A Multimedia Event Detection Method based on Dual-Path Iterative Synergistic Fusion for Modal Information Asymmetry

Gengchen Liu*, Tao Sun*†, Xiaoyu Wang*, Zhi Yang*, Jiahui Liu*, Zimeng Xu*
Qilu University of Technology (Shandong Academy of Sciences)
Jinan, China
10431230213@stu.qlu.edu.cn, sunt@qlu.edu.cn, {10431230054,10431230191,10431240239,10431240211}@stu.qlu.edu.cn

***Abstract*—Multimedia Event Detection (MED), as a fundamental task in multimodal understanding, plays a vital role in downstream applications such as information retrieval and situational awareness. However, current MED approaches often suffer from two key limitations: (1) insufficient exploitation of the rich semantics embedded in visual data, particularly in social media contexts where textual descriptions are brief, ambiguous, or noisy; and (2) limited cross-modal interaction, typically relying on shallow fusion strategies that fail to capture fine-grained alignments between modalities. To address these challenges, this paper proposes a novel Dual-Path Synergy Multimodal Model (DPSMM), which incorporates two core modules: the Vision-Language Synergy Enhancer (VLSE) and the Dual-Iterative Attentive Fusion (DIAF). VLSE enriches image representations by generating verb-centric captions and extracting local textual cues through OCR, thereby transforming implicit action semantics into explicit textual signals while filtering noise. DIAF then performs bidirectional, multi-round attention to progressively align and refine multimodal features. Together, these modules enhance semantic consistency across modalities and improve robustness under weakly aligned or noisy inputs. Experimental results on the M2E2 benchmark demonstrate that DPSMM achieves superior performance compared to existing state-of-the-art models, with a 3.6% increase in key evidence recall, and a 1.5% gain in overall F1 score, validating the effectiveness of the proposed model in capturing deep multimodal semantics for robust event detection.**

***Index Terms*—Multimedia Event Detection, Multimodal Representation Learning, Vision-Language Synergy**

I. Introduction

Social media has become a core channel for real-time information, but its explosive growth poses significant challenges for event detection due to the massive influx of heterogeneous data. The information conveyed by text and images is often complementary, illustrated in Figure 1. Pure text models may misinterpret ambiguous words (e.g., "died" implying Life-Die instead of Transport-Movement), while image-only models might focus on isolated visual patterns. Multimodal methods bridge this gap by aligning structured semantics, yet they still struggle with complex user-generated content.

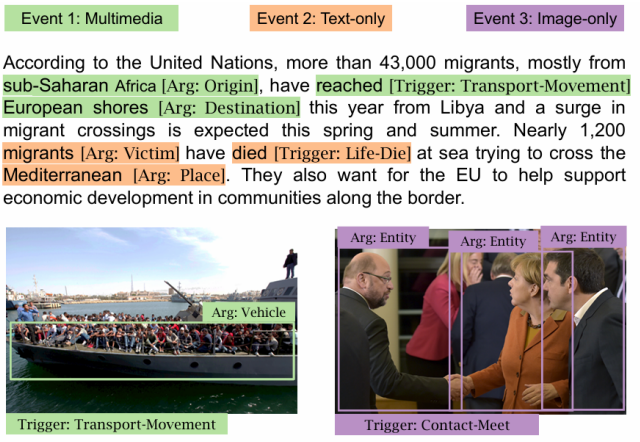

Fig. 1. Detection results from different modalities. The text-only model predicts a Life-Die event (orange region), while the image-only model identifies a Contact-Meet event (purple region). The multimodal model correctly detects a Transport-Movement event (green region) by integrating textual and visual cues.

For example, in the rally event from the M2E2 dataset shown in Figure 2, the posted text includes words like "gathering" and "conflict," while the visual information depicts a group of people holding a "stop occupation" banner. Existing methods struggle to capture the crucial details of implied verb semantics in the image, and the accuracy of event detection heavily relies on identifying verb trigger words. Additionally, for semantically similar but event-type-different words in the text (such as "rally" and "conflict"), traditional event detection models often suffer from weight distribution imbalances, leading to imprecise matching between visual and textual information, which results in incorrect event classification and misidentification of a rally event as a conflict event. This phenomenon reveals the fundamental flaws in current technologies regarding cross-modal semantic collaboration and dynamic adaptation.

