



2025 International Conference on Intelligent Computing

July 26-29, Ningbo, China

<https://www.ic-icc.cn/2025/index.php>

PAF-DM: Proposal Alignment Framework for Multimodal Event Extraction via Dynamic Masking

Hengrui Song, Chun Yuan^(✉)

Tsinghua Shenzhen International Graduate School, China

yuanc@sz.tsinghua.edu.cn

Abstract. Multimodal event extraction (MEE) aims to detect and classify event triggers and arguments by integrating information from both text and images. A major challenge in this task lies in the limited availability of annotated multimodal data, prompting the use of cross-modal data augmentation to synthesize missing modalities. However, such synthetic data often contains noise, which may harm rather than help learning. In this work, we propose a novel framework, PAF-DM, that systematically improves the utilization of synthetic data for MEE. Our approach enhances model robustness by introducing selectively filtering unreliable synthetic signals through a dynamic masking mechanism with fine-grained cross-modal alignment. Experiments on the M2E2 benchmark show that PAF-DM achieves state-of-the-art performance, with +1.1% and +0.9% F1 improvements on the event detection (ED) and argument extraction (EAE) tasks, respectively. These results demonstrate the effectiveness of principled synthetic data integration in multimodal event extraction.

Keywords: Multimodal Event Extraction, Cross-modal Data Augmentation, Proposal, Dynamic Masking.

1 Introduction

In real-world applications, event extraction (EE) systems are increasingly expected to handle multimodal inputs—such as text and images—to obtain more comprehensive and accurate event information. Multimodal event extraction (MEE)[1] extends traditional EE by identifying event triggers and arguments across modalities (**Fig. 1**). While multimodal approaches offer the potential to leverage complementary information from different sources, they rely heavily on high-quality, modality-aligned annotations, which are expensive and labor-intensive to collect at scale.

To alleviate this data scarcity, recent advances in MEE have increasingly turned to cross-modal data augmentation. A common strategy is to synthesize missing modalities—such as generating images from text or vice versa—so that unimodal examples can be converted into pseudo-multimodal training samples[2-3]. This enables the model to benefit from multimodal signals even when only one modality is originally available, and has become a promising direction for expanding training coverage without requiring additional manual annotation.


<p>Operation Inherent Resolve, as the U.S.-led coalition's military operation is called, began airstrikes against Islamic State targets in Iraq in June 2014. The air campaign was expanded in September 2014 to battle the militant group in Syria. FILE - Smoke rises over the Syrian city of Kobani, following a US led coalition airstrike, seen from outside Suruc, on the Turkey-Syria border Monday, Nov. 10, 2014.</p>				
Text Argument		Image Argument	Trigger	Event type
Attacker	Place	Target		
coalition	Kobani	house	airstrike	Attack


Fig. 1. Multimodal Event Extraction.

However, effectively leveraging synthetic data remains a major challenge. Generated images and texts are often noisy, containing hallucinated objects or irrelevant descriptions that misalign with real-world event semantics (**Fig. 2**). These artifacts may not only confuse alignment mechanisms but also introduce distribution shifts that hurt generalization. Moreover, existing methods typically apply synthetic data directly into training pipelines without assessing its reliability or relevance to the target event, making models prone to overfitting on misleading signals.

(a) Argument role errors in synthetic images

How do you feel though seeing pictures of civilians injured in the bombings	Argument role
	Victim

↓




Argument role: Instrument ✗

(b) Event type errors in synthetic images

We are not expecting to drive into Baghdad suddenly and seize it in a coup de mains, or anything like that.	Event type
	Transport

↓



Event type: Demonstration ✗

Fig. 2. Examples of hallucinations and artifact problems in text-to-image generation models.

To address these challenges, we first observe that noisy content in synthetic data—such as hallucinated phrases or misaligned visual regions—is often concentrated in specific segments rather than uniformly spread. This suggests that aligning modalities at a coarse level may obscure useful signals. Focusing on alignment at the proposal level, where candidate event triggers and visual objects are first identified, allows the model to isolate and attend to more relevant and reliable cross-modal correspondences.

