Evaluation Pitfalls and Challenges in Multimedia Event Extraction

Philipp Seeberger, Steffen Freisinger, Tobias Bocklet, Korbinian Riedhammer
Technische Hochschule Nürnberg Georg Simon Ohm
{philipp.seeberger,steffen.freisinger,tobias.bocklet,korbinian.riedhammer}@th-nuernberg.de

Abstract

Multimedia event extraction aims to jointly identify events and their arguments across multiple modalities, such as text and images, to support more comprehensive event understanding. While recent work reports steady and substantial progress, the reliability and comparability of these results critically depend on consistent and rigorous evaluation. In this work, we present the first systematic analysis of evaluation pitfalls in multimedia event extraction and identify three major sources of issues: inconsistent data processing, inconsistent task assumptions, and overly relaxed evaluation settings. We demonstrate, through a series of controlled experiments under a strict evaluation framework, that minor evaluation choices can cause large performance variations and lead to overestimation of a model's ability to ground real-world events across modalities. Our findings highlight the need for comparable evaluation standards and encourage a shift toward more rigorous evaluation in multimedia event extraction.[1]

1 Introduction

Event extraction (EE) is a fundamental task in natural language processing and information extraction, aiming to identify, structure, and organize event-related knowledge from documents (Ahn, 2006). While the majority of existing EE research has focused on texts (Peng et al., 2023a; Huang et al., 2024), recent work has increasingly explored the integration of additional modalities (Sun et al., 2024; Zhang et al., 2024). This shift is motivated by the growing prevalence of multimodal content in contemporary news media and online platforms (Li et al., 2020), where images, videos, and audio provide complementary information that can support more accurate and comprehensive event understanding.

Prior research has investigated EE or closely related tasks within individual modalities (Yatskar et al., 2016; Wadden et al., 2019; Sadhu et al., 2021; Wang et al., 2024), including text, images, video, and audio, or has leveraged cross-modal cues to address specific challenges such as ambiguity (Zhang et al., 2017; Tong et al., 2020). However, evaluation mostly remains restricted to a single target modality. Multimedia event extraction (MEE) has recently attracted attention and adopts a holistic view by jointly extracting and evaluating events across multiple modalities, typically combining textual and visual inputs (Li et al., 2020; Chen et al., 2021; Sanders et al., 2024). Despite this progress, existing MEE benchmarks remain limited and evaluation challenging, most notably due to annotation scarcity, lack of train splits, and increased evaluation complexity inherent to multimodal settings (Li et al., 2020; Sanders et al., 2024).

Prior studies have demonstrated that even traditional textual EE suffers from substantial and often overlooked evaluation challenges (Zheng et al., 2021; Peng et al., 2023b; Huang et al., 2024), making it prone to hidden pitfalls. These include discrepancies in data and task assumptions as well as metric design choices that can distort model comparisons and fail to reflect real-world performance (Huang et al., 2024). Crucially, extending textual EE to the multimodal setting not only inherits existing evaluation issues but also introduces additional pitfalls. These arise from factors such as data scarcity, heterogeneous modalities, and multi-stage pipelines commonly employed in MEE (Li et al., 2020; Liu et al., 2022; Du et al., 2023; Cao et al., 2025). As a consequence, inconsistent and under-specified evaluation settings can easily emerge, posing a potential obstacle to reliably assessing progress in MEE research.

Motivated by concerns about the reliability and comparability of current evaluations, this work systematically investigates hidden pitfalls and chal-

[1] https://github.com/seebergerph/StrictEval

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30885–30900

lenges in MEE evaluation, with the goal of raising awareness and encouraging a shift toward more rigorous evaluation practices. Through an in-depth analysis of the widely used M2E2 benchmark, we first identify three major categories with several issues: inconsistent data processing, inconsistent task assumptions, and relaxed evaluation settings. Building on this analysis, we introduce a more rigorous evaluation framework, STRICTEVAL, and use it to examine how hidden pitfalls influence reported performance. Finally, we show that minor experimental design choices can substantially affect evaluation outcomes.

In summary, our contributions are twofold: (1) We conduct a systematic analysis of evaluation pitfalls and challenges in MEE and propose a more rigorous evaluation framework (STRICTEVAL). (2) We systematically quantify how hidden evaluation pitfalls affect reported performance and reevaluate recent MEE approaches to highlight limitations.

2 Background and Related Work

2.1 Background

Textual EE is commonly formulated as a two-stage pipeline (Ahn, 2006) consisting of event detection (ED) and event argument extraction (EAE). Event detection aims to identify event mentions, typically grounded to trigger spans, and classify them into predefined event types. Event argument extraction focuses on identifying argument spans and assigning them semantic roles conditioned on the detected event mentions. Analogously, visual EE decomposes the task into detecting events grounded in images and linking their associated semantic roles to visual regions, such as objects (Pratt et al., 2020). Building on these two research directions, MEE integrates textual and visual information to jointly extract events and their arguments across modalities (Li et al., 2020; Chen et al., 2021). This multimodal integration introduces the additional subtask of cross-modal event coreference resolution, with the aim to unify event mentions from different modalities that refer to the same real-world event into a coherent multimedia event representation (see Figure 1).

2.2 Related Work

Multimedia Event Extraction Benchmarks While most EE benchmarks focus exclusively on text (Walker, Christopher et al., 2006; Song et al., 2015; Wang et al., 2020; Huang et al., 2024), early multimodal extensions augment textual datasets with images, but evaluation remains limited to textual events (Zhang et al., 2017; Tong et al., 2020). To overcome unimodal limitations, Li et al. (2020) introduce the first MEE benchmark, M2E2, which evaluates event and argument extraction for both texts and images. In addition, M2E2 includes cross-modal event coreferences, analogous to cross-document coreferences in text (Nath et al., 2024). Subsequent work extend to images and videos, such as VM2E2 (Chen et al., 2021), CMMEvent (Liu et al., 2025b), TVEE (Wang et al., 2023), and MultiVENT-G (Sanders et al., 2024). More recently, Zhang et al. (2024) propose a comprehensive benchmark covering textual, visual, and audio inputs by integrating datasets such as M2E2 and ACE with recorded speech. However, only M2E2 and MultiVENT-G publicly release complete data while other benchmarks are still closed-source (Wang et al., 2023; Liu et al., 2025b) or lack critical annotations (Chen et al., 2021; Zhang et al., 2024). Moreover, complex annotation formats, abundance of train splits, and missing evaluation scripts further hinder reliable benchmarking.

