CVRH: Cross-modal Variational Role Hypergraph Network via Semantic Enhancement for Multi-modal Event Argument Extraction

Bangze Pan[1], Yang Li[2*], Ruili Pu[1], Suge Wang[1,3*], Jian Liao[1,3], Jianxing Zheng[1,3], Xiaoli Li[4], Deyu Li[1,3]

[1]School of Computer and Information Technology, Shanxi University, China
[2]School of Finance, Shanxi University of Finance and Economics, China
[3]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, China
[4]Information Systems Technology and Design, Singapore University of Technology and Design, Singapore

Correspondence: liyang@sxufe.edu.cn,wsg@sxu.edu.cn

Abstract

Multi-modal **E**vent **A**rgument **E**xtraction task (MEAE) aims to extract all arguments related to a specific event from multiple modalities and identify their corresponding roles. Existing methods focus on weakly alignment of uni-modal representations and generatively data augmentation techniques. However, these methods ignore the potential impact of event role information on MEAE. To address this problem, we propose a **C**ross-modal **V**ariational **R**ole **H**ypergraph Network via Semantic Enhancement (CVRH). Unlike previous approaches, CVRH centers on event role information and designs a variational role hyperedge via semantic enhancement, which constructs a role hypergraph for event arguments within multi-modal documents. It explicitly modeling the high-order role correlations among cross-modal arguments in a document. Furthermore, CVRH introduces a modal shared encoder based on differential transformer, which effectively learns shared semantic representations across modalities and enhances the independence of argument representations. On the M2E2 benchmark, experimental results show that CVRH achieves a 6.9% improvement in F1-score on the MEAE compared to current state-of-the-art methods.

1 Introduction

Event **E**xtraction (EE) aims to identify all events and their corresponding arguments from unstructured data. With the rapid growth of multimedia information, the presentation of news media events often includes multiple forms of expression such as text and images, which has led to the development of **M**ulti-modal **E**vent **E**xtraction (MEE) technology. **M**ulti-modal **E**vent **A**rgument **E**xtraction (MEAE) is an essential subtask of MEE, which aims to jointly extract all arguments related to a specific event through multi-modal information and identify their corresponding roles. As shown in Figure 1, the multi-modal document contains a multi-modal event "**Justice:Arrest-Jail**", which comprises four roles. And each role have arguments with single or multiple modalities, respectively.

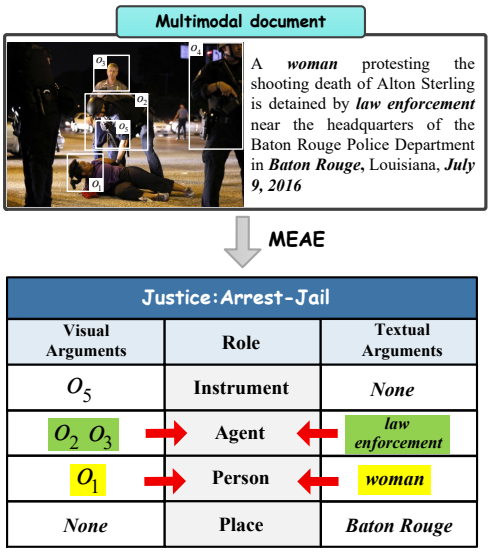


Figure 1: Example of multimodal event argument extraction on M2E2

Event roles are abstract types of arguments, and there are often potential associations between different arguments which play the same role. Modeling the role semantic association can effectively improve the accuracy of event argument extraction (Liu et al., 2023; Wan et al., 2025). As illustrated in Figure 1, although object detection can accurately identify four visual entities of the "*people*" category in the image, due to the lack of detailed event context in the image, existing methods will inevitably face the problem of argument confusion when distinguishing the visual entities corresponding to the roles "*Agent*" and "*Person*". And by introducing text arguments "*law enforcement*" and "*woman*"

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 19565–19575

which play the roles "*Agent*" and "*Person*", the model acquires discriminative semantics, thereby effectively distinguishing the feature distinctions among the four visual entities. It demonstrates the complexity of MEAE task. However, most existing methods for MEAE focus on weak alignment based on unimodal event types (Liu et al., 2024; Seeberger et al., 2024; Cui et al., 2025) and data augmentation based on image or text generation and cross-task learning (Du et al., 2023; Sun et al., 2024). These methods ignore the **cross-modal argument relationship** and potential impact of **event role information** on MEAE. And how to effectively **establish potential correlations** among cross-modal arguments is a major challenge for MEAE.

