** OneIE (Lin et al., 2020) trained on ACE (Walker et al., 2006) and MAVEN (Wang et al., 2020), and CEDAR (Li et al., 2023) trained on GLEN. We extend these text-only systems to video via CG-DETR (Moon et al., 2026) and to audio via Whisper (Radford et al., 2023) word-level transcripts.

**Zero-shot LLMs.** Open-source (Qwen (Yang et al., 2025; Bai et al., 2025a,b; Chu et al., 2024), Llama (Grattafiori et al., 2024), InternVL (Wang et al., 2025a)) and closed-source (GPT-4o (Hurst et al., 2024), Gemini-2.5-Flash (Comanici et al., 2025)) models under detailed prompting.

**Fine-tuned variants.** We fine-tune Qwen2.5-Omni-7B in two configurations: *AutoregQwen* applies a standard autoregressive next-token objective. *SpanQwen* keeps the same backbone but introduces a span prediction head, inspired by UniVTG (Lin et al., 2023), inserted after the visual and audio towers. This design decouples temporal localization from language generation, directly addressing the known failure of autoregressive models to regress precise timestamps. Fine-tuning details are in Appendix D.

**Coreference settings.** In the *GT* setting, de-

cisions are made over ground-truth events and phrases. In the *end-to-end* setting, each system extracts its own events and resolves coreference over its own event predictions. We additionally report AutoEventAnnotator's silver labels against gold as an empirical upper bound (details in Appendix R).

| Model | Text | | Video | | Audio | |
|---|---|---|---|---|---|---|
| | Human F1 | AI F1 | Human F1 | AI F1 | Human F1 | AI F1 |
| OneIE (ACE) | 3.82 | 4.55 | 8.19 | 8.77 | 13.95 | 12.36 |
| OneIE (MAVEN) | 5.31 | 6.84 | 12.30 | 11.68 | 15.67 | 18.18 |
| CEDAR (GLEN) | 27.22 | 26.45 | 15.70 | 15.43 | 16.79 | 17.81 |
| Llama3.3-70B | 45.24 | 43.11 | ✗ | ✗ | ✗ | ✗ |
| Qwen3-30B | 44.20 | 40.93 | ✗ | ✗ | ✗ | ✗ |
| InternVL3.5-38B | 20.31 | 28.77 | 14.22 | 15.21 | ✗ | ✗ |
| Qwen2.5-VL-7B | 17.33 | 14.67 | 5.17 | 3.94 | ✗ | ✗ |
| Qwen2.5-VL-72B | 20.56 | 22.32 | 21.05 | 18.26 | ✗ | ✗ |
| Qwen3-VL-30B | 18.97 | 33.74 | 7.26 | 17.10 | ✗ | ✗ |
| Qwen2-Audio-7B | 3.22 | 4.81 | ✗ | ✗ | 0.00 | 0.00 |
| Qwen2.5-Omni-7B | 5.76 | 8.90 | 0.00 | 0.00 | 0.00 | 0.00 |
| Qwen3-Omni-30B | 26.44 | 38.53 | 1.12 | 1.92 | 0.00 | 0.00 |
| GPT-4o | 39.87 | 43.29 | 20.00 | 14.12 | 0.00 | 0.00 |
| Gemini2.5-Flash | 45.19 | 44.67 | 33.01 | 26.50 | 0.00 | 0.00 |
| Qwen2.5-Omni-7B (AutoregQwen) | **67.53** | **72.46** | 4.26 | 3.17 | 6.29 | 3.42 |
| Qwen2.5-Omni-7B (SpanQwen) | 63.77 | 68.41 | **35.13** | **32.21** | **19.25** | **20.57** |
| AutoEventAnnotator | 91.51 | 90.23 | 80.08 | 78.68 | 84.21 | 82.44 |

Table 5: Comparison of event extraction and localization performance in both human and AI-assisted sets. Video and audio F1 scores are averaged over temporal IoU threshold $\in \{0.30, 0.35, ..., 0.60\}$. We also list the performance of our annotator tool against the gold set.

### 7.2 Result Analysis

*Zero-shot foundation models fail at video and audio temporal localization.* **Off-the-shelf LLMs and VLMs extract text events reasonably but collapse on video and audio, and fine-tuning narrows but doesn't close the gap.** On text events, closed-source models reach 45 Human F1, with open-source models following closely. On video and audio, the picture is starkly different: every non-fine-tuned LLM scores 0.00 on audio on both sets, and only Gemini exceeds 25 F1 on video (33.01 Human, 26.50 AI; Table 5). This pattern aligns with a well-known limitation of multimodal LLMs: alignment training optimizes "what happens" but not "when it happens," producing temporal hallucinations and timestamp drift in vision (Wu et al., 2025; Hannan et al., 2025) and audio (Yao et al., 2025) settings. Autoregressive fine-tuning of Qwen2.5-Omni-7B on silver data confirms that text gains do not transfer to localization: AutoregQwen reaches 72.46 AI F1 on text but only 3.17 (video) and 3.42 (audio) on the same set.

| Experiment Details | Model | Human | | AI | |
|---|---|---|---|---|---|
| | | F1 | CoNLL F1 | F1 | CoNLL F1 |
| Using GT Events and Details | Llama3.3-70B | 61.47 | **53.18** | 59.78 | 50.22 |
| | Qwen3-30B | 57.81 | 50.67 | 59.24 | 49.15 |
| | GPT-4o | **62.04** | 49.92 | **60.12** | **51.23** |
| | Gemini2.5-Flash | 61.33 | 51.44 | 59.18 | 48.55 |
| | Qwen2.5-Omni-7B | 51.07 | 29.88 | 48.28 | 26.73 |
| | Qwen2.5-Omni-7B (AutoregQwen) | 59.82 | 45.91 | 57.60 | 47.70 |
| | Qwen2.5-Omni-7B (SpanQwen) | 58.97 | 47.31 | 56.29 | 46.13 |
| Using Predicted Events and Details (End-to-end Evaluation) | GPT-4o | 12.27 | 15.93 | 12.55 | 16.81 |
| | Gemini2.5-Flash | 16.41 | 19.55 | 15.18 | 20.42 |
| | Qwen2.5-Omni-7B | 4.55 | 9.55 | 4.18 | 8.92 |
| | Qwen3-Omni-30B | 8.55 | 13.71 | 9.18 | 12.44 |
| | Qwen2.5-Omni-7B (AutoregQwen) | 18.92 | 23.41 | 19.27 | 24.88 |
| | Qwen2.5-Omni-7B (SpanQwen) | **51.27** | **56.18** | **49.49** | **55.05** |
| | AutoEventAnnotator | 63.91 | 61.77 | 65.42 | 63.11 |

Table 6: Performance of event coreference resolution. We evaluate two settings: (i) using ground-truth events and phrases, and (ii) end-to-end evaluation using model-predicted events and phrases. Macro F1 and CoNLL-F1 are reported as overall metrics across all modality pairs combined. Performance breakdown by different modality types alongside additional metrics is provided in Appendix S.

*Silver labels provide an effective training signal.* **Despite being LLM-generated, AutoEventAnnotator's silver labels translate into large downstream gains for both event extraction and coreference.** Fine-tuning Qwen2.5-Omni-7B on the silver set yields large gains across all modalities in SpanQwen: text AI F1 jumps from 8.90 to over 68.41, audio from 0.00 to 20.57, and video from 0.00 to 32.21, with comparable gains on the Human set. The improvements carry through to coreference: GT-setting AI CoNLL-F1 rises from 26.73 to 46.13 (Table 6).

*End-to-end coreference is gated by video and audio extraction.* **Coreference quality tracks phrase quality, so when video and audio events are missed or hallucinated, the entire cascade collapses.** Under the GT setting, every method performs coreference competitively, with Human CoNLL-F1 in the 29–53 range across many non-fine-tuned models. In end-to-end, the same approaches collapse: GPT-4o falls from 49.92 to 15.93 Human CoNLL-F1, Gemini from 51.44 to 19.55, with similar drops on the AI set (Table 6). End-to-end coreference rises if and only if video and audio extraction rise with it.

| Model | Macro F1 per Modality Pair (Human Set) | | | | | | Macro F1 per Modality Pair (AI Set) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T–T | T–V | T–A | V–V | V–A | A–A | T–T | T–V | T–A | V–V | V–A | A–A |
| GPT-4o | 18.2 | 6.8 | 0.0 | 3.2 | 0.0 | 0.0 | 22.4 | 8.2 | 0.0 | 4.3 | 0.0 | 0.0 |
| Gemini2.5-Flash | 23.8 | 9.4 | 0.0 | 5.8 | 0.0 | 0.0 | 28.4 | 12.8 | 0.0 | 7.8 | 0.0 | 0.0 |
| Qwen2.5-Omni-7B | 7.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Qwen3-Omni-30B | 11.0 | 0.6 | 0.0 | 0.5 | 0.0 | 0.0 | 12.7 | 0.8 | 0.0 | 0.4 | 0.0 | 0.0 |
| Qwen2.5-Omni-7B (AutoregQwen) | 25.9 | 2.7 | 2.9 | 1.2 | 2.0 | 2.7 | 27.9 | 2.4 | 3.1 | 1.2 | 1.8 | 2.3 |
| Qwen2.5-Omni-7B (SpanQwen) | 56.8 | 52.3 | 43.6 | 44.5 | 42.1 | 41.9 | 53.8 | 50.5 | 49.5 | 42.0 | 46.4 | 43.9 |

Table 7: Detailed end-to-end coreference breakdown (Macro F1 per modality pair) on the Human and AI-assisted test sets. All results use model-predicted events. Pairs with no extracted events score 0.00 (shaded). T, V, A denote text, video, and audio.

*The end-to-end coreference collapse is a per-modality cascade, not a coreference-reasoning failure.* **A link can only be scored when both of its events are extracted, so a modality with zero extraction wipes out every pair type that involves it, no matter how good the model's coreference ability is.** Table 7 makes this concrete. Every model with a 0.00 F1 score on audio extraction (GPT-4o, Gemini-2.5-Flash, Qwen2.5-Omni-7B) scores exactly 0.00 on all three audio pairs (T–A, V–A, A–A), even though the same models handle these pairs well when given the GT phrases (Table 6 and Appendix S). The residual performance is almost entirely text–text. Qwen2.5-Omni-7B fails on both video and audio, so it keeps only T–T (6.81 AI F1), and its other five pairs drop to 0.00. AutoregQwen shows the bottleneck most sharply. It tops text event extraction (72.46 AI F1), yet its video event (3.17 AI F1) and audio event (3.42 AI F1) extraction is catastrophic, so its end-to-end Human CoNLL-F1 falls to 23.41, and every cross-modal pair has an F1 score below 4. Coreference performance recovers only as far as a model's event extraction reaches across modalities. SpanQwen is the only fine-tuned model that localizes all three modalities, and it is the only one with moderately good scores on all six pairs (19.8–56.8 Human coreference F1). The bottleneck is upstream extraction, not coreference reasoning. Appendix S gives the full breakdown.

*Explicit span prediction is the decisive design choice for video and audio.* **Replacing next-token decoding with a span regression head closes most of the video/audio gap.** SpanQwen and AutoregQwen share the same backbone and decoder, differing only in how temporal spans are produced. On the AI set this difference yields a $10\times$ improvement on video (3.17 $\rightarrow$ 32.21 F1), a $6\times$ improvement on audio (3.42 $\rightarrow$ 20.57 F1), and an end-to-end CoNLL-F1 jump from 24.88 to 55.05 (Tables 5, 6); the Human set mirrors the trend. AutoregQwen retains a small lead on text extraction (72.46 vs. 68.41 AI F1), confirming that next-token prediction is adequate when temporal regression is not required, but when the model must localize in continuous time, span-based prediction provides a better lever than alignment or fine-tuning alone. Improving span localization further is, therefore, the highest-leverage direction for closing the remaining gap to AutoEventAnnotator's upper bound.

## 8 Conclusion

We introduce M$^3$EC, a large-scale benchmark for multimodal, multidocument event extraction and coreference across text, video, and audio. By organizing events within news clusters and modeling cross-modal links, it captures real-world event structure beyond prior datasets. We also present AutoEventAnnotator, a scalable framework that generates high-quality labels.

Experiments show that multimodal event understanding remains a challenge, especially in end-to-end settings. However, fine-tuning on our dataset yields consistent improvements, highlighting its practical value. We hope M$^3$EC will support future research on unified multimodal event and coreference understanding.

## Limitations

Processing full multimodal document clusters jointly remains computationally challenging due to the large volume of tokens across text, video,

and audio. As a result, our pipeline processes documents and events incrementally rather than reasoning over an entire cluster in a single pass, which may limit global context integration. Addressing this would require compression techniques capable of encoding an entire cluster jointly.

Our video event representations rely on editor-provided storylines to enrich entity information associated with visual events. While such metadata is commonly available in curated news video platforms, it may not be present in all real-world video sources. In settings where storylines are absent, the quality of entity assignment for video events may degrade. This reflects a broader limitation of current vision–language models, which still struggle to reliably extract fine-grained named entities directly from video content.

Finally, our experiments are limited to English-language sources. The AutoEventAnnotator pipeline and fine-tuned models are language-dependent, and extending M$^3$EC to non-English news sources would require multilingual LLMs and speech recognition models capable of cross-lingual event linking, a direction we leave to future work.

Because M$^3$EC is built from publicly available news, we cannot rule out that some article text appeared in the pretraining corpora of the models we evaluate. Their training data is undisclosed, so a definitive document-level contamination analysis is not possible. We note, however, that exposure to news content is not exposure to the supervision M$^3$EC evaluates: trigger boundaries, entity-enriched phrases, temporal intervals, cluster membership, and coreference links are newly produced and do not appear in the publicly available source articles. We therefore treat zero-shot text models as contemporary baselines rather than evidence of generalization to unseen news, and we note that zero-shot performance stays far below the human-verified and AutoEventAnnotator results, collapsing on video and audio localization where textual familiarity cannot help.

## Acknowledgments

We acknowledge Advanced Research Computing (ARC) at Virginia Tech for providing computational resources and technical support that contributed to the reported results. We also thank the anonymous reviewers for their constructive feedback, which helped improve the paper.

## References

Alibaba Group. 2024. Qwen LICENSE AGREEMENT(qwen-2024). https://scancode-licensedb.aboutcode.org/qwen-2024.html. Accessed: October 2025.

Hammad Ayyubi, Christopher Thomas, Lovish Chum, Rahul Lokesh, Long Chen, Yulei Niu, Xudong Lin, Xuande Feng, Jaywon Koo, Sounak Ray, and 1 others. 2024. Beyond Grounding: Extracting Fine-Grained Event Hierarchies Across Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17664–17672.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Coen Bron and Joep Kerbosch. 1973. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. 2021. Joint Multimedia Event Extraction from Video and Article. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 74–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-Sentence Argument Linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Google for Developers. 2025. Google Custom Search JSON API. `https://developers.google.com/custom-search/v1/overview`. Accessed: 2025-10-07.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. 2024. Context-Guided Spatio-Temporal Video Grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18330–18339.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. 2025. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19012–19022.

Kung-Hsiang Huang and Nanyun Peng. 2021. Document-level Event Extraction with Efficient End-to-end Learning of Cross-event Dependencies. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 36–47, Virtual. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o System Card. *arXiv preprint arXiv:2410.21276*.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Samuel Kriman and Heng Ji. 2021. Joint Detection and Coreference Resolution of Entities and Events with Document-level Context Aggregation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 174–179, Online. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media Structured Common Space for Multimedia Event Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

Sha Li, Qiusi Zhan, Kathryn Conger, Martha Palmer, Heng Ji, and Jiawei Han. 2023. GLEN: General-Purpose Event Detection for Thousands of Types. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2823–2838, Singapore. Association for Computational Linguistics.

Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. 2025. VidEvent: A Large Dataset for Understanding Dynamic Evolution of Events in Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5128–5136.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. UniVTG: Towards Unified Video-Language Temporal Grounding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2782–2792.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A Joint Neural Model for Information Extraction with Global Features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Linguistic Data Consortium. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events, Version 5.4.3*. Linguistic Data Consortium, Philadelphia, PA. Accessed: 2025-10-04.

Zhaoyu Liu, Kan Jiang, Murong Ma, Zhe Hou, Yun Lin, and Jin Song Dong. 2025. $F^3$Set: Towards analyzing fast, frequent, and fine-grained events from videos. In *International Conference on Learning Representations*, volume 2025, pages 10566–10580.

Jing Lu and Vincent Ng. 2021. Span-Based Event Coreference Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13489–13497.

Giuliano Martinelli, Bruno Gatti, and Roberto Navigli. 2025. xCoRe: Cross-context Coreference Resolution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34264–34278, Suzhou, China. Association for Computational Linguistics.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph

Grishman. 2004. The NomBank Project: An Interim Report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 24–31, Boston, Massachusetts, USA. Association for Computational Linguistics.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2015. Overview of TAC KBP 2015 Event Nugget Track. *Text Analysis Conference (TAC)*.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. 2017. Events Detection, Coreference and Sequencing: What's next? Overview of the TAC KBP 2017 Event Track. In *Text Analysis Conference (TAC)*.

WonJun Moon, Sangeek Hyun, Subeen Lee, and Jae-Pil Heo. 2026. Correlation-guided calibration of query dependency for video temporal grounding. *Pattern Recognition*, 174:112984.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. 2022. MEE: A Novel Multilingual Event Extraction Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual Semantic Role Labeling for Video Understanding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5585–5596.

Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, and Benjamin Van Durme. 2024. Grounding Partially-Defined Events in Multimodal Data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15905–15927, Miami, Florida, USA. Association for Computational Linguistics.

Reuben Tan, Bryan A. Plummer, Kate Saenko, JP Lewis, Avneesh Sud, and Thomas Leung. 2022. NewsStories: Illustrating Articles with Visual Summaries. In *Computer Vision – ECCV 2022*, pages 644–661, Cham. Springer Nature Switzerland.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, and 1 others. 2026. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. Web Download. Release Date: February 15, 2006. Languages: Mandarin Chinese, Standard Arabic, English. Applications: automatic content extraction.

Bin Wang, Meishan Zhang, Hao Fei, Yu Zhao, Bobo Li, Shengqiong Wu, Wei Ji, and Min Zhang. 2024. SpeechEE: A Novel Benchmark for Speech Event Extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 10449–10458, New York, NY, USA. Association for Computing Machinery.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A Unified Large-scale Dataset for Event Coreference, Temporal, Causal, and Subevent Relation Extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. 2025b. Time-R1: Post-Training Large

Vision Language Model for Temporal Video Grounding. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 83330–83364. Curran Associates, Inc.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. 2025. Number it: Temporal Grounding Videos like Flipping Manga. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13754–13765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Rundong Cheng, Lingrui Mei, Baolong Bi, Zhen Xiong, and Xueqi Cheng. 2025. Not in Sync: Unveiling Temporal Bias in Audio Chat Models. *arXiv preprint arXiv:2510.12185*.

Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. 2024. Recognizing Everything from All Modalities at Once: Grounded Multimodal Universal Information Extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14498–14511, Bangkok, Thailand. Association for Computational Linguistics.

## A Ethics Statement

The dataset that we used as a source is already public (Ayyubi et al., 2024). Additionally, we are not releasing any video or audio files with the dataset. However, we'll share the URLs of the files that we have worked on. The annotations on event extraction and coreference resolution have been performed using open-source LLMs or multimodal LLMs from the Qwen family. As per Section 4 of Qwen's license agreement (Alibaba Group, 2024), we are allowed to distribute the generated content for non-commercial use.

Our dataset was constructed by collecting publicly available news content from approximately 14,000 online media outlets. All data were accessed without circumvention of paywalls or authentication systems, and only publicly accessible pages were included. We store and use the data solely for non-commercial academic research under fair-use principles. Any content removed or restricted by the original publisher will be excluded from public release upon request.

Our use of external artifacts is consistent with their intended purpose. We used publicly released models (GPT-4o, Gemini2.5-Flash, Qwen2-Audio, Qwen3-30B, Claude Opus 4.7) strictly within their research or evaluation licenses. Our newly created $M^3EC$ dataset is derived from publicly accessible news videos, audio, and articles under fair-use principles. We release only metadata and annotations, not full copyrighted content, and specify that the dataset is intended **solely for non-commercial academic research**, consistent with the access conditions of the original data sources.

Large language models (including ChatGPT, GPT-4, and Gemini) were used as writing and coding assistants during this research, primarily for text editing, code debugging, and improving readability. All conceptual design choices, analyses, and final writing decisions were made by the human authors.

## B Potential Risks

Our dataset and models operate on publicly available YouTube videos and corresponding articles. While all data are collected from open sources under fair-use principles, videos may include identifiable individuals or copyrighted media. We release only references (URLs) rather than raw media to mitigate this risk.

The dataset and models could, in theory, be mis-

used for automated event surveillance. We discourage such applications and intend this resource solely for academic research in multimodal event understanding.

Our data are derived from publicly available news videos, audio, and articles, which may include identifiable individuals appearing in public contexts. We do not release any raw media or personal metadata; instead, we share only URLs and derived annotations. All data are used under fair-use principles for academic research. No private or non-public information is included in the released dataset.

## C Resource Budget

We report model sizes for all open models used (Qwen3-30B, Qwen2.5-VL-7B, Qwen2.5-VL-72B, Qwen3-VL-30B, Llama3.3-70B, InternVL3.5-58B, Qwen2-Audio-7B, Qwen2.5-Omni-7B, Qwen3-Omni-30B). GPT-4o and Gemini-2.5-Flash were accessed via their official APIs; their internal architectures and compute budgets are not publicly disclosed. We use NVIDIA H200 GPUs for inference. The approximate inference time for the entire silver label generation process is around 500 hours. Additionally, we fine-tuned Qwen2.5-Omni-7B for 1 epoch on the training dataset using 4 NVIDIA H200 GPUs. For AutoregQwen, 1 epoch takes approximately 8 hours; for SpanQwen, approximately 40 hours. The API bills for using Claude Opus 4.7 for coreference annotation generation for the AI-assisted set were around 200 USD.

## D Fine-tuning Setup

**AutoregQwen.** For text, the input is the full news article and the output is the set of event triggers alongside their corresponding phrases; token-level start and end indices are extracted for each trigger to enable localization. For video and audio, rather than requiring the model to regress raw timestamps, which is a known limitation of decoder-only LLMs, we discretize the full duration into 100 equal-width bins and train the model to predict the start and end bin index for each event. For coreference resolution, two event phrases are provided alongside a set of coreference rules, and the model produces a binary label (yes/no).

**SpanQwen.** The text and coreference formulations are identical to AutoregQwen. For video and audio, temporal localization is instead delegated to a span prediction module, inspired by UniVTG (Lin et al., 2023), inserted directly after the visual and audio towers of the Qwen Omni model. Given the encoded token sequence, the module predicts a set of event spans. Each predicted span is then processed independently in the decoder: event rules and contextual information encoded by the text encoder are concatenated with the original span representation from the visual/audio tower and passed to the decoder, which generates only the event phrase. Temporal localization is handled entirely by the span head, decoupling it from token generation.

**Training Configuration.** Both models are fine-tuned for one epoch on the silver training set using QLoRA (Dettmers et al., 2023) with 4-bit quantization, LoRA rank 8, and scaling factor $\alpha = 16$, on 4 NVIDIA H200 GPUs.

| Dataset | Text | | Video | | Audio | | Joint | |
|---|---|---|---|---|---|---|---|---|
| | #docs | #events | #docs | #events | #docs | #events | #docs | #events |
| ACE 2005 (Walker et al., 2006) | 599 | 5459 | ✗ | ✗ | ✗ | ✗ | 599 | 5459 |
| TAC KBP 2017 (Mitamura et al., 2017) | 167 | 3455 | ✗ | ✗ | ✗ | ✗ | 167 | 3455 |
| MAVEN (Wang et al., 2020) | 4480 | 118732 | ✗ | ✗ | ✗ | ✗ | 4480 | 118732 |
| MAVEN-ERE (Wang et al., 2022) | 4480 | 118732 | ✗ | ✗ | ✗ | ✗ | 4480 | 118732 |
| RAMS (Ebner et al., 2020) | 3993 | 9124 | ✗ | ✗ | ✗ | ✗ | 3993 | 9124 |
| MEE (Pouran Ben Veyseh et al., 2022) | 31226 (segs) | 50011 | ✗ | ✗ | ✗ | ✗ | 31226 (segs) | 50011 |
| ECB+ (Cybulska and Vossen, 2014) | 982 | 6841 | ✗ | ✗ | ✗ | ✗ | 982 | 6841 |
| GLEN (Li et al., 2023) | 6224 | 205045 | ✗ | ✗ | ✗ | ✗ | 6224 | 205045 |
| $M^2E^2$ (Li et al., 2020) | 245 | 1297 | 1014 (Images) | 391 | ✗ | ✗ | 1259 | 1688 |
| $VM^2E^2$ (Chen et al., 2021) | 860 | 4164 | 860 | 2702 | ✗ | ✗ | 1720 | 6866 |
| MultiHiEve (Ayyubi et al., 2024) | 526 | 2984 | 526 | 2529 | ✗ | ✗ | 1052 | 5513 |
| VidSitu (Sadhu et al., 2021) | ✗ | ✗ | 3037 | 146100 | ✗ | ✗ | 3037 | 146100 |
| VidEvent (Liang et al., 2025) | ✗ | ✗ | 1110 | 23989 | ✗ | ✗ | 1110 | 23989 |
| $F^3$Set (Liu et al., 2025) | ✗ | ✗ | 114 | 42846 | ✗ | ✗ | 114 | 42846 |
| SpeechEE (Wang et al., 2024) | ✗ | ✗ | ✗ | ✗ | 54825 | 130K | 54825 | 130000 |
| NewsStories (Tan et al., 2022) | **931679** | ✗ | **333357** | ✗ | **333357** | ✗ | **1.59 M** | ✗ |
| Ours ($M^3EC$) | 71051 | **2.98 M** | 55251 | **375 K** | 30216 | **259 K** | 156 K | **3.61 M** |
| $M^3EC$ Test (Human) | 14 | **532** | 30 | **180** | 23 | **171** | 67 | **883** |
| $M^3EC$ Test (AI) | 22 | **823** | 33 | **219** | 23 | **204** | 78 | **1246** |

Table 8: Comparison of scale and modality coverage across major event extraction datasets. Missing modalities or unavailable statistics are marked with ✗.