Multimedia Event Extraction Early approaches focus on cross-modal correlations and align visual and textual representations using large-scale unlabeled news corpora (e.g., VOA) (Chen et al., 2021; Liu et al., 2022, 2024), often in combination with contrastive learning objectives (Li et al., 2020, 2022, 2023). Subsequent studies explore complementary directions, including augmenting training data with synthetically generated image–text pairs (Du et al., 2023), designing sophisticated multi-grained fusion mechanisms (Wang et al., 2025; Liu et al., 2025a), or leveraging multi-task learning with pseudo labeling strategies (Cao et al., 2025). Other works narrow their focus to specific subtasks, such as ED (Sun et al., 2023) or EAE (Seeberger et al., 2024). With recent advances in multimodal large language models (MLLMs), several instruction-following approaches have been proposed to enable more universal information extraction (Sun et al., 2024; Zhang et al., 2024; Yuan et al., 2025; Chen et al., 2025; Yu et al., 2025). However, most of these methods primarily work with given image–text pairs and do not explicitly address broader MEE settings such as cross-modal event coreference resolution (Sun et al., 2024; Yuan et al., 2025; Chen et al., 2025). Notably, the majority of existing approaches are evaluated on the

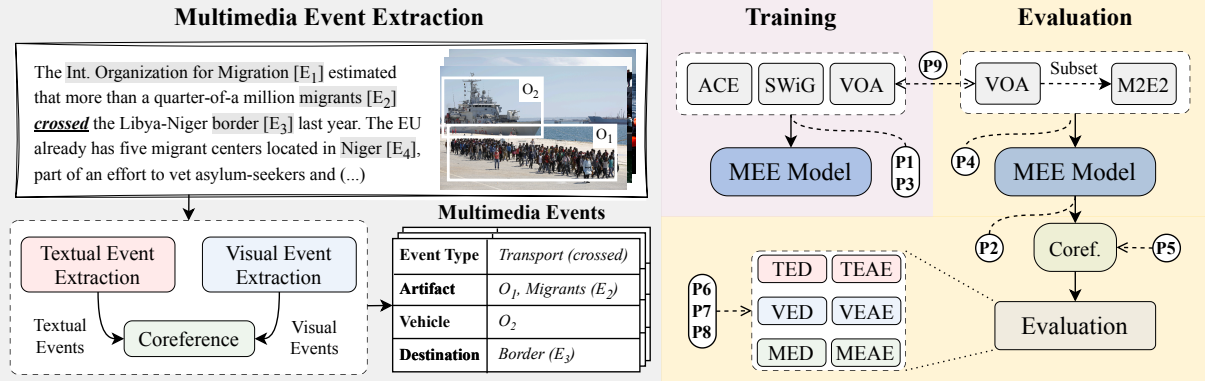

Figure 1: Overview of the M2E2 multimedia event extraction pipeline. The example illustrates a *Transport* event grounded in text and image. P markers indicate stages at which pitfalls occur. TED, TEAE, VED, VEAE, MED, and MEAE denote the textual, visual, and multimedia event detection and argument extraction subtasks, respectively.

M2E2 benchmark, underscoring its role in advancing MEE research. Despite substantial progress, prior methods adopt diverse task formulations and evaluation protocols, which hinders fair comparison across different modeling approaches.

Evaluation Pitfalls Recent studies have highlighted numerous issues in the evaluation of textual EE models, including inconsistent data assumptions, processing steps, output space discrepancies, and relaxed evaluation metrics (Zheng et al., 2021; Peng et al., 2023b,a; Huang et al., 2024). While these works clearly highlight significant differences in textual EE benchmarking, issues in the evaluation of visual and multimedia EE remains relatively underexplored.

3 Pitfalls and Challenges in Evaluation

Motivated by evaluation concerns for MEE, we first present our investigation setup (§3.1) and systematic analysis to identify evaluation issues (§3.2). We then provide a detailed analysis of common pitfalls across three major categories: data processing (§3.3), task assumptions (§3.4), and relaxed evaluation settings (§3.5). Lastly, building on the insights from this analysis, we introduce StrictEval as rigorous evaluation framework to address hidden pitfalls (§3.7).

3.1 Preliminaries

To examine the sources and impact of evaluation issues in MEE, we adopt the M2E2 benchmark (Li et al., 2020). Our choice is motivated by two main considerations: (1) M2E2 is publicly available and, to the best of our knowledge, the most widely used benchmark in MEE research. (2) As discussed in §2.2, alternative benchmarks are often incomplete (Chen et al., 2021; Zhang et al., 2024) or not accessible (Wang et al., 2023; Liu et al., 2025b).

M2E2 Dataset In Figure 1, we show the complete task and involved components. The M2E2 benchmark comprises 6,167 sentences and 1,014 images from 245 multimedia news documents collected from 108k Voice of America (VOA) documents. Overall, the events cover 8 event types and 15 argument roles, with 1297 textual and 391 visual events. Thereby, there exist 309 multimedia events which are coreferenced by 192 textual and 203 visual events. As no training data exists, the benchmark adopts ACE (Walker, Christopher et al., 2006) for textual and imSitu (Yatskar et al., 2016), with object groundings from SWiG (Pratt et al., 2020), for visual EE training. Annotation mappings to the M2E2 schema are provided by the original work (Li et al., 2020).

M2E2 Evaluation Following Li et al. (2020), evaluation is conducted separately for textual, visual, and multimedia EE with precision (P), recall (R), and F1 for the subtasks ED and EAE. In textual EE, an event mention is correct if its type and trigger offsets match the reference, while arguments must additionally match argument offsets and role types. For visual EE, a visual event mention is correct if its type matches the reference image, and a visual argument if its event type, role label, and bounding box match a reference argument with IoU > 0.5. Lastly, a multimedia event mention is correct if its event type and trigger offsets (or image) match the reference trigger (or image). The inherited textual and visual arguments are evaluated using the same criteria as in the textual and

visual modality. However, in our preliminary analysis we observe inconsistencies in the evaluation criteria for multimedia events, which we discuss in detail in §3.4.

3.2 Systematic Analysis

We collect peer-reviewed MEE studies evaluated on M2E2 published between 2020 and 2025 through keyword and citation-based searches, resulting in 18 articles across multiple venues (e.g., ACL, ACM, AAAI). Of these, we analyze 15 studies and exclude three (Moghimifar et al., 2023; Zhang et al., 2024; Xing et al., 2025) due to reliance on custom train-test splits or newly introduced evaluation metrics, which hinder direct comparison. The complete set of methods and reported evaluation scores is summarized in Table 5. For each study, we review the article, supplementary materials, and, when available, its public codebase, with particular attention on the training and evaluation stages (see Figure 1). We focus on data processing, experimental setups, and evaluation protocols (Peng et al., 2023b; Huang et al., 2024). Through this analysis, we uncover three major categories of evaluation issues: inconsistent data processing, inconsistent task assumptions, and overly relaxed evaluation settings. These issues largely stem from the inherent complexity of MEE, which relies on external training datasets, multi-stage pipelines, and heterogeneous modalities. Nevertheless, such inconsistencies can lead to unfair comparisons and performance estimates that do not reflect real-world scenarios.

3.3 Inconsistent Data Processing

Due to the absence of standardized preprocessing and the reliance on external training datasets, we observe substantial variation in data assumptions across studies. These differences include training set construction, preprocessing, and postprocessing procedures.

[P1] Train Size Discrepancies As described in §3.1, M2E2 relies on external datasets (e.g., ACE and SWiG) for training, which provide predefined train, development, and test splits. While the original M2E2 benchmark uses only train splits, subsequent work often incorporates other sets as additional training data. Consequently, models are optimized on differing numbers of samples (e.g., 75k vs. 100k images for SWiG).

[P2] Oracle Trigger Refinement Due to distributional annotation differences between ACE and M2E2, most reported evaluation scores applies a postprocessing step that adjusts textual ED predictions using M2E2 ground truth annotations. For example, this script removes predicted event mentions with *deadly* as the trigger span. However, many studies do not clearly specify this experimental setting or report results exclusively with postprocessing applied. We argue that ground truth-based postprocessing does not reflect real-world conditions.

[P3] Verb Mapping Refinement Because the label ontologies of SWiG and M2E2 differ, the original authors provide a verb and role mapping to align SWiG annotations with the M2E2 schema. We notice that some studies adopt refined versions of this mapping. For example, one refined mapping aligns 73 verbs rather than the original 67 verbs to the M2E2 event types. Such discrepancies in label alignment can lead to performance differences driven more by data engineering than by modeling improvements.