To address this challenge, we propose a **C**ross-modal **V**ariational **R**ole **H**ypergraph Network via Semantic Enhancement (CVRH), which captures role semantic associations among event arguments across modalities by utilizing hypergraphs. Specifically, we first design a differential Transformer-based modality-shared encoder. It effectively learns a unified representation for cross-modal arguments and enhances the independence of representations between different arguments. On the other hand, we construct a multi-modal role hypergraph centered on event roles for all arguments in each multi-modal document. Within the role hypergraph, we introduce variational role hyperedges via semantic enhancement, where each hyperedge representation integrates LLM-based role interpretation information and a dynamic variational vector. It effectively captures the potential correlations among cross-modal event arguments. Our contributions can be summarized as follows:

- We propose a Cross-modal Variational Role Hypergraph Network via Semantic Enhancement framework for MEAE, which fuses LLM-based role interpretation information and learnable dynamic variational vectors to effectively mine the implicit semantic correlation among cross-modal arguments.
- We design a cross-modal variational role hypergraph network, which models high-order argument relationships, effectively bridges the multi-modal argument associations based on event roles.
- We conduct extensive experiments on the M2E2 benchmark, and the results show that compared to existing state-of-the-art methods, CVRH achieves a 6.9% F1 improvement for multi-modal event argument extraction.

2 Related Work

2.1 Multi-modal event argument extraction

Traditional event argument extraction (Wei et al., 2022; Wang et al., 2024, 2025a) mostly focuses on single modes such as text, image or video. Although some works (Yu et al., 2020; Zheng et al., 2021) introduce image information, their output results are still presented as text.

To solve this challenge, (Li et al., 2020; Liu et al., 2022, 2024; Cui et al., 2025) employs the concept of cross-modal weak alignment to map representations from different modalities into a unified space. (Du et al., 2023) utilizes data augmentation strategies to generate some unimodal data for further enhancing cross-modal correspondence learning. (Li et al., 2022; Seeberger et al., 2024) using a fixed event framework and event template to enhance the effectiveness of MEAE. (Sun et al., 2024) uses instruction tuning techniques to enable large language models to learn general knowledge between different multi-modal information extraction tasks. (Liu et al., 2025b) focuses on visual events and uses semantic relationship filling to capture the relationships between events. However, these methods focus merely on the alignment of unimodal vector representations, while neglecting the impact of event structural information, such as event roles, on MEAE.

2.2 Hypergraph network

Hypergraph networks have received widespread attention in various fields due to their excellent ability to model and update high-order complex relationships between nodes (Jeong et al., 2022; Cheng et al., 2024; Liu et al., 2025a; Luo et al., 2025). In the multi-modal field, (Zeng et al., 2023) proposes a degree free hypergraph that addresses the challenges of heterogeneous modalities and data fusion. (Xu et al., 2025a) adopts a diversity hyperedge guided approach to extend multi-modal hate detection. (Li et al., 2025a) constructs a multi-modal hypergraph for images and sentences, effectively improving the effectiveness of multi-modal relationship extraction. However, in these methods, the associative relationships within the hypergraph are fixed input priors. Such predefined static associations prevent the graph structure to dynamically evolving and adapting based on data and tasks.

3 Methodology

We propose the **C**ross-modal **V**ariational **R**ole **H**ypergraph Network via Semantic Enhancement (CVRH) framework, as illustrated in Figure 2. First, we employ a differential Transformer-based multimodal encoder to map the candidate arguments representations from different modalities into a unified space. Then, we construct a cross-modal variational role hypergraph for multi-modal document to effectively aggregate and update the representations of argument nodes and role hyperedges. Finally, we determine the corresponding role for each candidate argument to obtain the MEAE results.

