# DTPD-MEE: A Dual-path Teacher and Pseudo-label Distillation Framework for Multimodal Event Extraction

Yuanye He*†, Shuhao Hu*†, Jin Wang‡, Ji Xiang*, Xiaobo Guo*, Qingfei Zhao*†
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
†School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
‡School of Information Science and Engineering, Yunnan University, Yunnan, China
{heyuanye,hushuhao}@iie.ac.cn, wangjin@ynu.edu.cn

***Abstract*—Multimodal Event Extraction (MEE) is critical for holistic semantic understanding but is currently severe bottlenecked by the extreme scarcity of high-quality annotated multimodal datasets. To address this challenge, we propose DTPD-MEE, a novel Dual-path Teacher and Pseudo-label Distillation framework that shifts the focus from architectural complexity to a data-centric knowledge transfer paradigm. Our approach adopts a "divide-and-conquer, then unify" strategy: first, two specialist teacher models are trained on large-scale unimodal corpora (text-only and image-only) to capture deep modality-specific expertise. These teachers then function as automated annotators for vast unlabeled multimodal data. Through a rigorous cross-modal consistency check filter, we successfully construct a high-quality synthetic dataset that is $3\times$ larger than the standard M2E2 benchmark. Finally, a unified student model inherits this distilled expertise through sequence-level distillation. Experimental results demonstrate that DTPD-MEE exhibits outstanding performance on the M2E2 benchmark, notably improving the multimedia argument extraction F1-score by 7.3 points. Our framework effectively demonstrates how to unlock the potential of unlabeled multimodal data by bridging the gap between unimodal abundance and multimodal scarcity.**

***Index Terms*—Multimodal Event Extraction, Knowledge Distillation, Large Vision-Language Models**

## I. Introduction

Event Extraction, a fundamental task in information retrieval, has traditionally focused on single-modality data such as text [1], [2], images [3], or videos [4], [5]. However, relying on a single modality can lead to an incomplete understanding, as critical event information is often distributed across visual and textual channels. For instance, it has been observed that up to 33% of visual arguments in multimedia news from the Voice of America (VOA) may be absent from the accompanying text, making a strong case for Multimodal Event Extraction (MEE) [6]. Unfortunately, training the powerful end-to-end models required for MEE demands vast amounts of supervised data, which presents a critical bottleneck for the field, as creating large-scale, high-quality annotated datasets is exceedingly time-consuming and costly.

Existing approaches for Multimodal Event Extraction (MEE) can be broadly categorized into two main streams: those focusing on cross-modal alignment and those dedicated to designing efficient extraction architectures. The former aims to bridge the semantic gap between heterogeneous modalities, exemplified by pioneering works like WASE [6] and advanced alignment frameworks such as CLIP-EVENT [7] and UniCL [8]. The latter concentrates on architectural sophistication to maximize extraction capability, ranging from the multi-grained inference mechanisms in MGIM [9] to recent large model-based paradigms like MMUTF [10] and UMIE [11]. However, both categories face inherent bottlenecks. Alignment-based methods are often constrained by weak supervision from a single modality, which prevents them from capturing true cross-modal semantics. Meanwhile, architecture-centric models, despite their sophisticated designs, consistently struggle to achieve optimal performance due to the critical scarcity of high-quality training data. To mitigate such data shortages, synthesizing training samples has emerged as a promising solution in many low-resource domains. Following this trend, CAMEL [12] attempts to tackle MEE data scarcity through generative data augmentation. Nevertheless, a critical flaw remains, since the synthetic data is typically derived from an incomplete, single-modality perspective, it suffers from severe distribution shift compared to real-world multimedia events. Consequently, none of these prior techniques have effectively resolved the fundamental challenge of data scarcity in MEE.

The severity of this data scarcity is starkly evident in current benchmarks. Despite its wide application, the standard M2E2 dataset [6] contains only 309 fully annotated multimodal instances. This extreme paucity stands in sharp contrast to the abundance of high-quality resources available for single-modality tasks: the text-only ACE2005 corpus [13] provides 16,531 annotated event mentions, while the SWiG dataset [14] offers 126,102 images with grounded situation labels. Furthermore, vast amounts of unlabeled multimodal data remain untapped. This huge disparity motivates a critical research question: can we unlock the potential of these abundant but disjoint unimodal resources to compensate for the lack of multimodal annotations?