3.4 Inconsistent Task Assumptions

We identify inconsistencies in task assumptions. For instance, some studies evaluate events only present in texts and images (Li et al., 2020; Liu et al., 2022; Du et al., 2023), while others focus exclusively on multimedia events (Sun et al., 2024; Yuan et al., 2025; Chen et al., 2025) (cf. Table 5). Similarly, some methods filter test data to exclude samples without events, whereas others evaluate on the full set. Consequently, reported results are often not directly comparable.

[P4] Test Subset Selection Recent works restrict test-time predictions to sentences or images containing at least one event. This filtering reduces the number of sentences from 6167 to 1086 and images from 1014 to 391, while other work evaluate on the full set. Moreover, methods focusing on the multimedia events task often only evaluate on 309 image-text pairs derived from the event coreference annotations, further reducing the test set to 192 sentences and 203 images, respectively.

[P5] MEE Task Discrepancies Due to the absence of standardized evaluation scripts for multimedia events, follow up work has adopted different task definitions. The original M2E2 benchmark considers a multimedia event correct if either the textual or visual event matches the reference and treats cross-modal coreference as a separate task.

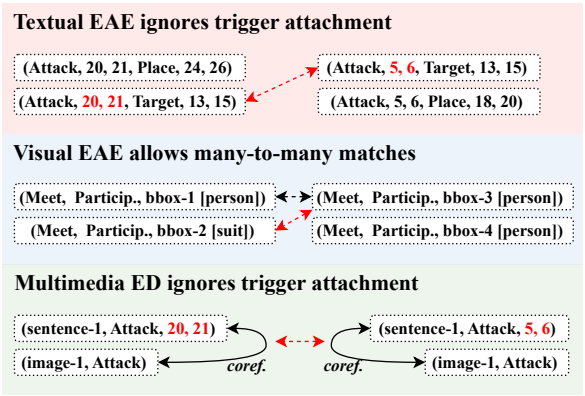

Figure 2: Illustration of relaxed evaluation settings. Red edges denote predictions that are incorrectly counted as correct, while red colored text indicates ignored trigger attachments (offsets such as *20, 21*). We eliminate these issues in StrictEval.

More recent work, however, introduce a stricter setting that additionally requires correct event coreference prediction as additional attachment. In contrast, some methods assume gold coreference links and evaluate only aligned image-text pairs. Despite these substantial differences in task formulation, results are often compared directly.

3.5 Relaxed Evaluation Settings

Similar to observations in TextEE (Huang et al., 2024), we find that MEE evaluation metrics are often imprecise due to relaxed matching criteria or missing structural constraints. In Figure 2, we illustrate common issues and discuss them below:

[P6] Relaxed Textual Evaluation Some studies ignore trigger offsets during textual EAE evaluation. When multiple events of the same type appear in a sentence, this relaxation allows arguments to be matched to any event of that type, potentially inflating reported performance such as discussed by Huang et al. (2024).

[P7] Relaxed Visual Evaluation Most prior work employs a many-to-many matching that considers a visual argument correct if it has the correct role and its IoU exceeds a threshold. However, this approach allows multiple predictions to match one gold argument (see Figure 4), effectively rewarding recall-oriented models for visual EAE. To address this issue, we propose a one-to-one strategy via bipartite matching, that penalizes redundant predictions (see A.2).

[P8] Relaxed Multimedia Evaluation We also observe missing attachment constraints in multimedia ED evaluation. For instance, ignoring trigger offsets allows incorrect event predictions to be counted as correct at the sentence level, rather than requiring span-level accuracy (see Figure 2). This relaxed assumption can substantially overestimate real-world performance.

Setting	P1	P2	P3	P4	P5	P6	P7	P8
StrictEval	✗	✗	✗	✗	PC	✗	✗	✗
Evaluation Setups								
Setup-1	✗	✗	✗	?	?	✗	✓	?
Setup-2	✓	✓	✓	?	?	✗	✓	?
Setup-3	✓	✓	✓	✗	PC	✗	✓	✓
Setup-4	✓	✓	✓	✗	PC	✓	✓	✓
Setup-5	✗	✓	✗	✓	PC	✗	✓	✓
Setup-6	✗	✗	?	✓	GC	✗	?	?

Table 1: Evaluation settings used in recent work. Px indicates that a specific setting is used, while ? denotes that the setting is unspecified. PC evaluates with prediced coreference resolution, whereas GC uses gold coreferences.

3.6 Data Leakage

[P9] Test Data Leakage M2E2 forms a subset of the collected VOA samples and several works utilize this image-caption dataset during training. We find that some of these works include images and captions from test documents in their training data. Although no ground-truth annotations are used, exposure to test images or captions can still result in information leakage.

3.7 StrictEval

Our analysis demonstrates that prior MEE evaluations vary widely in data processing, task assumptions, and matching criteria. To systematize these discrepancies, we annotate each study with the evaluation pitfalls it exhibits (P1–P9) in Figure 1 and cluster them into six distinct setups (see Table 1). Based on this analysis, we propose StrictEval, a more rigorous evaluation framework that eliminates all identified pitfalls. StrictEval enforces (1) consistent data usage without oracle postprocessing or test exposure, (2) a clearly specified task definition evaluated on the full benchmark, and (3) strict matching criteria that preserve structural constraints across textual, visual, and multimedia predictions. As shown in Table 1, StrictEval represents the strictest setup with the goal of providing a reproducible evaluation framework designed to more faithfully reflect real-world performance.

	Setting	ED				EAE			
		P	R	F1	ΔF1	P	R	F1	ΔF1
Textual EE	StrictEval	25.8 ±0.6	75.3 ±1.2	38.4 ±0.6	-	14.2 ±0.3	48.4 ±0.9	21.9 ±0.3	-
	w/ [P1] train with dev	26.5 ±0.4	73.6 ±0.7	39.0 ±0.5	+0.6	15.0 ±0.4	47.8 ±0.5	22.8 ±0.5	+0.9
	w/ [P2] trig. refinement	41.6 ±1.3	70.4 ±1.2	52.3 ±1.0	+13.9	20.9 ±0.6	45.9 ±0.9	28.7 ±0.6	+6.8
	w/ [P4] eval text subset	57.8 ±0.6	77.5 ±2.0	66.2 ±0.4	+27.8	24.2 ±0.4	50.1 ±1.3	32.6 ±0.2	+10.7
	w/ [P6] eval EAE relaxed	-	-	-	-	19.4 ±0.3	52.0 ±1.1	28.3 ±0.3	+6.4
Visual EE	StrictEval	52.6 ±1.1	29.1 ±2.7	37.4 ±2.2	-	20.7 ±0.8	11.9 ±0.9	15.1 ±0.8	-
	w/ [P1] train with dev	50.5 ±1.3	31.4 ±1.5	38.7 ±1.5	+1.3	18.2 ±1.6	12.2 ±1.1	14.6 ±1.3	-0.5
	w/ [P3] verbs refinement	57.6 ±1.5	36.2 ±1.6	44.4 ±0.8	+7.0	22.9 ±1.7	15.2 ±0.4	18.2 ±0.6	+3.1
	w/ [P5] eval image subset	70.2 ±0.5	64.8 ±0.2	67.4 ±0.3	+30.0	29.1 ±0.6	30.1 ±0.4	29.6 ±0.5	+14.5
	w/ [P7] eval EAE relaxed	-	-	-	-	21.0 ±0.8	12.2 ±1.0	15.4 ±0.8	+0.3

Table 2: Unimodal evaluation results of Single Task models under different setups (averaged over 3 runs). Starting from the StrictEval setting, each identified issue (P) is applied independently. The EAE relaxed settings only affect the EAE performance. ΔF1 denotes the absolute difference to StrictEval.