In addition, not all synthetic data is equally informative. Some proposals provide helpful training signals, while others introduce misleading patterns. Relying equally on all content risks overfitting to these artifacts. To address this, we introduce a dynamic masking mechanism that adaptively filters out low-confidence proposals during training, enabling the model to emphasize more trustworthy information.

Building on these principles, we propose PAF-DM, a Proposal Alignment Framework with Dynamic Masking, designed to fully exploit synthetic data while minimizing its adverse effects. PAF-DM introduces two key components: (1) Proposal-level Alignment: Instead of aligning entire modalities, PAF-DM first extracts high-confidence event trigger and object proposals from both text and image using confidence-guided modules. Cross-modal interactions are then performed at the proposal level via self-attention, ensuring more accurate and event-relevant alignment even in the presence of noisy inputs. (2) Dynamic Masking: To handle the unreliability of synthetic data, PAF-DM employs a dynamic masking strategy that identifies and suppresses low-confidence proposals during training. This allows the model to reduce its dependence on noisy or hallucinated content and focus on learning from more reliable signals.

Our contributions are summarized as follows:

- We design a fine-grained, proposal-based multimodal alignment mechanism that enhances the fusion of event-relevant information across text and image modalities.
- We propose a novel dynamic masking strategy that adaptively adjusts the influence of synthetic data during training, improving the robustness and generalization ability of the model.
- We conduct extensive experiments on the M2E2 benchmark dataset, achieving +1.1% and +0.9% F1 improvements on the ED and EAE tasks, respectively. These results demonstrate the effectiveness and superiority of our proposed method.

2 Related work

2.1 Visual Event Extraction

Visual event extraction aims to detect event types and their semantic arguments directly from images or videos. A pioneering line of work is Situation Recognition (SR)[4], which frames this task as predicting a verb (e.g., “riding”) along with associated roles (e.g., agent, place). While SR captures high-level semantics, it lacks explicit grounding between roles and image regions. To bridge this gap, the SWiG dataset[5] extends SR into Grounded Situation Recognition, adding bounding boxes for each argument to enable joint detection and localization.

Following this, methods like GSRTR[6] and SituFormer[7] apply Transformer architectures to improve global reasoning over visual features and role dependencies. Zhao et al[8] further incorporate relational structures through graph-based modeling, enhancing the coherence of event-role predictions. Despite these advances, most approaches remain limited to the visual domain and rely on static image-text templates or closed vocabularies, restricting scalability and generalization. Moreover, they are unable to leverage textual context, which is often critical for resolving ambiguity in visual scenes.

2.2 Multimodal Event Extraction

Multimodal Event Extraction (MEE) incorporates both visual and textual modalities to achieve more accurate and complete event understanding. M2E2[1] demonstrates that visual context can enhance textual event extraction, especially when text is ambiguous or sparse. However, multimodal training data with fine-grained event annotations is scarce, and in many scenarios, one modality may be missing—leading to the modality-missing problem. This challenges model robustness and generalization.

To address this, some works adopt cross-modal contrastive learning[9-10] to align representations of paired images and texts, enabling knowledge transfer between modalities. Others tackle missing-modality scenarios by retrieving or generating auxiliary inputs[11-12], e.g., generating pseudo-captions from images. Yet, these solutions often introduce noise, and the alignment tends to happen at the global or sentence level, failing to capture fine-grained correspondences between textual triggers / arguments and visual objects.

Moreover, most existing MEE approaches treat event extraction as a monolithic classification task, lacking an intermediate proposal stage. This limits interpretability and makes it difficult to adapt models to diverse event types or open-domain settings. Fine-grained cross-modal alignment—especially at the object-token level—remains under-explored but is essential for accurate multimodal event understanding.

3 Method

As shown in **Fig. 3**, our PAF-DM framework introduces two key designs: (1) Proposal-Level Alignment: Instead of holistic modality fusion, we align text and visual features through candidate event triggers (text) and spatial objects (image) via cross-modal self-attention, enabling precise event-aware correlation modeling (Section 3.1). (2) Dynamic Masking: A curriculum masking mechanism progressively filters low-confidence synthetic content during training, reducing hallucination dependency while amplifying reliable multimodal signals (Section 3.2).