To address this question, we tackle the challenge from

Corresponding authors: Shuhao Hu and Jin Wang.

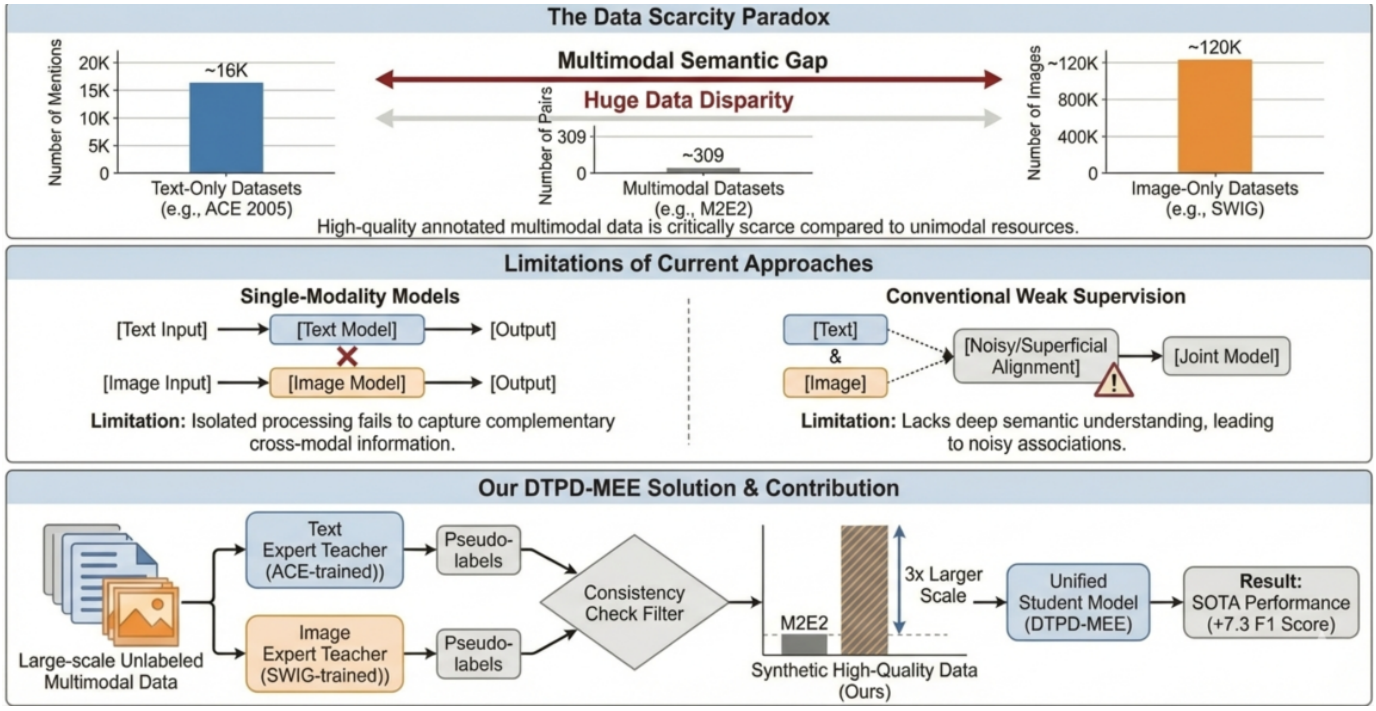


Fig. 1. Motivation of DTPD-MEE: (a) The significant data disparity between abundant unimodal datasets and scarce multimodal annotations. (b) Limitations of existing single-modality and weak-supervision methods in capturing true cross-modal semantics. (c) Our framework bridges this gap by distilling knowledge from unimodal experts to synthesize a high-quality multimodal corpus.