4 Experiments and Analysis

The analysis in §3 reveals notable discrepancies and pitfalls in MEE evaluation, raising concerns about the extent to which evaluation design choices influence reported performance. Starting with StrictEval, we conduct a series of controlled experiments in which each evaluation factor is examined in isolation.

4.1 Experimental Setup

MEE Model We adopt Single Task models to avoid complex architectural choices which ensures that performance differences stem from evaluation setups rather than model capacity. We train independent single-task models for each textual and visual subtask. Textual and visual ED are implemented as token-level and image-level classification models. For textual and visual EAE, we follow (Liu et al., 2022; Du et al., 2023; Cao et al., 2025) and classify ground-truth textual entities and detected visual objects (via YOLO) into argument roles. Multimedia events are constructed through an event coreference resolution step (Du et al., 2023), described below. Further implementation details are provided in Appendix A.3 and A.4.

Event Coreference Resolution Following prior work (Liu et al., 2022; Du et al., 2023; Seeberger et al., 2024; Cao et al., 2025), we perform event coreference resolution by computing CLIP-based similarity scores between image–sentence pairs (Radford et al., 2021). We note that StrictEval itself does not introduce a new coreference model, but instead relies on coreference predictions from existing approaches, which we compare under a unified evaluation setting. A textual and a visual event are merged when they share the same predicted event type and their image-text similarity exceeds a threshold of 20. We adopt this hyperparameter following (Du et al., 2023; Seeberger et al., 2024; Cao et al., 2025). The resulting multimedia event inherits all associated arguments from both modalities.

Evaluation Metrics As outlined in §3.1, we report micro-averaged P, R, and F1. Unlike (Li et al., 2020), we follow subsequent work and require multimedia events to match event mentions and their coreference links (Liu et al., 2022; Du et al., 2023; Seeberger et al., 2024; Cao et al., 2025; Wang et al., 2025). Unless otherwise specified, all experiments use our introduced evaluation framework StrictEval..

4.2 Unimodal Evaluation and Analysis

In this section, we focus on unimodal EE results. This allows us to isolate modality-specific behavior before examining the extraction of multimedia events. An overview of the unimodal results is presented in Table 2 and an extended analysis in Appendix A.6.

Impact of Data Processing Data processing choices lead to substantial performance variations. Applying the oracle trigger refinement step [P2] and verb refinements [P3] exhibits improvements of up to +13.9 F1 for textual and +7.0 F1 for visual ED. In contrast, incorporating the development sets of ACE and SWiG during training [P1] results in only marginal gains (up to +1.3 F1). These processing decisions also yield improvements of up to

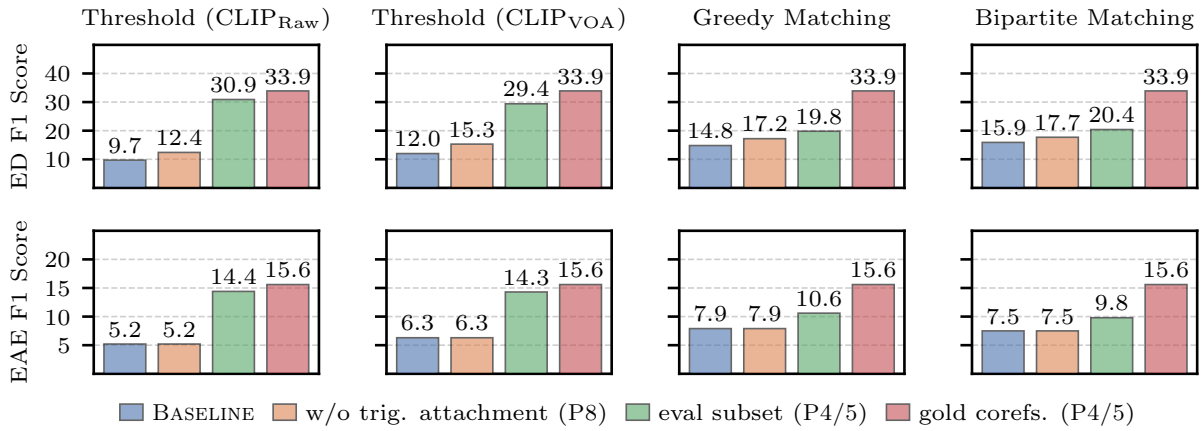

Figure 3: Multimedia ED and EAE scores using different coreference resolution techniques. Threshold Matching uses CLIP-based similarity with a threshold of 20. CLIP$_{Raw}$ denotes the pretrained model and CLIP$_{VOA}$ is further fine-tuned on the VOA image-caption dataset. Greedy Matching and Bipartite Matching follows Liu et al. (2022).

+6.8 and +3.1 F1 for textual and visual EAE. Importantly, these discrepancies extend to multimedia ED and EAE since both metrics depend on textual and visual predictions. These findings underscore the importance of consistent data processing, as new state-of-the-art results may not solely reflect advances in modeling techniques.

Impact of Task Assumptions Inconsistent task assumptions, particularly test subset selection [P4], significantly boosts unimodal EE evaluation scores. Restricting training and evaluation to texts and images that contain at least one event instance increases F1 scores of up to +27.8 and +30.0 for textual and visual ED, respectively, and improves EAE scores by up to +14.5. These gains are primarily driven by higher precision, as no-event instances are excluded from evaluation. Notably, this issue propagates to downstream multimedia EE and is further amplified when evaluating on smaller subsets, such as the 309 gold event coreference pairs [P4]. However, these results do not reflect real-world performance, where sentences and images without any targeted events are prevalent.

Impact of Relaxed Evaluation We further examine the effect of relaxed evaluation settings for textual and visual EAE, as described in §3.5. For textual EAE, ignoring the trigger span attachments [P6] introduces discrepancies of up to 6.4 F1. For visual EAE, many-to-many matching [P7] yields only marginal improvements (+0.3 F1), however, this highly depends on the object detector and confidence threshold. As shown in Table 6, object detectors trained on OpenImages often predict overlapping object categories (e.g., *person* and *suit*), which are counted as multiple correct predictions under relaxed evaluation (e.g., +8.0 F1 for YOLO-X OI and $\tau = 0.5$). This highlights that relaxed evaluation metrics can inflate the actual quality of EAE.

Setting	P	R	F1	ΔF1
STRICTEVAL	6.2	22.4	9.7	-
w/ train event subsets	4.3	53.6	8.0	-1.7
STRICTEVAL + eval subset	50.4	22.4	30.9	-
w/ train event subsets	53.1	53.6	53.4	+22.2

Table 3: Multimeda ED scores with and without training only on samples with at least one event instance.

4.3 Multimedia Evaluation and Analysis

In this section, we empirically analyze how varying task assumptions and evaluation settings affect multimedia EE results. Real-world applications typically do not provide oracle image-text pairs or prior knowledge about whether a text or image contains a multimedia event.

Experimental Setup To study the impact of task discrepancies [P5] and missing trigger attachments [P8], we evaluate three event coreference resolution strategies: threshold-based, greedy, and bipartite matching approaches proposed in prior work (Liu et al., 2022; Du et al., 2023; Cao et al., 2025). We further compare evaluations on the full dataset against subsets restricted to samples containing at least one multimedia event, and analyze the effect of applying the same subset selection during training.