## E Comparison of existing datasets against $M^3EC$

Tables 8 and 9 reveal two broad patterns in the prior landscape: a persistent text-centric bias, and a disconnect between document scale and annotation richness.

| Dataset | #Clusters | #docs | | | #coref | | | | | | Joint | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Text | Video | Audio | text-text | text-video | text-audio | video-video | video-audio | audio-audio | #docs | #coref |
| TAC KBP 2017 (Mitamura et al., 2017) | ✗ | 167 | ✗ | ✗ | Not mentioned | ✗ | ✗ | ✗ | ✗ | ✗ | 167 | Not Mentioned |
| MAVEN-ERE (Wang et al., 2022) | ✗ | 4480 | ✗ | ✗ | 1290050 | ✗ | ✗ | ✗ | ✗ | ✗ | 4480 | 1290050 |
| ECB+ (Cybulska and Vossen, 2014) | ✗ | 982 | ✗ | ✗ | ∼19000 | ✗ | ✗ | ✗ | ✗ | ✗ | 982 | ∼19000 |
| $M^2E^2$ (Li et al., 2020) | ✗ | 245 | 1014 (Images) | ✗ | ✗ | 309 (text-image) | ✗ | ✗ | ✗ | ✗ | 1259 | 309 |
| $VM^2E^2$ (Chen et al., 2021) | ✗ | 860 | 860 | ✗ | ✗ | 852 | ✗ | ✗ | ✗ | ✗ | 1720 | 852 |
| MultiHiEve (Ayyubi et al., 2024) | ✗ | 526 | 526 | ✗ | ✗ | 4601 | ✗ | ✗ | ✗ | ✗ | 1052 | 4601 |
| VidEvent (Liang et al., 2025) | ✗ | ✗ | 1110 | ✗ | ✗ | ✗ | ✗ | 17525 | ✗ | ✗ | 1110 | 17525 |
| Ours ($M^3EC$) | 20384 | **71051** | **55251** | **30216** | **41.2 M** | **1.67 M** | **4.1 M** | **480 K** | **2.31 M** | **3.48 M** | **156518** | **53.24 M** |
| ($M^3EC$) Test - Human | 5 | 14 | 30 | 23 | 2955 | 136 | 305 | 276 | 367 | 420 | 67 | 4459 |
| ($M^3EC$) Test - AI | 10 | 22 | 33 | 23 | 1128 | 120 | 80 | 107 | 116 | 111 | 78 | 1662 |

Table 9: Comparison of existing coreference resolution dataset against $M^3EC$. Missing modality combinations are marked with ✗.

**Text-centricity persists even in multimodal benchmarks.** Datasets that extend beyond text ($VM^2E^2$, MultiHiEve, $M^2E^2$) treat it as a fixed text–video pair, while SpeechEE and VidEvent address audio and video entirely in isolation. The coreference table reflects this directly: text–audio and video–audio coreference are entirely absent from all prior work, and video–video coreference exists only in VidEvent (17,525 links) without any cross-modal counterpart. This is not merely a coverage gap: models trained on any prior dataset cannot learn joint reasoning across heterogeneous modalities by design.

**Document scale does not imply annotation richness.** NewsStories, the largest dataset by document count (1.59M documents), contains no event annotations, since it was designed for visual–text representation learning. VidSitu similarly provides 146K video event annotations without coreference. $M^3EC$ is smaller in raw document count but provides event annotations and coreference links across all modalities and pair types, making it directly usable for event understanding models.

**$M^3EC$ expands scale across every dimension.** $M^3EC$ contains 20,384 clusters spanning 156K documents, 3.61M event mentions, and 53.24M coreference links. By comparison, the largest text event dataset (GLEN, 205K events) covers fewer events than our text modality alone (2.98M, 14x more), while our video and audio event counts individually exceed every prior single-modality counterpart. The coreference scale is similarly disproportionate: $M^3EC$'s text–text links (41.2M) alone are 32 times larger than MAVEN-ERE's (1.29M), the largest prior coreference dataset.

## F Key Differences of $M^3EC$ with existing works

Table 10 compares $M^3EC$ with representative multimodal event datasets and highlights key differences in modality coverage, clustering strategy, and event coreference support. While several prior datasets incorporate multiple modalities, most are limited to predefined document pairs or single-document settings and do not support multidocument event reasoning. In addition, many benchmarks either lack event-level annotations for certain modalities or focus primarily on grounding rather than event coreference. In contrast, $M^3EC$ organizes data into tightly connected news clusters and explicitly models event mentions and coreference links across text, video, and audio. These differences enable $M^3EC$ to better capture the structure of real-world multimodal news events and support comprehensive evaluation of multimodal event understanding.

## G Video-Centric Cluster Construction

Constructing reliable news clusters is a central step in building our multimodal multidocument dataset. Our approach adopts a **video-centric clustering strategy**, where visually rich news videos serve as the anchor modality around which related multimodal content is organized. This design is motivated by three observations. First, news videos typically capture concrete real-world actions and therefore provide strong evidence of events. Second, each video naturally includes an associated audio track, enabling two modalities to be obtained from a single source. Finally, once clusters are

| Dataset | Key Differences from M$^3$EC |
|---|---|
| NewsStories (Tan et al., 2022) | Provides clusters of news articles, videos, and images, but does not include event-level annotations or event coreference links. Clusters are often loosely formed, with many videos consisting of studio reports or interview-style content, and may group documents describing similar but distinct real-world events. |
| REAMO (Zhang et al., 2024) | Focuses primarily on multimodal grounding rather than cross-document or cross-modal event coreference. Parts of the audio modality are synthetically generated via text-to-speech, and audio is not decomposed into speech and background events. The benchmark is not explicitly multidocument. |
| M$^2$E$^2$ (Li et al., 2020) | Covers events in text and images but does not include video or audio modalities, nor multidocument news clusters for event coreference resolution. |
| VM$^2$E$^2$ (Chen et al., 2021), MultiHiEve (Ayyubi et al., 2024) | Study multimodal event coreference between videos and paired text articles, but assume predefined video–article pairs with aligned event flow. These datasets do not support multidocument clusters or audio modalities. |
| Multivent-Grounded (Sanders et al., 2024) | Does not provide event annotations for audio, either for speech or background sounds, and is restricted to a small, fixed set of event types. In contrast, M$^3$EC adopts ACE-inspired open-ended event identification and explicitly models audio events. |
| xCoRe (Martinelli et al., 2025) | Addresses cross-context coreference across long and multiple text documents, but is limited to entity coreference in the text modality only. It does not support event coreference, video, or audio, and requires a gold topic structure to partition documents for cross-document evaluation. Temporal event localization and cross-modal coreference pairs (text–video, text–audio, video–audio) are outside its scope. |

Table 10: Comparison between M$^3$EC and related multimodal event datasets, highlighting differences in modality coverage, clustering, and event coreference support.

formed around videos, related news articles can be retrieved via web search, whereas large-scale video retrieval from the web is not feasible in the reverse direction. A summary of the steps of this process is shown in Figure 2

### G.1 Step (1): Event and Entity Extraction

We build on the MultiHiEve dataset (Ayyubi et al., 2024), where each entry contains a news video paired with an editor-written YouTube description referred to as a *storyline*. Storylines summarize the main events depicted in the video and frequently include explicit entity mentions such as people, organizations, and locations, making them particularly useful for aligning visual events with textual context.

For each video, we use an LLM (Qwen3-30B) to extract candidate event triggers from the storyline following ACE-style trigger identification rules (Linguistic Data Consortium, 2005) (Prompt V.16). For each detected trigger, we run the model again to extract relevant entities such as actors, locations, and objects (Prompt V.17). To avoid extracting dense or less informative events, we retain only those that refer to the central events of the video context (Prompt V.18), yielding at most six events per storyline. Each event, together with its associated entities, is then converted into a short natural-language phrase (Prompt V.19), such as *"Police arrested a suspect in New York"*, forming a compact and semantically meaningful event representation.

### G.2 Step (2): Source Dataset and Initial Filtering

The original MultiHiEve dataset contains approximately 400K videos. We retain roughly 330K videos that include valid and sufficiently descriptive storylines. To ensure that retained videos contain observable real-world events, we apply LLM-based filtering to remove non-eventive content such as studio commentary, interviews without actions, podcasts, and generic news reporting segments. This step ensures that the remaining videos contain visually meaningful actions suitable for event extraction.

Ultimately, this process produces approximately 1.8M event mentions across the retained videos, confirming that the dataset contains rich and eventful content.

### G.3 Step (3): Event Similarity and Graph Construction

To identify relationships between videos, we match events across the dataset using a two-stage similarity procedure.

First, each event phrase is converted into a dense embedding using BAAI BGE embeddings (Chen et al., 2024). For every event, we retrieve the top-10 nearest neighbors from events in other videos, generating approximately 18M candidate matching event pairs. After removing duplicates, roughly

12M unique candidate pairs remain.

Second, we verify these candidate pairs using LLM-based semantic comparison (Prompt V.20). The model determines whether two event phrases refer to the same real-world event by considering their associated entities and the contextual information from the storylines. Pairs judged to describe the same event are retained as verified matches.

Verified matches are then represented as a weighted graph, where each node corresponds to a video and each edge indicates that the two connected videos share at least one verified matching event. Edge weights reflect the number of verified matches between the two videos. To ensure that only strong relationships are preserved, we retain only edges whose weight exceeds 3.

### G.4 Step (4): Clique Detection and Cluster Formation

Given the weighted video graph, we identify tightly connected groups of videos using the Bron–Kerbosch algorithm (Bron and Kerbosch, 1973), which enumerates maximal cliques in the graph. Each clique represents a set of videos in which every pair shares verified event matches.

Each discovered clique forms a *video cluster* representing a coherent set of videos that describe the same real-world event or a closely related sequence of events. Because edges are established only through verified event matches, this procedure produces clusters with strong semantic consistency and event-level alignment. To further improve cluster quality, we perform an additional LLM-based verification step in which the storylines of all videos within a cluster are jointly provided to the model. The model determines whether any video is inconsistent with the cluster's overall event context (Prompt V.21), and inconsistent clusters are removed. This final step ensures that the resulting clusters are tightly related and semantically coherent.

Audio tracks are extracted directly from the clustered videos. Non-English files and files containing no acoustic events are discarded.

### G.5 Step (5): Retrieval of Related News Articles

After constructing video clusters, we expand each cluster by retrieving related news articles from the web. For each video, we generate search queries from its storyline using the extracted entities and key event phrases. We also compile a list of prominent English-language news domains spanning multiple countries, which are provided to the Google Search API as priority sources during retrieval. Retrieved articles are not restricted to these domains, but articles from the prioritized sources are returned preferentially when multiple relevant results are available.

Each retrieved article is then validated using an LLM-based relevance check (Prompt V.22), in which the model compares the article content against the video storyline to determine whether both describe the same real-world event. Articles that pass this verification step are added to the corresponding cluster.

This video-centric strategy produces tightly connected multimodal clusters consisting of videos, audio tracks, and related news articles. Datasets that rely on weak alignment signals or metadata matching, such as NewsStories, tend to produce loosely related clusters that are less suitable for fine-grained event extraction and coreference analysis.

## H Shortcomings of NewsStories Dataset in Multimodal Multidocument Event Extraction and Coreference Resolution

| | **NewsStories** | **M$^3$EC** |
|---|---|---|
| #Clusters | 50 | 50 |
| #Articles | 164 | 137 |
| #Videos | 146 | 118 |
| #Clusters w/ Inconsistent Articles (% of original clusters) | 14 (28%) | 2 (4%) |
| #Clusters w/ Inconsistent Videos (% of original clusters) | 4 (8%) | 2 (4%) |
| #Non-eventful Videos (% of original videos) | 78 (53.4%) | 14 (11.8%) |

Table 11: Comparison between NewsStories and M$^3$EC on cluster consistency and video eventfulness. NewsStories contains substantially more non-eventful videos and clusters with inconsistent articles, while M$^3$EC exhibits tighter cluster coherence and richer visual event content.

NewsStories is the closest existing resource to our setting as it provides news clusters that include multiple modalities such as text, image, video, and audio. However, it was primarily designed for visual–text representation learning rather than event-level modeling. By design, the image–text links in NewsStories are only intended to be *illustrative* rather than literal descriptions of the same event. As a result, the connections between modalities are often weak at the event level, which limits

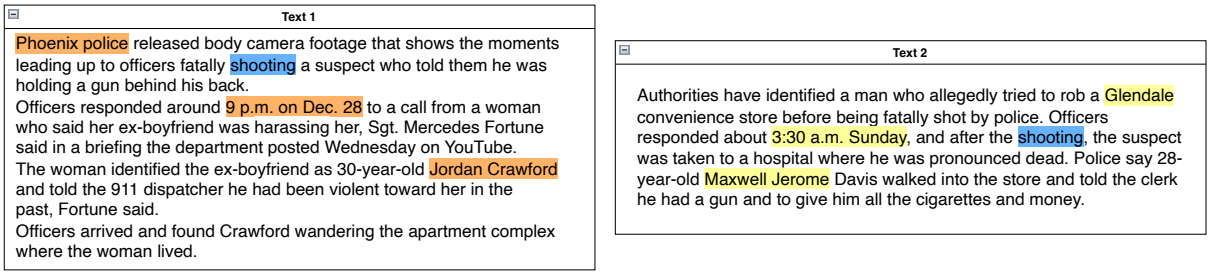



Figure 4: Sample of two news articles from the same cluster in NewsStories Dataset. Although both of them talk about *shooting* event, these two articles should not be in the same cluster as they are two completely different instances of shooting.

the dataset's usefulness for tasks such as event extraction or event coreference resolution. In addition, story clusters in NewsStories are constructed through agglomerative clustering over learned document embeddings followed by merging based on named entities. While this approach groups semantically related content, it can also merge documents describing distinct real-world incidents that share similar themes (e.g., two shootings occurring in different cities). An example of this issue is shown in Figure 4, where two articles discussing separate shooting incidents are placed in the same cluster.

The dataset authors themselves report residual noise in the clustering process. In their evaluation, only 82% of sampled clusters were judged to contain mutually related articles, and only 76% of image sets were considered relevant to the associated story. This indicates that a substantial fraction of clusters still contain loosely related or off-topic items. Furthermore, although NewsStories collects videos, its alignment objectives focus on text–image relationships; the videos are included as auxiliary data but are not explicitly used for cross-modal alignment. Consequently, many videos contain studio-style reporting segments or static scenes that lack clear visual actions or event dynamics. Figure 5(a) illustrates typical examples of such videos.

Our dataset addresses these limitations through a video-centric cluster construction strategy that prioritizes visually eventful content. As shown in Figure 5(b), videos in M$^3$EC frequently contain observable actions, object interactions, and scene dynamics that correspond directly to real-world events. This difference is further quantified in Table 11. We randomly pick 50 clusters from both datasets and note our observations. While both datasets contain the same number of clusters in this comparison subset, NewsStories exhibits substantially higher noise levels: 28% of clusters contain inconsistent articles compared to only 4% in M$^3$EC, and over half of the videos (53.4%) lack meaningful visual events. In contrast, only 11.8% of videos in M$^3$EC are non-eventful, reflecting our deliberate filtering of studio-style or commentary-based content. These results demonstrate that M$^3$EC produces tighter clusters with stronger event-level alignment across modalities.

Finally, NewsStories provides only coarse relevance labels between articles and images, without annotations for event triggers, event arguments, or event coreference. This lack of event-level supervision makes the dataset unsuitable as a starting point for benchmarking multimodal event extraction and cross-document event coreference, which are the primary goals of our work.

## I News Article Collection and Source Distribution

We compile an initial list of approximately 2,000 English-language news domains spanning multiple countries and use this list in conjunction with video storyline queries to retrieve related news articles via the Google Search API. The API prioritizes sources from the provided domain list, but may also return results from additional sources when no suitable matches are found. In the current version of the dataset, we collect ∼70,000 news articles originating from approximately 14,000 distinct news sources.

We observe that roughly 46% of the collected articles are sourced from a small set of frequently occurring outlets. Table 12 reports the top contributing news sources and their relative proportions in the dataset.

Although a comprehensive bias analysis is beyond the scope of this work, the source distribution indicates that a substantial portion of the dataset is




(a) Sample of videos from different clusters in NewsStories dataset.




(b) Sample of videos from different clusters in our dataset ($M^3EC$)

Figure 5: Comparison of video samples from NewsStories and our dataset. Videos in NewsStories often lack observable visual actions, events, or meaningful object motion, whereas $M^3EC$ contains videos with richer and more explicit visual event content.

| News Source | Percentage (%) | News Source | Percentage (%) |
|---|---|---|---|
| en.wikipedia.org | 7.45 | theguardian.com | 5.56 |
| bbc.com | 5.22 | nytimes.com | 5.20 |
| aljazeera.com | 3.30 | npr.org | 2.71 |
| nbcnews.com | 2.52 | cnn.com | 2.47 |
| pbs.org | 2.30 | voanews.com | 2.19 |
| cbsnews.com | 2.10 | abcnews.go.com | 1.94 |
| latimes.com | 1.10 | reddit.com | 0.90 |
| dw.com | 0.85 | | |

Table 12: Top contributing news sources in the dataset and their relative proportions.

derived from media outlets primarily based in the United States, the United Kingdom, Qatar, and Germany. Consequently, models trained on this dataset may inherit geographic or editorial biases associated with these regions, such as Western-centric framing of international events or variations in reporting norms and political perspectives.

# J Event and Coreference Annotation Generation using AutoEventAnnotator

## J.1 Text Event Extraction

Text documents are processed using a multi-stage pipeline that identifies event triggers, enriches them with contextual entities, and constructs concise event descriptions. The process is designed to capture event semantics while preserving precise trigger localization in the original article.

**Step (1): Trigger Identification with ACE Rules.** Event triggers are extracted at the sentence level using a large language model (Qwen3-30B) guided by ACE-style event trigger identification rules (see Prompt V.2). For each article, the document is first segmented into sentences using a sentence-aware splitter. Each sentence is then processed independently while the entire article is provided as context. The model identifies candidate triggers that correspond to concrete real-world actions, occurrences, or resulting states. Triggers may appear as verbs, event-denoting nouns, or resultative adjectives.

**Step (2): Trigger Filtering.** The set of candidate triggers produced by the language model may include generic or non-eventive words. To reduce such noise, we filter the extracted triggers using lexical resources from NomBank, PropBank, and VerbNet. These resources provide large collections of predicates commonly associated with events. The filtering step does not impose a fixed event ontology; it is used only to remove clearly non-eventive candidates while retaining open-domain event expressions.

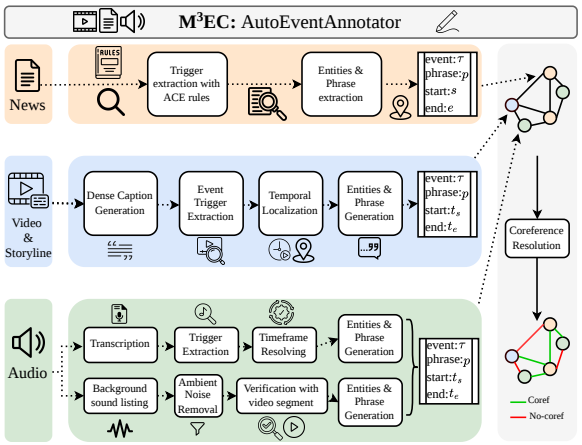



Figure 6: Overview of the AutoEventAnnotator pipeline for multimodal event extraction and coreference annotation across text, video, and audio modalities. Each modality follows a dedicated extraction pathway to produce structured event representations, which are then jointly processed for coreference resolution.

**Step (3): Entity Extraction.** For each retained trigger, we extract the entities that participate in or characterize the event. The language model analyzes the entire article and identifies entities such as actors, affected objects, locations, time expressions, causes, instruments, and results (see Prompt V.3). Entities are selected based on their direct semantic connection to the event rather than their proximity to the trigger sentence. This step provides global contextual grounding for the event and allows information appearing elsewhere in the article to be incorporated.

**Step (4): Event Phrase Extraction and Trigger Localization.** After entity extraction, the language model generates a short descriptive phrase summarizing the event (Prompt V.4). The phrase

is centered on the trigger and integrates relevant entities that clarify the event context, typically containing no more than 15 words. The purpose of this step is to create a normalized event description that captures the essential meaning of the event while remaining distinct from other events in the cluster. Sentence-aware tokenization is then used to compute exact character offsets of each trigger and verify that the annotated span matches the original trigger text in the document, ensuring precise localization for downstream coreference annotation.

### J.2 Video Event Extraction

Video event extraction is performed through a multi-stage pipeline that combines video-level captioning, language-based event extraction, temporal grounding, cross-modal verification, and event consolidation. The goal is to produce temporally localized video events enriched with trigger words, contextual entities, and short descriptive phrases.

**Step (1): Dense Caption Generation.** For each video, we first generate a global textual description using a video-language model. This description is produced from the full video and serves as a compact summary of the visible scenes, actions, participants, and environment. Because the output is generated from visual input, it captures the sequence of observable events but often lacks reliable named entities or detailed background context.

**Step (2): Event Trigger Extraction and Filtering.** We apply an LLM-based trigger extraction step to the full video description following the same general principles used in the text pipeline, but with video-specific constraints so that only *visually grounded, time-bounded events* are retained. We keep triggers that correspond to visible actions, interactions, movements, or state changes, such as gathering, rescuing, collapsing, or moving, and explicitly exclude non-eventive expressions such as camera operations, narration-related phrases, static scene descriptions, or editorial interpretations (see Prompt V.5). As in the text pipeline, we additionally apply lexical filtering using NomBank, PropBank, and VerbNet.

**Step (3): Query Construction for Temporal Localization.** Each retained trigger is converted into a short, visually grounded query phrase using an LLM (Prompt V.6). These queries describe the corresponding visible event compactly while preserving the core action and relevant visual participants. Since the moment retrieval model operates on visual-language alignment rather than named-entity matching, the query generation step avoids named entities and instead emphasizes directly observable content such as people, objects, actions, and scene context.

**Step (4): Temporal Localization with Moment Retrieval.** We use the generated query for each event to retrieve candidate temporal segments from the video using CG-DETR (Moon et al., 2026). The model predicts the start and end time of the segment that best matches the query. This step provides an initial temporal localization for each extracted event, but may also introduce noisy or weakly aligned segments, especially when the event description is visually ambiguous or when multiple similar actions occur in the same video.

**Step (5): Segment-Level Verification.** To improve precision, for each predicted segment we extract the corresponding video clip and generate a new segment-level description using the same video-language model. We then compare the localized segment description with the original query, while providing the full-video description as context, and use an LLM to determine whether the retrieved segment is semantically aligned with the intended event. Segments that do not match the event query are discarded, removing false positives from moment retrieval.

**Step (6): Entity Extraction and Event Phrase Synthesis.** After a segment is verified, we enrich the event with contextual entities. Since raw video descriptions often omit named entities, dates, or broader event context, an LLM jointly reads the localized video description and the editor-provided storyline to extract entities directly tied to the event, including participants, locations, times, causes, and results (Prompt V.7). Events for which the alignment between video and textual context is too weak to yield reliable entities are discarded. Using the verified trigger, localized description, and extracted entities, we then generate a short, entity-enriched event phrase that summarizes the grounded event (similar to Prompt V.4).

**Step (7): Merging Overlapping Event Predictions.** Because multiple extracted triggers may correspond to the same underlying video moment, we perform a final consolidation step. Overlapping predictions with at least 90% temporal IoU

are merged into a single event instance. We aggregate the triggers, entities, and phrases from all overlapping predictions and use an LLM to produce a unified representation with a generalized trigger, a deduplicated entity set, and a single final phrase (Prompts V.8, V.9, V.10).

### J.3 Audio Event Extraction

For audio documents, we extract two complementary types of events: *speech events* and *background sound events*. Speech events correspond to semantically meaningful spoken actions or statements, while background sound events capture non-verbal acoustic signals such as impacts, alarms, or crowd reactions.

#### J.3.1 Speech Event Extraction

**Step (1): Transcription.** We transcribe each audio track using the Whisper automatic speech recognition model (Radford et al., 2023), which produces a full textual transcription together with word-level timestamps, allowing each spoken token to be temporally aligned with the audio signal.

**Step (2): Event Trigger Extraction and Filtering.** We apply the same LLM-based event trigger extraction pipeline used for text articles (Prompt V.2) to the transcription. Candidate triggers are extracted using ACE-style rules and filtered using NomBank, PropBank, and VerbNet to remove non-eventive expressions.

**Step (3): Entity Extraction and Phrase Generation.** We extract relevant entities from the transcription using the same procedure as the text pipeline (Prompt V.3), and generate a short entity-enriched event phrase that summarizes the spoken event while preserving its original semantic meaning (Prompt V.4).

**Step (4): Timeframe Resolving.** Since a single spoken trigger word is too short to represent a meaningful temporal event segment, we prompt an LLM to identify the minimal contiguous sequence of words from the transcription that contains the trigger and fully expresses the event context (Prompt V.11). The start and end timestamps of this phrase define the event interval $(t_s, t_e)$.