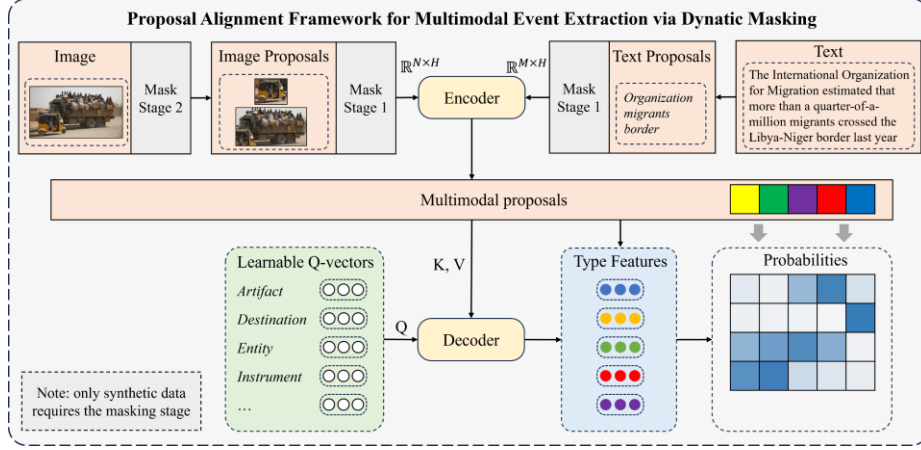


Fig. 3. Overview of PAF-DM.

3.1 Proposal-Level Alignment

This section provides a detailed explanation of the proposed fine-grained modality alignment strategy. Existing works generally align multimodal data using image-text pairs through methods like contrastive learning to bring related data closer and push unrelated data further apart. However, these methods are typically independent of events and perform coarse-grained alignment only at the text and image levels. To align shared event-relevant elements between modalities, this study introduces a proposal-based fine-grained modality alignment method. This method consists of two components: proposal generation and modality alignment. The proposal generation module utilizes a variable-confidence proposal strategy to generate keyword and object proposals with varying confidence levels for text and image modalities, improving the coverage of real keywords and objects. The modality alignment module, using the Q-Former (Query-based Transformer) structure, gradually aligns and integrates the information from both modalities, ensuring that each modality's information can contribute fully to the task, thereby enhancing performance in multimodal event extraction.

First, the proposal generation module is introduced. The proposal generation for the text modality follows the sequence labeling methods to predict the probability of each word in the text being classified as a BIO tag. Based on a given confidence threshold, the text keyword proposals and their corresponding confidence scores are extracted. It is worth noting that, we additionally store the confidence of each proposal to provide a basis for the subsequent dynamic masking strategy. The feature vector for each candidate proposal is obtained by performing average pooling on the last hidden layer outputs of the sequence representation at the word positions it spans. The confidence is calculated as the mean probability of the words in the candidate proposal belonging to the respective type. For example, for a candidate proposal $T_k = [i_s, i_e]$, the feature vector X_k is given by the following equation:

$$X_k = \text{AvgPool}([x_{i_s}, x_{i_s+1}, \dots, x_{i_e}]) \quad (1)$$

The confidence $Conf(T_k)$ for the proposal T_k is:

$$Conf(T_k) = \text{AvgPool}([p_{i_s}, p_{i_s+1}, \dots, p_{i_e}]) \quad (2)$$

Where x_i represents the last hidden state vector of the sequence labeling model, and p_i is the probability of a word being classified with a BIO label.