a data-centric perspective by proposing DTPD-MEE, a novel framework for **D**ual-path **T**eacher and **P**seudo-label **D**istillation based **M**ultimodal **E**vent **E**xtraction. Our approach leverages a teacher-student paradigm to transfer knowledge from data-rich unimodal domains. The methodology first involves training two expert unimodal teacher models—one for textual events on corpora like ACE2005 [13] and one for visual events on datasets like SWiG [14]. These specialists then serve as annotators, generating high-quality pseudo-labels for unlabeled multimodal data. Through this process, we construct a high-quality synthetic dataset that is $3\times$ larger than the existing M2E2 benchmark. Subsequently, this pseudo-labeled corpus trains a unified student model via knowledge distillation. Our framework reframes the problem from designing complex models for scarce data to creating a high-quality training corpus via knowledge transfer. We validate our approach on the public M2E2 benchmark, where DTPD-MEE exhibits outstanding performance. Notably, it boosts the F1-score for multimedia argument extraction by a significant 7.3 points over the previous best, showcasing the effectiveness of our strategy.

In summary, our main contributions are as follows:

- We propose DTPD-MEE, a novel framework for Multimodal Event Extraction that addresses the critical bottleneck of annotated data scarcity through a dual-path teacher and pseudo-label distillation approach.
- We tackle the Multimodal Event Extraction task from a data-centric perspective, constructing a high-quality synthetic dataset that is $3\times$ larger than the widely used M2E2 benchmark. We further empirically demonstrate the effectiveness of this generated dataset in mitigating the data scarcity problem.
- Extensive evaluations on the public M2E2 benchmark demonstrate that DTPD-MEE exhibits outstanding performance, notably improving the multimedia argument extraction F1-score by a significant 7.3 points.

## II. Related Work

In this section, we review the development of event extraction from unimodal to multimodal perspectives and discuss recent paradigms designed to address the data scarcity bottleneck.

### A. Unimodal Event Extraction

Event Extraction (EE) has traditionally been investigated within single-modality contexts. In the textual domain, early research focused on refining extraction through cross-document inference [1] and joint extraction models utilizing structured prediction with global features [15]. These efforts primarily relied on annotated corpora like ACE 2005 [13]. In the visual domain, the task is often framed as situation recognition [3], which involves visual semantic role labeling to identify activities and their participants, often utilizing datasets like SWiG [14]. Additionally, extensive work has been conducted on event detection in videos using datasets like UCF101 [4] and EventNet [5]. However, relying on a single modality can lead to incomplete event understanding,