#### J.3.2 Background Sound Event Extraction

**Step (1): Audio Segmentation and Sound Listing.** Each audio track is segmented into fixed-length windows of approximately 10 seconds. For each segment, we use the Qwen2-Audio model to generate textual descriptions of the dominant sounds present.

**Step (2): Ambient Noise Removal.** We prompt an LLM (Prompt V.13) to remove ambient and non-eventive sounds such as wind, background music, or microphone artifacts, retaining only causal, time-bounded events such as impacts, mechanical actions, alarms, or crowd reactions.

**Step (3): Visual Verification and Label Normalization.** We apply a verification stage in which retained sounds are checked against the corresponding video segment and storyline to remove hallucinated detections (Prompt V.12). Since similar sounds may be described differently across segments (e.g., "clapping" versus "applause"), we additionally normalize sound labels to concise canonical forms using LLM-based reasoning, ensuring consistent representation across the dataset.

**Step (4): Merging Adjacent Segments.** Temporally adjacent windows that share the same normalized sound label are merged into a single continuous event interval $(t_s, t_e)$, producing a consolidated temporal span for each background sound event.

**Step (5): Entity Extraction and Phrase Generation.** We prompt an LLM to combine the sound label, the video segment description, and the accompanying storyline to construct a concise, entity-enriched event phrase describing the sound event in context (Prompt V.14).

Both speech and background sound events are represented using a unified structure consisting of a temporal interval, an event trigger or sound label, and a short descriptive phrase enriched with contextual entities. These temporally grounded audio events are later integrated with text and video events during multimodal coreference annotation.

### J.4 Event Coreference Annotation

After event mentions are extracted from all documents within a cluster, we construct cross-document and cross-modal coreference links. Each event mention is represented as a structured tuple containing the event trigger, the event phrase $p$, associated entities, and contextual information derived from the source document.

Given a cluster containing events from multiple documents and modalities, our goal is to determine whether two event mentions refer to the same real-world event instance. Event pairs can arise from six modality combinations: *text–text*, *text–video*,

*text–audio*, *video–video*, *video–audio*, and *audio–audio*, originating from the same document, different documents of the same modality, or different documents across modalities.

**LLM-based Coreference Verification.** Each event pair is verified using an LLM that performs structured coreference reasoning (Prompt V.15). For each pair, the model is provided with the event trigger, the entity-enriched event phrase, the surrounding document context, and modality-specific information such as moment queries or storyline descriptions. The model reasons about the precise action, participating entities, location, and temporal context before outputting a binary coreference verdict. To ensure high precision, the prompt enforces strict rejection criteria: pairs are considered non-coreferent if they describe different actions within the same broader situation, occur in different incidents, or contain conflicting entity or location information. The model is instructed to return a negative decision whenever uncertainty exists.

A high-level overview of this pipeline is provided in Figure 6. A full example of annotated cluster is shown at Figure 7.

## K Human Annotation

The written instruction given to the human annotator for event extraction and coreference resolution is provided in Prompt V.1.

| Modality | Before (Auto) | After (Human) |
|---|---|---|
| Text | 538 | 532 |
| Video | 215 | 180 |
| Audio | 185 | 171 |
| **Total** | 938 | 883 |

Table 13: Event annotation counts before and after human verification on the human-annotated split.

| Link Type | Before (Auto) | After (Human) |
|---|---|---|
| Text–Text | 4931 | 2955 |
| Text–Video | 248 | 136 |
| Text–Audio | 558 | 305 |
| Video–Video | 413 | 276 |
| Video–Audio | 532 | 367 |
| Audio–Audio | 595 | 420 |
| **Total** | 7277 | 4459 |

Table 14: Coreference link counts before and after human verification on the human-annotated split.

To quantify the impact of human verification under dense annotation, we compare the number of event mentions and coreference links before and after the human annotation process. The pre-annotation counts correspond to events and links automatically generated by the AutoEventAnnotator. Human annotators exhaustively verify, remove spurious predictions, and correct boundaries and links, resulting in a smaller but higher-quality set of annotations. We report statistics separately for event extraction and event coreference. Table 13 and Table 14 contain counts of events and coreferences before and after the human annotation process, respectively. All post-annotation counts are taken from the M$^3$EC Dense split, while pre-annotation counts are estimated from the corresponding automatic outputs over the same clusters.

## L Evaluation

### L.1 Event Extraction Evaluation

Each modality is evaluated independently at the news-cluster level. Let

$$\mathcal{C} = \{D_1, D_2, \ldots, D_N\}$$

denote a news cluster. Let $E^{pred}$ and $E^{gold}$ denote the sets of predicted and gold event mentions for a document (or modality-specific subset), respectively.

Each event mention is associated with an event phrase $p(e)$ describing its semantics. To verify semantic consistency between predicted and gold events, we use an LLM-based binary judge

$$\mathrm{Sem}(p_i^{pred}, p_j^{gold}) \in \{0, 1\},$$

which returns 1 if the two phrases are judged to refer to the same real-world event and 0 otherwise. If the LLM output is malformed or does not contain a valid decision, we set $\mathrm{Sem} = 0$.

**Text Events.** For text documents, an event is represented by a contiguous trigger span, written as $(b, d)$, where $b$ and $d$ are the start and end token indices of the trigger. A predicted event with trigger span

$$e_i^{pred} = (b_i^{pred}, d_i^{pred})$$

matches a gold event with trigger span

$$e_j^{gold} = (b_j^{gold}, d_j^{gold})$$

if and only if both span boundaries match exactly and the event phrases are semantically consistent:

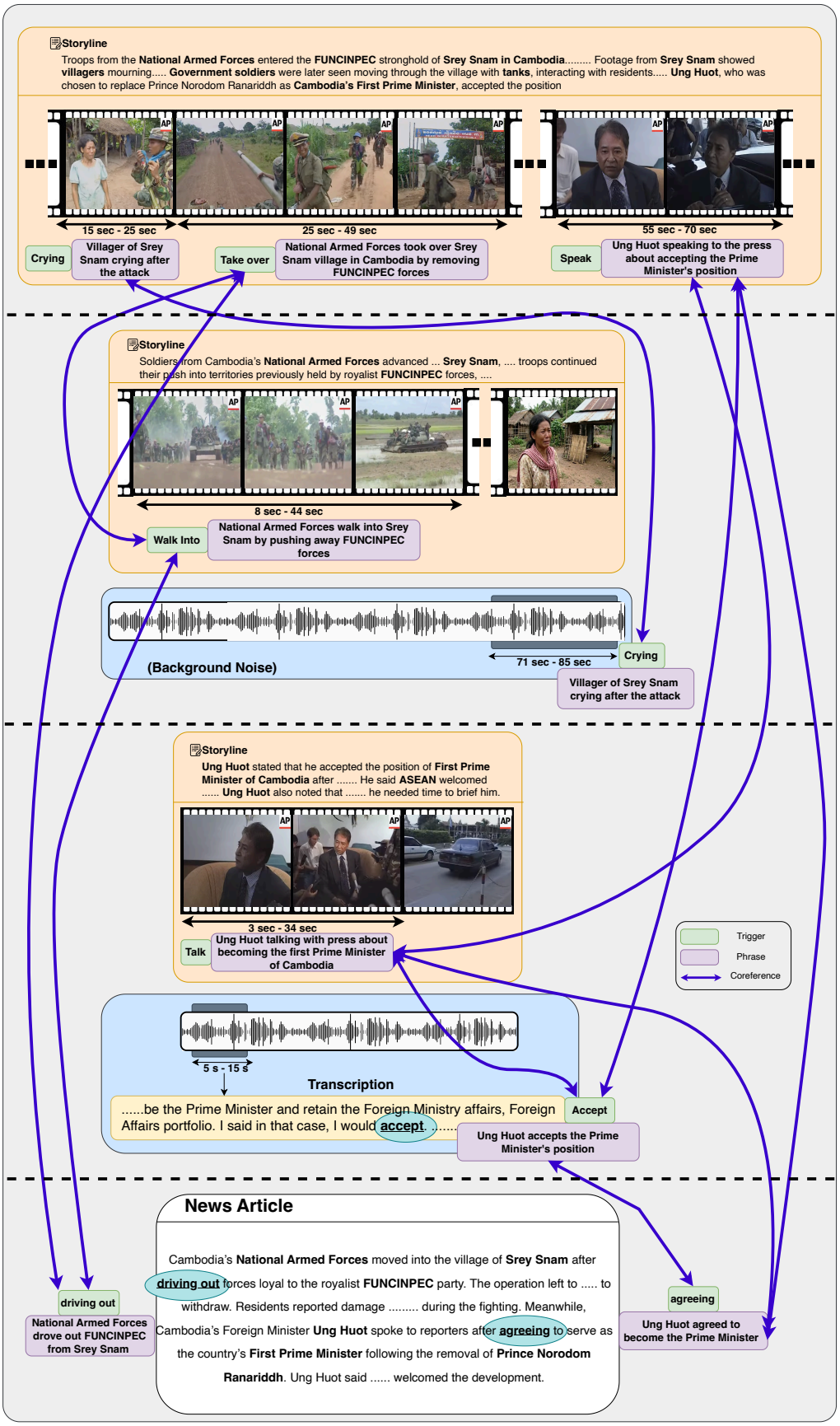



Figure 7: A sample cluster of the M$^3$EC dataset containing event mentions across text, video, and audio modalities with various coreference links.

$$\text{match}(e_i^{pred}, e_j^{gold}) = \mathbb{I}\Big[b_i^{pred} = b_j^{gold}\Big] \cdot \mathbb{I}\Big[d_i^{pred} = d_j^{gold}\Big] \cdot \text{Sem}\Big(p(e_i^{pred}),\, p(e_j^{gold})\Big)$$

**Video and Audio Events.** For video and audio documents, an event is represented by a temporal interval

$$e = (t_s, t_e),$$

and each event has an associated textual phrase $p(e)$ describing its semantics, where $t_s$ and $t_e$ are the start and end temporal indices of the trigger. Temporal overlap between a predicted event $e_i^{pred}$ and a gold event $e_j^{gold}$ is measured using Intersection-over-Union (IoU):

$$\text{IoU}(e_i^{pred}, e_j^{gold}) = \frac{\max(0, \min(t_e^{pred}, t_e^{gold}) - \max(t_s^{pred}, t_s^{gold}))}{\max(t_e^{pred}, t_e^{gold}) - \min(t_s^{pred}, t_s^{gold})}.$$

Given an IoU threshold $t$, a predicted event $e_i^{pred}$ is *eligible* to match a gold event $e_j^{gold}$ if:

$$\text{Elig}_t(e_i^{pred}, e_j^{gold}) = \mathbb{I}\big[\text{IoU}(e_i^{pred}, e_j^{gold}) \geq t\big] \cdot \text{Sem}\big(p(e_i^{pred}), p(e_j^{gold})\big).$$

**Gold-Centric Greedy Matching.** We perform one-to-one greedy matching in a gold-centric manner. Let $U$ be the set of unmatched predictions. For each gold event $e_j^{gold}$, we select:

$$i^\star = \arg\max_{i \in U} \text{IoU}(e_i^{pred}, e_j^{gold}) \quad \text{s.t. } \text{Elig}_t(e_i^{pred}, e_j^{gold}) = 1.$$

If such $i^\star$ exists, we match $e_{i^\star}^{pred}$ to $e_j^{gold}$ and remove $i^\star$ from $U$. Let $TP_t$ be the number of matched gold events under threshold $t$. Then:

$$FP_t = |E^{pred}| - TP_t, \qquad FN_t = |E^{gold}| - TP_t.$$

**Metrics.** Finally, we report event-level F1:

$$F1_t = \frac{2 \cdot TP_t}{2 \cdot TP_t + FP_t + FN_t}.$$

**Multi-IoU Averaging (Video/Audio).** For video and audio, we evaluate over:

$$\mathcal{T} = \{0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60\},$$

and report the mean score:

$$\overline{F1_t} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} F1_t,$$

**Aggregation.** All metrics are computed per cluster and averaged across clusters to obtain dataset-level results. If $\mathcal{Q}$ is the set of all clusters in the dataset,

$$F1_{overall} = \frac{1}{|\mathcal{Q}|} \sum_{k=1}^{|\mathcal{Q}|} \overline{F1_t}^{(k)}.$$

### L.2 Coreference Evaluation

We formulate event coreference as a pairwise binary classification task. Given the set of all event pairs within a cluster

$$\mathcal{P} = \{(e_i, e_j) : i < j\},$$

each pair is assigned a binary label $y_{ij} \in \{0, 1\}$, where $y_{ij} = 1$ indicates that $e_i$ and $e_j$ are coreferent and $y_{ij} = 0$ otherwise. Let $\hat{y}_{ij} \in \{0, 1\}$ denote the predicted label.

**Per-Class Statistics.** For each class $c \in \{0, 1\}$, we compute:

$$\text{TP}_c = \big|\{(e_i, e_j) \in \mathcal{P} : y_{ij} = c \ \wedge \ \hat{y}_{ij} = c\}\big|,$$
$$\text{FP}_c = \big|\{(e_i, e_j) \in \mathcal{P} : y_{ij} \neq c \ \wedge \ \hat{y}_{ij} = c\}\big|,$$
$$\text{FN}_c = \big|\{(e_i, e_j) \in \mathcal{P} : y_{ij} = c \ \wedge \ \hat{y}_{ij} \neq c\}\big|.$$

**Per-Class Precision, Recall, and F1.**

$$P_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c},$$
$$R_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c},$$
$$F1_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}.$$

**Macro F1.** The macro F1 score is the unweighted mean of the per-class F1 scores across both the coreferent ($c = 1$) and non-coreferent ($c = 0$) classes:

$$\text{Macro-F1} = \frac{1}{2} \sum_{c \in \{0,1\}} F1_c.$$

**CoNLL-F1.** We additionally report CoNLL-F1 (Pradhan et al., 2012), a standard cluster-level metric that averages the F1 scores of three complementary scoring algorithms: MUC, $B^3$, and $\text{CEAF}_e$. These components capture link-based overlap, mention-level agreement, and entity-level alignment, respectively, together providing a balanced assessment of global cluster structure. We refer the reader to Pradhan et al. (2012) for the complete formulation.

**Aggregation.** All coreference metrics are computed per cluster and averaged across clusters to obtain dataset-level results:

$$\text{Score}_{\text{overall}} = \frac{1}{|\mathcal{C}|} \sum_{k=1}^{|\mathcal{C}|} \text{Score}^{(k)}.$$

## M Per Cluster Inter Annotator Agreement (IAA) Score on Human Test Set

In the **human annotated set**, we provide the annotators with labels generated by AutoEventAnnotator to verify exhaustively. The annotation tool supports adding, deleting, and modifying both events and coreference links. We label six clusters (C1–C6) using six annotators (A–F), where each annotator independently labels exactly two clusters and each cluster receives annotations from exactly two distinct annotators, ensuring full cross-coverage without annotator overlap within any single cluster. We compute inter-annotator agreement (IAA) using the same metrics as our experiments (Appendix L): event-level F1 for event extraction and macro F1 for coreference resolution. We retain only clusters where both scores exceed 0.80.

| Cluster | Annotators | Event F1 | Coref Macro F1 | Status |
|---|---|---|---|---|
| C1 | A, B | 0.87 | 0.83 | ✓ |
| C2 | C, D | 0.91 | 0.85 | ✓ |
| C3 | E, F | 0.84 | 0.81 | ✓ |
| C4 | A, C | 0.89 | 0.84 | ✓ |
| C5 | B, E | 0.86 | 0.82 | ✓ |
| C6 | D, F | 0.83 | **0.76** | ✗ |

Table 15: Inter-annotator agreement per cluster. Each cluster is independently annotated by two annotators (A–F), each of whom labels exactly two clusters. Clusters with both Event F1 and Coreference Macro F1 above 0.80 are retained. C6 is excluded due to low coreference agreement.

Table 15 reports per-cluster IAA scores alongside annotator assignments. Event extraction agreement is consistently higher than coreference agreement across all clusters, reflecting the greater subjectivity involved in judging whether two events across different modalities refer to the same real-world occurrence. Five of six clusters meet both thresholds. Cluster C6, annotated by D and F, is excluded: despite adequate event extraction agreement (0.83), its coreference macro F1 falls to 0.76, indicating systematic disagreement on cross-modal link decisions. This results in five final dense clusters retained for evaluation.

## N Evaluation of Semantic Judge

In our evaluation pipeline, the semantic judge plays a central role: given a predicted event phrase and a gold event phrase, it determines whether both describe the same real-world event. This enforces that models do not merely localize events but also capture their semantic content accurately. Since manual verification of phrase-level semantic equivalence is infeasible at the scale of our benchmark, we use Qwen3-30B as an LLM judge for this purpose.

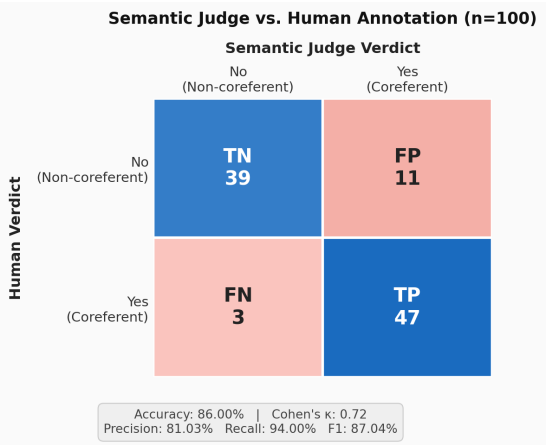



Figure 8: Confusion matrix of semantic judge.

*The semantic judge aligns strongly with human verdicts but exhibits a systematic over-prediction tendency.* **Across 100 randomly (50 matching, 50 non-matching) sampled phrase pairs, the judge achieves 86% accuracy and Cohen's $\kappa = 0.72$ with human annotators, indicating substantial agreement.** As shown in Figure 8, the judge correctly identifies 47 of 50 true coreferent pairs (recall 94.0%) but mistakenly marks 11 non-coreferent pairs as coreferent, yielding a precision of 81.0% and F1 of 87.0%. False positives, cases where the judge affirms a match that humans reject, constitute the dominant error mode (11 vs. 3 false negatives), reflecting a known tendency of LLMs to favour semantic similarity over strict event identity. Despite this bias, the high recall ensures that true semantic matches are rarely missed, and the overall agreement level supports the reliability of Qwen3-30B as a semantic judge for large-scale evaluation.

## O Robustness of the Semantic Judge

Our evaluation uses an LLM as a semantic judge. This judge decides whether a predicted event

| Model | Text F1 | | | Video F1 | | | Audio F1 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Qwen3 30B | Gemma4 31B | Llama3.3 70B | Qwen3 30B | Gemma4 31B | Llama3.3 70B | Qwen3 30B | Gemma4 31B | Llama3.3 70B |
| Llama-3.3 70B | 43.11 | 44.29 | 43.87 | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Qwen3-30B | 40.93 | 42.50 | 40.95 | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Qwen2.5-VL-7B | 14.67 | 17.23 | 15.03 | 3.94 | 7.82 | 6.54 | ✗ | ✗ | ✗ |
| Qwen3-VL-30B | 33.74 | 34.44 | 34.98 | 17.10 | 19.56 | 19.24 | ✗ | ✗ | ✗ |
| Qwen2-Audio-7B | 4.81 | 9.33 | 8.75 | ✗ | ✗ | ✗ | 0.00 | 2.90 | 2.84 |
| Qwen2.5-Omni-7B | 8.90 | 9.52 | 9.88 | 0.00 | 6.52 | 6.52 | 0.00 | 3.85 | 3.85 |
| Qwen3-Omni-30B | 38.53 | 39.81 | 38.97 | 1.92 | 3.38 | 3.38 | 0.00 | 0.00 | 0.00 |
| GPT-4o | 43.29 | 44.26 | 43.55 | 14.12 | 17.43 | 17.02 | 0.00 | 4.03 | 4.87 |
| Gemini-2.5-Flash | 44.67 | 46.33 | 45.48 | 26.50 | 29.33 | 27.24 | 0.00 | 2.87 | 2.50 |
| Qwen2.5-Omni-7B (AutoregQwen) | 72.46 | 72.87 | 72.93 | 3.17 | 1.89 | 1.65 | 3.42 | 6.36 | 6.14 |
| Qwen2.5-Omni-7B (SpanQwen) | 68.41 | 70.11 | 69.55 | 32.21 | 37.92 | 34.98 | 20.57 | 20.44 | 20.70 |

Table 16: Event extraction F1 on the AI test set under three independent evaluation-time semantic judges: Qwen3-30B, Gemma 4 31B, and Llama3.3-70B. Benchmark labels are held fixed; only the evaluation-time judge is varied. Video and audio F1 are averaged over temporal IoU thresholds $\in \{0.30, 0.35, \ldots, 0.60\}$. ✗ denotes modalities the model does not support. Absolute scores shift across judges, but system rankings remain stable.

phrase and a gold event phrase describe the same real-world event (Section 5). A natural concern is whether our conclusions depend on the specific model used as the judge. To test this, we re-evaluated all systems on the event extraction task using three independent open-weight judges from three different model families: Qwen3-30B (Yang et al., 2025) (our original judge), Gemma4-31B (Team et al., 2026), and Llama3.3-70B (Grattafiori et al., 2024). All three are open-weight, so the evaluation is fully reproducible. We held the benchmark labels fixed and varied only the evaluation-time judge. This isolates the effect of the judge on the measured scores.

Two patterns emerge from Table 16. First, absolute scores shift only modestly with the judge. Without very few exceptions, Gemma4 and Llama3.3 assign slightly higher scores than Qwen3-30B. This gap is largest on video and audio. It shows that our original Qwen3-30B judge is the most conservative of the three. The scores reported in the main paper are therefore not the product of a lenient evaluator.

Second, the ranking of systems is stable across all three judges. On text, every judge ranks the same top systems in the same order. Although Qwen3-30B ranks GPT-4o higher than LLamma, the difference is very small. AutoregQwen and SpanQwen are strongest, followed by Gemini-2.5-Flash. On video and audio, all three judges rank SpanQwen first by a wide margin. The systems that follow it are also the same across judges. The only differences appear among systems whose scores cluster near zero. There, small absolute changes cause minor reordering. These reorderings do not affect any of our conclusions. The relative standing of the best systems is preserved. The overall difficulty ordering across modalities is also preserved.

Taken together, these results show that the semantic judge affects absolute scores but not the comparative conclusions drawn from the benchmark.

## P Claude Opus 4.7 Coreference Judge Validation

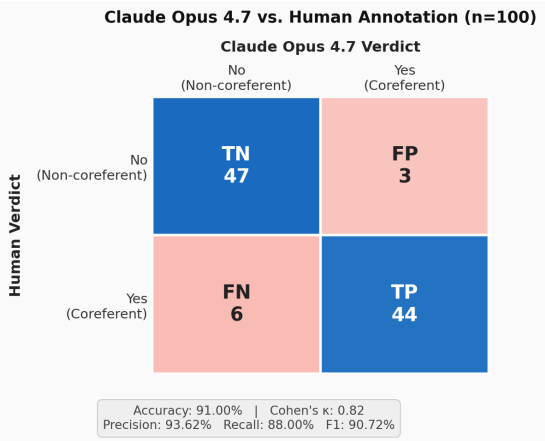



Figure 9: Confusion matrix of Claude Opus 4.7 as coreference link evaluator.

For the AI-assisted annotation set, coreference labels for all within-cluster event pairs are generated by Claude Opus 4.7. As with the semantic judge, we validate its alignment with human decisions on 100 randomly sampled (50 coreferent, 50

non-coreferent) event pairs.

*Claude Opus 4.7 achieves near-perfect agreement with human annotators, with a conservative bias toward rejecting coreference.* **Across 100 sampled pairs, Claude Opus 4.7 attains 91% accuracy and Cohen's $\kappa = 0.82$, placing it in the almost-perfect agreement range.** As shown in Figure 9, the model correctly identifies 44 of 50 true coreferent pairs and 47 of 50 non-coreferent pairs. Unlike the semantic judge, whose dominant error is over-prediction, Claude Opus 4.7 exhibits the opposite bias: false negatives outnumber false positives (6 vs. 3), indicating a tendency to reject coreference when evidence is ambiguous. This yields high precision (93.6%) at a slight cost to recall (88.0%), with an overall F1 of 90.7%. The conservative bias has a bounded effect on annotation quality, as missed coreferent links represent a small fraction of the total pairs and do not systematically distort cluster structure.

To directly assess how much human correction an Opus 4.7-labeled cluster requires, we randomly selected one cluster from the AI-assisted set and manually re-verified all of its coreference links. For each link asserted by Opus 4.7, a human annotator recorded whether it was kept, rejected, or newly added (Table 17).

| Pair | Asserted | Kept | Rej. | Added | Final |
|---|---|---|---|---|---|
| T–T | 102 | 96 | 6 | 7 | 103 |
| T–V | 10 | 9 | 1 | 2 | 11 |
| T–A | 7 | 6 | 1 | 2 | 8 |
| V–V | 12 | 9 | 3 | 3 | 12 |
| V–A | 9 | 8 | 1 | 2 | 10 |
| A–A | 11 | 9 | 2 | 3 | 12 |
| **Total** | **151** | **137** | **14** | **19** | **156** |

Table 17: Manual re-verification of Claude Opus 4.7 coreference links on one AI-assisted cluster. **Asserted**: links proposed by Opus 4.7. **Kept/Rej.**: asserted links a human retained or removed. **Added**: correct links the human added that Opus missed. **Final**: the human-verified link set.