For the image modality, the proposal generation process is divided into two sub-tasks: event detection and event argument extraction. In event detection, there is exactly one event proposal per image, which is the entire image itself, and its confidence is set to 1. In the event argument extraction task, a target detection model is used to detect all potential target objects in the image, with each target object being treated as an argument proposal, and its confidence is the confidence of the object detection bounding box. This study uses the CLIP model to encode the input image to obtain its feature representation. The feature of the proposal is the average pooling of the feature vectors of all patches it occupies, as shown in the following equation:

$$X_k = \text{AvgPool}(\text{CLIP}(\text{patches}(I_k))) \quad (3)$$

For each synthetic modality candidate proposal, the cosine similarity between its feature vector and that of every candidate proposal from the corresponding real modality is computed, and this similarity is used as a weight to compute the weighted average confidence. The adjusted confidence is then calculated by multiplying the similarity-weighted average confidence with the similarity of the synthetic modality proposal, as shown in the following equation:

$$\widehat{Conf}(S_j) = \frac{\sum_{i=1}^N \cos(X_{S_j}, X_{R_i}) \cdot Conf(R_i)}{\sum_{i=1}^N \cos(X_{S_j}, X_{R_i})} \cdot Conf(S_j) \quad (4)$$

Where R represents the real modality proposals, S represents the synthetic modality proposals, X represents the feature vectors of the proposals, and $Conf(\cdot)$ represents the initial confidence of a proposal, while $\widehat{Conf}(\cdot)$ represents the adjusted confidence.

The modality alignment module introduces the Q-Former structure, which uses an attention mechanism to fine-tune the alignment of text and image proposals, enabling the fine-grained fusion of multimodal information. The Q-Former consists of the following components: (1) Type Query Vectors: To effectively align event information across modalities, this study assigns a learnable query vector q for each event type (in event detection) or argument role (in event argument extraction). These query vectors are optimized during training and can learn the feature representations of shared event types or argument roles across both text and image modalities, guiding the alignment of similar object proposals from the two modalities. (2) Multimodal Encoder-Decoder: To fuse object features across modalities, this study uses a Transformer encoder structure with self-attention to integrate feature vectors of real modality and masked synthetic modality proposals. The text and image modality proposal features are input together, as shown in the following equation:

$$H = \text{Encoder}([X_{T1}, X_{T2}, \dots, X_{TN}, X_{I1}, X_{I2}, \dots, X_{IM}]) \quad (5)$$

Where X_T and X_I represent the features of text and image proposals, and N and M are the number of text and image proposals, respectively. The encoder represents the encoder structure in the Transformer.

3.2 Dynamic Masking

As discussed above, in MEE, cross-modal data augmentation has been adopted in previous work to alleviate the issue of missing modality in training data. However, such approaches often suffer from distributional discrepancies between synthetic and real data, thereby undermining model inference on real-world inputs. Particularly, current text-to-image and image-to-text generation models still struggle with hallucinations and artifacts, which may cause the synthesized content to deviate from the semantics of the original data, introducing undesirable noise into model training.

To address this challenge, we propose a Dynamic Masking strategy that adaptively modulates masking policies to mitigate the negative impact of synthetic data, thereby improving the robustness and generalization. The key idea is to flexibly mask certain features or proposals in the synthetic modality, thus reducing the model's reliance on synthetic distributions while retaining useful cross-modal information.

As shown in **Fig. 4**, our dynamic masking approach adjusts three aspects during training: (1) Masking Target Adjustment. We design a two-stage masking process. In the first stage, the model masks proposal objects from the synthetic modality to suppress noisy or irrelevant cross-modal proposals. In the second stage, patch-level features within image proposals of the synthetic modality are masked, which helps reduce the influence of low-level synthetic artifacts. (2) Masking Ratio Adjustment. In the first stage, the masking ratio of synthetic proposals gradually decreases to encourage the model to first learn from unimodal real data, then gradually incorporate multimodal information. Conversely, in the second stage, the masking ratio of patch features progressively increases, which helps the model move away from overfitting to the synthetic modality and focus more on learning generalizable representations from real data. The dynamic masking ratio at step t is defined by:

$$\alpha(t) = \alpha_{init} - t \cdot \frac{\alpha_{init} - \alpha_{final}}{t_{total}} \quad (6)$$

where α_{init} and α_{final} are the initial and final masking ratios, t is the current training step, and t_{total} is the total number of training steps.

(3) Proposal Selection for Masking. In the first stage, we select proposals to mask based on their confidence scores, prioritizing the masking of low-confidence proposals. The masking selection strategy is defined as:

$$\hat{S}(t) = \{s \in S \mid \text{Conf}(s) > \text{Quantile}(\{\text{Conf}(s)\}, \alpha(t))\} \quad (7)$$

where $\hat{S}(t)$ denotes the set of remaining proposals after masking, and *Quantile* is the $\alpha(t)$ -quantile function applied to the confidence scores.