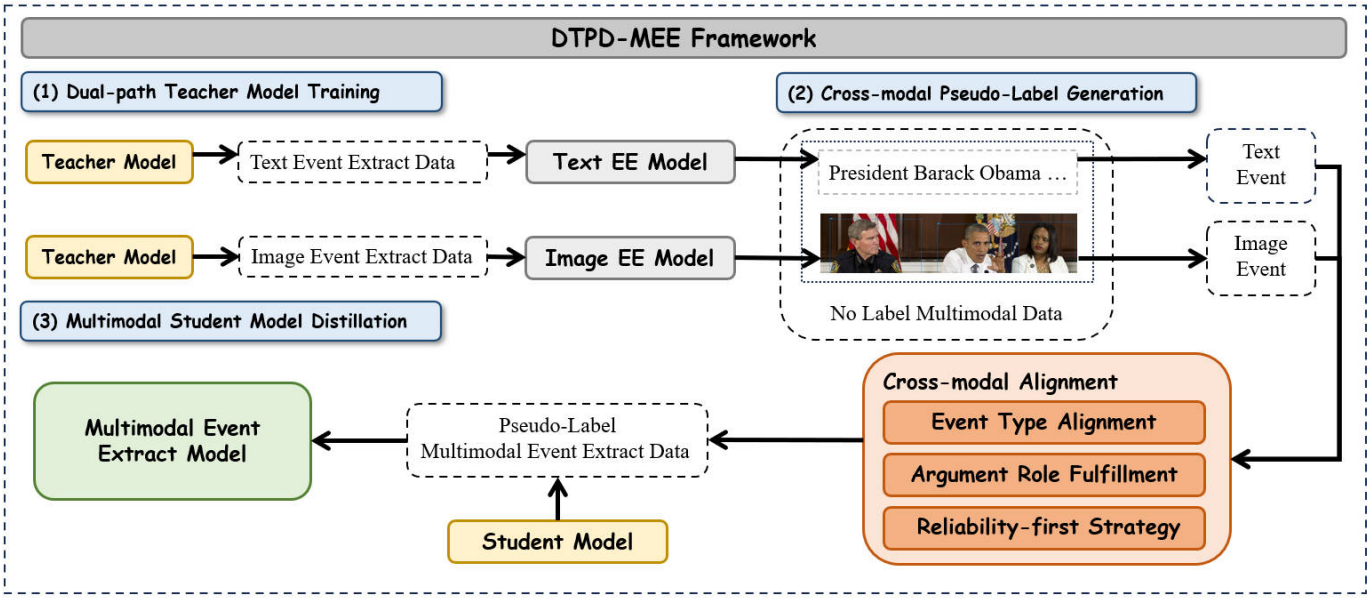


Fig. 2. The overall framework of DTPD-MEE.

as critical event information is often distributed across visual and textual channels.

### B. Multimodal Event Extraction (MEE)

To address the limitations of unimodal event extraction, MEE aims to extract structured event information from both text and images. The WASE framework [6] established a foundational weak supervision paradigm by learning a shared embedding space for cross-media modeling. Subsequent research has advanced on two main fronts: cross-modal alignment and architectural sophistication. CLIP-EVENT [7] leveraged pre-trained structures to connect textual triggers with image regions, while UniCL [8] introduced a unified contrastive learning objective. Furthermore, the Multi-grained Gradual Inference Model (MGIM) [9] proposed a hierarchical approach to refine multimodal representations through multi-grained reasoning.

The field currently faces a critical bottleneck due to the high cost of manual annotation for multimodal data. Recent advances in large pre-trained models have inspired new paradigms to circumvent this issue. UMIE [11] explored instruction tuning for unified information extraction, while MMUTF [10] reframed argument extraction as a unified template-filling task. To tackle data scarcity directly, CAMEL [12] utilized generative data augmentation through synthetic images and captions.

Our proposed DTPD-MEE framework diverges from these methods by adopting a data-centric perspective. By utilizing large pre-trained models such as the Qwen [16] and Qwen-VL [17] series, we transfer specialized knowledge from unimodal experts to a unified student model. This approach effectively unlocks the potential of large-scale unlabeled multimodal data without the need for manual labels.

## III. Method

The core objective of DTPD-MEE is to bridge the multimodal data gap by leveraging mature unimodal knowledge to supervise the learning of a multimodal system. We propose a three-stage pipeline: (1) **Knowledge Acquisition**, where unimodal teachers are trained as domain experts; (2) **Knowledge Fusion and Filtering**, which synthesizes high-fidelity pseudo-labels; and (3) **Knowledge Transfer**, where a student model inherits this distilled expertise. The architecture is detailed in Fig. 2.

### A. Dual-path Teacher Model Training

Based on the preceding analysis, while high-quality annotated multimodal data is critically scarce, text-only and image-only event extraction tasks possess an abundance of high-quality available data. To fully leverage these unimodal resources to assist multimodal event extraction, we first construct and train two parallel, specialist teacher models, $T_{text}$ and $T_{vision}$, each optimized for deep event understanding ability within its respective modality.

**Text Teacher Model ($T_{text}$):** We leverage a Large Language Model (LLM), specifically Qwen2.5-7B, as the backbone. The model is fine-tuned on $\mathcal{D}_{text}$ (ACE2005) using an instruction-following paradigm. Given a sentence $t$, $T_{text}$ is trained to generate a structured JSON-like sequence:

$$E_{text} = \{e_{type}, e_{trigger}, \{(r_i, s_i)\}_{i=1}^{n}\} \tag{1}$$

where $e_{type}$ is the event category, $e_{trigger}$ is the anchor word, and $(r_i, s_i)$ represents the $i$-th argument role and its corresponding text span. This fine-tuning enables the teacher to capture complex linguistic nuances in event structures.

**Visual Teacher Model ($T_{vision}$):** For visual event understanding, we employ Qwen2.5-VL-7B, a vision-language

model. Trained on SWiG ($\mathcal{D}_{vision}$), the model performs grounded situation recognition. For an image $i$, it identifies the primary action and its participants:

$$E_{vision} = \{v_{type}, \{(r_j, b_j)\}_{j=1}^{m}\} \tag{2}$$

where $v_{type}$ is the action class and $b_j$ is the spatial bounding box for the $j$-th argument $r_j$. This path ensures the framework can localize event components in the visual channel.