Of the 151 links Opus 4.7 asserted, the human annotator kept 137 unchanged (90.7%), rejected 14, and added 19 that Opus had missed. Computed over all links in this cluster, this gives precision 90.7%, recall 87.8%, and F1 89.3%, closely matching the held-out validation above (precision 93.6%, recall 88.0%, F1 90.7%). The final human-verified count (156) is nearly identical to the Opus-asserted count (151), and the corrections are dominated by additions rather than removals (19 added vs. 14 rejected), reflecting the same conservative bias. The re-verification therefore required only a small number of corrections and confirms that Opus 4.7's coreference judgments on the AI-assisted set are close to human-level.

# Q In-depth Analysis of AutoEventAnnotator Intermediate Stages

## Q.1 Analysis of Text Modalities

The text event extraction pipeline identifies candidate triggers using ACE-style rules, followed by lexical filtering and entity-enriched phrase generation (Appendix J.1). To reach the final gold standard of text events in the dense split (532 events), human annotators performed the following corrections on the initial silver labels generated by AutoEventAnnotator (538 events):

- **Added Events:** 14 events (2.6% of final) were missing from the auto-generated set and added by annotators.
- **Removed Events:** 20 events (3.7% of initial) were identified as spurious or non-eventive and removed.
- **Rewritten Phrases:** 29 events (5.4% of final) required corrections to the semantic phrases to ensure precise alignment with the article context.

The analysis reflects the high precision of the LLM-based text extraction while necessitating human intervention to improve recall and boundary accuracy.

## Q.2 Analysis of Video Modalities

Video extraction is the most complex pipeline, involving global description generation, moment retrieval, and segment-level verification (Appendix J.2).

**Visual Recognition Accuracy:** Before temporal grounding, we evaluate how many real-world events were captured in the initial global video descriptions. Of the 180 final verified video events, 153 (85%) were successfully identified in the initial dense description generated by Qwen3-VL-8B. However, the initial extraction also produced 86 hallucinated events (actions not present in the video, human storyline, or camera motion events). Among the 86 hallucinated events, 24 of these were removed subsequently during the **Segment-Level**

**Verification** and **Merging Overlapping Event Predictions** stages. Note that trigger-level (after Step (2) of video event extraction) and event-level counts (after Step (4) and onward) are not directly comparable: a single trigger may ground to multiple temporal segments in Step (4) of video event extraction, and segments are later filtered and merged, so the number of event instances changes across stages.

**Pipeline Performance:** We measure performance at each critical stage: initial retrieval via `CG-DETR` (Moon et al., 2026), segment-level VLM verification, and overlapping interval merging.

| Pipeline Stage | Precision | Recall | F1 Score |
|---|---|---|---|
| Raw Moment Retrieval (CG-DETR) | 62.15 | 85.67 | 70.86 |
| + Segment-Level Verification | 67.32 | 84.40 | 76.68 |
| + Interval Merging (Final) | 73.49 | 87.78 | **80.08** |

Table 18: Intermediate performance of the video event extraction pipeline on the dense split.

**Human Correction Statistics:** To reach the final gold video events (180 events) from silver predictions generated by AutoEventAnnotator (215 events), we had to make the following changes:

- **Added Events:** 11 events (6.1% of final) added by humans.
- **Removed Events:** 46 events (21.4% of initial) removed due to hallucinations.
- **Rewritten Phrases:** 11 events (6.1% of final) required phrase modifications.

### Q.3 Analysis of Audio Modalities

The audio pipeline handles speech events (via transcription) and background (BG) sounds independently (Appendix J.3).

**Transcription Analysis:** For speech transcription, we employ `Whisper-large-v3`, an open-source model that enables scalable and cost-effective inference across large datasets. As ground-truth transcriptions are not available, we approximate transcription quality through model-based comparison. Specifically, we sample 50 audio segments uniformly from the dataset and obtain reference transcriptions using AssemblyAI's Universal-2 model, a strong commercial ASR system.

Treating the Universal-2 outputs as pseudo ground truth, we evaluate the Whisper-generated transcriptions and report a Word Error Rate (WER) of 5.3% and a Character Error Rate (CER) of 4.4%. These results indicate that Whisper-large-v3 produces high-quality transcriptions suitable for downstream event extraction tasks.

**Background Sound (BG) Filtering:** For background sound event detection, we apply a two-stage filtering pipeline on top of the outputs generated by the `Qwen2-Audio-7B` model. The first stage performs LLM-based noise removal to eliminate non-event or irrelevant acoustic descriptions. The second stage applies video segment-based verification to ensure that the detected audio events are temporally and contextually aligned with the corresponding video content.

Table 19 reports the performance across three stages: (i) raw model outputs, (ii) after LLM-based filtering, and (iii) after segment-level verification. This progressive evaluation highlights the effectiveness of each filtering stage in improving the quality of detected background sound events.

| Stage | Precision | Recall | F1 Score |
|---|---|---|---|
| Raw Qwen2-Audio Output | 45.20 | 85.73 | 60.27 |
| + LLM Ambient Noise Filter | 68.43 | 81.21 | 75.88 |
| + Visual Verification (Final) | 84.50 | 80.30 | **82.34** |

Table 19: Impact of filtering stages on Background sound detection.

**Human Correction Statistics:** Combined speech and background labels totaled 185 initial silver labels, reduced to 171 verified gold events:

- **Added Events:** 1 event (0.6% of final) added by humans.
- **Removed Events:** 15 events (8.1% of initial) removed, primarily due to non-eventive ambient noise or transcription hallucinations.
- **Rewritten Phrases:** 18 events (10.5% of final) required human rewriting.

## R Anchoring of silver labels in human (gold) set construction

### R.1 Anchoring Analysis

Our released human set is built by having annotators exhaustively verify the silver labels from AutoEventAnnotator. This anchors the human set to the AutoEventAnnotator's output. A verification-based

| Mod. | B | M | B\S | S\B | PM | S | F1 |
|---|---|---|---|---|---|---|---|
| Text | 218 | 196 | 10 | 11 | 12 | 219 | 89.70 |
| Video | 75 | 62 | 7 | 18 | 6 | 86 | 77.02 |
| Audio | 70 | 59 | 4 | 7 | 7 | 73 | 82.52 |
| **All** | **363** | **317** | **21** | **36** | **25** | **378** | **85.56** |

Table 20: Event extraction performance, silver pipeline (S) against blind annotation (B). **Mod.**: modality. **B**, **S**: number of events in the blind and silver sets. **M**: matched events. **B\S**: events in B but not S. **S\B**: events in S but not B. **PM**: phrase mismatch. F1 measures the silver set against the independent blind reference.

| Mod. | B | M | B\V | V\B | PM | V | F1 |
|---|---|---|---|---|---|---|---|
| Text | 218 | 210 | 4 | 2 | 4 | 216 | 96.77 |
| Video | 75 | 70 | 3 | 1 | 2 | 73 | 94.59 |
| Audio | 70 | 67 | 2 | 1 | 1 | 69 | 96.40 |
| **All** | **363** | **347** | **9** | **4** | **7** | **358** | **96.26** |

Table 21: Event extraction performance, released gold (V) against blind annotation (B). **Mod.**: modality. **B**, **V**: number of events in the blind and released-gold sets. **M**: matched events. **B\V**: events in B but not V. **V\B**: events in V but not B. **PM**: phrase mismatch. This measures how far our verification-based gold sits from independent, from-scratch annotation.

process can make recall hard to measure. Annotators who start from a populated list may add fewer missing events than annotators who start from nothing. This could make the pipeline's agreement with the gold set look higher than it truly is.

Our annotation interface already supports adding, removing, and editing events and coreference links, and two annotators labeling the same cluster reach above 0.80 agreement for the retained clusters. Neither of these fully removes the anchoring concern, since both annotators still begin from a pipeline-populated list. To measure the size of this effect directly, we ran a blind annotation study. In this study, annotators labeled clusters completely from scratch, with no pipeline labels shown.

We annotated two clusters for event extraction and one cluster for coreference resolution in this blind setting. We use the same evaluation setup stated in Section 5. Coreference is far more exhaustive to annotate blind, which is why we completed one cluster for that stage. We define three label sets. **S** is the silver set from AutoEventAnnotator. **V** is the human-verified set built by editing S, which is our released gold or human set. **B** is the blind from-scratch human set.

Two comparisons matter. **S vs B** measures the silver pipeline against a fully independent reference. And **B vs V** measures whether our released gold differs from what independent, from-scratch annotation would have produced. The second comparison speaks directly to the anchoring concern.

**Event extraction.** Against a fully independent reference, the AutoEventAnnotator pipeline reaches 85.56 overall event F1 (Table 20). Video is the weakest modality at 77.02. This matches the hallucination pattern we already reported in Table 3 and Appendix Q. More important for the anchoring concern, our released gold (V) agrees closely with blind annotation (B) at 96.26 F1 (Table 21). Only 9 events across the three modalities were found by blind annotation but missed during verification. A single blind study cannot fully rule out anchoring. Still, the result suggests its effect is small. The event set is largely the same whether an annotator starts from the pipeline's output or from a blank slate.

| Pair | B | M | B\S | S\B | S | P | R | F1 |
|---|---|---|---|---|---|---|---|---|
| T–T | 650 | 520 | 130 | 500 | 1020 | 50.98 | 80.00 | 62.28 |
| T–V | 34 | 23 | 11 | 29 | 52 | 44.23 | 67.65 | 53.49 |
| T–A | 72 | 49 | 23 | 69 | 118 | 41.53 | 68.06 | 51.58 |
| V–V | 62 | 46 | 16 | 42 | 88 | 52.27 | 74.19 | 61.33 |
| V–A | 82 | 65 | 17 | 45 | 110 | 59.09 | 79.27 | 67.71 |
| A–A | 94 | 80 | 14 | 45 | 125 | 64.00 | 85.11 | 73.06 |
| **All** | **994** | **783** | **211** | **730** | **1513** | **51.75** | **78.77** | **62.47** |

Table 22: Coreference resolution performance (one cluster), silver pipeline (S) against blind annotation (B). **Pair**: modality pair. **B**, **S**: number of links in the blind and silver sets. **M**: matched links. **B\S**: links in B but not S. **S\B**: links in S but not B. **P**, **R**, **F1**: precision, recall, and F1 of the silver positive-class links against the independent blind reference.

**Coreference.** The S-vs-B comparison shows high recall (78.77) but lower precision (51.75) against independent set (Table 22). This matches the over-linking tendency we already reported in Table 14. Our released gold (V) again agrees closely with blind annotation (B) at 84.11 F1 (Table 23).

**Summary.** These results characterize the anchoring effect. The released gold (Human Set) shows substantial agreement with independent from-scratch annotation, with 96.26 F1 for events and 84.11 F1 for coreference. We also report the AutoEventAnnotator's direct quality against an independent reference. It is 85.56 event F1 and 62.47 coreference F1. The coreference study covers a single cluster, so we present it as indicative rather than definitive.

| Pair | B | M | B\V | V\B | V | P | R | F1 |
|---|---|---|---|---|---|---|---|---|
| T–T | 650 | 540 | 110 | 85 | 625 | 86.40 | 83.08 | 84.71 |
| T–V | 34 | 28 | 6 | 2 | 30 | 93.33 | 82.35 | 87.50 |
| T–A | 72 | 60 | 12 | 6 | 66 | 90.91 | 83.33 | 86.96 |
| V–V | 62 | 52 | 10 | 6 | 58 | 89.66 | 83.87 | 86.67 |
| V–A | 82 | 66 | 16 | 11 | 77 | 85.71 | 80.49 | 83.02 |
| A–A | 94 | 69 | 25 | 19 | 88 | 78.41 | 73.40 | 75.82 |
| **All** | **994** | **815** | **179** | **129** | **944** | **86.33** | **81.99** | **84.11** |

Table 23: Coreference resolution performance (one cluster), released gold (V) against blind annotation (B). Columns as in Table 22, with V replacing S: **B\V** and **V\B** are the links unique to each set. This measures how far our verification-based coreference gold sits from independent, from-scratch annotation.

| Mod. | B | Miss | Hall. | Temp. | Ev. Err. | Ph. Err. | S |
|---|---|---|---|---|---|---|---|
| Text | 115 | 4 | 8 | — | 12 | 7 | 119 |
| Video | 33 | 3 | 10 | 2 | 15 | 4 | 40 |
| Audio | 38 | 2 | 5 | 3 | 10 | 2 | 41 |
| **All** | **186** | **9** | **23** | **5** | **37** | **13** | **200** |

Table 24: Event errors by originating stage, silver (S) against blind (B) on the single blind-coreference cluster. **B**, **S**: event counts. **Miss**: events missed by the pipeline. **Hall.**: hallucinated events. **Temp.**: temporal-boundary mismatches. **Ev. Err.**: total event-identification errors (Miss + Hall. + Temp.). **Ph. Err.**: phrase errors on matched events.

## R.2 Error Propagation from Extraction to Coreference

AutoEventAnnotator produces labels in multiple LLM-based stages, so a natural question is how errors from earlier stages (event identification and phrase generation) propagate to the final coreference labels. We answer this using the blind set **B** for the single cluster that has fully blind coreference annotation. This analysis is therefore restricted to that one cluster; its event counts are a subset of Table 20, which aggregates both blind event-extraction clusters.

**Event errors by originating stage.** We first classify every event discrepancy between the silver set (S) and the blind set (B) by the stage it originates from (Table 24). Missed and hallucinated events exist on only one side, whereas temporal and phrase errors involve events present in both S and B. The pattern matches our broader results in Appendix Q: text has few event-identification errors, video carries the largest hallucination burden, and audio is comparatively reliable apart from some temporal-alignment errors.

| Pair | B-only | | | | S-only | | | |
|---|---|---|---|---|---|---|---|---|
| | Tot. | Ev. | Ph. | Res. | Tot. | Ev. | Ph. | Res. |
| T–T | 130 | 22 | 9 | 99 | 500 | 78 | 24 | 398 |
| T–V | 11 | 4 | 3 | 4 | 29 | 7 | 7 | 15 |
| T–A | 23 | 5 | 3 | 15 | 69 | 14 | 7 | 48 |
| V–V | 16 | 6 | 6 | 4 | 42 | 13 | 12 | 17 |
| V–A | 17 | 6 | 5 | 6 | 45 | 13 | 12 | 20 |
| A–A | 14 | 4 | 2 | 8 | 45 | 10 | 5 | 30 |
| **All** | **211** | **47** | **28** | **136** | **730** | **135** | **67** | **528** |

Table 25: Coreference link disagreements decomposed by upstream cause, silver (S) against blind (B) on one cluster. **B-only**: links in B but not S; **S-only**: links in S but not B. **Tot.**: total disagreeing links. **Ev.**: attributable to an event-identification error at an endpoint. **Ph.**: attributable to a phrase error at an endpoint. **Res.**: residual coreference errors, where both endpoints are correct and only the pairwise decision is wrong.

**Propagation to coreference.** To trace how these upstream errors affect coreference, we take the link disagreements between silver and blind annotation (the B-only and S-only links of Table 22) and attribute each to a single upstream cause using the precedence order Event Error > Phrase Error > Residual Coreference. A link is assigned to Event Error if either endpoint has an event error, to Phrase Error if either endpoint has a phrase error, and to Residual Coreference otherwise (both endpoints correct, only the pairwise verdict wrong). Because each link is counted once, the decomposition preserves the B-only and S-only totals of Table 22 exactly.

Aggregating across both sides, event-identification errors account for 182 of the 941 link disagreements (19.3%), phrase-generation errors for 95 (10.1%), and the pairwise coreference decision for the remaining 664 (70.6%). Two observations follow. First, upstream errors do propagate: because events in a dense cluster are highly interconnected, event-level and phrase errors together affect 277 links (29.4% of the 941 disagreements), which quantifies the amplification effect. Second, and more important for benchmark validity, the majority of disagreements (70.6%) do not stem from upstream errors at all, but from the pairwise coreference decision on correctly extracted and phrased events. This is consistent with the over-linking we already report (silver links reduced from 7,277 to 4,459 during verification): most spurious links are coreference false positives (528 of 730 silver-only links), not artifacts of faulty extraction. Upstream event and phrase quality is therefore the smaller contributor; the dominant and

| Setting | Model | Macro F1 per Modality Pair | | | | | | Link-Level (Overall) | | | Cluster-Level (Overall) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T–T | T–V | T–A | V–V | V–A | A–A | P | R | F1 | MUC | $B^3$ | $CEAF_e$ | CoNLL-F1 |
| Using GT Events and Details | Llama3.3-70B | 62.5 | 57.8 | 59.4 | 63.2 | 58.7 | 60.1 | 60.23 | 74.88 | 61.47 | 63.82 | 51.24 | 44.48 | 53.18 |
| | Qwen3-30B | 58.7 | 53.4 | 55.9 | 59.1 | 54.2 | 56.8 | 53.94 | 80.22 | 57.81 | 61.17 | 48.83 | 42.01 | 50.67 |
| | GPT-4o | 63.2 | 58.1 | 61.4 | 63.8 | 59.2 | 61.7 | 61.15 | 78.33 | 62.04 | 60.44 | 47.58 | 41.74 | 49.92 |
| | Gemini2.5-Flash | 62.4 | 57.2 | 60.8 | 62.9 | 58.3 | 61.0 | 58.77 | 82.01 | 61.33 | 62.11 | 49.27 | 42.94 | 51.44 |
| | Qwen2.5-Omni-7B | 52.1 | 46.8 | 49.3 | 52.8 | 47.5 | 50.2 | 55.18 | 82.94 | 51.07 | 41.87 | 27.33 | 20.44 | 29.88 |
| | Qwen2.5-Omni-7B (AutoregQwen) | 60.8 | 55.3 | 58.4 | 61.2 | 56.1 | 58.9 | 58.63 | 73.41 | 59.82 | 56.42 | 43.81 | 37.50 | 45.91 |
| | Qwen2.5-Omni-7B (SpanQwen) | 59.9 | 54.2 | 57.6 | 60.3 | 55.4 | 58.1 | 55.42 | 74.88 | 58.97 | 57.88 | 44.72 | 39.33 | 47.31 |
| Using Predicted Events (End-to-end) | GPT-4o | 18.2 | 6.8 | 0.0 | 3.2 | 0.0 | 0.0 | 10.92 | 15.41 | 12.27 | 22.14 | 15.22 | 10.43 | 15.93 |
| | Gemini2.5-Flash | 23.8 | 9.4 | 0.0 | 5.8 | 0.0 | 0.0 | 15.18 | 17.22 | 16.41 | 27.48 | 18.73 | 12.44 | 19.55 |
| | Qwen2.5-Omni-7B | 7.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 4.21 | 6.83 | 4.55 | 13.84 | 9.12 | 5.69 | 9.55 |
| | Qwen3-Omni-30B | 11.0 | 0.6 | 0.0 | 0.5 | 0.0 | 0.0 | 7.42 | 12.18 | 8.55 | 19.42 | 13.17 | 8.54 | 13.71 |
| | Qwen2.5-Omni-7B (AutoregQwen) | 25.9 | 2.7 | 2.9 | 1.2 | 2.0 | 2.7 | 17.83 | 23.51 | 18.92 | 31.74 | 22.18 | 16.31 | 23.41 |
| | Qwen2.5-Omni-7B (SpanQwen) | 56.8 | 52.3 | 43.6 | 44.5 | 42.1 | 41.9 | 51.93 | 62.81 | 51.27 | 66.82 | 52.91 | 48.81 | 56.18 |
| | AutoEventAnnotator | 66.2 | 54.8 | 60.2 | 58.1 | 55.4 | 62.3 | 64.82 | 72.55 | 63.91 | 72.44 | 58.12 | 54.75 | 61.77 |

Table 26: Detailed coreference breakdown on the Human test set. Macro F1 per modality pair is a link-level binary score (coreferent / non-coreferent). Macro P, R, F1 are aggregate link-level metrics across all pairs. MUC, $B^3$, and $CEAF_e$ are cluster-level metrics whose average gives CoNLL-F1. In the end-to-end setting, pairs with no extracted events score 0.00. T, V, A denote text, video, and audio.

most improvable factor is the coreference decision itself.

## S Breakdown of End-to-End Coreference Collapse

A persistent finding in our experiments is the sharp drop in coreference performance between the GT and end-to-end (E2E) settings (Table 26, 27). This appendix explains *why* the collapse occurs and shows that its magnitude is mainly predicted by per-modality event extraction quality (Table 5).

**The Cascade Principle.** End-to-end coreference operates on model-predicted events. A coreference link between two events can only be evaluated if both events were successfully extracted and aligned to their gold counterparts. Consequently, if a model fails to extract events in a given modality, every coreference pair type that involves that modality automatically receives zero credit, no matter how strong the model's underlying coreference reasoning is.

**Empirical Evidence of the Cascade.** The per-pair breakdown in Table 26 and 27 matches this prediction precisely:

- Every model with audio extraction F1 of 0.00 (GPT-4o, Gemini-2.5-Flash, Qwen2.5-Omni-7B) scores *exactly* 0.00 on T–A, V–A, and A–A end-to-end pairs, despite achieving 48–62 CoNLL-F1 in the GT setting.
- Qwen2.5-Omni-7B, which fails on both video and audio (0.00 / 0.00), retains only T–T at 6.81 (AI) in end-to-end; the remaining five pair types all collapse to 0.00.
- GPT-4o and Gemini, with partial video extraction (14.12 / 26.50 AI F1) but no audio, recover marginal T–V scores (8.21 / 12.78) and weak V–V (4.31 / 7.82), but stay at exactly 0.00 on every audio-involving pair.
- AutoregQwen, with strong text extraction (72.46 AI F1) but near-zero video (3.17) and audio (3.42), reproduces the same pattern at a slightly higher floor: T–T at 27.85, while all cross-modal and non-text pairs remain below 4.

**Recovery Requires Cross-Modal Extraction.** The only systems that achieve non-trivial end-to-end coreference across all six pair types are those that extract events reliably across all three modalities. SpanQwen (text 68.41, video 32.21, audio 20.57 AI F1) produces non-zero scores on every pair type, ranging from 42.02 (V–V) to 53.80 (T–T). AutoEventAnnotator, with the best extraction (90.23 / 78.68 / 82.44 AI F1), achieves the most balanced per-pair performance (56.31–67.85). These systems empirically confirm the cascade: once extraction is reliable across modalities, end-to-end coreference recovers proportionally.

**Implication.** End-to-end coreference is not bottlenecked by coreference reasoning itself but by the upstream ability to surface candidate events from each modality. This is consistent with the GT-setting numbers in the same tables, where even models with poor extraction perform respectably

| Setting | Model | Macro F1 per Modality Pair | | | | | | Link-Level (Overall) | | | Cluster-Level (Overall) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T–T | T–V | T–A | V–V | V–A | A–A | P | R | F1 | MUC | $B^3$ | $CEAF_e$ | CoNLL-F1 |
| Using GT Events and Details | Llama3.3-70B | 62.2 | 55.4 | 52.2 | 64.2 | 58.9 | 55.3 | 57.01 | 77.96 | 59.78 | 64.21 | 43.83 | 42.62 | 50.22 |
| | Qwen3-30B | 61.6 | 55.0 | 51.7 | 63.6 | 58.7 | 54.8 | 56.17 | 78.05 | 59.24 | 63.70 | 42.81 | 40.95 | 49.15 |
| | GPT-4o | 62.5 | 55.8 | 52.4 | 64.6 | 59.6 | 55.5 | 58.32 | 80.11 | 60.12 | 65.18 | 44.78 | 43.73 | 51.23 |
| | Gemini2.5-Flash | 61.5 | 54.8 | 51.6 | 63.4 | 58.6 | 54.8 | 56.11 | 79.58 | 59.18 | 62.30 | 42.07 | 41.28 | 48.55 |
| | Qwen2.5-Omni-7B | 49.5 | 48.6 | 43.5 | 57.7 | 44.0 | 41.6 | 52.29 | 85.41 | 48.28 | 59.13 | 14.75 | 6.32 | 26.73 |
| | Qwen2.5-Omni-7B (AutoregQwen) | 59.5 | 49.8 | 53.2 | 59.7 | 53.4 | 55.2 | 55.18 | 76.73 | 57.60 | 64.25 | 40.51 | 38.35 | 47.70 |
| | Qwen2.5-Omni-7B (SpanQwen) | 58.8 | 48.9 | 51.4 | 58.6 | 52.3 | 53.9 | 56.23 | 77.73 | 56.29 | 61.43 | 39.61 | 37.35 | 46.13 |
| Using Predicted Events (End-to-end) | GPT-4o | 22.4 | 8.2 | 0.0 | 4.3 | 0.0 | 0.0 | 11.83 | 14.22 | 12.55 | 17.08 | 18.65 | 14.70 | 16.81 |
| | Gemini2.5-Flash | 28.4 | 12.8 | 0.0 | 7.8 | 0.0 | 0.0 | 14.27 | 16.83 | 15.18 | 20.51 | 22.41 | 18.34 | 20.42 |
| | Qwen2.5-Omni-7B | 6.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.42 | 7.18 | 4.18 | 8.94 | 9.79 | 8.03 | 8.92 |
| | Qwen3-Omni-30B | 12.7 | 0.8 | 0.0 | 0.4 | 0.0 | 0.0 | 8.31 | 11.27 | 9.18 | 12.49 | 13.66 | 11.17 | 12.44 |
| | Qwen2.5-Omni-7B (AutoregQwen) | 27.9 | 2.4 | 3.1 | 1.2 | 1.8 | 2.3 | 18.42 | 22.11 | 19.27 | 25.01 | 27.31 | 22.32 | 24.88 |
| | Qwen2.5-Omni-7B (SpanQwen) | 53.8 | 50.5 | 49.5 | 42.0 | 46.4 | 43.9 | 50.26 | 64.42 | 49.49 | 51.69 | 60.31 | 53.16 | 55.05 |
| | AutoEventAnnotator | 67.9 | 56.3 | 61.7 | 59.4 | 56.9 | 63.8 | 66.31 | 71.18 | 65.42 | 63.43 | 69.21 | 56.69 | 63.11 |

Table 27: Detailed coreference breakdown on the AI-assisted test set. Macro F1 per modality pair is a link-level binary score. P, R, F1 are aggregate link-level metrics across all pairs. MUC, $B^3$, and $CEAF_e$ are cluster-level metrics whose average gives CoNLL-F1. In the end-to-end setting, pairs with no extracted events score 0.00. T, V, A denote text, video, and audio.

on coreference (Human CoNLL-F1 in the 29–53 range for non-fine-tuned models), confirming that linguistic capacity for coreference is present; what fails is extraction. The practical takeaway is that improvements in per-modality event extraction translate into end-to-end coreference gains.