Through this tri-level dynamic masking strategy, our model learns to retain high-quality synthetic information while filtering out hallucinated or low-confidence

proposals, effectively reducing over-dependence on synthetic distributions and enhancing the model’s tolerance to noise. The training Procedure is illustrated in **Fig. 4** (c). We first separately train the model using real image and text data in the first masking stage, adopting modality-specific learning rates—a technique proven effective in prior work[13]. After this separation, we perform joint training using mixed modalities to mitigate distributional divergence between different data sources.

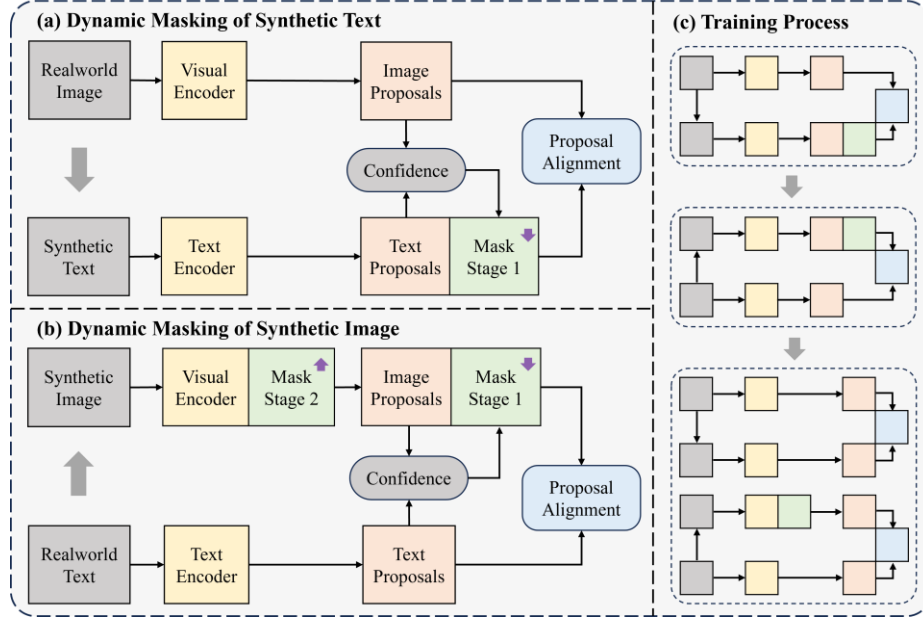


Fig. 4. Dynamic Masking and Training Process.

4 Experiment

4.1 Experiment Setup

Datasets and Metrics. In line with previous studies, we conduct our evaluations on the MultiMedia Event Extraction (M2E2) benchmark[1], which integrates the widely used ACE2005 dataset[14] from the NLP domain with the imSitu dataset[4] from computer vision. M2E2 offers a unified multimodal event taxonomy and establishes fine-grained alignment between textual and visual modalities. As M2E2 does not include official training data, we follow prior approaches[9,12] by utilizing ACE2005 and imSitu for model training. The ACE2005 text corpus covers 33 event types, 8 of which overlap with those defined in M2E2. The imSitu dataset includes annotations for 504 situations and 1,788 semantic roles. Among these, 98 imSitu situation types are mapped to the 8 shared event types in M2E2, based on a predefined alignment.

Metrics. We follow standard evaluation protocols established in prior work[9,12], and report F1 scores for the following subtasks: Event Detection (ED) — a predicted event is considered correct if the identified trigger span in the text exactly matches a gold span, and the predicted event type is accurate. Event Argument Extraction (EAE) — an argument prediction is counted as correct if it satisfies both span-level or region-level alignment and role correctness: (i) the textual argument span exactly matches a ground-truth span, or (ii) the predicted visual bounding box achieves an Intersection over Union (IoU) above a specified threshold with the corresponding ground-truth box; and in both cases, (iii) the assigned argument role must be correct.