		ORIGINAL						STRICTEVAL						Δ(STRICT - ORIG)	
		ED			EAE			ED			EAE			ED	EAE
	Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	ΔF1	ΔF1
Textual EE	SINGLE TASK (ours)	-	-	-	-	-	-	25.8	75.3	38.4	14.2	48.4	21.9	-	-
	CAMEL	43.5	68.7	53.3	23.5	41.8	30.1	26.8	72.8	39.1	16.1	36.1	22.3	-14.2	-7.8
	MMUTF	-	-	-	31.5	46.8	37.7	-	-	-	15.9	51.7	24.3	-	-13.4
	X-MTL	44.2	65.8	52.9	29.9	38.9	33.8	26.6	70.0	38.5	20.2	41.7	27.2	-14.4	-6.6
Visual EE	SINGLE TASK (ours)	-	-	-	-	-	-	52.6	29.1	37.4	20.7	11.9	15.1	-	-
	CAMEL	66.7	46.5	54.8	27.7	19.8	23.1	55.6	33.0	41.4	21.7	11.5	15.0	-13.4	-8.1
	MMUTF	-	-	-	27.3	15.8	20.0	-	-	-	23.4	10.4	14.4	-	-5.6
	X-MTL	71.7	69.8	70.7	31.1	29.6	30.3	28.2	69.8	40.2	14.1	28.8	18.9	-30.5	-11.4
Multi. EE	SINGLE TASK (ours)	-	-	-	-	-	-	6.2	22.4	9.7	3.4	10.8	5.2	-	-
	CAMEL	56.2	42.4	48.3	29.4	24.0	26.5	6.9	27.0	11.0	3.8	11.0	5.7	-37.3	-20.8
	MMUTF	-	-	-	37.0	18.7	24.9	-	-	-	4.1	9.1	5.6	-	-19.3
	X-MTL	75.5	57.9	65.6	36.7	40.6	38.6	4.9	43.3	8.8	2.9	20.0	5.0	-56.8	-33.6
	SSGPF	53.7	70.7	61.0	15.8	12.6	14.1	10.3	30.9	15.4	2.8	4.2	3.3	-45.6	-10.8

Table 4: Experimental results on the M2E2 benchmark under the original and our STRICTEVAL evaluation settings. In line with Seeberger et al. (2024), the MMUTF model results are based on event predictions from CAMEL.

Impact on Multimedia Results Figure 3 presents the multimedia evaluation results. Consistent with textual evaluation, ignoring trigger offsets for textual events leads to an drop of up to 3.3 F1 for ED, indicating correct sentence-level predictions with incorrect trigger spans. Larger differences arise across task setups. Restricting evaluation to samples containing multimedia events leads to gains of up to +21.2 F1 for threshold-based methods, although such prior knowledge would not be available at deployment. Applying the same subset selection during training further amplifies this effect and reveals opposing trends between full and subset evaluations (see Table 3). Under realistic conditions, bipartite matching performs best, whereas threshold-based methods outperform it on the filtered subset (30.9 vs. 20.4 F1), attributed to increased recall. Beyond multimedia ED, we also observe similar patterns for EAE (bottom row of Figure 3), which is connected to ED performance due to the pipeline setup. Finally, experiments using gold image-text coreference annotations show the highest scores, highlighting that subset-based evaluation can substantially overestimate real-world performance.

5 Consistent Evaluation

Our analysis reveals several limitations in current MEE evaluation practices, however, so far ignores the combination of hidden pitfalls and advanced modeling techniques. Therefore, we reproduce and reevaluate recent methods (Du et al., 2023; Seeberger et al., 2024; Cao et al., 2025; Chen et al., 2025)[2] under their original evaluation settings and compare them with our proposed framework STRICTEVAL. Notably, SSGPF supports only multimedia evaluation and assumes given image-text pairs. Reproduction details are provided in Appendix A.5.

Reevaluation Results and Discrepancies The results in Table 4 highlight discrepancies between the ORIGINAL and STRICTEVAL setting. First, evaluation scores change substantially in absolute performance levels. Specifically, textual and visual ED scores differ significantly, primarily due to trigger postprocessing [P2], verb refinement [P3], and test subset selection [P4] (cf. Table 2). This indicates that minor implementation choices can dominate reported gains. Second, removing these pitfalls consistently leads to lower EAE performance due to the pipeline setup. This can overestimate advances in multimedia EAE, which are often driven by increased ED performance. Third, we observe most degradation in multimedia ED and EAE, with drops of up to 56.8 and 33.6 F1, respectively. When neither test subset selection nor gold event coreference annotations are used, precision decreases dramatically, revealing that cross-modal event coreference resolution remains the largest challenge in a more realistic setting. Further intuition about these low scores and performance drops is provided in Appendix A.6.

[2]Our selection focuses on methods with complete publicly available code and instructions, except for SSGPF as MLLM-based method.

Implications for Evaluation and Future Work Overall, these findings underscore the need for standardized and transparent evaluation protocols. We encourage the research community to report detailed evaluation choices, avoid reliance on gold annotations, and adopt unified evaluation pipelines to ensure fair comparison and reproducibility across MEE methods Our analysis further indicates that progress in multimedia event extraction requires a stronger focus on accurate cross-modal event coreference and semantic alignment, which remain underexplored in recent work. Finally, future advances would benefit from more comprehensive multimodal datasets with explicit coreference annotations, high-quality training data, and standardized splits to support robust and comparable evaluation.

6 Conclusion

In this work, we present the first systematic analysis of evaluation pitfalls and challenges in MEE and reveal substantial gaps between reported performance and a model's actual ability to ground events across textual and visual modalities. Our analysis of the M2E2 benchmark uncovers three major sources of discrepancies: inconsistent data processing, inconsistent task assumptions, and overly relaxed evaluation settings. To address this, we propose the evaluation framework StrictEval, which enforces strict evaluation constraints for more challenging evaluation. Controlled experiments show that minor evaluation choices can significantly affect performance, highlighting that cross-modal event coreference resolution and precise MEE remain open challenges.

Limitations

In this work, we focus on analyzing evaluation pitfalls primarily based on the M2E2 benchmark, which, despite being the most widely used public benchmark for MEE, represents only a subset of possible settings. As a result, some identified issues and recommended practices may not fully generalize to newer benchmarks, annotation schemes, or domains beyond news media. Extending our analysis to additional datasets, modalities (e.g., videos or audio), and task formulations remains an important direction for future work. Moreover, our empirical analysis relies on a relatively simple MEE model and a limited set of reproduced recent methods. While this design choice allows us to isolate the impact of evaluation decisions, it does not capture the full diversity of modeling approaches, particularly recent instruction-following MLLMs. Although we expect the identified evaluation pitfalls to persist across architectures, their quantitative impact may vary with different modeling paradigms. Finally, our proposed evaluation framework adopts stricter evaluation settings and result into substantially lower absolute performance scores. While this may hinder direct comparisons with prior work, our goal is to promote more realistic and transparent evaluation that better reflects the real-world challenges of MEE.

Ethical Considerations

The results reported in this paper are intended to improve evaluation transparency in MEE and should not be interpreted as implying misconduct by prior work. The identified evaluation pitfalls are subtle and often related to underspecified benchmark evaluation standards, making them easy to be overlooked. Therefore, the motivation of this work is to raise awareness about these issues and to promote more reliable evaluation practices.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Jianwei Cao, Yanli Hu, Zhen Tan, and Xiang Zhao. 2025. Cross-modal multi-task learning for multimedia event extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11454–11462.

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.

Xinrong Chen, Xiang Yuan, Haochen Li, Hang Yang, Guanyu Wang, Weiping Li, and Tong Mo. 2025. Stepwise schema-guided prompting framework with parameter efficient instruction tuning for multimedia event extraction. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zilin Du, Yunxin Li, Xu Guo, Yidan Sun, and Boyang Li. 2023. Training multimedia event extraction with generated images and captions. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 5504–5513, New York, NY, USA. Association for Computing Machinery.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12804–12825, Bangkok, Thailand. Association for Computational Linguistics.