## T Qualitative Error Analysis

Figure 7 visualizes a sample cluster with its triggers, phrases, and coreference links. We use the same cluster to illustrate the error types that can arise in practice. We focus on video and audio, since these are the weakest modalities in our experiments, and we analyze the output of SpanQwen. The examples are drawn from two publicly available videos of the cluster shown in Figure 7 (and their audio tracks):

- Video 1: Cambodia – Advance of CPP troops
- Video 2: Cambodia – Ranariddh's successor

**Localization error (Video 1).**

- **Phrase:** Injured man covered with a blanket in Srey Snam, Cambodia, after the rocket attack on 15 July 1997.
- **Predicted time:** 4s–14s **GT time:** 11s–15s
- **Reason:** The trigger and entities are correct, but the temporal boundary is misaligned. The error is confined to the localization stage.

**Correct event (Video 2).**

- **Phrase:** Ung Huot addressing the political transition of Cambodia as new prime minister in Phnom Penh on 16 July 1997.
- **Time:** 2s–27s

**Missing-entity error in phrase generation, and duplicate events (Video 2).**

- **Predicted phrase:** Ung Huot speaks as Cambodia's new first prime minister in Phnom Penh about political transition.
- **GT phrase:** Ung Huot speaks as Cambodia's new first prime minister in Phnom Penh on 16 July 1997, addressing political transition.
- **Time:** 11s–29s
- **Reason:** The event is correctly understood, but the phrase omits a temporally individuating entity (the date). Because coreference is decided over phrases, this omission can affect a downstream link. Separately, when an event continues for a long time, the model sometimes predicts duplicate events.

**Correct event (Audio 1).**

- **Phrase:** National Armed Forces launched rockets on Srey Snam, Cambodia, capturing the FUNCINPEC stronghold on 16 July 1997.
- **Speech excerpt:** "National Armed Forces loyal to Cambodia's People's Party, the CPP, launched a rocket attack."
- **Time:** 2s–9s

**Coreference error from underspecified phrases (Video 1).** Two distinct events receive the same phrase and are wrongly linked.

- **Event 1 (14s–19s).** Predicted phrase: *Woman crying in Srey Snam after the attack.* GT phrase: *Woman beside an injured person crying in Srey Snam after the attack.*
- **Event 2 (20s–25s).** Predicted phrase: *Woman crying in Srey Snam after the attack.* GT phrase: *Woman crying in Srey Snam standing next to a soldier after the attack.*
- **Reason:** These are two separate events - in one, a woman mourns beside an injured or dead person; in the other, a woman cries while standing next to a soldier. The pipeline generated the same underspecified phrase for both, and the coreference stage incorrectly linked them. Had the distinguishing details been present in the phrases before coreference, this error would have been avoided.

## U Event Annotation Tool

Figure 10, 11, 12, 13 include snippets from our web-based event annotation tool.

## V Prompts

All the LLM prompts used in this dataset are provided in V.1–V.23.

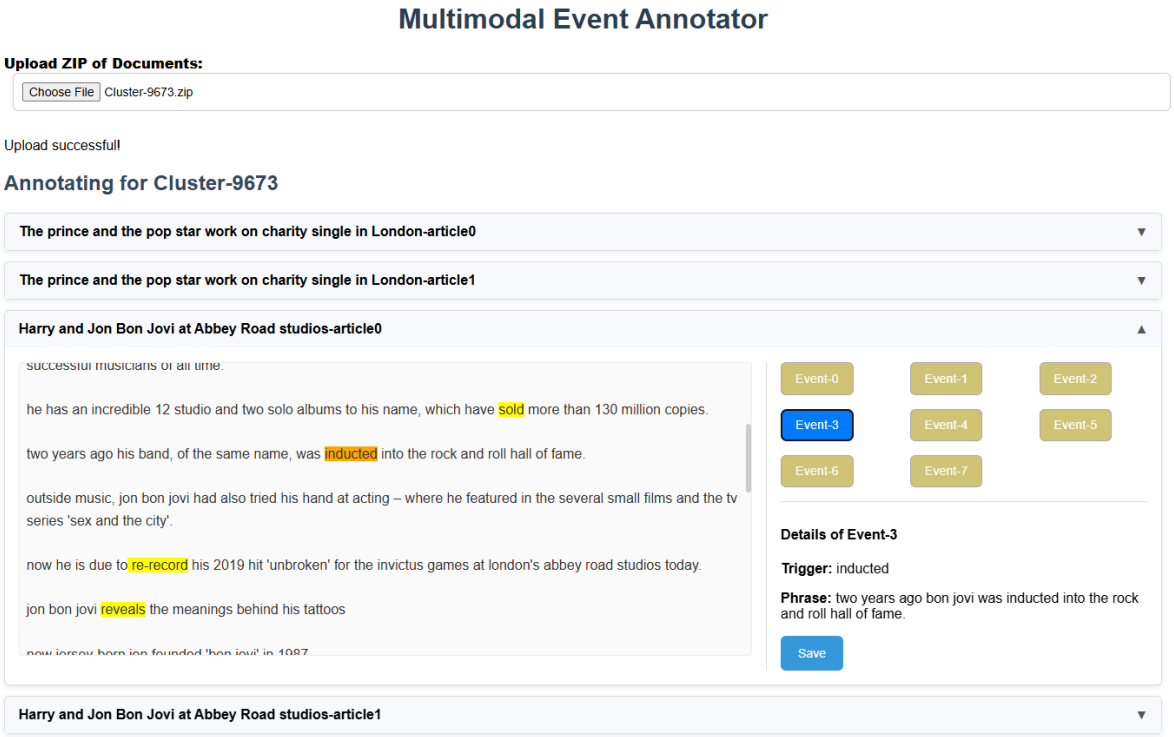



Figure 10: Annotation tool for text events.

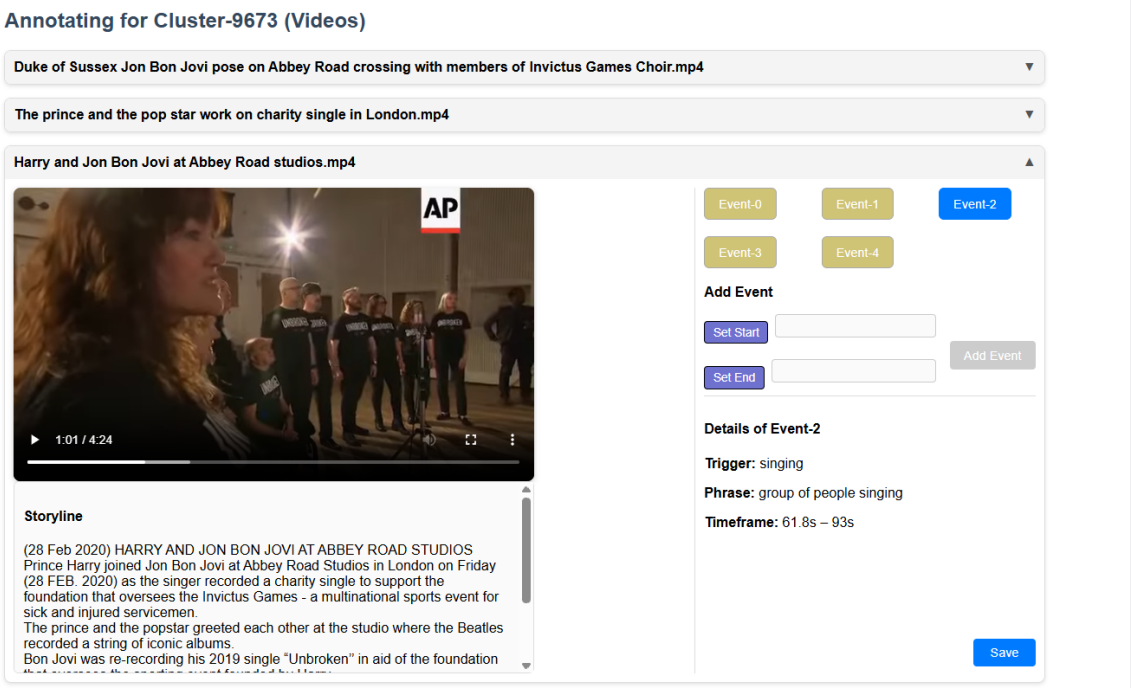



Figure 11: Annotation tool for video events.

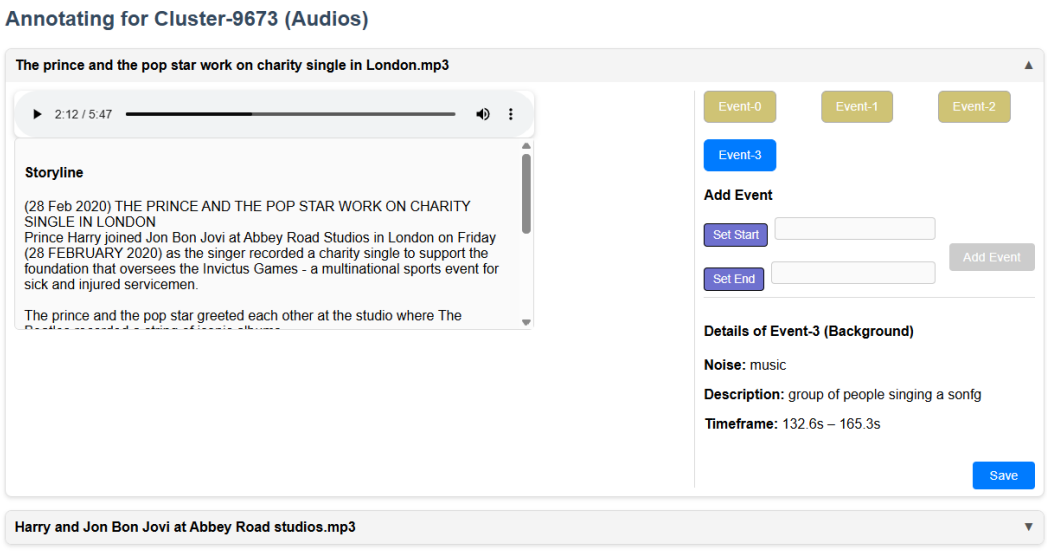



Figure 12: Annotation tool for audio events.

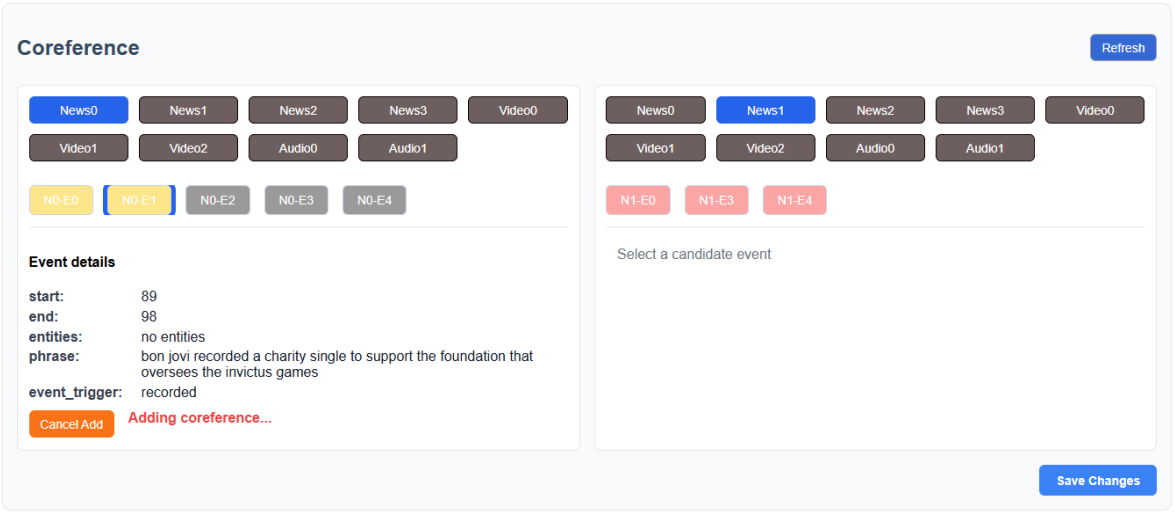



Figure 13: Annotation tool for coreferencing.

## Prompt V.1: Instruction for human annotation

**Event Trigger Identification**
An *event trigger* is the smallest word or phrase that explicitly indicates that a real-world event has occurred, was attempted, or resulted in a new state. Event triggers typically correspond to verbs, event-denoting nouns, or result-state expressions.

**Trigger Selection Rules**

1. **Prefer the main action word.** Select the word that directly expresses the event. For example, "He *died* yesterday" → *died*; "He was *arrested*" → *arrested*.
2. **Use event-denoting nouns when appropriate.** If a noun itself names the event, select the noun rather than an associated light verb. For example, "They held a *meeting*" → *meeting*.
3. **Result states count as events.** Words that express a resulting condition caused by an event should be tagged. For example, "Several people were *injured*" → *injured*.
4. **Verb–particle constructions.** If the verb and particle form a contiguous unit, tag both; if split, tag only the verb. For example, "step *down*" → *step down*.
5. **Ignore generic or habitual statements.** Only tag specific, time-bounded events, not general truths or habitual actions.
6. **Light verb constructions.** When a weak verb (e.g., *make*, *take*, *give*) appears with an event noun, select the noun. For example, "made an *arrest*" → *arrest*.
7. **Multiple events require multiple triggers.** If a sentence describes multiple distinct events, each should receive its own trigger.
8. **Minimality.** Always select the smallest possible span that uniquely expresses the event.

**What Not to Tag**
Do not tag entity roles, objects, locations, vague references, hypothetical or negated events, or continuous background states without a clear temporal boundary.

**Video and Audio Events**
The same trigger identification rules apply to all modalities. In addition, video and audio events must be aligned to a specific temporal interval.

- **Video events** must correspond to visible, time-bounded actions or state changes (e.g., a collision, movement, appearance, or disappearance).
- **Audio speech events** correspond to time-bounded acts of speaking or communicative actions (e.g., announcing, chanting, testifying).
- **Background audio events** correspond to distinct, recognizable sound events with a clear onset (e.g., explosion, gunshots, sirens), but not continuous ambient noise.

If an event cannot be confidently aligned to a specific time interval, it should not be annotated.

**Event Coreference Decisions**
Two event mentions corefer only if they refer to the same real-world micro-event.

1. **Same micro-action.** The actions must be identical at the incident level, allowing paraphrases but not different stages or roles.
2. **Same real-world incident.** Events must belong to the same reported occurrence, not merely the same topic.
3. **Entity consistency.** Missing details are acceptable, but conflicting entities, times, or locations indicate non-coreference.
4. **Conservative linking.** When there is uncertainty, do not assign a coreference link.

## Prompt V.2: Prompt for Event Trigger Identification from Text Articles

**System**
You are an expert in event trigger identification. I have an article, and I want to figure out event triggers from that article. You will be given the whole article as context and a specific sentence as input.
Your task is to find the event trigger from the sentence. While doing so, you will also need to output the reasoning why you think a word/phrase you selected was an event trigger.
Follow these rules carefully.

---

**DEFINITION**

- An event trigger is the minimal word or phrase that most clearly expresses that an EVENT occurred, was attempted, planned, or resulted in a new state.
- Triggers may be:
  - Main verbs: "attacked", "resigned"
  - Event-denoting nouns: "explosion", "meeting", "merger"
  - Adjectives/participles denoting result states: "injured", "dead", "indicted"

---

**RULES**

1. MAIN VERB PREFERENCE
   - Use the main verb of the event.
   - Ignore auxiliaries/modals/support verbs ("be", "have", "would", "try", "continue").

   ```
   Example: "He died yesterday." (Trigger =
       "died")
   Example: "He was arrested." (Trigger = "
       arrested", not "was")
   ```

2. VERB + NOUN CONSTRUCTIONS (Stand-Alone Noun Rule)
   - If the noun can stand alone to denote the event, use the noun.

"They held a meeting." (Trigger = "meeting")
"Hamas launched an attack." (Trigger = "attack")
"They held" (verb alone doesn't fully express the event)

3. VERB + ADJECTIVE/PARTICIPLE RESULTATIVES (Stand-Alone Adjective Rule)
   - If the adjective/participle denotes the resulting state alone, use it.

"The crash left 20 dead." (Trigger = "dead")
"The blast left dozens injured." (Trigger = "injured")

4. VERB-PARTICLE CONSTRUCTIONS
   - If contiguous, tag verb+particle together.

"He will step down tomorrow." (Trigger = "step down")

   - If split, use the main verb.

"He will take the company over." (Trigger = "take")

5. GENERIC STATEMENTS
   - Do not mark generic or habitual facts that describe what usually happens.

Examples: "People die in wars" -> no trigger
"Children go to school daily" -> no trigger

Only tag specific, time-anchored events
("Two soldiers died in the blast" -> trigger = "died")

6. LIGHT VERB CONSTRUCTIONS
   - If a light verb (make, take, give, have, do) combines with an eventive noun, use the noun.

Examples: "made an arrest" -> trigger = "arrest"
"had a meeting" -> trigger = "meeting"

If the noun is non-eventive ("made a cake"), use the verb ("made") instead.

7. MULTIPLE TRIGGERS IN ONE SENTENCE

"The attack killed seven and injured twenty."
(Three events: attack, kill, injure)

8. INVALID TRIGGERS
   - Reject entity/role nouns: "attacker", "nominee"
   - Reject instruments/places: "bomb", "hospital"
   - Reject entity heads: "the injured" (here "injured" = entity, not event)
   - Reject vague anaphors ("this event", "it") unless clearly coreferent with a prior explicit event.

9. NEGATION AND MODALITY

Examples: "did not attack" -> no trigger
"might resign" -> no trigger

Only tag events that are actually asserted or realized in the text.

10. MULTIPLE OR NO TRIGGERS
    - There can be multiple triggers in a sentence, or none at all.
    - If someone states/says something, find the triggers within that as well.

11. NO MODIFICATION RULE
    - Do not modify any word or phrase to make it look like a trigger.
    - Keep the sentence exactly as it is.

12. MINIMALITY
    - Keep the trigger as small as possible. In most cases, the trigger will be a single word.

---

**INPUT FORMAT**

<article>
...
(The whole article that you'll use as your context to understand what we are talking about)
</article>

<sentence>
...
(A sentence or phrase taken from the article. You'll need to find the event triggers here.)
</sentence>

---

**OUTPUT FORMAT**
**TRIGGERS**

- The answer of trigger identification needs to be wrapped within the <answer> </answer> tag.
- If the sentence doesn't contain any trigger, output

  <answer> no trigger </answer>

- If the sentence contains one or multiple triggers, wrap them inside <trigger> </trigger> tags within the <answer> </answer> tag.

  Example:

  <answer> <trigger-1> first event trigger 1</trigger-1>
  <trigger-2> second event trigger </trigger-2> </answer>

- You cannot change any word or phrase before tagging it.

- Also, mark the trigger in the input sentence where the trigger word appears.

  Example:

  ```
  Input:
  "At the upcoming Monterey Car Week 2015,
       Aston Martin is set to bring a
       selection of its best models"

  Trigger: "bring"

  Output sentence:

  At the upcoming Monterey Car Week 2015,
       Aston Martin is set to <event>bring</
       event> a selection of its best models
  ```

- This output should be wrapped within <sentence> </sentence> tag.

**REASON**

- Output the reasoning behind your event trigger decision (max 100 words).
- Wrap it within <reason> </reason> tag.
- Don't just restate the rules; provide meaningful reasoning grounded in the sentence context.

---

**EXAMPLE**

```
Input:

<article>

Margot Kidder, best known for playing reporter
     Lois Lane opposite Christopher Reeve in
     the "Superman" film franchise of the late
     1970s and early 1980s, has died at her
     Montana home. She was 69. Kidder's death
     was first reported by TMZ but confirmed by
     the Franzen Davis funeral home in
     Livingston, Montana, and later by her
     agent. "The actress and activist passed
     away on Sunday, May 13th, 2018 at her
     home," the funeral home said. A cause of
     death was not given.

</article>

<sentence>

Kidder's death was first reported by TMZ but
     confirmed by the Franzen Davis funeral
     home in Livingston, Montana, and later by
     her agent.

</sentence>

Output:

<answer> <trigger-1> death </trigger-1> <
     trigger-2> reported </trigger-2> <trigger
     -3> confirmed </trigger-3> </answer>

<sentence>
Kidder's <event>death</event> was first <event
     >reported</event> by TMZ but <event>
     confirmed</event> by the Franzen Davis
     funeral home in Livingston, Montana, and
     later by her agent.
</sentence>

<reason>
"Death" is a valid trigger as it denotes the
     resulting state and clearly expresses
     Kidder's passing. "Reported" and "
     confirmed" denote two different actions
     about the news of Kidder's death. Hence,
     they are valid triggers.
</reason>
```

## Prompt V.3: Prompt for Entity Extraction for Event Triggers of Text Articles

**Task Description**
You are given a full news article where one event trigger has already been annotated. The trigger is wrapped in <event> </event> tags.
Your task is to extract **all entities directly related to that event trigger**.

---

**Core Goal**
Identify every entity that participates in, causes, enables, is affected by, or results from the event denoted by the trigger. You must capture **all roles** (actor, object, cause, result, time, place) if they can be inferred from the article.

---

**Entity Extraction Rules**
**(1) Include:**

- **Actors / Participants:** people, organizations, or entities performing or experiencing the event.
- **Objects / Patients:** the targets or entities affected by the event.
- **Time / Location:** if explicitly stated or implied nearby.
- **Causes / Instruments / Results:** if they clearly relate to this event.
- **Contextual enablers:** large-scale contexts (institutions, conditions, environments) **if they directly influence this event**.
- Look **across the entire article** — not just the event's sentence — to identify who did what, when, where, and why.

**Typically, each event will have 3–6 entities of mixed types (e.g., actor, object, time, place, cause).**
**Do not stop after identifying only one entity.**

**(2) Exclude:**

- Generic or static descriptions, unrelated background details, pronouns, or repeated mentions not causally tied to the event.
- Hypothetical or comparative statements.
- Long phrases — keep entities short, specific, and factual.

**(3) Step-by-step reasoning:**

1. Identify the actor and object directly tied to the trigger.
2. Check earlier and later sentences for time, cause, or effect related to this same event.
3. Include inferred but clear entities (e.g., "Olympic Games," "Los Angeles," "gold medal," "age 17").
4. If the event implies a known context (e.g., competition, location, organization), include it concisely.

---

**Input Format**

```
<article>
The full article containing one <event> tag.
</article>

<event>
The event trigger (the word inside <event> ...
     </event>).
</event>
```

---

**Output Format**

Wrap the comma-separated list of entities inside <entity> </entity> tags.
Provide a concise reasoning inside <reason> </reason> (limit 60 words).

```
Example Output:

<entity>Actor, Object, Time, Place, Cause,
    Result</entity>

<reason>
The entities capture who performed the event,
    what was affected,
and when/where it occurred.
</reason>
```

## Prompt V.4: Prompt for Event Phrase Generation for Text Article

**Task Description**

You are given a full news article where the event trigger has already been annotated. The trigger is wrapped with tags in the form <event> EVENT TRIGGER </event>.
You'll have the whole article as your context. You'll also be given the entities (e.g., actors, locations, causes, instruments, time, results, etc.) related to the event trigger as input.
Your job is to generate one short, informative sentence (or sub-sentence) for that event trigger.

---

**Instructions**

1. Use the event trigger properly:
   - Use the trigger word as the core of the sentence.
   - The trigger should be the main focus of your sentence.
   - You must include the trigger event in the sentence.
   - Include the suitable entities that will be provided for that event.
2. Keep the sentences:
   - Concise: don't make them too long.
   - Expressive and informative: highlight the meaning of the event.
   - Distinct: Try to make the sentence stand apart. It should include distinctive information related to that trigger so the sentence is clearly separable from other such sentences.
3. You may also:
   - Add additional relevant entities or context from the article if it helps make the sentence more informative.

---

**Input Format**

```
<article> </article>
(You will be given an article as input. There
    will be only one <event> </event> tag in
    the whole article. The word/phrase inside
     the article will be the event trigger.
First, you'll have to look for the event
    trigger inside the <event> </event> tag
    and then attempt to infer its entity.)

<event> </event>
(Even though the trigger will be annotated
    inside the article, you will be given
    this as separate input as well.)

<entity> </entity>
(all the entities related to this event
    trigger)

<reason> </reason>
(the reason why each entity was chosen)
```

---

**Output Format**

- You will have to output the sentence, following the rule.
- It is not mandatory to include all the entities from <entity> </entity> input to make the sentence or sub-sentence. You can take the ones that are more suitable to make the sentence unique, distinct, and expressive.
- Don't make the sentence more than 20 words in any way.
- Output the sentence wrapped within <sentence> </sentence> tag.