Implementation Details. We adopt cross-modal data augmentation strategies based on both text-to-image and image-to-text generation, following the approach in[12]. Specifically, for each event in ACE2005, we generate 3 to 7 images at a resolution of 512×512, while for each image in imSitu, we generate 1 to 3 corresponding textual descriptions. To ensure fair comparison with existing state-of-the-art baselines[9,12], we use BERT-Large[15] as the text encoder and CLIP[16] as the visual encoder. For training, we set the batch size to 64 and a learning rate of 1e-4 for the visual modality, and a batch size of 16 and learning rate of 1e-4 for the textual modality. We employ the AdamW optimizer[17] with a weight decay of 0.01, and apply a cosine learning rate decay schedule. The training process includes 5 epochs for visual modality, 3 epochs for textual modality, followed by 2 epochs of joint training. The maximum input length for textual sequences is set to 200 tokens.

4.2 Results and Analysis

This subsection presents a comprehensive evaluation of the proposed method. First, we compare the performance of our approach against several state-of-the-art (SOTA) baselines. Next, we conduct ablation studies and experiments with varying masking rates to validate the effectiveness of the two key modules. Finally, we investigate the impact of the number of synthetic images and texts generated through cross-modal augmentation to assess the robustness of our method.

Main Results. Table 1 shows the performance comparison between our method and other SOTA baselines on the M2E2 benchmark. The results demonstrate that our method achieves superior performance in multimodal event extraction, with F1 scores improving by 1.1% and 0.9% on the ED and EAE tasks, respectively, compared to the best-performing baseline. Our approach also outperforms the baselines in unimodal settings, achieving gains of 0.7%/0.6% (ED/EAE) in the textual modality and 1.2% (ED) in the visual modality. These results clearly demonstrate the effectiveness of our method in improving multimodal event extraction.

Table 1. F1 scores (%) of PAF-DM and baselines on M2E2 dataset. We bold the best result and underline the second-best.

	Model	Text-Only		Image-Only		Multimodal	
		ED	EAE	ED	EAE	ED	EAE
Text	JMEE[21]	48.7	25.3	-	-	38.1	15.8
	GAIL[22]	47.9	26.1	-	-	37.3	16.4
	WASE[10]	48.2	24.9	-	-	36.7	15.7
	UniCL[9]	52.6	29.4	-	-	-	-
	MGIM[20]	48.8	26.7	-	-	-	-
Image	WASE[10]	-	-	38.7	11.2	24.1	4.9
	UniCL[9]	-	-	56.3	14.5	-	-
	MGIM[20]	-	-	54.9	12.8	-	-
Multimodal	Flat[18]	46.1	24.0	35.8	7.6	42.5	16.1
	WASE[10]	50.6	26.4	49.9	11.9	50.8	19.2
	UniCL[9]	53.7	30.7	57.6	15.2	53.4	23.4
	CLIP-Event[19]	-	-	-	-	52.7	17.1
	CAMEL[12]	55.4	31.1	<u>58.5</u>	24.4	<u>57.5</u>	<u>33.2</u>
	MGIM[20]	<u>55.8</u>	<u>31.2</u>	<u>58.5</u>	17.8	55.6	24.6
	PAF-DM (Ours)	56.5	31.8	59.7	<u>23.2</u>	58.7	34.3

Ablation study. Table 2 presents the results of the ablation study. When removing the Q-Former structure (w/o learnable Q-vector), the F1 scores drop by 2.8% and 1.2% on ED and EAE, respectively. Excluding the first-stage mask-based generation of proposal objects (w/o mask stage 1) leads to drops of 1.9% and 1.8%, while removing the second-stage masked image synthesis (w/o mask stage 2) causes performance drops of 2.3% and 2.2%. Finally, removing the dynamic masking mechanism entirely (w/o mask) results in the largest declines of 2.9% and 3.1%. These findings confirm the effectiveness of the fine-grained modality alignment and dynamic masking modules.

Table 2. Ablation results.