H. W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Jiaqi Li, Chuanyi Zhang, Miaozeng Du, Dehai Min, Yongrui Chen, and Guilin Qi. 2023. Three stream based multi-level event contrastive learning for text-video event extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1666–1676, Singapore. Association for Computational Linguistics.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16399–16408.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Jian Liu, Yufeng Chen, and Jinan Xu. 2022. Multimedia event extraction from news with a unified contrastive learning framework. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1945–1953, New York, NY, USA. Association for Computing Machinery.

Maofu Liu, Zhenyi Hu, Bingying Zhou, Huijun Hu, Chen Qiu, and Xiaokang Zhang. 2025a. Cross-modal event extraction based on adaptive feature selection and semantic-aware graph. *Knowledge-Based Systems*, 326:114038.

Maofu Liu, Bingying Zhou, Huijun Hu, Chen Qiu, and Xiaokang Zhang. 2025b. Cross-modal event extraction via visual event grounding and semantic relation filling. *Information Processing Management*, 62(3):104027.

Yang Liu, Fang Liu, Licheng Jiao, Qianyue Bao, Long Sun, Shuo Li, Lingling Li, and Xu Liu. 2024. Multi-grained gradual inference model for multimedia event extraction. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10507–10520.

Farhad Moghimifar, Fatemeh Shiri, Van Nguyen, Yuan-Fang Li, and Gholamreza Haffari. 2023. Theia: Weakly supervised multimodal event extraction from incomplete data. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–145, Nusa Dua, Bali. Association for Computational Linguistics.

Abhijnan Nath, Huma Jamil, Shafiuddin Rehan Ahmed, George Arthur Baker, Rahul Ghosh, James H. Martin, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. Multimodal cross-document event coreference resolution using linear semantic transfer and mixed-modality ensembles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11901–11916, Torino, Italia. ELRA and ICCL.

Hao Peng, Xiaozhi Wang, Feng Yao, Zimu Wang, Chuzhao Zhu, Kaisheng Zeng, Lei Hou, and Juanzi Li. 2023a. OmniEvent: A comprehensive, fair, and easy-to-use toolkit for event understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 508–517, Singapore. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu, and Weixing Shen. 2023b. The devil is in the details: On the pitfalls of event extraction evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9206–9227, Toronto, Canada. Association for Computational Linguistics.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded Situation Recognition. In Andrea Vedaldi, Horst Bischof,

Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12349, pages 314–332. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*. Version Number: 1.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Philipp Seeberger, Dominik Wagner, and Korbinian Riedhammer. 2024. MMUTF: Multimodal multimedia event argument extraction with unified template filling. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6539–6548, Miami, Florida, USA. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. 2024. Umie: Unified multimodal information extraction with instruction tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19062–19070.

Yuxuan Sun, Kai Zhang, and Yu Su. 2023. Multimodal question answering for unified information extraction. *Preprint*, arXiv:2310.03017.

Meihan Tong, Shuai Wang, Yixin Cao, Bin Xu, Juanzi Li, Lei Hou, and Tat-Seng Chua. 2020. Image enhanced event detection in news articles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9040–9047.

Rejin Varghese and Sambath M. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6.

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.

Walker, Christopher, Strassel, Stephanie, Medero, Julie, and Maeda, Kazuaki. 2006. ACE 2005 Multilingual Training Corpus. Artwork Size: 1572864 KB Pages: 1572864 KB.

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.

Shuo Wang, Meizhi Ju, Yunyan Zhang, Yefeng Zheng, Meng Wang, and Guilin Qi. 2023. Cross-modal contrastive learning for event extraction. In *Database Systems for Advanced Applications*, pages 699–715, Cham. Springer Nature Switzerland.

Xiaoyu Wang, Tao Sun, Gengchen Liu, Zhi Yang, Jiahui Liu, and Zimeng Xu. 2025. Mgfsg-ee: A method based on multi-grained fusion and scene graph enhancement for event extraction. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 3103–3112, New York, NY, USA. Association for Computing Machinery.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fuyu Xing, Zimu Wang, Wei Wang, and Haiyang Zhang. 2025. Benchmarking and improving LVLMs on event extraction from multimedia documents. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 734–742, Hanoi, Vietnam. Association for Computational Linguistics.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*.

Jiaao Yu, Yijing Lin, Zhipeng Gao, Xuesong Qiu, and Lanlan Rui. 2025. Multimedia event extraction with LLM knowledge editing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4116–4124, Suzhou, China. Association for Computational Linguistics.

Li Yuan, Yi Cai, Xudong Shen, Qing Li, Qingbao Huang, Zikun Deng, and Tao Wang. 2025. Collaborative multi-lora experts with achievement-based multi-tasks loss for unified multimodal information extraction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 6940–6948. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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.

Tongtao Zhang, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph Ellis, Lifu Huang, Wei Liu, Heng Ji, and Shih-Fu Chang. 2017. Improving event extraction via multimodal integration. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 270–278, New York, NY, USA. Association for Computing Machinery.

Yuhui Zhang, Yongxiu Xu, Minghao Tang, Xinkui Lin, Yubin Wang, Hongbo Xu, and Gaopeng Gou. 2025. Rda: Regularized domain adaptation for multimedia event extraction. In *Advanced Intelligent Computing Technology and Applications*, pages 308–319, Singapore. Springer Nature Singapore.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2021. Revisiting the evaluation of end-to-end event extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4609–4617, Online. Association for Computational Linguistics.

A Appendix

A.1 Articles for Systematic Analysis

Table 5 presents a comprehensive list of studies included in our systematic analysis, along with their reported evaluation scores. These works cover diverse modeling paradigms, ranging from early pipeline-based approaches to recent instruction-following and MLLMs.

	Textual		Visual		Multimedia	
Method	ED	EAE	ED	EAE	ED	EAE
WASE_{A} (2020)	48.1	30.1	42.8	10.3	49.1	19.9
WASE_{O} (2020)	50.6	26.4	49.9	11.9	50.8	19.2
UniCL (2022)	53.7	30.7	57.6	15.2	53.4	23.4
$\text{CLIP}_{\text{Event}}$ (2022)	-	-	52.7	17.1	-	-
CAMEL (2023)	55.4	31.1	58.5	24.4	57.5	33.2
UMIE (2024)	-	-	-	-	62.1	24.5
MGIM (2024)	55.8	31.2	58.5	17.8	55.6	24.6
MMUTF (2024)	55.5	38.2	57.0	20.9	54.6	27.4
RDA (2025)	56.6	33.4	60.3	26.1	58.7	34.6
VEGSRF (2025b)	-	-	-	-	53.9	25.3
AFSSAG (2025a)	-	-	-	-	54.4	26.5
C-LoRAE (2025)	-	-	-	-	63.5	32.6
X-MTL (2025)	56.6	36.0	71.7	32.2	66.2	41.4
MGFSG (2025)	58.5	31.9	61.8	20.3	57.9	27.4
MLLM (2025)	76.0	39.3	65.3	27.3	87.4	44.3
SSGPF (2025)	-	-	-	-	65.7	36.0

Table 5: Reported F1 scores of methods on M2E2.

A.2 One-to-One Matching Strategy

As described in §3.5, visual arguments are initially evaluated using a many-to-many matching strategy. In this setting, multiple predicted arguments may be matched to the same ground-truth argument (many-to-one) and a single predicted argument may match multiple ground-truth arguments (one-to-many). While this approach captures all overlapping predictions, it can inflate evaluation scores and does not enforce a strict mapping between predicted and gold arguments.