### B. Cross-modal Pseudo-Label Generation

With expert teacher models $T_{text}$ and $T_{vision}$ independently trained, we employ them as automated annotators to process the large-scale unlabeled VOA corpus $\mathcal{D}_{unlabeled}$. However, since these teachers are modality-specific, their raw predictions $E_{text} = T_{text}(t)$ and $E_{vision} = T_{vision}(i)$ are often disjoint and potentially noisy due to inherent modality gaps and model-specific biases. To synthesize high-fidelity multimodal supervision, we design a rigorous two-stage **Cross-modal Consistency Check** to filter noise and ensure semantic alignment.

**Hybrid Semantic Taxonomy Alignment:** The primary obstacle to knowledge fusion is the heterogeneous taxonomies: ACE2005 contains 33 event types, while SWiG defines 504 verbs. To bridge this gap, we construct a mapping dictionary $\mathcal{M}$ through a multi-step hybrid pipeline. First, we perform **Initial Candidate Retrieval** by computing the semantic similarity between each textual event type and all visual verbs using Sentence-BERT embeddings. This process represents each category as a dense vector, allowing us to rank potential matches based on their cosine similarity scores. Second, to handle linguistic nuances that automated methods might miss, we implement **Expert-led Manual Refinement**. Domain experts review the top-ranking candidates for each textual event to resolve polysemy and ensure cross-dataset precision. For example, the visual verb *shooting* is selectively mapped to either *Conflict:Attack* or *Justice:Execute* depending on the specific event schema, rather than relying on a naive one-to-one match. Finally, we established accurate event type mappings from text event extraction datasets ACE2005 and image event extraction datasets SWiG to multimodal event extraction datasets M2E2.

**Structural Argument Integrity Verification:** For aligned event pairs, we further enforce an **Argument Role Fulfillment** constraint to ensure the synthesized label is structurally complete and informative. Based on the event schema $\mathcal{S}$ of the aligned type, we identify a set of core roles $R$, such as *Attacker*, *Target*, or *Place*. The unified pseudo-label $E_{pseudo}$ is generated by a conditional fusion of unimodal arguments:

$$\begin{aligned} E_{pseudo} = &\text{Merge}(e_{text}, e_{vision}) \\ &\text{s.t.} \quad |\{r \in R \mid a_r \in (e_{text} \cup e_{vision})\}| \geq \tau \end{aligned} \tag{3}$$

where $\tau$ is a role coverage threshold. During the merging process, we resolve modality conflicts using a **Reliability-first Strategy**. Specifically, when both the text teacher $T_{text}$ and the vision teacher $T_{vision}$ predict argument candidates (i.e., a textual span and a visual bounding box) for the same event role, we assess their cross-modal alignment using a lightweight CLIP model. We compute the cosine similarity between the embeddings of the textual span and the visual region features. By establishing a similarity threshold $t$, argument pairs with scores falling below $t$ are identified as semantic conflicts (indicating a mismatch between the linguistic entity and the visual object) and are strictly filtered out to ensure the reliability of our data.

This filtering process is intentionally highly selective. As evidenced by our dataset statistics, only 978 high-quality image-text pairs were distilled from the initial 11,214 VOA candidates. This drastic reduction represents a deliberate design choice: prioritizing the purity and consensus of the distillation signal over sheer data quantity. This ensures that the student model learns robust cross-modal correlations rather than fitting to the noise inherent in original unimodal predictions.

### C. Multimodal Student Model Distillation

The final stage of our framework involves distilling the synthesized knowledge from the teacher-filtered corpus $D_{pseudo}$ into a unified student model $S$. To ensure the student possesses the requisite architectural capacity for end-to-end MEE, we employ Qwen2.5-VL-7B as the base model. This model acts as a mapping function $S(i, t) \rightarrow \hat{E}_{multi}$, which directly translates an interleaved image-text pair into a structured event representation.