Model	Text-Only		Image-Only		Multimodal	
	ED	EAE	ED	EAE	ED	EAE
PAF-DM	56.5	31.8	59.7	23.2	58.7	34.3
-w/o learnable Q-vector	54.6	<u>30.2</u>	55.8	20.9	55.9	<u>33.1</u>
-w/o mask stage 1	<u>55.7</u>	29.7	<u>56.7</u>	<u>21.9</u>	<u>56.8</u>	32.5
-w/o mask stage 2	54.6	30.1	56.2	21.3	56.4	32.1
-w/o mask	54.1	29.3	55.5	20.7	55.8	31.2

Table 3 shows the results without or with other generation models. We adopt the BLIPv2 and Stable Diffusion 2-1 to replace our original image-to-text and text-to-image generation models, respectively. The results demonstrate that different generation models lead to performance fluctuations across various metrics, and no single model consistently outperforms others on all evaluation dimensions. This suggests a degree of model dependency as well as metric complementarity among these models, also verifies the robustness of our proposed framework. Despite this variability, removing the generation process leads to a significant performance drop, underscoring the effectiveness and necessity of this component.

Table 3. Results without or with other generation models.

Model		Text-Only		Image-Only		Multimodal	
		ED	EAE	ED	EAE	ED	EAE
PAF-DM		56.5	<u>31.8</u>	59.7	23.2	<u>58.7</u>	34.3
Synthetic Text	BLIPv2[23]	<u>56.2</u>	32.1	<u>59.4</u>	22.6	58.1	<u>33.8</u>
	none	55.9	29.9	56.2	21.1	54.2	29.5
Synthetic Images	Stable Diffusion 2-1[24]	55.8	29.7	58.9	<u>22.9</u>	59.2	<u>33.8</u>
	none	54.0	28.5	57.3	21.9	54.5	30.4

Table 4 reports the results under different masking rate ranges. The best performance is achieved when the masking rate varies between 20% and 50%. As the masking rate increases, the model performance drops significantly due to insufficient utilization of synthetic modality data, which hampers the alignment and fusion of complete modality features. Conversely, reducing the masking rate slightly degrades performance, as low masking fails to suppress erroneous proposals and low-level features from synthetic data, making the model overly dependent on them. Nonetheless, the model still retains decent multimodal understanding capabilities.

Table 4. Impact of mask rate range.

Mask Rate Range	ED	EAE
0%-30%	<u>57.9</u>	<u>33.6</u>
20%-50%	58.7	34.3
40%-70%	57.1	32.9

Impact of synthetic data. **Table 5** explores the effect of varying the number of synthetic images per real text input. The results show that adding or removing two generated images has a minor impact on performance. This is mainly because the number of proposals generated from images is typically small, especially for the ED task where one image yields only a single proposal. Another reason is that real text descriptions tend to be more informative, and thus have a larger influence in multimodal learning, making the performance less sensitive to changes in synthetic images.

Table 5. Impact of synthetic image number.

Number of Synthetic Images	ED	EAE
3	57.9	<u>33.8</u>
5	58.7	34.3
7	<u>58.2</u>	33.4

In contrast, **Table 6** shows that increasing the number of synthetic text descriptions leads to a more significant drop in performance. Even adding one synthetic sentence causes more degradation than varying image numbers; adding two synthetic texts further deteriorates results. This is because synthetic texts contribute more proposals and thus exert a stronger influence during training. Moreover, text generated from image-to-text models typically describes visual content, while real ACE2005 texts are news articles, resulting in substantial domain shifts.

Table 6. Impact of synthetic text number.

Number of Synthetic Texts	ED	EAE
0%-30%	58.7	34.3
20%-50%	<u>58.0</u>	<u>33.5</u>
40%-70%	57.4	33.1

To further illustrate the adverse effect of excessive synthetic text, we provide examples of incorrect but highly specific captions generated by image-to-text models in **Fig. 5**. In the left example, a “hat” is mistakenly described as a “banana,” while in the right example, a hallucinated event of “driving a truck” is introduced. These descriptions are easily extracted as arguments or used to infer event types, thereby misleading the model during training. Therefore, excessive use of synthetic textual data can significantly impair the model’s reasoning over real-world texts.