Although the event detection capabilities of existing multi-

This work was supported by the Shandong Provincial Natural Science Foundation General Project (ZR2023MG069); the Shandong Province Science and Technology Small and Medium-sized Enterprise Innovation Capacity Enhancement Project (2023TSGC0212) and the Key Project of Undergraduate Teaching Reform Research of Shandong Province (Project No. Z2024165).

*Also with Key Laboratory of Computing Power Network and Information Security, Ministry of Education; Shandong Computer Science Center (National Supercomputer Center in Jinan); Shandong Engineering Research Center of Big Data Applied Technology; Faculty of Computer Science and Technology; Shandong Provincial Key Laboratory of Industrial Network and Information System Security; Shandong Fundamental Research Center for Computer Science, Jinan, China.

†Corresponding author: Tao Sun (also with Jinan Key Laboratory of Low-Code Research).

Image	
Text	**People take part in a rally in support of Azerbaijan over the conflict in the breakaway Nagorno-Karabakh region, outside the Armenian embassy in Kyiv, Ukraine, April 8, 2016.**

Fig. 2. An example of a Demonstrate event from the M2E2 dataset. The textual context's ambiguity (triggering "conflict") is rectified by complementary visual evidence (e.g., protesters with signs), illustrating the inherent semantic imbalance in social media posts.

modal models have been validated on some publicly available datasets, several notable challenges still exist for multimodal event detection tasks, summarized as follows:

1) Cross-modal trigger word ambiguity: A single visual scene may correspond to multiple event types (e.g., "crowd gathering" could be linked to "rally," "protest," or "celebration"). Relying solely on textual descriptions makes it challenging to accurately identify trigger words.

2) Implicit expression of action semantics in images: The implicit nature of trigger words within visual features makes them challenging to represent directly. Since images lack explicit trigger word labels, actions must be inferred from elements such as scenes, objects, and text (e.g., "holding a sign" implies a "protest" action).

3) The failure of traditional coarse-grained cross-modal alignment mechanisms: Text and images often carry complementary information asymmetrically, creating a "visual wealth - textual poverty" cognitive gap. However, mainstream models rely on global feature alignment, overlooking local semantic connections. Traditional methods (such as feature concatenation and average pooling) use fixed weights or simple linear combinations, making it difficult to adjust the importance of textual words based on image content.

To tackle these challenges, this paper proposes the Dual-Path Synergy Multimodal Model (**DPSMM**):

· To address challenges 1) and 2) , this paper introduces the Vision-Language Synergy Enhancer (**VLSE**), which generates verb-oriented structured descriptions and performs cross-modal alignment. Its aim is to deepen the utilization of key event information, suppress noise interference, enhance the purity of visual information, and improve the model's accuracy and robustness in event detection.

· To address challenge 3) , this paper introduces the Dual-Iterative Attentive Fusion Module (**DIAF**), which enhances the model's ability to adapt to complex social media data through fine-grained alignment. It focuses on high-value, cross-modal consistent information, filters out single-modal noise, and enables the model to more precisely capture the essence of events, avoiding classification biases caused by modality fragmentation or noise interference.

II. Current Research

In this section, we present a thorough summary of the existing research, concentrating on three principal domains: event detection, image description, and multimodal fusion.

A. Multimedia Event Detection and Modal Alignment

Traditional ED primarily focused on unimodal text using CNNs or GCNs [1]–[5]. Multimodal integration advanced this by grounding textual triggers in visual contexts [6]–[10]. Recently, Generative MEE frameworks leverage Large Multimodal Models (LMMs) for structural reasoning [11], [12], though they still face challenges with domain-specific social media noise.

B. Visual-Semantic Enrichment via Textual Proxies

Image captioning in MED transforms implicit visual signals into natural language proxies [13]–[16]. Our work aligns with strategies using generated captions [17] and LLM-based augmentation [18], [19] to internalize rich visual semantics and enhance cross-modal interaction.

C. Cross-Modal Interaction and Synergistic Fusion

Fusion strategies have progressed from feature concatenation [20], [21] to attention-driven interactions [22]–[25]. Recent models employ multi-granularity reasoning [26] and knowledge editing [27]–[31]. Unlike these approaches, we focus on dynamic, bidirectional attention to achieve fine-grained semantic alignment and noise suppression.