3.1 Cross-modal shared representation learning

Transformer (Vaswani et al., 2017) has demonstrated powerful capabilities in handling multimodal data. Our proposed cross-modal shared representation learning model is based on the Transformer framework. It specifically includes three parts: single modal encoder, candidate argument extraction, and differential transformer-based modal shared encoder. It effectively maps candidate argument features from different modalities into a unified space, while enhancing the discriminability of semantically similar candidate argument representations.

Single modal encoder: For a text t or an image m, we obtain their text representation H^t and image representation H^m as:

$$H^t = \text{Text-Encoder}(t) \tag{1}$$

$$H^m = \text{Vision-Encoder}(m) \tag{2}$$

Candidate argument extraction: In order to obtain candidate arguments representation in different modalities, we follow the (Du et al., 2023; Seeberger et al., 2024), using existing entity recognition models and object detection models to extract text entities and image objects respectively, thus obtaining a set of text candidate argument representations $\bar{H}^{e_n t} = \left\{\bar{H}_i^{e_n t}\right\}$ and an image candidate argument representation set $\bar{H}^{e_n m} = \left\{\bar{H}_i^{e_n m}\right\}$, where $\bar{H}_i^{e_n t}$ represents the i-th candidate argument representation of text t, $\bar{H}_i^{e_n m}$ represents the i-th candidate argument representation of image m.

Differential transformer-based modal shared encoder: Candidate arguments within a document often include semantically similar entity pairs, these highly similar entities can exacerbate model confusion and affect the results of MEAE. To address the representational heterogeneity across modalities and improve the discriminability of semantically similar arguments, we design a differential Transformer-based modality-shared encoder **(DiffTrans)**. This method employs a differential Transformer-based framework (Ye et al., 2025) as the main architecture of the modality-shared encoder, enabling it to adaptively capture the representational distinctiveness of different entities within specific contexts. Specifically, given an input $\bar{x}$, the specific operations of the modality-shared encoder on $\bar{x}$ are as follows:

$$[\bar{x}_{q1}, \bar{x}_{q2}] = \text{Split}(W_q \bar{x}) \tag{3}$$

$$[\bar{x}_{k1}, \bar{x}_{k2}] = \text{Split}(W_k \bar{x}) \tag{4}$$

$$x = \text{DiffTrans}(\bar{x}) = \left(\text{Softmax}\left(\frac{\bar{x}_{q1} \cdot \bar{x}_{k1}^{\top}}{\sqrt{d}}\right) - \lambda \cdot \text{Softmax}\left(\frac{\bar{x}_{q2} \cdot \bar{x}_{k2}^{\top}}{\sqrt{d}}\right)\right) \cdot W_v \bar{x} \tag{5}$$

where $W_q \in \mathbb{R}^{d \times 2d}$, $W_k \in \mathbb{R}^{d \times 2d}$ and $W_v \in \mathbb{R}^{d \times d}$ are learnable parameters, λ is a learnable scalar, $Split$ is a vector segmentation function.

We input each candidate argument representation independently into the encoder, and finally obtain a set of modal shared text candidate argument representations set $H^{e_n t}$ and an image candidate argument representation set $H^{e_n m}$ as:

$$H^{e_n t}; H^{e_n m} = \text{DiffTrans}(\bar{H}^{e_n t}; \bar{H}^{e_n m}) \tag{6}$$

3.2 Variational Role Hypergraph Construction and Updating

Event roles are high-order type abstractions of arguments. By using roles as a medium, the potential structured dependencies among different arguments can be modeled effectively. Therefore, we assign event roles as the hyperedges, introduce a dynamic role hypergraph structure that clusters cross-modal argument nodes around event role hyperedges. It mainly consists of the following two components: role hypergraph initialization, nodes and hyperedges updating.