**Fig. 5.** Error-Prone Text Data from Image-to-Text Generation.

5 Conclusion

In this work, we proposed PAF-DM, a novel MEE approach based on proposal alignment and dynamic masking. We designed a fine-grained modality alignment module

tailored for MEE, enabling the precise fusion and alignment of event-relevant elements across text and image modalities. To mitigate the model's reliance on synthetic data, we also designed a three-dimensional dynamic masking strategy. This strategy prevents the model from overfitting to low-quality synthetic proposals containing hallucinations or artifacts, and gradually reduces its dependence on synthetic modality features. Detailed experiments on M2E2 benchmark show that PAF-DM achieves state-of-the-art performance. Extensive ablation and comparative experiments further validated the effectiveness of the proposed modules. Overall, we offer a new perspective on leveraging cross-modal data augmentation to synthesize missing modality data and aim to inspire future innovations in the field of multimodal learning.

Acknowledgments. This work was supported by the National Key R\&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032), and Beijing Key Lab of Networked Multimedia.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Li M, Zareian A, Zeng Q, et al. Cross-media Structured Common Space for Multimedia Event Extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2557-2568. (2020)
2. Luan Y, Wadden D, He L, et al. A general framework for information extraction using dynamic span graphs. 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). pp. 3036-3046. (2019)
3. Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684-10695. (2022)
4. Yatskar M, Zettlemoyer L, Farhadi A. Situation recognition: Visual semantic role labeling for image understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5534-5542. (2016)
5. Pratt S, Yatskar M, Weihs L, et al. Grounded Situation Recognition. In: Proceedings of the European Conference on Computer Vision. pp. 314-332. (2020)
6. Cho J, Yoon Y, Lee H, et al. Grounded situation recognition with transformers. In: Proceedings of the British Machine Vision Conference (BMVC), (2021).
7. Wei M, Chen L, Ji W, et al. Rethinking the two-stage framework for grounded situation recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. 36(3): pp. 2651-2658. (2022)
8. Zhao Y, Fei H, Cao Y, et al. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5281-5291. (2023)
9. Liu J, Chen Y, Xu J. Multimedia event extraction from news with a unified contrastive learning framework. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1945-1953. (2022)

10. Li M, Zareian A, Zeng Q, et al. Cross-media Structured Common Space for Multimedia Event Extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2557-2568. (2020)
11. Tong M, Wang S, Cao Y, et al. Image enhanced event detection in news articles. In: Proceedings of the AAAI Conference on Artificial Intelligence. 34(05): pp. 9040-9047. (2020)
12. Du Z, Li Y, Guo X, et al. Training multimedia event extraction with generated images and captions. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5504-5513. (2023)
13. Wang W, Tran D, Feiszli M. What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12695-12705. (2020)
14. Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ace) program-tasks, data, and evaluation. *Lrec*. 2(1): pp. 837-840. (2004)
15. Devlin J, Chang M W, Lee K, et al. 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). pp. 4171-4186. (2019)
16. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International conference on machine learning. PmLR, pp. 8748-8763. (2021)
17. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, (2017).
18. Wang S, Ju M, Zhang Y, et al. Cross-modal contrastive learning for event extraction. In: Proceedings of the International Conference on Database Systems for Advanced Applications. Cham: Springer Nature Switzerland, pp. 699-715. (2023)
19. Li M, Xu R, Wang S, et al. Clip-event: Connecting text and images with event structures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16420-16429. (2022)
20. Liu Y, Liu F, Jiao L, et al. Multi-Grained Gradual Inference Model for Multimedia Event Extraction. In: Proceedings of the IEEE Transactions on Circuits and Systems for Video Technology, 34(10): pp. 10507-10520. (2024)
21. Liu X, Luo Z, Huang H. Jointly multiple events extraction via attention-based graph information aggregation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. pp. 1247-1256. (2018)
22. Zhang T, Ji H. Event extraction with generative adversarial imitation learning. *arXiv preprint arXiv:1804.07881*, (2018).
23. Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the International conference on machine learning. PMLR: pp. 19730-19742, (2023).
24. Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition: pp. 10684-10695. (2022).