III. Methods

This section introduces the Dual-Path Synergy Multimodal Model (DPSMM), illustrated in Figure 3. DPSMM addresses the modal information asymmetry in social media event detection by combining structured semantic enhancement with a bidirectional, dual-round attention mechanism. Specifically, the Vision–Language Synergy Enhancer (VLSE) augments visual features with explicit textual cues, and the Dual-Iterative Attentive Fusion (DIAF) progressively refines cross-modal alignment for robust event understanding.

A. Problem Definition

The event detection task is framed as a multi-class, single-label classification problem. Given a multimodal dataset $D = \{(S_i, I_i, y_i)\}_{i=1}^{N}$, where $S_i =< w_1, w_2, ..., w_n >$is the text sequence, I_i is the associated image, and $y_i \in \{1, ..., k\}$ represents the event category label (with 8 categories in M2E2). The goal of the model is to learn the mapping function $f{:}(S, I) \rightarrow y$ that minimizes the classification loss:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k} \log(\hat{y}_{i,k}) \tag{1}$$

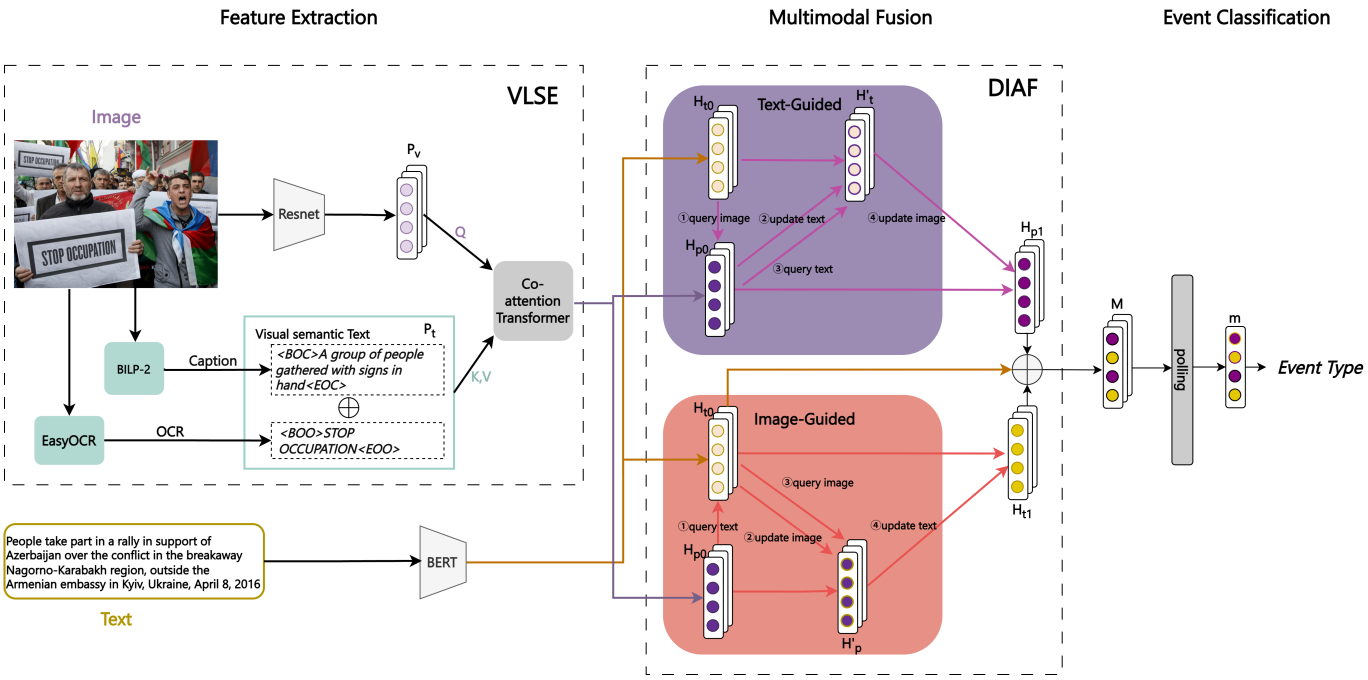

Fig. 3. An overview of the proposed Dual-Path Synergy Multimodal Model (DPSMM). VLSE enriches image semantics through verb-oriented captioning and OCR-based extraction, addressing the asymmetric information density between rich visual content and sparse social media text. These features are then processed by DIAF via two rounds of bidirectional attention to capture fine-grained cross-modal interactions. Finally, a classification head maps the fused representation to event types, ensuring robust detection in complex contexts.