One-to-One Matching To address this limitation, we adopt a one-to-one matching strategy based on the Hungarian algorithm (Kuhn, 1955). Predicted and ground-truth arguments are modeled as nodes in a bipartite graph and a globally optimal matching is computed between the two sets. This formulation guarantees that each predicted argument is matched to at most one ground-truth argument, yielding a stricter and more interpretable evaluation.

Empirical Impact Analysis In Table 6, we show the impact for object detectors YOLOv8 (Redmon et al., 2016; Varghese and M., 2024) and Faster R-CNN (Ren et al., 2015), trained on COCO (Lin et al., 2014) and OpenImages (Kuznetsova et al., 2020), respectively. Models trained on COCO show robust performance under the stricter matching strategy, whereas results on OpenImages exhibit substantial differences. For example, we observe an F1 increase of +8.0 for YOLO-X OI at $\tau = 0.5$. This behavior can be attributed to overlapping object categories in OpenImages (e.g., *person* and *suit*) which are otherwise counted as multiple correct predictions under many-to-many matching (see Figure 4).

			many-to-many			one-to-one		
Detector	τ	**#Pred**	P	R	F1	P	R	F1
YOLO-X CC	0.8	1617	32.5	36.8	34.5	31.5	35.7	33.5
YOLO-X CC	0.5	3162	20.9	46.3	28.8	20.1	44.4	27.6
YOLO-X CC	0.1	5067	14.6	51.9	22.8	13.4	47.7	21.0
YOLO-X OI	0.8	267	51.7	9.7	16.3	45.7	8.5	14.4
YOLO-X OI	0.5	2349	26.5	43.6	33.0	20.1	33.0	25.0
YOLO-X OI	0.1	5917	17.8	73.7	28.7	11.1	45.8	17.8
FR-CNN OI	0.8	191	36.7	4.9	8.6	35.6	4.7	8.4
FR-CNN OI	0.5	1037	37.1	26.9	31.2	30.3	21.9	25.5
FR-CNN OI	0.1	4086	28.8	82.4	42.7	12.7	36.3	18.8

Table 6: Visual EAE scores using ground truth events with the many-to-many (original) and one-to-one (ours) evaluation. Models labeled CC and OI correspond to object detectors trained on COCO (80 classes) and OpenImages (600 classes), respectively. The parameter τ denotes the chosen minimum confidence threshold for each object.

A.3 Proposed Single Task Models

In this section, we describe the Single Task models used in the experiments reported in §4. Each subtask is modeled independently, including textual ED / EAE, visual ED / EAE, and Event Coreference Resolution.

Textual Event Extraction For textual ED and EAE, we employ BERT (Devlin et al., 2019) as the text encoder. Given an input sentence s, BERT produces contextualized subtoken representations, which are mean-pooled to obtain token-level embeddings. Textual ED is formulated as a sequence labeling task where each token is classified into an event type using a linear classifier. For textual EAE, following Li et al. (2020), we assume gold entity mentions and perform role classification by mean-pooling the subtoken representations of each entity mention. Textual ED and EAE are trained as separate models using cross-entropy loss.

Visual Event Extraction For visual ED and EAE, we adopt the CLIP (Radford et al., 2021) vision encoder. Given an input image i, CLIP produces global and patch-level representations. Visual ED is treated as image-level classification by feeding the [CLS] token representation into a linear classifier. For visual EAE, we first detect objects using an offline detector (Cao et al., 2025). Patch representations corresponding to each object are mean-pooled to form object embeddings, which are then classified into argument roles using a linear layer. Analogous to the textual tasks, Visual ED and EAE are trained independently with cross-entropy loss.

Multimedia Event Extraction As described in §4.1, event coreference resolution between textual and visual events is performed using CLIP-based similarity scores. Following prior work (Du et al., 2023; Seeberger et al., 2024; Cao et al., 2025), we construct a multimedia event when a text–image pair shares the same predicted event type and their similarity score exceeds a threshold of 20. The resulting multimedia event aggregates all associated textual and visual arguments. In addition to this heuristic approach, we also report results obtained using greedy and bipartite matching strategies (Liu et al., 2022).

Training We train all models using a learning rate of 1×10^{-5} for encoder parameters and 1×10^{-4} for classifier parameters. Textual models are trained with a batch size of 16 for 20 epochs, while visual models use a batch size of 64 and are trained for 10 epochs. Model performance is evaluated at the end of each epoch on the ACE development set for textual tasks and the SWiG development set for visual tasks. We select the checkpoint with the best development set performance for final evaluation.

A.4 Implementation Details

All models are implemented using the *Transformers* (Wolf et al., 2020) (v4.55.0) library in conjunction with *PyTorch* (v2.8.0). Unless otherwise specified, we use *bert-base-uncased* (Devlin et al., 2019) with 222M parameters as text encoder and *clip-vit-base-patch16* (Radford et al., 2021) with 85M parameters as vision encoder. Object detections are obtained using YOLOv8 (Varghese and M., 2024)[3] trained on COCO and detections with confidence scores below 0.8 are discarded. All experiments are conducted on NVIDIA A100 GPUs within a single compute node running CUDA 12.3. We run each experiment with three random seeds and report the average performance across runs.

A.5 Reproduction Details

In this section, we provide detailed reproduction information for the models discussed in §5, along with potential explanations for any observed differences in performance scores.

CAMEL We use the official released code[4] provided by Du et al. (2023). Our reproduced F1 scores largely align with the reported results, except for multimedia ED (48.3 vs. 57.5 F1) and EAE (26.5 vs 33.2 F1). We attribute these discrepancies primarily to the absence of balanced visual ED training in our reproduction that reduced recall in favor of precision.

MMUTF We use the official released code[5] provided by Seeberger et al. (2024). Following the original paper, which mainly focuses on EAE, we use ED predictions from CAMEL. Our reproduced scores closely match the reported results with minor decreases in multimedia ED and EAE, which we also attribute to the lower recall of CAMEL predictions.

X-MTL We use the official released code[6] provided by Cao et al. (2025). Our reproduced results show only minor deviations from the originally reported results (e.g., 56.6 vs. 52.9 F1 for textual ED). We believe these differences are related to inconsistencies in the pseudo-labeled VOA image-caption dataset, caused by broken links during dataset construction.

SSGPF We re-implement SSGPF using *LLaVA-v1.5-7B* as described in the paper[7] (Chen et al., 2025). The model assumes aligned image-sentence pairs and requires manually written event and role descriptions. We obtain comparable performance on multimedia ED (65.7 vs. 61.0 F1) but observe a substantial drop on EAE (36.0 vs. 14.1 F1), which we attribute to implementation differences in EAE evaluation, visual grounding (SEEM), and role descriptions. For StrictEval, we use the proposed fine-tuned cross-modal retrieval model to construct image-sentence pairs.

[3] https://docs.ultralytics.com/models/yolov8

[4] https://github.com/ZILIN003/CAMEL

[5] https://github.com/seebergerph/MMUTF

[6] https://github.com/aoine-dev/X-MTL

[7] https://github.com/MartinYuanNJU/SSGPF

A.6 Analysis of StrictEval

We base our experiments on the human-annotated M2E2 dataset (Inter-Annotator Agreement of 81.2%) (Li et al., 2020) and remove relaxed assumptions used in prior work to ensure more consistent comparison. The resulting substantially lower scores are primarily due to stricter evaluation protocols that expose false positives otherwise ignored, which we detail next. Using our proposed Single Task models, we provide further intuition for these effects.