---

**Example**

```
Input:

<article>

Satellite images from a NOAA (National Oceanic
     and Atmospheric Administration)
    satellite are capturing the massive smoke
     plumes rising over wildfires burning in
    Northern
```

```
California. A Northern California wildfire
    that is now the third-largest in state
    history had burned for weeks, mostly in
    remote wildland areas with few people,
    before
it roared through the little mountain
    community of Greenville, driven by
    shifting winds and bone-dry vegetation.
    Now the town is in ashes after hot, dry,
    gusty
weather drove the fire through the Gold Rush-
    era Sierra Nevada community of about
    1,000. The blaze <event>incinerated</
    event> much of the
downtown that included wooden buildings more
    than a century old.

</article>

<event> incinerated </event>

<entity> Actor: Northern California Fire,
    downtown buildings, Location: Greenville,
     Extent: much of downtown </entity>

<reason> The trigger incinerated refers to the
     town being demolished by the fire. We
    can infer the main actor was the Northern
     California Fire that ravaged through the
     town of Greenville and burned much of
    its downtown buildings. So these entities
     make the most sense if I infer globally.
     </reason>

Output:

<sentence> The blazing fire incinerated much
    of downtown Greenville, destroying century
    -old wooden buildings. </sentence>
```

## Prompt V.5: Prompt for Event Trigger Extraction from Full Video Descriptions

**Task Description**
You are an expert in identifying **event triggers from full video descriptions**. You will be given the entire textual description of a video (produced by a video-language model). Your task is to find **all event triggers** that describe *visible, time-bounded actions or state changes* that occur anywhere in the video.

---

**DEFINITION**
A **video event trigger** is the minimal word or phrase that expresses a **visually perceivable event** — something that *happens*, *changes*, *moves*, *interacts*, or *transforms* in the scene.
An event occurs **only if there is a visible state change** in the video. Each trigger must correspond to a distinct, visually observable moment or action in the video timeline.

---

**EXAMPLES OF VISUAL STATE CHANGES**
The following are valid event triggers.

```
"A car crashes into a pole." >> Trigger = "
    crashes"
"Smoke rises from the building." >> Trigger =
    "rises"
"Firefighters rescue people." >> Trigger = "
    rescue"
"The building collapses." >> Trigger = "
    collapses"
"Crowds gather in the square." >> Trigger = "
    gather"
"The crowd disperses." >> Trigger = "disperses
    "
"A man picks up a bag." >> Trigger = "picks up
    "
"Fire starts spreading." >> Trigger = "
    spreading"
```

---

**NON-EVENTIVE / INVALID CASES**
The following are not event triggers

```
**Camera or perception verbs**: "shows", "
    appears to show", "focuses", "cuts to", "
    we see".
**Static descriptions**: "is standing", "is
    wearing", "is visible".
**Scene-setting context**: "It is nighttime."
    "The stadium is full."
**Generic/habitual facts**: "People usually
    gather here."
**Editorial interpretations**: "It seems
    chaotic.", "The situation looks tense."
```

---

**RULES**

1. **VISUAL STATE RULE (Highest Priority)**
   - Select only triggers that correspond to observable, dynamic, or state-changing actions visible in the video.
   - A visual event occurs when a person, object, or scene changes — moves, interacts, appears, disappears, or transforms.
   - Exclude static or purely descriptive conditions such as posture, appearance, or environmental facts.
   - Example (event): The building **collapses**. (trigger = "collapses")
   - Example (not event): The building **is large**. ("is" describes a state, not a change)
   - Example (event): Smoke **rises** from the building. (trigger = "rises")
   - Example (not event): Smoke **is visible** above the building. ("is visible" is static)
2. **WHOLE-DESCRIPTION COVERAGE**
   - Read the entire video description as one coherent sequence of scenes.
   - Identify all distinct events that occur anywhere in the description, even if spread across multiple sentences.
   - Each event must represent a unique state change or distinct visual action.
   - Capture a comprehensive summary of everything that visibly happens in the video.
3. **DEDUPLICATION AND UNIQUENESS**
   - If the same event (same action, same participants) appears repeatedly, include it only once for that scene.

- Example: People are seen **walking** in groups. People are **heading** toward different areas. (Only one event trigger kept: "walking")
- If the same event continues across multiple sentences describing a single continuous activity, record it once per scene.
- Separate distinct events if they differ in action type, participants, or outcome.

4. **MAIN VERB PREFERENCE**
   - Choose the main lexical verb that expresses the event.
   - Ignore auxiliaries, helpers, and tense markers such as "is", "was", or "have been".
   - Example: The crowd **cheered** after the speech. (trigger = "cheered")
   - Example (not event): The crowd **was cheering** after the speech. ("was" is auxiliary)
5. **VERB + PARTICLE / PHRASAL VERBS**
   - Keep the entire phrasal verb if both parts together express the action.
   - Example: The soldier **set off** a flare. (trigger = "set off")
   - Example: The official **stepped down** from his role. (trigger = "stepped down")
   - Do not split such phrases into partial triggers like "set" or "off" separately.
6. **EVENTIVE NOUNS AND ADJECTIVES**
   - Accept nouns or adjectives only if they directly imply a visible event or result.
   - Example (eventive noun): The **explosion** damaged several cars. (trigger = "explosion")
   - Example (result adjective): Several people were **injured**. (trigger = "injured")
   - Exclude static nouns such as "man", "woman", "city" or identifiers like "attacker", "victim".
   - Exclude material or descriptive properties such as "metallic", "bright", "large".
   - If a visible change is described implicitly (for example, "a damaged building"), infer the causative result as an implicit event trigger ("damaged").
7. **EXCLUDE CAMERA OR NARRATION ACTIONS**
   - Do not mark any actions belonging to the camera, narrator, or viewer.
   - Exclude all cinematographic and perceptual verbs including: "shows", "pans", "zooms", "cuts", "captures", "focuses", "frames", "records", "transitions", "reveals", "depicts", "appears to show", "is seen", "we see", "camera follows", "camera pans", "camera zooms", "camera captures".
   - Example (not event): The camera **zooms** in on the speaker. ("zooms" is a camera action)
   - Example (not event): The video **shows** a man walking. ("shows" belongs to the camera)
   - Only include events that occur within the scene itself, not how the video is filmed.
8. **STATIC CONFIGURATION RULE**
   - Do not mark postural or positional verbs such as "standing", "sitting", "seated", "lying", or "waiting" unless they describe a transition into or out of that state.
   - Example (event): He **stood up** from his chair. (trigger = "stood up")
   - Example (not event): He **is standing** near the wall. ("is standing" is a static state)
   - Example (event): The injured man **fell** to the ground. (trigger = "fell")
9. **PROGRESSIVE ACTIONS**
   - Keep continuous or ongoing dynamic actions that reflect real motion or process, such as "is running", "is fighting", "is burning".
   - Exclude actions that describe someone remaining still without change, such as "is sitting", "is standing", "is waiting".
   - If an action clearly involves motion or transformation, it qualifies as an event trigger.
10. **GENERIC OR HYPOTHETICAL STATEMENTS**
    - Reject statements that are habitual, potential, or hypothetical rather than observed.
    - Example (event): People **protested** in the capital. (trigger = "protested")
    - Example (not event): People **might protest** later. ("might protest" is hypothetical)
    - Example (not event): People **usually gather** here. ("usually gather" is habitual)
11. **SCENE-LEVEL DEDUPLICATION**
    - Within a single scene, if the same continuous event (for example, a speech, a march, or a protest) is described in multiple sentences, record it only once.
    - Example: The man **speaks** into a microphone. He continues **speaking** passionately to the crowd. (Single event trigger retained: "speaking")
12. **NO MODIFICATION RULE**
    - Do not modify, paraphrase, or generalize any words in the description.
    - Use the exact surface text as it appears in the input.
    - Example (event): "is **running**" remains "running" as the trigger. Do not replace or normalize.
13. **MINIMALITY**
    - Keep each trigger span short and precise, usually one word or a short verb phrase.
    - Do not include modifiers, subjects, or extra descriptive context.
    - Example (correct): The building **collapsed**. (trigger = "collapsed")
    - Example (incorrect): The **building collapsed suddenly**. (includes unnecessary context)
14. **SCENE TRANSITION EVENTS**

- Scene-level transitions such as "video opens", "scene shifts", "focus returns", or "video ends" are not events and should not be marked.
- Example (not event): The **video opens** with a crowd. ("video opens" is a structural phrase)
- Example (not event): The **camera pans** to the crowd. ("camera pans" is cinematographic)
- Only include events that describe visible changes within the world of the scene itself.

15. **IMPLICIT STATE-CHANGE EVENTS**

- If a visible change or result is implied without an explicit action verb, infer the event trigger from that state change.
- Example (event): The building is **damaged** and windows are broken. (trigger = "damaged")
- Example (event): The area appears **destroyed** after the blast. (trigger = "destroyed")

---

**INPUT FORMAT**

```
<video_description>
(The entire video description text)
</video_description>
```

---

**OUTPUT FORMAT**
**TRIGGERS:**

- Wrap the full list in <answer> </answer> tags.
- If there are no valid visual events, output <answer> no trigger </answer>.
- Otherwise, wrap each identified trigger in <trigger> </trigger> tags inside <answer>.

Example:

```
<answer> <trigger-1> crashes </trigger-1> <
    trigger-2> catches fire </trigger-2> <
    trigger-3> evacuate </trigger-3> </answer>
```

**MARKED DESCRIPTION:**

- Reprint the entire video description, marking all detected triggers with <trigger> </trigger> tags.

Example:

```
"A car <trigger>crashes</trigger> into a wall
    and <trigger>catches fire</trigger>.
    Firefighters <trigger>arrive</trigger> to
     <trigger>extinguish</trigger> the flames
    ."
```

Put this inside <description> </description> tag.
**REASON:**

- Explain (less than 250 words) why these triggers were selected.
- Describe how each expresses a visible change or action, and why others were excluded.
- Put it inside <reason> </reason> tag.

---

**EXAMPLE**

```
Input:
<video_description>
The video depicts a group of individuals
    gathered around a voting booth during an
    election event. The setting appears to be
     indoors, with a yellow wall in the
    background. A man wearing a white hat and
     traditional attire is seen interacting
    with the camera crew, gesturing towards a
     blue and white ballot box placed on a
    table. He seems to be explaining or
    demonstrating something related to the
    voting process. Other individuals,
    including security personnel and other
    attendees, are present in the scene, some
     holding cameras and recording the event.
     The atmosphere suggests a formal and
    organized event, likely a political or
    civic function.
</video_description>

Output:
<answer>
<trigger-1> gathered </trigger-1> <trigger-2>
    interacting </trigger-2> <trigger-3>
    gesturing </trigger-3> <trigger-4>
    explaining </trigger-4> <trigger-5>
    demonstrating </trigger-5> <trigger-6>
    holding </trigger-6> <trigger-7>
    recording </trigger-7>
</answer>

<description>
The video depicts a group of individuals <
    trigger>gathered</trigger> around a
    voting booth during an election event.
    The setting appears to be indoors, with a
     yellow wall in the background. A man
    wearing a white hat and traditional
    attire is seen <trigger>interacting</
    trigger> with the camera crew, <trigger>
    gesturing</trigger> towards a blue and
    white ballot box placed on a table. He
    seems to be <trigger>explaining</trigger>
     or <trigger>demonstrating</trigger>
    something related to the voting process.
    Other individuals, including security
    personnel and other attendees, are
    present in the scene, some <trigger>
    holding</trigger> cameras and <trigger>
    recording</trigger> the event. The
    atmosphere suggests a formal and
    organized event, likely a political or
    civic function.
</description>

<reason>
Each selected trigger corresponds to a **
    visible, dynamic action** occurring in
    the video: "gathered" (people moving
    together), "interacting" (social
    engagement), "gesturing" (body motion), "
    explaining/demonstrating" (speech and
    hand movement linked to instruction), and
     "holding/recording" (physical actions by
     camera operators). These express actual *
    state changes or movements* observable in
     the footage. Static or descriptive
    phrases like "appears to be indoors" and
```

```
    "suggests a formal event" were excluded,
    as they convey context but not events.
</reason>
```

---

**EVALUATION CRITERIA**
When deciding on triggers, prioritize in this order:

1. Visual state change or motion (highest priority)
2. Action or interaction involving actors or objects
3. Temporal distinctness — avoid duplicate mentions of the same event
4. Relevance to visible events in the video

## Prompt V.6: Prompt for Query Generation for Video Event Grounding

**Task**
You are an expert in multimodal video understanding and temporal event grounding. You are given a video description where one event trigger is annotated using `<event> ... </event>` tags. Your task is to generate a **short, visually grounded query phrase** that best describes the visible action or scene corresponding to the annotated event.

---

**Objective**
Create a short, clear phrase (not necessarily a full sentence) that a vision-language model like CG-DETR could use to localize this event in a video. The phrase must describe *what is visibly happening* — the action, key participants, and setting — in no more than **10 words**.

---

**Guidelines**

1. Use **the event trigger** naturally in the phrase.
2. Describe **only visible and concrete details** (people, actions, objects, places).
3. You may omit articles ("a", "the") if needed — focus on clarity and brevity.
4. The output can be:
   - a **short sentence**, or
   - a **sub-sentence or meaningful phrase**
5. Do not include any named entities in the query. A named entity is a specific real-world name or identifier mentioned in text — such as a person, organization, or location. Example: In "Barack Obama visited Paris," the named entities are Barack Obama (PERSON) and Paris (LOCATION). You can't include named entities in the query.
6. Keep it **under 10 words** and **visually localizable**.
7. Output must be wrapped inside `<query> </query>` tags.

---

**Input Format**

```
<description>
........ Video description annotated with one
    event trigger.............
</description>

<event>
The event which was annotated inside the video
     description
</event>
```

---

**Output Format**

```
just output the final query within <query> </
    query> tag. Follow all the rules given
    above.
```

---

**Example 1**

```
Input:
<description>
The video depicts a group of individuals <
    event>walking</event> through what
    appears to be an indoor corridor or
    hallway.
</description>

Output:
<query> people walking through indoor corridor
     </query>
```

---

**Example 2**

```
Input:
<description>
A man in a suit is being <event>escorted</
    event> by two uniformed officers through
    a crowd of photographers outside a
    building at night.
</description>

Output:
<query> officers escort man through crowd at
    night </query>
```

---

**Example 3**

```
Input:
<description>
Two individuals dressed in military camouflage
     <event>walking</event> side by side near
     a helicopter on an airfield.
</description>

Output:
<query> soldiers walking near helicopter on
    airfield </query>
```

---

**Example 4**

```
Input:
<description>
A group of photographers and journalists are <
    event>gathered</event> near a metal
    barricade outside a lit building.
</description>

Output:
<query> journalists gathered near barricade
    outside building </query>
```

---

**Example 5**

```
Input:
<description>
A group of travelers is <event>moving</event>
    through a hallway, each carrying bags and
     backpacks.
</description>

Output:
<query> travelers moving through hallway with
    bags </query>
```

---

**Example 6**

```
Input:
<description>
A helicopter is seen <event>taking off</event>
     from a dusty airfield surrounded by
    military vehicles.
</description>

Output:
<query> helicopter taking off from dusty
    airfield </query>
```

---

**Example 7**

```
Input:
<description>
Barack Obama is being <event>escorted</event>
    by two uniformed officers through a crowd
     of photographers outside a building at
    night.
</description>

Output:
<query> officers escort man through crowd at
    night </query>
```

---

**Your Turn**
Generate the query for the following description:

## Prompt V.7: Prompt for Cross-Modal Entity Extraction from Video Events

**Task Description**
You are given a video description where one event trigger has already been annotated. The trigger is wrapped in `<event> </event>` tags inside the video description. You are also given a related news article found alongside the video.
Your task is to extract **only the key entities directly related to that specific event trigger** by combining contextual evidence from **both the video description and the news article**.

---

**Core Goal**
Identify entities that **participate in, cause, enable, are affected by, or result from** the event indicated by the trigger. You must reason across both inputs — the video description (visual context) and the news article (textual context) — to ensure the extracted entities are **visually grounded** and **semantically supported** by both whenever possible.
If you are unsure about alignment or the entities cannot be reliably associated with the event, output:

```
<entity> No Entity </entity>
```

---

**Entity Extraction Rules**
**(1) Include ONLY**
Entities that satisfy one or more of the following:

- **Actors / Participants:** people, organizations, objects, or forces that perform or experience the event.
- **Objects / Patients:** things directly affected by the event.
- **Time / Location:** explicitly stated or inferable from the inputs.
- **Causes / Instruments / Results:** entities that cause or result from the event.
- **Contextual enablers:** large-scale situations (e.g., wars, disasters, institutions, crises) that have a clear causal or defining influence on the event.
- Entities must belong to the **same time frame** as the event or be **immediate causes or effects**.
- Each entity should contribute to answering *who did what, where, when, why, or how*.

**(2) Cross-Modal Alignment**

- Look for **corresponding or supporting information** between the video description and the news article.
- If both mention similar scenes, actions, or participants, assume they refer to the same real-world event.
- Prefer entities confirmed or supported by **both** inputs.
- If only one source mentions a detail but it clearly aligns with the event, include it cautiously.
- Do **not** add entities from the article that relate to other events or background context.

**(3) Ambiguity Handling**

- If there is **no clear alignment** between the event in the video description and any event in the article, or if you cannot confidently identify matching participants, output `<entity> No Entity </entity>`.
- Do not guess, infer loosely, or assume based on general similarity.
- If multiple people or objects could match but you can't confirm which one is correct, do not extract ambiguous entities.

**(4) Exclude**

- Generic or persistent attributes (e.g., shapes, colors, materials, or generic visual features).
- Long descriptive phrases — keep entities short, precise, and specific.
- Pronouns — replace them with the full referent inferred globally.
- Irrelevant or background entities not directly connected to the event trigger.

- Broad impacts or general effects not described as part of the specific event.
- Historical or hypothetical entities unrelated to the current event.
- Include entities that are very related to the event trigger itself. Don't include entities that are important for global context understanding, but don't help with the event itself.

**(5) Think Sentence + Context**

- Start from the sentence containing the <event> tag in the video description.
- Expand outward to other sentences only if they help identify who performs or experiences the event, where, when, or why it occurs.
- Use the article as supplementary context to **disambiguate, name, or enrich** entities seen in the video.

**(6) Specificity & Compactness**

- Keep entity mentions short and meaningful (e.g., "firefighters", "Greenville", "collapsed bridge").
- Use explicit type markers when appropriate, such as:
  - Actor:
  - Object:
  - Location:
  - Time:
  - Cause:
  - Instrument:
  - Result:
- Avoid redundancy or listing the same entity multiple times under different forms.

---

**Input Format**

```
<description> ... </description>
(Video description containing a single <event>
     ... </event> tag.)

<event>
The event trigger which is annotated inside
    the description.
</event>

<article> ... </article>
(Related news article providing textual
    context.)
```

---

**Output Format**

```
Output the entities inside:
<entity> ... </entity>

Separate multiple entities by commas.

If no confident entities can be extracted,
    output:
<entity> No Entity </entity>

Then, provide the reasoning within:
<reason> ... </reason>

Limit the reasoning to within 100 words.
```

---

**Example1**

```
Input:
<description>
A group of young women dressed in purple choir
     robes <event>sing</event> inside a room,
     each holding sheet music
and standing in rows, focused on their
    performance. Bookshelves filled with
    books suggest it might be a library or
    study area.
</description>
<event> sing </event>
<article>
VOICE-OVER SCRIPT:
GIRL POWER HITS ALL THE RIGHT NOTES HERE IN
    CANTERBURY, ENGLAND...
(SOT: Abby Cox, chorister: ++standing
    alongside two other choristers, 16-year-
    old Poppy Braddy (left) and 12-year-old
    Chloe Chawner (centre)++)
"I think that's an amazing thought in the back
     of your mind that no girl has sung in
    this cathedral for over an amazingly long
     period of time. It's such an honor to be
     chosen."
THAT AMAZINGLY LONG PERIOD OF TIME IS ABOUT A
    THOUSAND YEARS. EVER SINCE THE DARK AGES,
     THE CANTERBURY CATHEDRAL HAS HOSTED AN
    ALL-BOYS CHOIR.
BUT NO GIRLS ALLOWED. UNTIL SATURDAY, WHEN
    ABOUT SIXTEEN OF THEM WERE SCHEDULED TO
    MAKE THEIR BIG DEBUT.
IT MARKS THE FIRST TIME IN HISTORY THAT AN ALL
    -GIRLS CHOIR WILL SING IN THE CHURCH.
(SOT: David Newsholme, director of girls'
    choir)
"Well, obviously they all wanted the
    opportunity to sing in this magnificent
    building. They felt rightly that this was
     going to be an historic event and they
    wanted to be a part of that."
AS THE MOTHER CHURCH OF THE WORLD'S 80 MILLION
     ANGLICANS, THE CHANGE IS SIGNIFICANT.
WHILE SOME FEAR THAT THE CANTERBURY CATHEDRAL'
    S ANCIENT TRADITIONS WILL BE LOST, THE
    DEAN ISN'T TOO WORRIED.
</article>
Output:
<entity> Actor: girls in purple choir robes,
    Location: CANTERBURY CATHEDRAL, date:
    Saturday, directed by: David Newsholme </
    entity>
<reason> The video description mentioned a
    girls' choir being performed. The article
     also states the same. So it can be
    inferred that the events are aligned. Now
     the video description says the girls
    were singing while wearing purple robes.
    Also, the information about the place,
    date, and director of this singing event
    was taken from the news article. </reason>
```

---

**Example2**

```
Input:

<description>
```

```
A woman in a pink dress <event>stands</event>
    in front of a movie poster for "A Kid
    Like Jake" and smiles, engaged in
    conversation with two men beside her.
A woman dressed in a black sequined outfit
    poses on a red carpet at an awards
    ceremony. She smiles and adjusts her
    jacket, preparing for a photo opportunity.

</description>

<event> stands </event>

<article>
29 Jul 2018) OCTAVIA SPENCER TO STAR AS BLACK
    HAIRCARE MOGUL MADAM WALKER
Octavia Spencer is bringing the story of black
     haircare mogul Madam C.J. Walker to
    television.
Netflix said Sunday (29 JULY 2018) that
    Spencer will produce and star in a
    limited series about the outsized life of
     Sarah Breedlove, who was known
    professionally as Walker.
The eight-episode drama is based on the book "
    On Her Own Ground" and includes
    basketball star LeBron James as a producer
    .
Walker became one of America's first self-made
     female millionaires by creating and
    marketing hair products for African-
    Americans at the turn of the 20th century.

</article>

Output:
<entity> woman in a pink dress </entity>
<reason> The description mentions two women.
    One in a pink dress and another one in a
    black dress. "Stands" event is related to
     the woman in a pink dress. However, the
    news article only mentions one woman -
    Octavia Spencer. There are no details
    about her outfit in the news article.
    Hence, it is impossible to understand
    between the woman in the pink dress and
    the woman in the black dress who Octavia
    Spencer is. So we can't extract any named
     entity from the articles. </reason>
```

## Prompt V.8: Prompt for Trigger Meaning Merging

```
You are an expert task follower. Your task is
    to merge different lists of entries into
    one. You will be given a list of event
    triggers. Event triggers are basically
one/two words that express the root of an
    event or action. You will be given a set
    of event triggers, and your task is to
    output a merged meaning of all those
    triggers.
Like what all those triggers have in common.
    Please use fewer than 8 words to express
    it.
wrap your output within <output> </output> tag.


Example:
Input:
[
    "stands",
    "featuring",
    "bright",
    "gesturing",
    "pointing",
    "maintaining",
    "shift",
    "remains"
],
Output:
<output> express how participants are
    positioned, or physically engaging </
    output>
```

## Prompt V.9: Prompt for Entity List Merging

```
You are an expert task follower. Your task is
    to merge different lists of entities into
     one. You will be given a list of
    entities/objects or named entities. You
    will
be given a set of different entities, and your
     task is to output all the unique
    entities. If some entities appear more
    than once, you just need to put them once
     in the final output.
wrap your output within <output> </output> tag.


Example:
Input:
[
    "group of six young men, location: Teen
    Choice 2019 red carpet event, event: Teen
     Choice Awards 2019, label: FOX, sponsor:
     JAX, performer: Wonho (Hoseok Lee),
    former band: Monsta X",
    "Wonho (Hoseok Lee), Monsta X, Highline
    Entertainment, Starship Entertainment,
    Teen Choice 2019, FOX, JAX",
    "Location: Teen Choice 2019 red carpet
    event, Group: six young men, Event: Teen
    Choice 2019, Brand: FOX, Brand: JAX,
    Visual: large yellow surfboard graphic",
    "Wonho (Hoseok Lee), Monsta X, Highline
    Entertainment, Starship Entertainment,
    Teen Choice 2019, FOX, JAX, April 10,
    2020",
    "Wonho (Hoseok Lee), Highline
    Entertainment, Starship Entertainment,
    Teen Choice 2019, FOX, JAX, blue carpet,
    red carpet event",
    "group of six young men, event: Teen
    Choice 2019, location: red carpet event,
    label: Highline Entertainment, artist:
    Wonho (Hoseok Lee), band: Monsta X",
    "Wonho (Hoseok Lee), Monsta X, Highline
    Entertainment, Starship Entertainment, 10
     April 2020",
    "Wonho (Hoseok Lee), Monsta X, Highline
    Entertainment, Starship Entertainment,
```

```
    Teen Choice 2019, FOX, JAX, April 10,
    2020"
],
Output:
<output> group of six young men, location:
    Teen Choice 2019 red carpet event, event:
     Teen Choice Awards 2019, label: FOX,
    sponsor: JAX, performer: Wonho (Hoseok
    Lee), former band: Monsta X, Monsta X,
    Highline Entertainment, Starship
    Entertainment, Teen Choice 2019, Visual:
    large yellow surfboard graphic, April 10,
     2020, blue carpet, red carpet event,
    artist: Wonho (Hoseok Lee), band: Monsta X

</output>
```

## Prompt V.10: Prompt for Sentence Merging

```
You are an expert task follower. Your task is
    to merge different lists of entries into
    one. You will be given a list of
    sentences. You just need to combine all
    the different dimensions of all sentences
     into one sentence and output it.
Your output sentence should not be more than
    20 words. And the final merged sentence
    should have some context overlap with
    each sentence of the list. Try to keep
    the most important context and named
    entities in the final output.
wrap your output within <output> </output> tag.


Example:
Input:
[
    "Six young men, including ex-Monsta X
    member Wonho, stand on the Teen Choice
    2019 red carpet with FOX and JAX sponsors
    ",
    "Wonho, former Monsta X member, is
    featured at Teen Choice 2019 event with
    FOX and JAX logos, marking his public
    appearance before joining Highline
    Entertainment",
    "The Teen Choice 2019 red carpet event
    glows bright with a colorful backdrop and
     a large yellow surfboard graphic",
    "Wonho of Monsta X gestures at Teen Choice
     2019 red carpet, marking his comeback
    after legal clearance in April 2020",
    "Wonho points at camera on blue carpet at
    Teen Choice 2019 event, signaling his
    return amid fan support",
    "Six young men maintain engagement at Teen
     Choice 2019 red carpet, representing K-
    pop's evolving scene post-Wonho's return",

    "Wonho of Monsta X shifts position at red
    carpet event following his return via
    Highline Entertainment on 10 April 2020",
    "Atmosphere remains lively at Teen Choice
    2019 red carpet, echoing Wonho's
    celebratory comeback on April 10, 2020"
],
Output:
<output> Wonho and six young men appeared at
    the Teen Choice 2019 red carpet with FOX,
     JAX, marking his comeback in 2020.
</output>
```

## Prompt V.11: Prompt for Speech Event Span Extraction

**Task Description**

You are given a full human speech transcription where the event trigger has already been annotated. The trigger is wrapped with tags in the form `<event> EVENT TRIGGER </event>`.