Where $\hat{y}_{i,k}$ represents the predicted probability by the model that the i-th sample belongs to class k.

B. Feature Extraction

This section provides a detailed explanation of how input text and images are processed in the feature extraction layer.

1) Image Feature Extraction

First, we resize the input image to 224×224 pixels and and extract visual features using ResNet: $P_v = Resnet(I)$;

Next, To expose implicit action semantics, we obtain **textual proxies** directly from the image:

Verb-centric captions (BLIP-2): generate concise action-focused descriptions (e.g., C_v="Crowd gathering on the street, holding signs.").

OCR evidence (EasyOCR): extract high-confidence text strings from banners or signs, with low-confidence or duplicate tokens removed. The OCR output is a set: $O_v = \{(s_i, b_i)\}$, where s_i represents the text content and b_i represents the bounding box coordinates.

Both sources are encoded with BERT, producing contextual representations. The combined output is denoted as:

$$P_t = Concat(C_v, O_v) \tag{2}$$

which represents the text features derived from the image.

2) Synergy step

The proxy-derived text features Pt are used to guide the visual embeddings Pv, producing enhanced visual representations:

$$Attn_{pt \to pv} = softmax\left(\frac{Q_v K_t^T}{\sqrt{d}}\right) \tag{3}$$

$$H_{p0} = \sigma(W(Attn_{pt \to pv}.V_t) + b) \tag{4}$$

3) Text Feature Extraction

In parallel, the original post text T is encoded by BERT, yielding:

$$H_{t0} = BERT(S) \in R^{30 \times 768} \tag{5}$$

This module enables the transformation from "visual wealth" to "textual wealth" through the techniques described above, providing high-purity, strongly semantically correlated multi-modal input for the subsequent dual-path attention mechanism, laying the groundwork for fine-grained event classification.

C. Multimodal Fusion

In this section, we provide a detailed explanation of the dual-path attention fusion method we propose. Since image and text information mutually influence each other, the focus area of the same image differs depending on the textual context. Similarly, the same word can convey different events under different visual descriptions. To address this, we designed the Dual-Iterative Attentive Fusion Module, as shown in Figure 4.

1) Text-Guided Visual Attention

In the first round, we use three fully connected layers to map the image representation H_{p0} to the first two inputs of the scaled dot-product attention module, and map the text representation H_{t0} to the third input, denoted as v_{image}, k_{image}, and q_{text}.

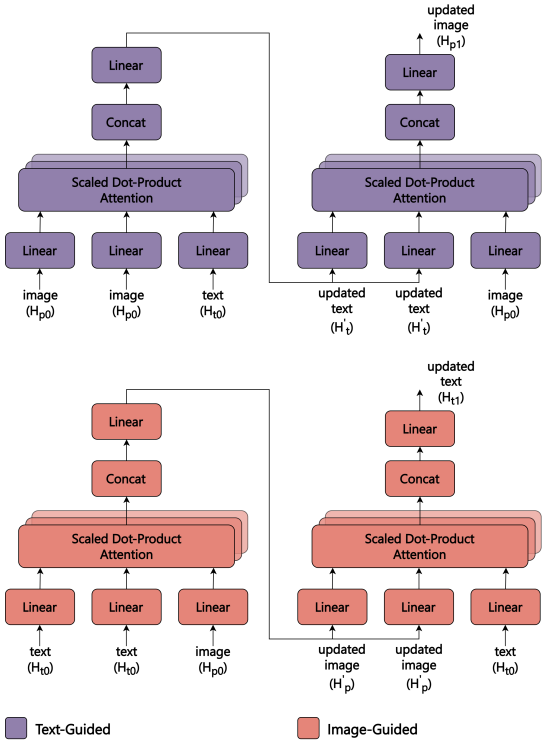

Fig. 4. The illustration of Dual-Iterative Attentive Fusion Module (DIAF). DIAF comprises two channels: a text-guided visual update (indicated by the purple part) and an image-guided textual update (indicated by the red part). Each channel performs two rounds of attention: the first reduces unimodal ambiguity, while the second refines cross-modal alignment, enabling iterative and fine-grained fusion of multimodal features.