Role hypergraph initialization: Let $G = (V, E, A)$ be a hypergraph, $V = \left\{v_1, v_2, ..., v_{|V|}\right\}$ is the set of multi-modal candidate argument nodes. $E = \left\{e_1, e_2, ..., v_{|E|}\right\}$ is the set of hyperedges connecting the multi-modal candidate arguments.

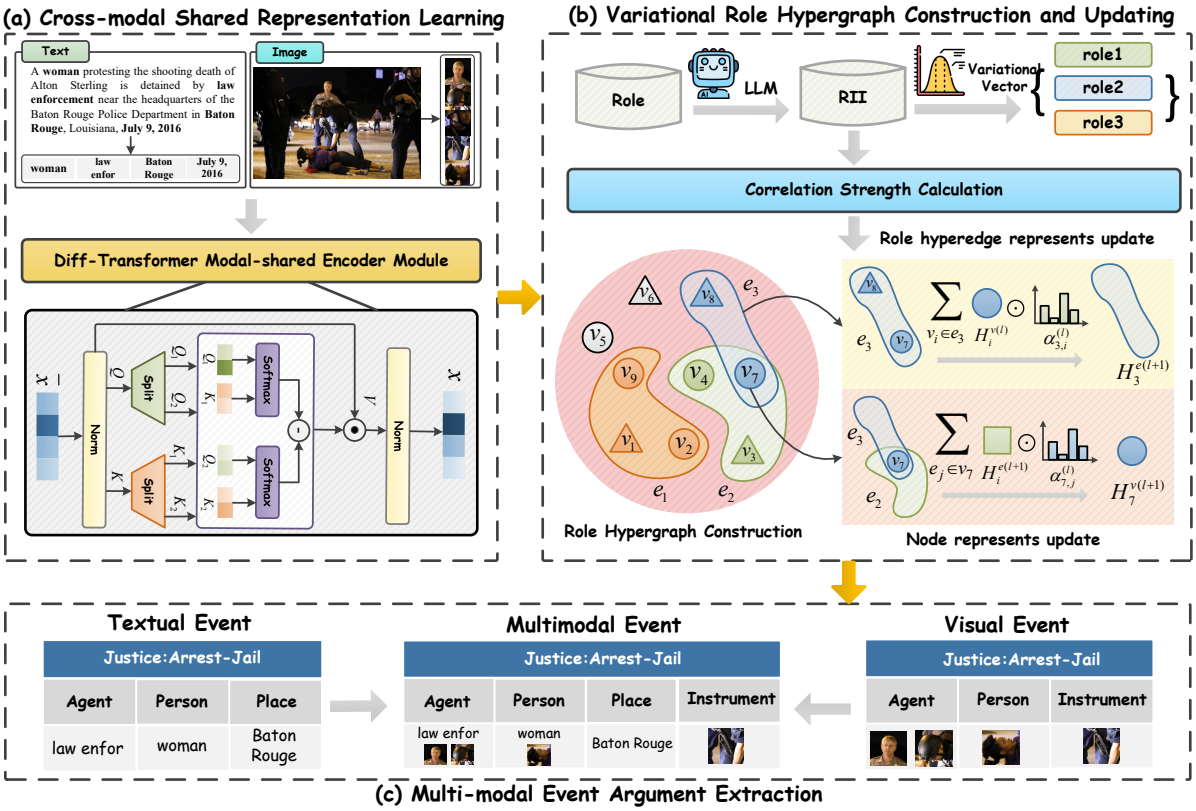


Figure 2: The overview of our proposed Cross-modal Variational Role Hypergraph Network via Semantic Enhancement framework.

$A \in \{0,1\}^{|V|\times|E|}$ denotes the hypergraph correlation matrix, which contains the relationship between hyperedges and nodes in hypergraph G.

For the multi-modal candidate argument node set $V = V^t \cup V^m$ in the hypergraph, where V^t represents the text modality node set, V^m represents the image modality node set. We use H^{ent} and H^{enm} obtained in Section 3.1 as the initialization representations for V^t and V^m, respectively, and obtaining the node initialization representation set $H^V = \{H_i^v\}_{i=1}^{|V|} = \{H_i^{ent}\}_{i=1}^{|H^{ent}|} \cup \{H_i^{enm}\}_{i=1}^{|H^{enm}|}$.