Unlike traditional knowledge distillation methods that rely on matching soft logit distributions—which is computationally prohibitive for large-scale generative models—we adopt **Sequence-level Distillation**. This approach reframes the distillation task as a conditional generation problem, where the student is supervised by the linearized, high-quality pseudo-labels $E_{pseudo}$ generated in the previous stage. The training objective is to minimize the sequence-level negative log-likelihood (NLL) loss over the pseudo-labeled corpus $D_{pseudo}$:

$$\mathcal{L}_{distill}(\theta) = -\sum_{(i,t)\in D_{pseudo}} \sum_{k=1}^{L} \log P(y_k \mid y_{<k}, i, t; \theta) \tag{4}$$

where $y = \{y_1, y_2, \ldots, y_L\}$ represents the tokens of the structured event description $E_{pseudo}$, and $\theta$ denotes the trainable parameters of the student model.

This distillation paradigm offers two primary advantages. First, by supervising the student with consensus labels that satisfy both textual and visual constraints, the objective compels the model to learn **Implicit Cross-modal Alignments**. For instance, the model must learn to associate a specific person's name mentioned in the text with the corresponding bounding box coordinates provided by the visual expert to minimize the loss. Second, we initialize the weights of $S$ using the parameters of the visual teacher $T_{vision}$ rather than a vanilla pre-trained model. This strategy not only accelerates convergence but also ensures that the student retains the strong visual grounding and situation recognition priors acquired during the visual teacher's specialized training phase. Consequently, the student model achieves robust MEE performance without ever being exposed to manually annotated multimodal data.

TABLE I
MAIN RESULT. **BOLD** NUMBERS INDICATE THE BEST EAE MODEL WHEREAS UNDERLINED METRICS DENOTE THE SECOND BEST.

| | **Textual Events** | | | | | | **Visual Events** | | | | | | **Multimedia Events** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Event Mention | | | Argument Role | | | Event Mention | | | Argument Role | | | Event Mention | | | Argument Role | | |
| **Model** | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| $FLAT_{ATT}$ | 34.2 | 63.2 | 44.4 | 20.1 | 27.1 | 23.1 | 27.1 | 57.3 | 36.7 | 4.3 | 8.9 | 5.8 | 33.9 | 59.8 | 42.2 | 12.9 | 17.6 | 14.9 |
| $FLAT_{OBJ}$ | 38.3 | 57.9 | 46.1 | 21.8 | 26.6 | 24.0 | 26.4 | 55.8 | 35.8 | 9.1 | 6.5 | 7.6 | 34.1 | 56.4 | 42.5 | 16.3 | 15.9 | 16.1 |
| $WASE_{ATT}$ | 37.6 | 66.8 | 48.1 | 27.5 | 33.2 | 30.1 | 32.3 | 63.4 | 42.8 | 9.7 | 11.1 | 10.3 | 38.2 | 67.1 | 49.1 | 18.6 | 21.6 | 19.9 |
| $WASE_{OBJ}$ | 42.8 | 61.9 | 50.6 | 23.5 | 30.3 | 26.4 | 43.1 | 59.2 | 49.9 | 14.5 | 10.1 | 11.9 | 43.0 | 62.1 | 50.8 | 19.5 | 18.9 | 19.2 |
| $CLIP_{EVENT}$ | - | - | - | - | - | - | 41.3 | **72.8** | 52.7 | 21.1 | 13.1 | 17.1 | - | - | - | - | - | - |
| UniCL | 49.1 | 59.2 | 53.7 | 27.8 | 34.3 | 30.7 | 54.6 | 60.9 | 57.6 | 16.9 | 13.8 | 15.2 | 44.1 | 67.7 | 53.4 | 24.3 | 22.6 | 23.4 |
| MGIM | 50.1 | 66.5 | **55.8** | 28.2 | 34.7 | 31.2 | 55.7 | 64.4 | 58.5 | 24.1 | 14.1 | 17.8 | 46.3 | **69.6** | 55.6 | 25.2 | 21.7 | 24.6 |
| UMIE | - | - | - | - | - | - | - | - | - | - | - | - | - | 62.1 | - | - | - | 24.5 |
| CAMEL | 45.1 | **71.8** | 55.4 | 24.8 | 41.8 | 31.1 | 52.1 | 66.8 | 58.5 | 21.4 | **28.4** | 24.4 | **55.6** | 59.5 | **57.5** | 31.4 | 35.1 | 33.2 |
| MMUTF | 48.5 | 65.0 | 55.5 | 33.6 | **44.2** | 38.2 | 55.1 | 59.1 | 57.0 | 23.6 | 18.8 | 20.9 | 47.9 | 63.4 | 54.6 | **39.9** | 20.8 | 27.4 |
| DTPD-MEE | **56.8** | 51.1 | 53.6 | **40.6** | 43.9 | **42.8** | **76.6** | 57.4 | **65.6** | **35.2** | 28.1 | **31.2** | 47.8 | 43.2 | 45.4 | 31.3 | **57.0** | **40.5** |