First, we calculate the attention α_{text} by using q_{text} to query k_{image}:

$$\alpha_{text} = softmax\left(\frac{q_{text}k_{image}^{T}}{\sqrt{d}}\right) \tag{6}$$

Next, we compute the dot product of the learned attention α_{text} with the first input v_{image} to obtain the text representation H'_t after the shallow interaction:

$$H'_t = \alpha_{text}.v_{image} \tag{7}$$

We express the calculations of formulas 6 and 7 as ω, thus the first round of interaction can be summarized as:

$$H'_t = \omega(H_{t0}, H_{p0}) \tag{8}$$

In the second round, the calibrated text H'_t serves as the query to reweight the visual features H_{p0}, yielding the refined representation:

$$H_{p1} = \omega(H_{p0}, H'_t) \tag{9}$$

This step deepens cross-modal associations by aligning text-informed semantics with visual evidence.

2) Image-Guided Textual Attention

In the symmetric branch, the enhanced visual features H_{p0} guide the refinement of the plain text features H_{t0}. Through two rounds of attention, the model produces enhanced textual features H_{t1}:

$$H_{t1} = \omega(H_{t0}, \omega(H_{p0}, H_{t0})) \tag{10}$$

3) Concatenation and Preliminary Projection

After obtaining the two enhanced features, we integrate the pure text representation back into the enhanced features to better preserve the original text semantics and prevent the vanishing of gradients in BERT parameters during training, as shown below:

$$M = H_{p1} \oplus H_{t1} \oplus H_{t0} \tag{11}$$

Where $\oplus$ denotes the concatenation operation.

D. Classifier

We use average pooling to obtain the final embedding:

$$m = avg - pooling(M) \tag{12}$$

A two-layer fully connected network with a standard ReLU activation function is used to project the fused features M into the target event category space:

$$\hat{y} = Softmax(Relu(mW_c + b_c)) \tag{13}$$

Where W_c and b_c are classifier parameters, and $\widehat{y}$ represents the probability vector for all event categories. Finally, we use the cross-entropy loss function as the objective function to optimize all parameters during training:

$$\mathcal{L}_C(y, \hat{y}) = -\sum_{i=1}^{|D_{tr}|} y_i \log(\hat{y}_i) \tag{14}$$

Where y_i represents the correct event category.

IV. Experiments

A. Dataset and Setup

We align event ontologies between text and image modalities using the ACE2005 dataset (comprising 33 text event types and 36 semantic roles) and the imSitu dataset (over 200 event categories, 504 verbs, and 1,788 roles) for training [11]. We evaluate our model on the unified M2E2 benchmark, which contains 245 documents annotated with 1,105 text-only, 188 image-only, and 385 multimodal events. Following standard practices in event detection, we use precision (P), recall (R), and F1-score (F1) to evaluate performance.

Baselines: We compare DPSMM against representative unimodal and multimodal baselines: UniCL [26], Flat and its variants (Flat$_{att}$, Flat$_{obj}$) [11], WASE and its variants (WASE$_{att}$, WASE$_{obj}$, WASE$^{I}{}_{att}$, WASE$^{I}{}_{obj}$) [11], VES-C [28], CLIP-Event [25], and MGIM [29].

All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU, where the online inference model (comprising ResNet, BERT, and DIAF) adds only 9.6M parameters (7% of backbone) ensuring an efficient real-time latency of 30ms per sample.

TABLE I
RESULTS OF DPSMM WITH OTHER MODELS ON THE MULTIMODAL EVENT DETECTION TASK.

Modality	Model	Multimedia Evaluation Event Trigger		
		P	**R**	**F1**
Text	WASE^T	41.2	33.1	36.7
	Unicl	49.1	59.2	53.7
Image	WASE$^I{}_{att}$	28.3	23.0	25.4
	WASE$^I{}_{obj}$	26.1	22.4	24.1
	DPSMM-I	31.7	29.4	30.5
Multimedia	VSE-C	33.3	48.2	39.3
	Flat$_{att}$	33.9	59.8	42.2
	Flat$_{obj}$	34.1	56.4	42.5
	WASE$_{att}$	38.2	67.1	49.1
	WASE$_{obj}$	43.0	62.1	50.8
	Unicl	44.1	67.7	53.4
	CLIP-Event	41.3	72.8	52.7
	MGIM	46.3	69.6	55.6
	Our DPSMM	**46.8**	**73.2**	**57.1**

B. Comparative Experiments

The results of comparing DPSMM with the baselines are shown in Table 1. As shown in Table 1, we can observe that our method performs better in multimedia evaluation scenarios than previous baseline models.