You'll have the whole transcription as your context. You'll have to read the whole transcription to understand the context of the event related to the trigger.

Now the event trigger itself is actually one or two words that don't convey any actual meaning. Instead, I am more interested in a very small phrase that conveys some meaning.

Your job is to select consecutive words before and after the trigger words, so that it turns into a very short phrase (that includes the trigger word/s)

The phrase has to be exactly taken out from the transcription using exact and consecutive words. No modification or rephrasing of the original transcription segment is allowed.

---

**Instructions**

1. Use the event trigger properly:
    - Use the trigger word as the core of the phrase.
    - The trigger should be the main focus of your phrase.
    - You must include the trigger event in the phrase.
2. Use the transcription:
    - You should use the trigger word and some consecutive words before or after it. You must take the words from the input transcription without altering or modifying it.
    - You have to form the phrase using the consecutive words before or after the trigger. You can't skip any words. You can't alter any words.
3. Keep the phrases-
    - Concise: don't make them too long.
    - Meaningful: the phrase must convey a meaning related to the event trigger .

---

**Input Format**

```
<transcription> 
```

**Output Format**

```
You will have to output the phrase, following
    the rule.
Don't make the phrase more than 12 words. The
    words have to be consecutive words that
    appear before and after the trigger word
    in the transcription. You are not allowed
     to modify any word or order.
Output the sentence wrapped within <phrase> </
    phrase> tags.
```

**Example1**

```
Input:
<transcription>
Satellite images from a NOAA (National Oceanic
     and Atmospheric Administration)
    satellite are capturing the massive smoke
     plumes rising over wildfires burning in
    Northern
California. A Northern California wildfire
    that is now the third-largest in state
    history had burned for weeks, mostly in
    remote wildland areas with few people,
    before
it roared through the little mountain
    community of Greenville, driven by
    shifting winds and bone-dry vegetation.
    Now the town is in ashes after hot, dry,
    gusty
weather drove the fire through the Gold Rush-
    era Sierra Nevada community of about
    1,000. The blaze <event>incinerated</
    event> much of the
downtown that included wooden buildings more
    than a century old.
```

defining crowd actions like "stadium chant" → keep as "chant" when central)

- Non-causal ambience: soundtrack, background score, ambient buzz, clock ticking.
- Natural ambience: wind, ocean waves, birds, rain, insects, forest sounds, unless they directly signal an event (e.g., "thunder clap")
- Generic noise artifacts: static, hum, distortion, microphone noise.
- Vehicle motion or signals: car horn, engine revving, braking, acceleration, plane or boat movement, etc.
- Passive engine states: engine idling, air conditioning, continuous vehicle hum, mechanical noise, mechanism, generic impact, etc.

---

**KEEP THESE (Eventive / Meaningful Sounds)**

- Physical impacts or contact: thud, bang, crash, clatter, knock, tap, etc.
- Explosions or weapon use: gunshot, explosion, artillery, blast, etc.
- Mechanical or object actions: door closing/opening, mechanism click, shutter, camera sound, etc.
- Emergency or signaling: siren, alarm, bell, whistle, etc.
- Tool or mechanical operation: drilling, hammering, typing, scraping, etc.
- Animal or non-verbal actions: bark, growl, gallop, flap (if relevant to the scene), etc.
- Crowd noise: clapping, laughing, etc.
- Other sounds that clearly align with visible events or causal changes in the video.

---

**SALIENCE RUBRIC (apply all)**

1. Causality: Does the sound indicate an action or state change (not just ambience)?
2. Event alignment: Does it match or clarify what the video/news describes?
3. Foregrounding: Is it likely foreground/attention-drawing, not constant wallpaper?
4. Specificity: Is it a concrete event (e.g., "door slam") vs vague ("background noise")?
5. Uniqueness: Does it help identify the moment (e.g., "gunshot") rather than generic "noise"?

If a sound satisfies most of the above, KEEP it. Otherwise, EXCLUDE.

---

**NORMALIZATION RULES**

- If a sound is kept, **shorten its name** to the most concise form possible.

  Example: "piano playing" → "piano", "car engine revving" → "engine revving".
- Avoid duplicates; if multiple mentions describe the same sound type, keep only one.
- The output must be a concise, comma-separated list of retained sound names.

---

**INPUT FORMAT**

```
<video desc>
A visual description of the video from where
    the audio has been taken.
Use this to assess the eventful-ness of the
    background noise.
</video desc>

<news>
The news article related to the video source.
Use it as a context to understand which
    background noise is important or not.
</news>

<background>
The background noises that you need to process
</background>
```

---

**OUTPUT FORMAT**

```
You need to provide your output wrapped around
     the <sound> </sound> tag.

If there are multiple salient background
    sounds, separate them by a comma.

Wrap the final comma-separated list in XML-
    style tags like this:
<sound>sound1, sound2, sound3</sound>

If no important or eventive sounds remain,
    output:
<sound>none</sound>
```

---

**REASONING STRATEGY (Implicit, don't output)**

1. Read the video description carefully — understand what events are occurring.
2. From the list of noises, understand which noises align with the context from the video description and news. Select those that align with these visual or causal events.
3. Discard everything else that's purely ambient or non-causal.
4. Normalize wording (short, lowercase nouns or short noun phrases).
5. Output the final list inside <sound> </sound> tags.

## Prompt V.13: Prompt for Eventive Sound Classification

You are an expert in distinguishing meaningful environmental sounds from background noises of a video. Your goal is to **filter and keep only the eventive or important background sounds**. You'll be given a sound. You have to output whether that sound is eventive or not.

---

**Eventive / Meaningful Sounds**

- Physical impacts or contact: thud, bang, crash, clatter, knock, etc.
- Explosions or weapon use: gunshot, explosion, artillery, blast, etc.
- Mechanical or object actions: shutter, camera sound, etc.
- Emergency or signaling: siren, alarm, bell, whistle, etc.
- Tool or mechanical operation: drilling, hammering, typing, etc.
- Animal or non-verbal actions: bark, growl, gallop, flap (if relevant to the scene), etc.
- Crowd noise: clapping, laughing, etc.
- Other sounds that clearly align with visible events or causal changes in the video.

---

**Non Eventive / Non Meaningful Sounds**

- Human sounds: dialogue, whispering, breathing, murmuring, coughing, tap, crowd hubbub etc. (Exception: clearly event-defining crowd actions like "stadium chant" $\rightarrow$ keep as "chant" when central)
- Non-causal ambience: soundtrack, background score, ambient buzz, clock ticking.
- Natural ambience: wind, ocean waves, birds, rain, insects, forest sounds, unless they directly signal an event (e.g., "thunder clap")
- Generic noise artifacts: static, hum, distortion, microphone noise, scrape, etc.
- Vehicle motion or signals: car horn, engine revving, braking, acceleration, plane or boat movement, etc.
- Passive engine states: engine idling, air conditioning, continuous vehicle hum, mechanical noise, mechanism, generic impact, scraping, door noise (opening/-closing/slamming) etc.

---

These are not exhaustive lists. You have to decide which sounds correspond to real, time-bounded, CAUSAL or ACTION-DENOTING events. Try to abide by the lists provided above, but reason yourself as well how a sound not present in the list would be classified.

---

**INPUT FORMAT**

`<sound> input sound description </sound>`

---

**OUTPUT FORMAT**

There are 2 parts

**1. verdict**

Decide whether the sound is eventive or not. The decision should be wrapped around the `<verdict> </verdict>` tag.

If the input sound is eventive, output

`<verdict> yes </verdict>`

otherwise

`<verdict> no </verdict>`

**2. Reason**

Also output the reason why you think the sound is eventful or not. If the input sound is not present on any of the lists, also include the reasoning why you think it was eventive or not.

Try to align the unseen sound with the most similar list of noises in the provided lists.

Output the reasoning within 30 words, and it must be wrapped around the `<reason> </reason>` tag.

## Prompt V.14: Prompt for Background Noise Event Sentence Generation

**Task Description**

You are given a background noise and the description of a video segment for the duration where the noise takes place. The video segment description doesn't generally include any entities. However, there is a news article that is associated with the whole video (not just the segment), which contains all the actions described in the video. It may not be fully consistent with the video progression, but the news article contains all the events (with entities) that are found in the video.

You need to understand the video segment description in light of the background noise. The background noise is our main item here. And based on the news article and the related segment description, you need to create a short sentence with entities (from the news article) to represent the description, highlighting the noise.

Your job is to generate one short, informative sentence (or sub-sentence) for that noise event.

---

**Instructions**

1. Use the noise event properly:
   - The noise word as the core of the sentence.
   - The noise should be the main focus of your sentence.
   - Include the suitable entities along with the video segment description to identify how the noise occurs.
   - Only use the related entities that make the sentence expressive
2. Keep the sentences:
   - Concise: don't make them too long.
   - Expressive and informative: highlight the meaning of the event.
   - Distinct: Try to make the sentence stand apart. It should include distinctive information related to that trigger so the sentence is clearly separable from other such sentences.
3. You may also:

- Add additional relevant entities or context from the video description or the news article (that is not present in the provided entities) if it helps make the sentence more informative.

---

**Input Format**

```
<noise> </noise>
(The noise description)

<segment desc> </segment desc>
(Video segment description where the noise
    takes place)

<news> </news>
(The news article related to the entire video
    (not just the segment).)
```

---

**Output Format**

```
You will have to output the sentence,
    following the rules.
Don't make the sentence more than 20 words in
    any way.
Output the sentence wrapped within <sentence>
    </sentence> tag.
```

---

**Example1**

```
Input:

<noise>
glass shatter
</noise>

<segment desc>
Police officers walk past a taped-off
    residential street with debris scattered
    near damaged row houses.
</segment desc>

<news>
A powerful gas explosion tore through a row of
     homes in west Baltimore early Monday
    morning, killing one resident and
    injuring several others.
Fire officials said the blast destroyed at
    least three houses on North Gilmor Street
     and damaged nearby buildings.
Baltimore Police and Fire Department crews
    responded within minutes, evacuating
    residents from surrounding blocks.
Mayor Brandon Scott said the city would assist
     displaced families while investigators
    determine the cause.
Utility crews shut off gas service to the area
     as emergency operations continued
    throughout the day.
</news>

Output:
<sentence>Glass shatter follows the Baltimore
    gas explosion as police inspect debris
    along North Gilmor Street.</sentence>
```

---

**Example2**

```
Input:

<noise>
crowd chant
</noise>

<segment desc>
A group of people stands behind police tape
    while officers monitor the scene near
    damaged homes.
</segment desc>

<news>
Residents gathered near the site of a deadly
    gas explosion in west Baltimore that
    destroyed multiple homes.
Community members expressed anger and grief
    over the loss of life and displacement of
     families.
Some residents called for accountability from
    city officials and utility companies.
Police maintained a perimeter along North
    Gilmor Street as the crowd grew
    throughout the afternoon.
City leaders promised a full investigation
    into the cause of the explosion.
</news>

Output:
<sentence>Crowd chant rises as Baltimore
    residents gather near North Gilmor Street
     after the fatal gas explosion.</sentence>
```

## Prompt V.15: Prompt for Multimodal Event Coreference Resolution

You are an expert event coreference resolver.
Your task is to decide whether TWO events refer to the SAME real-world event. One event may come from a TEXT news article. The other event may come from a VIDEO. This is NOT about scene similarity, topic similarity, or shared surroundings. You must only output YES if you are absolutely confident they describe the SAME real-world MICRO-INCIDENT.

---

**INPUT FORMAT (TEXT EVENT vs VIDEO EVENT)**
You will receive TWO events:

- One TEXT Event (from a news article)
- One VIDEO Event (from a video)

They DO NOT have the same structure.

---

**TEXT EVENT**
A Text Event contains:

1. `<event_trigger> ... </event_trigger>`
    - A short verb or phrase indicating the core action.
2. `<sentence> ... </sentence>`
    - A concise sentence or phrase with entities inserted.
    - Clearly states who did what, where, or when.
3. `<news> ... </news>`

- A full news article.
- The target event is explicitly marked with `<event> </event>`.
- This article is the PRIMARY source of truth for:
  - event identity
  - entities
  - location
  - time
  - real-world incident context

---

**VIDEO EVENT**
A Video Event contains:

1. `<event_trigger> ... </event_trigger>`
   - A short verb or phrase indicating the core action shown in the video.
2. `<video_sentence> ... </video_sentence>`
   - A concise sentence with relevant entities inserted (if identifiable).
   - Entities may be incomplete or implicit.
3. `<moment_query> ... </moment_query>`
   - A short natural-language query describing the event moment.
4. `<video_desc> ... </video_desc>`
   - A full description of the video.
   - The target event is explicitly marked with `<event> </event>`.
   - Describes how the event appears within the full video context.
5. `<news_article> ... </news_article>`
   - A multi-paragraph news article describing the real-world events related to the video.
   - This is the PRIMARY source of truth for event identity, entities, location, and time.
   - The article may not follow video order but contains all real events shown.

---

**HOW YOU SHOULD REASON**
You MUST reason in this order:
**STEP 1 — Understand each event independently**
Use event_trigger + sentence for text events and event_trigger + sentence + moment query for video events. Identify:

- What exact action happened?
- Who performed it?
- What object or target was involved?
- Where and when (if stated)?

**STEP 2 — Identify the MICRO-ACTION (CRITICAL)**
Focus on the precise physical or communicative action. Both Events are NOT the same if:

- One is speaking at a podium, another is chatting privately.
- One is handing flowers, another is throwing petals.
- One is announcing, another is reacting.

Even if they occur in the SAME place, SAME time, SAME ceremony, or contribute to the same bigger event – different micro-actions mean DIFFERENT events.
**STEP 3 — Anchor to the events' real-world news article**
Use news article to identify:

- The real-world incident being reported
- Location, date, cause, and key actors or more related and relevant entities

This step has the HIGHEST importance.
**STEP 4 — Compare the two events**
Output YES only if ALL conditions hold:

- The MICRO-ACTION is the same (allow paraphrase, not role change)
- The real-world incident is the SAME
- Entities are compatible (missing is OK, conflicting is NOT)
- Location and time do NOT contradict

---

**STRICT NO CONDITIONS (VERY IMPORTANT)**
You MUST output NO if ANY of the following is true:

- Different micro-actions (even within same ceremony or event)
- Different real-world incidents
- Different locations or times
- Conflicting entities
- Only generic similarity (e.g., two floods, two speeches, two greetings)
- You are uncertain or evidence is insufficient

When in doubt, ALWAYS say NO.

---

**OUTPUT FORMAT**

```
If two events corefer, or denote the same real
    -world event according to the above rules,
     output <verdict> yes </verdict>
else, output <verdict> no </verdict>

Then provide a brief justification for your
    answer (max 35 words):

<reason> your explanation </reason>
```

---

**EXAMPLES**

```
Example 1:
Text Event
<event_trigger>speaks</event_trigger>
<sentence>
Prime Minister Alexei Morozov delivers a
    formal address during the national
    independence ceremony in Minsk.
</sentence>
```

```
<news>
The national independence ceremony was held
    Sunday in Minsk, attended by senior
    government officials and foreign
    diplomats.
Prime Minister Alexei Morozov delivered a
    keynote address outlining economic
    reforms and national unity.
During the ceremony, Morozov <event>speaks</
    event> from the main podium while
    ministers and military officials stand
    behind him.
After the address, the formal program
    concluded and guests moved to a reception
     area.
</news>

Video Event
<event_trigger>talking</event_trigger>
<video_sentence>
The prime minister talks with guests after the
     independence ceremony.
</video_sentence>
<moment_query>
leader chatting informally with attendees
    after ceremony
</moment_query>
<video_desc>
After leaving the main stage, the prime
    minister is <event>talking</event> with
    diplomats and guests in a nearby
    reception area.
People stand in small groups exchanging
    greetings and casual remarks.
</video_desc>
<news_article>
The independence ceremony concluded with an
    informal reception.
Officials and invited guests interacted with
    Prime Minister Alexei Morozov following
    his official address.
These conversations were not part of the
    formal ceremony.
</news_article>

Output:
<verdict> no </verdict>
<reason>Both events occur at the same ceremony
    , but one is a formal public speech while
     the other is informal conversation,
    representing different micro-actions.</
    reason>

Example 2:
Text Event
<event_trigger>clearing</event_trigger>
<sentence>
Residents clear mud and debris from flood-
    damaged homes on the Greek island of Evia
     after record rainfall.
</sentence>
<news>
Severe flooding struck the Greek island of
    Evia on Monday following record rainfall.
Entire neighborhoods were inundated as
    floodwaters filled homes with mud and
    debris.
Emergency services reported extensive damage
    to residential buildings.
Residents were seen <event>clearing</event>
    mud from inside their homes using shovels
     and brooms as waters receded.
Authorities warned recovery efforts would take
     weeks.
</news>

Video Event
<event_trigger>cleaning</event_trigger>
<video_sentence>
Local residents clean mud and debris inside
    flood-damaged houses in Evia.
</video_sentence>
<moment_query>
people cleaning debris from flooded homes
    after storm
</moment_query>
<video_desc>
Inside damaged houses, residents are <event>
    cleaning</event> mud from floors and
    walls.
Buckets, shovels, and brooms are used as water
     stains and debris remain visible
    throughout the rooms.
</video_desc>
<news_article>
Heavy storms earlier this week caused flash
    flooding across Evia, Greece.
Floodwaters filled homes with mud and damaged
    furniture and personal belongings.
As conditions stabilized, residents began
    cleaning debris from inside their houses.
Local officials confirmed these cleanup
    efforts followed the same flooding
    incident.
</news_article>

Output:
<verdict> yes </verdict>
<reason>Both events describe residents
    removing flood debris inside homes in
    Evia following the same flooding incident,
     with matching actions, location, timing,
     and compatible entities.</reason>
```

## Prompt V.16: Prompt for Trigger Extraction from storyline

**System**
You are an expert in event trigger identification. I have an article, and I want to figure out event triggers from that article. You will be given the whole article as context and a specific sentence as input.
Your task is to find the event trigger from the sentence. While doing so, you will also need to output the reasoning why you think a word/phrase you selected was an event trigger. Follow these rules carefully.

---

**DEFINITION**
- An event trigger is the minimal word or phrase that most clearly expresses that an EVENT occurred, was attempted, planned, or resulted in a new state.
- Triggers may be:

  - Main verbs: "attacked", "resigned"
  - Event-denoting nouns: "explosion", "meeting", "merger"

- Adjectives/participles denoting result states: "injured", "dead", "indicted"

---

**RULES**

1. **Main verb preference**
   - Use the main verb of the event.
   - Ignore auxiliaries/modals/support verbs ("be", "have", "would", "try", "continue"). Example: "He died yesterday." (Trigger = "died")
   - "He was arrested." (Trigger = "arrested", not "was")

2. **Verb + Noun constructions (Stand-Alone Noun Rule)**
   - If the noun can stand alone to denote the event, use the noun.
     "They held a meeting." (Trigger = "meeting")
     "Hamas launched an attack." (Trigger = "attack")
     "They held" (verb alone doesn't fully express the event)

3. **Verb + Adjective/Participle resultatives (Stand-Alone Adjective Rule)**
   - If the adjective/participle denotes the resulting state alone, use it.
     "The crash left 20 dead." (Trigger = "dead")
     "The blast left dozens injured." (Trigger = "injured")

4. **Verb-Particle constructions**
   - If contiguous, tag verb+particle together.
     "He will step down tomorrow." (Trigger = step down)
   - If split, use the main verb.
     "He will take the company over." (Trigger = "take")

5. **Multiple triggers in one sentence**
   - Each event gets its own trigger.
     "The attack killed seven and injured twenty." (Three events: attack, kill, injure)

6. **Invalid triggers**
   - Reject entity/role nouns: "attacker", "nominee"
   - Reject instruments/places: "bomb", "hospital"
   - Reject entity heads: "the injured" (here "injured" = entity, not event)
   - Reject vague anaphors ("this event", "it") unless clearly coreferent with a prior explicit event.

7. There can be multiple triggers in a sentence. Also, there can be no triggers in a sentence. If someone states/says something, find the triggers within that as well.

8. Do not modify any word or phrase to make it look like a trigger. You need to keep the sentence as it is. Do not change any word or phrase.

9. Try to keep the trigger as small as possible. In the majority of cases, the trigger will only be a single word of the sentence.

---

**INPUT FORMAT**
<article>
...
...
(The whole article that you'll use as your context to understand what we are talking about)
</article>
<sentence>
....
(A sentence or phrase taken from the article. You'll need to find the event triggers here.)
</sentence>

---

**OUTPUT FORMAT**

- **Triggers :**
  - The answer of trigger identification needs to be wrapped within <answer> </answer> tag.
  - If the sentence doesn't contain any trigger, you should output <answer> no trigger </answer>
  - If the sentence contains one or multiple triggers, wrap them inside <trigger> </trigger> tag within the <answer> </answer> tag. For example, <answer> <trigger-1> first event trigger 1</trigger-1> <trigger-2> second event trigger </trigger-2></answer>. You can not change any word or phrase before putting it into the trigger tag. You must keep them as they are in the input sentence.
  - Also, you need to mark the trigger in the input sentence where the trigger word appears. While doing so, you cannot modify any word or phrase. For example, if the input sentence was "At the upcoming Monterey Car Week 2015, Aston Martin is set to bring a selection of its best models", and "bring" was supposed to be an event output "At the upcoming Monterey Car Week 2015, Aston Martin is set to <event>bring</event> a selection of its best models" This output should be wrapped within <sentence> </sentence> tag.

- **Reason:**
  - Output the reason behind your event trigger decision, within 100 words. The reason should be wrapped within <reason> </reason> tag. Don't just state any of the rules stated above as a reason. Do a deep analysis on why you think the word is an event trigger.

---

**Example**
Input:
<article>
Margot Kidder, best known for playing reporter Lois Lane opposite Christopher Reeve in the "Superman" film franchise of the late 1970s and early 1980s, has died at her Montana home. She was 69. Kidder's death was first reported by TMZ but confirmed by the Franzen Davis funeral home in Livingston, Montana, and later by her agent."The actress and activist passed away on Sunday, May 13th, 2018 at her home," the funeral home said. A cause of death was not given.
</article>
<sentence>
Kidder's death was first reported by TMZ but confirmed by the Franzen Davis funeral home in Livingston, Montana, and later by her agent.

</sentence>
Output:
<answer> <trigger-1> death </trigger-1> <trigger-2> reported </trigger-2> <trigger-3> confirmed </trigger-3> </answer>
<sentence>
Kidder's <event>death</event> was first <event>reported</event> by TMZ but <event>confirmed</event> by the Franzen Davis funeral home in Livingston, Montana, and later by her agent.
</sentence>
<reason>
"Death" is a valid trigger as it denotes the resulting state and clearly expresses Kidder's passing. "Reported" and "confirmed" denote two different actions about the news of Kidder's death. Hence, they are valid triggers.
</reason>

## Prompt V.17: Prompt for Entity Extraction from Storyline

**System**
You are given a full news article where the event trigger has already been annotated. The trigger is wrapped with tags in the form <event> EVENT TRIGGER </event>. Your task is to extract the key entities associated with that event trigger.

---

**Instructions**
When extracting entities:

1. Look at the entire article for context, not just the sentence where the trigger appears.
2. Entities can include:
    - Actors/participants (people, groups, organizations, objects that perform or undergo the event)
    - Locations
    - Time expressions
    - Causes, instruments, or results related to the event
    - Any other content that defines the event trigger and relates to it
3. Always pick the most specific and defining entities available.
4. Do not use pronouns (he, she, they, etc.) — replace them with the most specific entity you can infer globally from the whole article.
5. Avoid copying long sentence fragments. Focus on concise, meaningful entities.
6. You'll also need to output a reasoning on why you chose those entities. Why do you think those specific entities best describe the event?