TABLE II
DATASET STATISTIC

| **Modality** | **Dataset** | **Number** | **event type** |
|---|---|---|---|
| text | ACE2005 | 16531 | 33 |
| | M2E2 (text-only) | 1105 | 8 |
| image | SWiG | 126102 | 504 |
| | M2E2 (image-only) | 188 | 8 |
| multimodal | M2E2 (multimodal) | 309 | 8 |
| | VOA (w/o event label) | 11214 | 0 |
| | VOA (w pseudo-label) | 978 | 8 |

## IV. EXPERIMENTS

### A. Experimental Setup

*1) Datasets:* We evaluate our framework on the public Multimodal Event Extraction benchmark dataset M2E2 [6]. The teacher models are trained on the ACE2005 [13] (text-only) and SWiG [14] (image-only) datasets, respectively. The unlabeled data used for pseudo-label distillation comes from the VOA dataset [6], which contains 11,214 news image-text pairs. We successfully generated 978 pairs with high-quality pseudo-labels for training the student mode. The dataset statistics are shown in Table.II.

*2) Evaluation Metrics:* Following [15], we use standard Precision (Prec.), Recall (Rec.), and F1-score (F1) to evaluate performance on Event Mention and Argument Role identification. An event mention is correct only if its event type and trigger offset (or corresponding image) are correctly predicted. An event argument is correct if its event type, span (text offset or visual bounding box IoU $> 0.5$), and role label are all matched.

*3) Baselines:* We compare DTPD-MEE with eight strong baseline models, including early works like FLAT and WASE [6], and recent SOTA models such as CLIP-EVENT [7], UniCL [8], MGIM [9], UMIE [11], CAMEL [12], and MMUTF [10].

*4) Implementation Details:* For our experiments, we utilized models from the Qwen series. The Text Teacher $T_{text}$ was developed by fine-tuning the Qwen2.5-7B [16] model, while the Visual Teacher $T_{vision}$ and the final Student model $S$ were both based on the Qwen2.5-VL-7B [17] architecture. The teacher models were trained for 5 epochs with a batch size of 4 using standard hyperparameters. We used the AdamW optimizer with a cosine learning rate schedule and an initial learning rate of $2 \times 10^{-7}$ for the student model's distillation stage. All experiments were performed on a server equipped with four NVIDIA A800 GPUs, and the implementation was built upon PyTorch and the Hugging Face Transformers library.

### B. Main Results

As shown in Table.I, our proposed DTPD-MEE framework exhibits outstanding performance, with its most significant contributions concentrated in the challenging task of argument role extraction across all modalities. In the textual domain, our model achieves an F1 score of 42.8% for argument roles, substantially outperforming the previous best model, MMUTF (38.2%), by 4.6 points. The performance gains are even more pronounced in visual and multimedia contexts. For visual argument roles, DTPD-MEE attains an F1 score of 31.2%, surpassing the prior SOTA from CAMEL (24.4%) by a remarkable 6.8 points. This advantage widens further in the comprehensive multimedia setting, where our model's 40.5% F1 score represents a 7.3-point improvement over CAMEL (33.2%), driven by an exceptionally high recall of 57.0%. While the framework's primary strength lies in argument extraction, its robust performance is also evident in its SOTA result for Visual Event Mention (65.6% F1). These results demonstrate that the knowledge transferred via unimodal teachers and pseudo-label distillation equips the student model with superior capabilities for fine-grained, cross-modal event understanding.