Textual Evaluation In textual evaluation, oracle trigger refinement[8] and test subset selection remove a large number of incorrect predictions and negative samples. For example, the Single Task model predicts 3,969 text events, of which refinement removes 1,636 false positive events and 1,885 arguments prior to evaluation. Furthermore, our manual analysis suggests annotation discrepancies between ACE and M2E2, potentially contributing to the large number of false positive events.

Visual Evaluation Similarly, test subset selection excludes 623 images without annotated events, restricting evaluation to only 391 positive samples. As a result, false positives from the excluded images are not counted. For instance, the Single Task model predicts 88 false positive visual events and 341 arguments on these omitted images. Evaluating only on positive samples therefore inflates precision by ignoring predictions on negative images.

Multimedia Evaluation For multimedia evaluation, recent work often assumes access to ground truth image–sentence pairs or applies post-hoc filtering using gold sentence and image IDs (e.g., via coreference annotations). This substantially reduces the number of evaluated pairs and removes negative candidates. In contrast, StrictEval constructs and evaluates over all possible image–sentence pairs. Consequently, models must handle a much larger and noisier candidate space, leading to a significant increase in false positives. For our baseline models, post-hoc filtering excludes 1,118 false positive multimedia events, resulting into a substantial drop in precision and F1 scores.

[8] https://github.com/jianliu-ml/Multimedia-EE/blob/main/code/textualEE/refine_result.py

A.7 Additional Experimental Results

This section supplements the multimedia results presented in §4.3 by providing EAE scores along with precision and recall. We report evaluations for different coreference resolution strategies: threshold-based (Table 7 and Table 8), greedy matching (Table 9), and bipartite matching (Table 10).

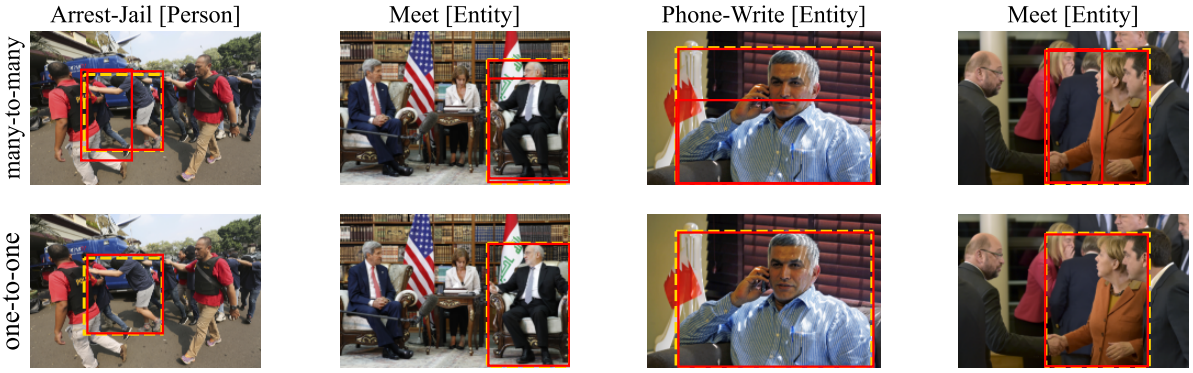

Figure 4: Qualitative error analysis of visual EAE. Gold rectangles denote ground-truth argument roles while red rectangles indicate correctly counted predictions. The top row illustrates a common failure case in which overlapping objects (often other persons, suits, or shirts) are incorrectly counted as matches. This inflates performance metrics such as recall. In contrast, the bottom row shows our proposed one-to-one version which alleviates these error cases.

		ED				EAE			
	Setting	P	R	F1	ΔF1	P	R	F1	ΔF1
Multi. EE	StrictEval	6.2 ±0.9	22.4 ±3.1	9.7 ±1.4	-	3.4 ±0.4	10.8 ±1.1	5.2 ±0.5	-
	w/ [P8] eval MED relaxed	8.4 ±1.2	23.7 ±3.3	12.4 ±1.6	+2.7	3.4 ±0.4	10.8 ±1.1	5.2 ±0.5	-
	w/ [P4/5] eval text/image subset	50.4 ±3.7	22.4 ±3.1	30.9 ±3.3	+21.2	22.9 ±0.9	10.8 ±1.1	14.4 ±1.0	+9.2
	w/ [P4/5] gold coreferences	66.5 ±2.5	22.8 ±3.2	33.9 ±3.8	+24.2	27.5 ±0.7	11.3 ±1.3	15.6 ±1.3	+10.4

Table 7: Evaluation results of MEE model with threshold-based coreference resolution (CLIP_{Raw}).

		ED				EAE			
	Setting	P	R	F1	ΔF1	P	R	F1	ΔF1
Multi. EE	StrictEval	8.6 ±1.1	20.4 ±2.8	12.0 ±1.5	-	4.6 ±0.5	10.2 ±1.0	6.3 ±0.5	-
	w/ [P8] eval MED relaxed	11.9 ±1.4	21.7 ±3.1	15.3 ±1.7	+3.3	4.6 ±0.5	10.2 ±1.0	6.3 ±0.5	-
	w/ [P4/5] eval text/image subset	52.9 ±2.6	20.4 ±2.8	29.4 ±3.3	+17.4	24.0 ±0.1	10.2 ±1.0	14.3 ±1.0	+8.0
	w/ [P4/5] gold coreferences	66.5 ±2.5	22.8 ±3.2	33.9 ±3.8	+21.9	27.5 ±0.7	11.3 ±1.3	15.6 ±1.3	+9.3

Table 8: Evaluation results of MEE model with threshold-based coreference resolution (CLIP_{VOA}).

		ED				EAE			
	Setting	P	R	F1	ΔF1	P	R	F1	ΔF1
Multi. EE	StrictEval	20.0 ±0.8	11.8 ±1.1	14.8 ±0.9	-	10.0 ±0.8	6.6 ±0.7	7.9 ±0.7	-
	w/ [P8] eval MED relaxed	26.8 ±1.0	12.7 ±1.3	17.2 ±1.2	+2.4	10.0 ±0.8	6.6 ±0.7	7.9 ±0.7	-
	w/ [P4/5] eval text/image subset	60.5 ±3.1	11.8 ±1.1	19.8 ±1.7	+5.0	27.1 ±1.1	6.6 ±0.7	10.6 ±1.0	+2.7
	w/ [P4/5] gold coreferences	66.5 ±2.5	22.8 ±3.2	33.9 ±3.8	+14.1	27.5 ±0.7	11.3 ±1.3	15.6 ±1.3	+7.7

Table 9: Evaluation results of MEE model with greedy matching coreference resolution.

		ED				EAE			
	Setting	P	R	F1	ΔF1	P	R	F1	ΔF1
Multi. EE	StrictEval	23.3 ±1.7	12.1 ±1.3	15.9 ±1.5	-	10.1 ±0.8	6.0 ±0.8	7.5 ±0.8	-
	w/ [P8] eval MED relaxed	29.0 ±1.2	12.8 ±1.5	17.7 ±1.6	+1.8	10.1 ±0.8	6.0 ±0.8	7.5 ±0.8	-
	w/ [P4/5] eval text/image subset	67.5 ±3.6	12.1 ±1.3	20.4 ±2.1	+4.5	26.4 ±1.4	6.0 ±0.8	9.8 ±1.2	+2.3
	w/ [P4/5] gold coreferences	66.5 ±2.5	22.8 ±3.2	33.9 ±3.8	+18.0	27.5 ±0.7	11.3 ±1.3	15.6 ±1.3	+8.1

Table 10: Evaluation results of MEE model with bipartite matching coreference resolution.