First, compared to single-modality models, our F1 score increases by 3.4%–20.4% over text-only models and 26.6%–33.0% over image-only models, validating the necessity of multimodal integration. Among multimodal models, our F1 score improves by 1.5%–17.8%. Notably, DPSMM outperforms the highly competitive MGIM by 1.5%. While MGIM benefits from its fine-grained alignment mechanism, our superior results confirm that our image semantic enhancement module more effectively resolves modal information asymmetry.

Second, text-only methods surprisingly perform better than image-only ones, despite the typical sparsity of social media text. We attribute this to the fact that M2E2 event labels are derived from the text-centric ACE dataset.

Finally, DPSMM-I significantly improves the F1 score over previous image-only methods, further validating our image semantic enhancement module. Notably, this variant's performance competes with less effective multimodal methods, underscoring the immense value of enriched visual data in event detection.

C. Ablation Experiments

To assess the effectiveness of the proposed modules, we compared DPSMM against three variants: w/o VLSE, w/o DIAF, and a fine-grained w/o Caption. As shown in Table II, removing VLSE and DIAF reduces the F1 score by 7.1% and 3.7%, respectively. Notably, the w/o caption variant outperforms w/o VLSE, indicating that while BLIP-2 significantly enriches visual semantics, the model also derives critical utility from OCR tokens and the DIAF mechanism. This confirms that VLSE is essential for suppressing noise and extracting meaningful visual semantics, yet the architecture remains robust even without generated descriptions.

TABLE II
ABLATION RESULTS OF THE DPSMM MODEL BY REMOVING THE VLSE AND DIAF MODULES, RESPECTIVELY.

Eval.	Method	Event Trigger		
		P	**R**	**F1**
Mul. ED	w/o VLSE	42.5	60.7	50
	w/o DIAF	44.2	67.4	53.4
	w/o Caption	43.8	64.1	52.1
	DPSMM	**46.8**	**73.2**	**57.1**

D. Variant Experiments

TABLE III
PERFORMANCE COMPARISON OF THE IMAGE-ONLY VARIANT DPSMM-I AGAINST OTHER IMAGE-ONLY MODELS.

Modality	Model	Image-Only Evaluation Event Trigger		
		P	**R**	**F1**
Image	WASE$^I{}_{att}$	29.7	61.9	40.1
	WASE$^I{}_{obj}$	28.6	59.2	38.7
	Unicl	50.8	63.2	56.3
	MGIM	45.7	58.3	54.9
	DPSMM-I	**53.5**	**65.4**	**58.8**

To comprehensively evaluate the effectiveness of our model in handling visual information, we introduce DPSMM-I, an image-only variant of DPSMM, and assess its performance on an image-only event detection task. As shown in Table III, our variant model DPSMM-I shows a 3.9% improvement in F1 score compared to the previously best-performing method. Moreover, the comparison with the previous MGIM model in this scenario further corroborates our hypothesis from the comparative experiments.

TABLE IV
EFFECTS OF DIFFERENT FUSION STRATEGIES ON THE DPSMM MODEL.

Eval.	Method	Event Trigger		
		P	**R**	**F1**
Mul. ED	DPSMM-concat	44.2	67.4	53.4
	DPSMM-Tguide	**47.4**	68	55.9
	DPSMM-Vguide	45.2	72.4	55.7
	DPSMM	46.8	**73.2**	**57.1**

To evaluate the DIAF module, we compared DPSMM against three variants (Table IV): DPSMM-concat (feature concatenation), DPSMM-Tguide (text-guided only), and DPSMM-Vguide (image-guided only). F1 scores decreased

by 3.7%, 1.2%, and 1.4% respectively, confirming that our dual-path design facilitates finer-grained interactions than simple concatenation. The similar performance between single-channel variants likely stems from the limited size and high cross-modal correlation of the M2E2 test set. Future work will involve further validation on more diverse datasets.

V. Conclusion

This paper proposes DPSMM, integrating the Vision-Language Synergy Enhancer (VLSE) and Dual-Iterative Attentive Fusion (DIAF) to advance multimodal event detection. By effectively enriching visual semantics and refining cross-modal alignment, DPSMM achieves state-of-the-art results on the M2E2 benchmark. Future research will address remaining challenges—such as multi-image posts, sarcasm, and OCR noise—through external knowledge and pragmatic reasoning, while exploring broader datasets to ensure cross-domain robustness. Overall, DPSMM demonstrates the value of structured cross-modal collaboration, providing a robust foundation for interpretable and generalized event detection.