---

**Input Format**
You will be given an article as input.
There will be only one <event> </event> tag in the whole article. The word/phrase inside the article will be the event trigger.
First, you'll have to look for the event trigger inside the <event> </event> tag and then attempt to infer its entity.

---

**Output Format:**
You will have to output the entities. If there are multiple entities related to that event trigger, separate each entity by a comma afterwards.
You must wrap all the comma-separated entities inside <entity> </entity> tag.
You'll need to output the reasoning behind your choice of entities inside the <reason> </reason> tag. Limit your reasoning within 100 words.

---

**Example:**
Input:
Satellite images from a NOAA (National Oceanic and Atmospheric Administration) satellite are capturing the massive smoke plumes rising over wildfires burning in Northern California. A Northern California wildfire that is now the third-largest in state history had burned for weeks, mostly in remote wildland areas with few people, before it roared through the little mountain community of Greenville, driven by shifting winds and bone-dry vegetation. Now the town is in ashes after hot, dry, gusty weather drove the fire through the Gold Rush-era Sierra Nevada community of about 1,000. The blaze **<event>incinerated</event>** much of the downtown that included wooden buildings more than a century old.
Output:
<entity> Actor: Northern California Fire, downtown buildings, Location: Greenville, Extent: much of downtown </entity>
<reason> The trigger incinerated refers to the town being demolished by the fire. We can infer the main actor was the Northern California Fire that ravaged through the town of Greenville and burned much of its downtown buildings. So these entities make the most sense if I infer globally. </reason>

## Prompt V.18: Prompt Important Event Selection

**System**
You are an intelligent agent that will work with event triggers from a text article. You will be given a text article, where the possible event triggers have already been marked. We will use a specific HTML type tag to denote the event triggers.
For example, the first event trigger will be written inside <event-1></event-1> tag and the n-th event will be written inside <event-n></event-n> tag. To be specific, In this sentence: Satellite images from a NOAA (National Oceanic and Atmospheric Administration) satellite are <event-1>capturing</event-1> the massive smoke plumes <event-2>rising</event-2> over wildfires <event-3>burning</event-3> in Northern California.
The event triggers are: 1. capturing, 2. rising, 3. burning.
Additionally, after the event trigger marked article, you will be given a list where you can find all the entities related to each event trigger that has already been mentioned in the article.
Now, not all the events are important to express the core theme of the article. You will analyze each event along with its entities with the context of the article and assess if that entity plays a vital role in expressing the main content/theme presented in the article. You'll only output 4-7 event triggers that are very, very important in the context of the article. You'll have to strictly follow the output format provided in the example below.

---

Example 1:

Input Article:
Satellite images from a NOAA (National Oceanic and Atmospheric Administration) satellite are <event-1>capturing</event-1> the massive smoke plumes <event-2>rising</event-2> over wildfires <event-3>burning</event-3> in Northern California. A Northern California wildfire that is now the third-largest in state history had <event-4>burned</event-4> for weeks, mostly in remote wildland areas with few people, before it <event-5>roared</event-5> through the little mountain community of Greenville, <event-6>driven</event-6> by shifting winds and bone-dry vegetation. Now the town is in ashes after hot, dry, gusty weather <event-7>drove</event-7> the fire through the Gold Rush-era Sierra Nevada community of about 1,000. The blaze <event-8>incinerated</event-8> much of the downtown that included wooden buildings more than a century old. The winds were expected to <event-9>calm</event-9> and <event-10>change</event-10> direction heading into the weekend, but that good news <event-11>came</event-11> too late for Gorman. She was <event-12>told</event-12> that her home <event-13>burned</event-13> down - but is waiting until she can <event-14>see</event-14> it with her own eyes to <event-15>believe</event-15> it's gone. The Dixie Fire, <event-16>named</event-16> for the road where it <event-17>started</event-17>, was still <event-18>raging</event-18> on Friday after <event-19>growing</event-19> overnight by 110 square miles (285 square kilometers), greater than the size of New York City. The fire remained at 35% <event-20>contained</event-20> Friday morning and now spans an area of 676 square miles (1,751 square kilometers). No <event-21>injuries</event-21> or <event-22>deaths</event-22> have been <event-23>reported</event-23> but the fire continued to <event-24>threaten</event-24> more than 10,000 homes. Officials have not yet <event-25>assessed</event-25> the number of <event-26>destroyed</event-26> buildings, but Plumas County Sheriff Todd Johns <event-27>estimated</event-27> on Thursday that "well over" 100 homes <event-28>burned</event-28> in and near the town. The three-week-old Dixie Fire was one of 100 active, large <event-29>fires</event-29> <event-30>burning</event-30> in 14 states, most in the West where historic drought has left lands parched and ripe for <event-31>ignition</event-31>. The fire's cause was under <event-32>investigation</event-32>, but the Pacific Gas & Electric utility has <event-33>said</event-33> it may have been <event-34>sparked</event-34> when a tree <event-35>fell</event-35> on one of the utility's power lines. On Thursday, the weather and towering smoke clouds <event-36>produced</event-36> by the fire's intense, erratic winds kept firefighters struggling to put out fires at shifting hot spots. Heat waves and historic drought tied to climate change have made wildfires harder to <event-37>fight</event-37> in the American West. Scientists say climate change has made the region much warmer and drier in the past 30 years and will continue to make weather more extreme and wildfires more frequent and destructive. The blaze <event-38>hit</event-38> Greenville from two angles and firefighters already were in the town trying to save it but first they had to <event-39>risk</event-39> their lives to save people who had <event-40>refused</event-40> to <event-41>evacuate</event-41> by <event-42>loading</event-42> people into cars to get them out, fire officials said. The flames also reached the town of Chester, northwest of Greenville, but crews managed to protect homes and businesses there, with only minor damage to one or two buildings, officials said. The fire was not far from the town of Paradise, which was largely destroyed in a 2018 wildfire sparked by PG&E equipment that killed 85 people, making it the nation's deadliest wildfire in at least a century. Gorman said she asks herself how another California town could be reduced to ashes.

Event List with their entities:
1. <event-1>capturing</event-1> - (Actor: NOAA satellite, smoke plumes, Location: Northern California)
2. <event-2>rising</event-2> - (Actor: smoke plumes, Cause: wildfires, Location: Northern California)
3. <event-3>burning</event-3> - (Actor: wildfires, Location: Northern California)
4. <event-4>burned</event-4> - (Actor: wildfire, Time: weeks, Location: Northern California, Extent: third-largest in state history)
5. <event-5>roared</event-5> - (Actor: wildfire, Location: Greenville, Cause: shifting winds, bone-dry vegetation)
6. <event-6>driven</event-6> - (Actor: wildfire, Cause: shifting winds, bone-dry vegetation)
7. <event-7>drove</event-7> - (Actor: hot, dry, gusty weather, wildfire, Location: Greenville, Result: town in ashes)
8. <event-8>incinerated</event-8> - (Actor: blaze, downtown buildings, Location: Greenville, Time: past, Extent: much of downtown)
9. <event-9>calm</event-9> - (Actor: winds, Time: weekend, Result: fire behavior change)
10. <event-10>change</event-10> - (Actor: winds, Time: weekend, Result: fire behavior change)
11. <event-11>came</event-11> - (Actor: Gorman, Time: too late)
12. <event-12>told</event-12> - (Actor: Gorman, Content: home burned down)
13. <event-13>burned</event-13> - (Actor: fire, Gorman's home, Location: Greenville)
14. <event-14>see</event-14> - (Actor: Gorman, burned home, Time: future expectation)
15. <event-15>believe</event-15> - (Actor: Gorman, Patient: burned home)
16. <event-16>named</event-16> - (Actor: Dixie Fire, Cause: road name)
17. <event-17>started</event-17> - (Actor: fire, Location: Dixie Road)
18. <event-18>raging</event-18> - (Actor: fire, Time: Friday)
19. <event-19>growing</event-19> - (Actor: fire Time: past night)
20. <event-20>contained</event-20> - (Actor: fire, Time: Friday morning)
21. <event-21>injuries</event-21> - (Actor: people, Time: Friday morning)
22. <event-22>deaths</event-22> - (Actor: people, Time: Friday morning)
23. <event-23>reported</event-23> - (Actor: officials, Time: Friday morning)
24. <event-24>threaten</event-24> - (Actor: fire, 10,000 homes)
25. <event-25>assessed</event-25> - (Actor: officials, number of destroyed buildings)
26. <event-26>destroyed</event-26> - (Actor: fire, buildings)
27. <event-27>estimated</event-27> - (Actor: Sheriff Todd Johns, number of homes burned, Time: Thursday)
28. <event-28>burned</event-28> - (Actor: fire, homes, Location: in and near Greenville)
29. <event-29>fires</event-29> - (Actor: multiple wildfires, Extent: 100+ active fires, Location: 14 states)

30. <event-30>burning</event-30> - (Actor: multiple fires, Location: 14 states, Cause: historic drought)
31. <event-31>ignition</event-31> - (Agent: dry lands, Cause: historic drought, Outcome: wildfires)
32. <event-32>investigation</event-32> - (Actor: officials, fire cause, Time: ongoing, Location: California)
33. <event-33>said</event-33> - (Actor: Pacific Gas & Electric, Content: possible cause of fire)
34. <event-34>sparked</event-34> - (Actor: fallen tree, Cause: hitting PG&E power line, Outcome: fire ignition)
35. <event-35>fell</event-35> - (Actor: tree, PG&E power line, Outcome: fire ignition)
36. <event-36>produced</event-36> - (Actor: fire, smoke clouds, Cause: intense, erratic winds)
37. <event-37>fight</event-37> - (Actor: firefighters, fire)
38. <event-38>hit</event-38> - (Actor: blaze, Location: Greenville)
39. <event-39>risk</event-39> - (Agent: firefighters, Cause: saving people)
40. <event-40>refused</event-40> - (Actor: residents, Location: Greenville)
41. <event-41>evacuate</event-41> - (Actor: residents, Location: Greenville)
42. <event-42>loading</event-42> - (Actor: firefighters, Patient: people, Outcome: escape fire)
Output:
4. <event-4>burned</event-4> - (Actor: wildfire, Time: weeks, Location: Northern California, Extent: third-largest in state history)
5. <event-5>roared</event-5> - (Actor: wildfire, Location: Greenville, Cause: shifting winds, bone-dry vegetation)
7. <event-7>drove</event-7> - (Actor: hot, dry, gusty weather, wildfire, Location: Greenville, Result: town in ashes)
8. <event-8>incinerated</event-8> - (Actor: blaze, downtown buildings, Location: Greenville, Time: past, Extent: much of downtown)
34. <event-34>sparked</event-34> - (Actor: fallen tree, Cause: hitting PG&E power line, Outcome: fire ignition)
39. <event-39>risk</event-39> - (Agent: firefighters, Cause: saving people)
Reasoning: The main theme of the article is about a large fire in northern California. It also talks about the severity and how it started. Based on this, the selected event triggers make the most sense.
Try to limit the extracted event within 4-7. Do not try to output all events, that's the worst.

## Prompt V.19: Prompt for Phrase Generation

**Task Description**
You are given a full news article where event trigger has already been annotated. The trigger is wrapped with tags in the form <event> EVENT TRIGGER </event>.
You'll have the whole article as your context. You'll also be given the entities(e.g., actors, locations, causes, instruments, time, results, etc.) related to the event trigger as input. Your job is to generate one short, informative sentence (or sub-sentence) for that event trigger

---

**Instructions**

1. Use the event trigger properly:
    - Use the trigger word as the core of the sentence.
    - The trigger should be the main focus of your sentence.
    - You must include the trigger event into the sentence.
    - Include the suitable entities that will be provided for that event.
2. Keep the sentences:
    - Concise: don't make them too long.
    - Expressive and informative: highlight the meaning of the event.
    - Distinct: Try to make the sentence stand apart. It should include distinctive information related to that trigger so the sentence is clearly separable from other such sentences.
3. You may also:

    Add additional relevant entities or context from the article if it helps make the sentence more informative.

---

**Input Format**

- <article> </article>

  (You will be given an article as input. There will be only one <event> </event> tag in the whole article. The word/phrase inside the article will be the event trigger. First, you'll have to look for the event trigger inside the <event> </event> tag and then attempt to infer its entity.)

- <event> </event>

  (even though the trigger will be annotated inside the article, you will be given this as separate input as well.)

- <entity> </entity>

  (all the entities related to this event trigger)

- <reason> </reason>

  (the reason why each entity was chosen)

---

**Output Format:**
You will have to output the sentence, following the rule.
It is not mandatory to include all the entities from <entity> </entity> input to make the sentence or sub-sentence. You can take the ones that are more suitable to make the sentence unique, distinct, and expressive.
Don't make the sentence more than 20 words in any way.
Output the sentence wrapped within <sentence> </sentence> tag.

---

**Example:**
Input:
<article>
Satellite images from a NOAA (National Oceanic and Atmospheric Administration) satellite are capturing the massive smoke plumes rising over wildfires burning in Northern California. A Northern California wildfire that is now the third-largest in state history had burned for weeks, mostly in remote wildland areas with few people, before it roared through the little mountain community of Greenville, driven by shifting winds and bone-dry vegetation. Now the town is in ashes after hot, dry, gusty weather drove the fire through the Gold Rush-era Sierra Nevada community of about 1,000. The blaze <event>incinerated</event> much of the downtown that included wooden buildings more than a century old.

</article>
<event> incinerated </event>
<entity>
Actor: Northern California Fire, downtown buildings, Location: Greenville, Extent: much of downtown
</entity>
<reason>
The trigger incinerated refers to the town being demolished by the fire. We can infer the main actor was the Northern California Fire that ravaged through the town of Greenville and burned much of its downtown buildings. So these entities make the most sense if I infer globally.
</reason>
Output:
<sentence> The blazing fire incinerated much of downtown Greenville, destroying century-old wooden buildings.
</sentence>

## Prompt V.20: Prompt For Checking if Two Event Mentions from Storyline are the same

You are given **two event candidates** (A and B). Each candidate includes: 1) a **context document** (the whole source text) where the event mention is **wrapped with <event>...</event>**; 2) the event **trigger** (string); 3) a list of **entities** relevant to that trigger.
**Goal**. Decide if A and B refer to the **same real-world event** (coreferential) or not.
**Decision policy (read carefully):**

- Consider **who/what**, **what happened**, **where**, and **when**. Use both explicit facts and clear inferences licensed by the context.
- Answer **YES** only if A and B plausibly denote the *same specific occurrence* (same participants or aliases, same incident, compatible time/place).
- If they are different incidents, different times/places, different participants, or only share type-level similarity (e.g., two protests in different cities), answer **NO**.
- If evidence is insufficient to conclude identity, answer **NO**.

**Step-by-step reasoning (concise):**

1. **Anchor extraction**. For A and B, list brief anchors: *Action*, *Time*, *Location*, *Participants/Entities*, and any *Unique details*.
2. **Alignment check**. Compare anchors across A and B; note matches and conflicts.
3. **Verdict**. Apply the decision policy above.

**Strict output format** (nothing else):

```
<reasoning>
[2–4 short lines: anchors for A and B,
key matches/conflicts, justification]
</reasoning>
<answer>yes|no</answer>
```

**Input format** (fill the placeholders):

```
[EVENT A]
Context:
{{DOC_A_WITH_EVENT_TAGS}}

Trigger: {{TRIGGER_A}}
Entities: {{ENTITIES_A}}

[EVENT B]
Context:
{{DOC_B_WITH_EVENT_TAGS}}

Trigger: {{TRIGGER_B}}
Entities: {{ENTITIES_B}}
```

---

[EVENT A] Context: Police evacuated Times Square at 8 p.m. after an **<event>explosion</event>** near 45th Street. Mayor Adams and NYPD reported minor injuries.
Trigger: explosion Entities: Times Square, 45th Street, NYPD, Mayor Adams, 8 p.m.
[EVENT B] Context: Witnesses heard a **<event>blast</event>** around 8 p.m. in Times Square; officers cleared the area and treated two people.
Trigger: blast Entities: Times Square, officers, 8 p.m., two injured
*Expected model output (format only):*

```
<reasoning>
A: explosion @8 p.m., Times Square, NYPD;
    injuries reported.
B: blast ~8 p.m., Times Square; officers; two
    injured.
Time/location match; action synonyms; injuries
     consistent; no conflicts.
</reasoning>
<answer>yes</answer>
```

## Prompt V.21: Prompt for Cluster Event Consistency Verification

**System**
You are an intelligent agent that analyzes clusters of news articles.
Each cluster contains multiple articles. All articles within a cluster are expected to describe at least one common real-world event. Your task is to verify whether this assumption is correct.

---

**Definition**

- An **event** is an action or occurrence that happened, such as a protest, a performance, an agreement, an arrest, a premiere, or a signing.
- An event is typically defined by the following components:
  - **Action:** what happened.
  - **Entities:** who was involved, where it happened, and when it happened.
- Not all entities must appear in every article. However, there must be sufficient overlap in the action and core entities to determine that the articles refer to the same underlying event.
- Articles mentioning the same people, organizations, or places but describing different events should **not** be considered a match.

---

**Instructions**

1. Carefully read all articles within the cluster.

2. Identify the event or events described in each article.

3. Compare the extracted events across all articles in the cluster.

4. Determine whether there exists **at least one event** that satisfies the following conditions:

   - The event appears in **every article** in the cluster.
   - The event matches in terms of the **action** and the **core entities** (participants, location, and/or time).

5. If such an event exists, the cluster is considered **consistent**.

6. If no common event exists across all articles, the cluster is considered **inconsistent**.

7. Provide a short reasoning explaining your decision.

---

**Input Format**
You will be given multiple articles belonging to a single cluster. Each article will be wrapped with tags in the following format:

```
<Article 1>
...
</Article 1>

<Article 2>
...
</Article 2>

<Article 3>
...
</Article 3>
```

All articles together form one cluster.

---

**Output Format (Strict)**

- Your final decision must be wrapped inside an **<answer>** tag.
  - If the cluster contains a common event across all articles:

    ```
    <answer> yes </answer>
    ```

  - If the cluster does not contain a common event across all articles:

    ```
    <answer> no </answer>
    ```

- Your explanation must be wrapped inside a **<reasoning>** tag.
  - The reasoning must be at most **150 words**.
  - It should reference the relevant events and entities across the articles.
  - Clearly explain why the cluster is consistent or inconsistent.

**Important:**
Your final output must contain **only** the answer and the reasoning in the required format. Do not output any additional text.

---

**Example**
**Input**

<Article 1>
'(29 Jul 2018) OCTAVIA SPENCER TO STAR AS BLACK HAIRCARE MOGUL MADAM WALKER Octavia Spencer is bringing the story of black haircare mogul Madam C.J. Walker to television. Netflix said Sunday (29 JULY 2018) that Spencer will produce and star in a limited series about the outsized life of Sarah Breedlove, who was known professionally as Walker. The eight-episode drama is based on the book "On Her Own Ground" and includes basketball star LeBron James as a producer. Walker became one of America's first self-made female millionaires by creating and marketing hair products for African-Americans at the turn of the 20th century. Netflix said the series will detail the hostility, rivalries and tumultuous personal life that marked Walker's life. Spencer is an Oscar-winning actress whose credits include "The Shape of Water," "Hidden Figures" and "The Help." She also recently began production on a drama series called "Are You Sleeping" for Apple TV. It's based on a novel of the same name and produced by Reese Witherspoon.
</Article 1>
<Article 2>
(7 Sep 2014) COSTER, SPENCER PREMIERE 'BLACK AND WHITE' IN TORONTO Kevin Costner and Octavia Spencer walked the red carpet Saturday (6 SEPT. 2014) in Toronto for their new movie, "Black and White." The film, written and directed by Mike Binder, tells the story of a grandfather played by Costner drawn into a legal battle over his mixed-race granddaughter that he raised when her mother died in childbirth. Costner was thankful that the festival accepted the film. "We really need to thank this festival for giving a little movie that doesn't have a platform a place to play. I think it has big implications," Costner said. The actor says he learned something from the film, and he hopes others do as well. "Not all movies can have a message in them; they simply can't, nor should they try. And this movie, I think without trying, kind of helped me as a person. I thought that if I make it, maybe it will help some other people too surround this argument that seems to trip us up when we're talking about racism," Costner said. Spencer was happy to be in Toronto. "This is an amazing film festival with really wonderful films, and the fact that our film is a part of this roster means a lot. And it's a really beautiful city," she said. The Oscar-winning actress also explained why her next project is a television series. "Why not? I think some of the most interesting roles for me have been in the form of a television script. I love the project that I'm working on, 'Red Band Society,' and I'm excited for people to see." "Black and White" does not have a release date.
</Article 2>
**Output**

```
<answer> no </answer>

<reasoning>
Although both articles mention Octavia Spencer
    , the contexts describe different events.
The first article discusses a television
    series about Madam C.J. Walker, while the
second article focuses on the premiere of the
    film "Black and White" at the Toronto
Film Festival. Since the actions and contexts
    differ, the articles do not describe
the same event.
</reasoning>
```

## Prompt V.22: Prompt for Pairwise News Event Verification

**System**
You are an expert in comparative news event verification. You will be given two articles:

- A **base article**, which is a verified and trustworthy news report.
- A **scraped article**, which is an automatically collected online article that may or may not describe the same event.

Your task is to determine whether the scraped article refers to the **same or a closely related real-world news event** as the base article.

---

**Guidelines for Comparison**

1. **Event-Level Alignment**
   - Determine whether both articles describe at least **one common real-world event**.
   - Two events are considered the same only if they align across the following dimensions:
     - **What happened:** the main action, occurrence, or incident.
     - **Who or what entities were involved:** individuals, organizations, or groups.
     - **When:** the same or a very similar time period.
     - **Where:** the same or closely related geographic location.
   - If wording is similar but the participants, time, or location differ significantly, treat them as **different events**.
2. **Partial vs. Complete Overlap**
   - Some scraped articles may describe the same event only partially (for example, focusing on the aftermath, investigation, or a specific participant).
   - Such articles can still be considered a match as long as they clearly refer to the **same underlying real-world event**.
   - Articles that describe entirely unrelated events, even if they belong to the same domain (e.g., two separate protests or accidents), should **not** be considered a match.
3. **Content Authenticity and Noise Handling**
   - Ignore short advertisements, promotional text, or unrelated web content that may appear within the scraped article.
   - Do not reject an article solely because of minor web noise. Focus on the main narrative of the article.
   - If the scraped article appears to be a **fictional story, poem, blog post, or book excerpt**, it should be **rejected** as it does not represent a real news event.
   - If the scraped article is a legitimate **corporate or organizational announcement** (e.g., product launch, merger, official statement), treat it as valid news content.
4. **Decision Logic**
   - The scraped article should be considered a **match** if:
     - It is a genuine news article, and
     - It shares at least one real-world event with the base article with consistency in what happened, who was involved, when, and where.
   - Otherwise, the scraped article should be considered **not a match**.

---

**Input Format**

```
<base article>
(Base article content)
</base article>

<scraped article>
(Scraped article content)
</scraped article>
```

---

**Output Format (Strict)**

- Your final decision must be wrapped inside the `<answer>` tag.
  - If the two articles describe the same or closely related real-world event:

    `<answer> yes </answer>`
  - If the articles describe different or unrelated events:

    `<answer> no </answer>`
- Your explanation must be wrapped inside the `<reason>` tag.
  - The explanation should be a brief reasoning (**maximum 150 words**).
  - It should describe why the scraped article does or does not correspond to the same real-world event as the base article.
  - The reasoning should reference overlaps or differences in event type, entities, time, and location.

Your final output must contain **only** the answer and the reasoning wrapped in the required tags.

---

**Notes**

- Focus on **factual alignment of events**, not stylistic similarity.
- Be conservative: answer **yes** only when the overlap of the real-world event is clear.
- Ignore superficial linguistic similarities that do not reflect the same underlying event.

## Prompt V.23: Prompt For Headline Generation

You are an intelligent agent that processes textual news articles. Your task is to identify the **2–3 most important entities** (person, organization, place, or time) and the **2–3**

**most important events or actions** from the article. Then, combine them to produce a **very short sub-sentence** summarizing the article's main focus.
**Instructions:**

- Select only the most central entities and events; ignore details, modifiers, or secondary information.
- Construct the sub-sentence by connecting the chosen words directly—avoid adding extra function words or unnecessary fillers.
- The result may be incomplete grammatically, but it should remain understandable and reflect the article's core meaning.
- You may use a few conjunctions (e.g., *and, for, but*) only if needed for readability.
- Do **not** include any explanation, reasoning, or intermediate steps—only output the final phrase.

**Expected output format:**

`<Entity/Event condensed phrase>`

**Examples:**
*Example 1* ## Article: Bianca Jagger, representatives from the European Parliament, the Council of Europe and other high-level activists gathered in Paris on Monday to announce a forthcoming congress against the death penalty. The first world congress for the abolition of the death penalty will be held in Strasbourg from June 21 to 23..... ## Output: Bianca Jagger announces congress in Strasbourg
*Example 2* ## Article: China and Mexico are advancing trade, investment, and technology cooperation as Chinese Vice President Zeng Qinghong continues his three-day visit... ## Output: China's Zeng Qinghong meets Vicente Fox for advancing cooperation and signing accords
*Example 3* ## Article: Police killed a 27-year-old man as he kicked, punched, and stomped a toddler to death near Modesto, California... ## Output: Modesto police officer kills man beating toddler, but child later dies