### C. Ablation Study

To validate the effectiveness of each key component in our framework, we conducted a comprehensive ablation study, with results presented in Table III. Removing the cross-modal consistency check (w/o cross-modal alignment) results in a

TABLE III
ABLATION STUDY

| | Multimedia Events | | | | | |
|---|---|---|---|---|---|---|
| | Event Mention | | | Argument Role | | |
| **Model** | P | R | F1 | P | R | F1 |
| DTPD-MEE | 47.8 | 43.2 | 45.4 | **31.3** | **57.0** | **40.5** |
| w/o cross-modal alignment | 45.7 | 41.4 | 43.4 | 29.9 | 55.4 | 38.2 |
| w/o fine-tune | 43.7 | 39.6 | 41.5 | 22.0 | 32.2 | 26.2 |
| fine-tune in ACE2005 | 47.9 | 42.1 | 44.8 | 30.7 | 52.2 | 38.7 |
| fine-tune in SWiG | **57.9** | **50.2** | **53.8** | 18.5 | 17.8 | 18.2 |

2.3 point drop in the Argument Role F1 score to 38.2, confirming that our alignment strategy is crucial for generating high-quality, high-confidence pseudo-labels. Furthermore, the model without any distillation (w/o fine-tune) exhibits the most significant performance degradation in argument extraction, with its F1 score plummeting to 26.2, which starkly indicates that our distillation process is indispensable for knowledge transfer. We also explored fine-tuning on single-modality datasets, revealing a critical modality bias when training only on visual data (fine-tune in SWiG). While this model achieves the highest F1 score for Event Mention (53.8), its ability to identify argument roles collapses to the lowest F1 score of 18.2. In contrast, fine-tuning solely on the text-rich ACE2005 dataset yields a much stronger argument role F1 of 38.7. Taken together, these results confirm the essential roles of both the cross-modal alignment for generating high-quality pseudo-labels and the distillation process for effectively fusing unimodal knowledge.

## V. CONCLUSION

In this paper, we proposed DTPD-MEE, a data-centric framework designed to overcome the scarcity of annotated data in Multimodal Event Extraction. By distilling knowledge from mature unimodal experts, we successfully constructed a high-quality synthetic dataset that is $3\times$ larger than the M2E2 benchmark. This scalable supervision enabled our model to achieve superior performance, significantly improving multimedia argument extraction by 7.3 F1 points. Our findings confirm that leveraging abundant unimodal resources is a highly effective strategy for low-resource multimodal understanding. Future work will extend this paradigm to more complex scenarios, such as video event extraction.

## REFERENCES


[1] Heng Ji and Ralph Grishman, "Refining event extraction through cross-document inference," in *Proceedings of ACL-08: HLT*, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, Eds., Columbus, Ohio, June 2008, pp. 254–262, Association for Computational Linguistics.

[2] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, and H. Wu, "Unified structure generation for universal information extraction," in Proc. 60th Annual Meeting of the Assoc. Comput. Linguistics (ACL), May 2022, pp. 5755–5772.

[3] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5534–5542.

[4] Khurram Soomro, Amir Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *ArXiv*, vol. abs/1212.0402, 2012.

[5] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang, "Eventnet: A large scale structured concept library for complex event detection in video," in *Proceedings of the 23rd ACM International Conference on Multimedia*, New York, NY, USA, 2015, MM '15, p. 471–480, Association for Computing Machinery.

[6] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang, "Cross-media structured common space for multimedia event extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 2557–2568, Association for Computational Linguistics.

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

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

[9] Yang Liu, Fang Liu, Licheng Jiao, Qianyue Bao, Long Sun, Shuo Li, Lingling Li, and Xu Liu, "Multi-grained gradual inference model for multimedia event extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 10507–10520, 2024.

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

[11] Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou, "Umie: Unified multimodal information extraction with instruction tuning," in *Proceedings of AAAI*, 2024.

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

[13] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda, "Ace 2005 multilingual training corpus," *Philadelphia: Linguistic Data Consortium*, 2006.

[14] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi, "Grounded situation recognition," *ArXiv*, vol. abs/2003.12058, 2020.

[15] Qi Li, Heng Ji, and Liang Huang, "Joint event extraction via structured prediction with global features," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Hinrich Schuetze, Pascale Fung, and Massimo Poesio, Eds., Sofia, Bulgaria, Aug. 2013, pp. 73–82, Association for Computational Linguistics.

[16] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu, "Qwen2.5 Technical Report," 2025.

[17] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.