References

[1] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 167–176.

[2] X. Liu, Z. Luo, and H.-Y. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 1247–1256.

[3] Y. Lin, H. Ji, F. Huang, and L. Wu, "A joint neural model for information extraction with global features," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7999–8009.

[4] T. Nguyen and R. Grishman, "Graph convolutional networks with argument-aware pooling for event detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[5] S. Duan, R. He, and W. Zhao, "Exploiting document level information to improve event detection via recurrent neural networks," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 352–361.

[6] M. Tong, S. Wang, Y. Cao, B. Xu, J. Li, L. Hou, and T.-S. Chua, "Image enhanced event detection in news articles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9040–9047.

[7] B. Chen, X. Lin, C. Thomas, M. Li, S. Yoshida, L. Chum, H. Ji, and S.-F. Chang, "Joint multimedia event extraction from video and article," *arXiv preprint arXiv:2109.12776*, 2021.

[8] J. Yu, J. Jiang, L. Yang, and R. Xia, "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer." Association for Computational Linguistics, 2020.

[9] Y. Lu, H. Lin, J. Xu, X. Han, J. Tang, A. Li, L. Sun, M. Liao, and S. Chen, "Text2event: Controllable sequence-to-structure generation for end-to-end event extraction," *arXiv preprint arXiv:2106.09232*, 2021.

[10] T. Zhang, S. Whitehead, H. Zhang, H. Li, J. Ellis, L. Huang, W. Liu, H. Ji, and S.-F. Chang, "Improving event extraction via multimodal integration," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 270–278.

[11] M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S.-F. Chang, "Cross-media structured common space for multimedia event extraction," *arXiv preprint arXiv:2005.02472*, 2020.

[12] S. Liu, J. Li, G. Zhao, Y. Zhang, X. Meng, F. R. Yu, X. Ji, and M. Li, "Eventgpt: Event stream understanding with multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 139–29 149.

[13] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[14] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[15] Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier, "Language models can see: Plugging visual controls in text generation," *arXiv preprint arXiv:2205.02655*, 2022.

[16] J. Fei, T. Wang, J. Zhang, Z. He, C. Wang, and F. Zheng, "Transferable decoding with visual entities for zero-shot image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3136–3146.

[17] Z. Du, Y. Li, X. Guo, Y. Sun, and B. Li, "Training multimedia event extraction with generated images and captions," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 5504–5513.

[18] Z. Meng, T. Liu, H. Zhang, K. Feng, and P. Zhao, "Cean: Contrastive event aggregation network with llm-based augmentation for event extraction," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 321–333.

[19] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[20] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6892–6899.

[21] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[22] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[23] J. Liu, Y. Chen, and J. Xu, "Multimedia event extraction from news with a unified contrastive learning framework," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1945–1953.

[24] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, and S. Wang, "Cross-modal contrastive learning for multimodal fake news detection," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 5696–5704.

[25] M. Li, R. Xu, S. Wang, L. Zhou, X. Lin, C. Zhu, M. Zeng, H. Ji, and S.-F. Chang, "Clip-event: Connecting text and images with event structures," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 420–16 429.

[26] Y. Liu, F. Liu, L. Jiao, Q. Bao, L. Sun, S. Li, L. Li, and X. Liu, "Multi-grained gradual inference model for multimedia event extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 10 507–10 520, 2024.

[27] Z. Lin, J. Xie, and Q. Li, "Multi-modal news event detection with external knowledge," *Information Processing & Management*, vol. 61, no. 3, p. 103697, 2024.

[28] F. Shi, J. Mao, T. Xiao, Y. Jiang, and J. Sun, "Learning visually-grounded semantics from contrastive adversarial samples," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 3715–3727.

[29] M. Abavisani, L. Wu, S. Hu, J. Tetreault, and A. Jaimes, "Multimodal categorization of crisis events in social media," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 679–14 689.

[30] L. Peng, S. Jian, Z. Kan, L. Qiao, and D. Li, "Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection," *Information Processing & Management*, vol. 61, no. 1, p. 103564, 2024.

[31] J. Yu, Y. Lin, Z. Gao, X. Qiu, and L. Rui, "Multimedia event extraction with llm knowledge editing," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 4116–4124.