For role hyperedges E in hypergraphs, given a role r_i, we first utilize an LLM to generate role interpretation information (RII) for this role from three perspectives: word meaning, usage scenarios, and synonyms. The generated RII is then manually reviewed to ensure its accuracy, the specific prompt information can be found in Appendix 2. Finally, we use an independent text encoder to encode the RII, obtaining the interpretation representation H_i^{RII} for role r_i:

$$H_i^{RII} = \text{Text-Encoder}(LLM(r_i)) \tag{7}$$

However, the quality of hyperedge representations that rely solely on role interpretation information is heavily constrained by the generation results of the LLM, which prevents the hyperedge representations from adaptively dynamically updating during iterations. To address this problem, we introduce a dynamic variational vector for the role hyperedges to enhancing the learnability of hyperedges and reinforcing their feature independence. Specifically, for role r_i, we randomly sample a dynamic tensor H_i^x and ensure it conforms to a standard normal distribution $\mathcal{N}(0,1)$). Then, we concatenate H_i^x with the interpretative representation H_i^{RII} to obtain the concatenated tensor H_i^c. Finally, we model the concatenated tensor using a Gaussian distribution:

$$H_i^e = \mu_i + \sigma_i \cdot H_i^c \tag{8}$$

where μ_i and σ_i represent the mean and variance of H_i^c, respectively. We employ H_i^e as the initialization representation of the role hyperedge for role r_i. Using the same method, we create initialization representations of role hyperedges for all roles in the dataset, resulting in the variational role hyperedge set $H^E = \{H_i^e\}_{i=1}^{|E|}$. This design strikes a balance between preserving the rich prior information of event roles and ensuring the dynamic

learnability of the representations, enabling the hyperedges to possess both strong semantics and high adaptability.

Nodes and hyperedges updating: To model the potential relationships among cross-modal arguments, we center on variational role hyperedges to guide and aggregate semantically relevant arguments from different modalities. Specifically, inspired by (Li et al., 2025b), we measure the initial correlation strength matrix $\bar{A} = \{\bar{a}_{i,j}\}^{|V|\times|E|}$ between nodes and hyperedges by employing a multi-head attention mechanism, with the calculation method as follows:

$$\bar{a}_{i,j} = \text{Softmax}\left(\frac{(W_v H_i^v)\cdot(W_e H_j^e)^\top}{\sqrt{d}}\right) \quad (9)$$

where $\bar{a}_{i,j}$ represents the correlation strength between the i-th node and the j-th hyperedge, W_v and W_e are learnable parameters, d represents the dimension.

Finally, we determine the specific structure of the hypergraph through a hyperparameter β and complete the construction of the hypergraph:

$$\bar{a}_{i,j} = \begin{cases} 0 & if \quad \bar{a}_{i,j} < \beta \\ \bar{a}_{i,j} & otherwise \end{cases} \quad (10)$$

Then, we update the representations of hyperedges and nodes through a multi-layer hypergraph attention mechanism. First, we update the information of hyperedges. Specifically, for any hyperedge e_j, we aggregate all node information within the hyperedge at layer l to update its representation. The representation $H_j^{e\ (l+1)}$ at layer $l+1$ is calculated as follows:

$$H_j^{e\ (l+1)} = \sigma\left(\sum_{v_i \in e_j} H_i^{v\ (l)} \odot \alpha_{j,i}^{(l)} \cdot W_x\right) + H_j^{e\ (l)} \quad (11)$$

$$\alpha_{j,i}^{(l)} = \text{Softmax}\left(\frac{\text{FFN}(H_j^{e\ (l)})\cdot\text{FFN}(H_i^{v\ (l)})^\top}{\sqrt{d}}\right) \quad (12)$$

where $H_j^{v\ (l)}$ is the representation of node v_i in the l-th layer, W_x is a learnable parameter, $\alpha_{j,i}^{(l)}$ denotes the association strength of node v_i with respect to hyperedge e_j at the l-th layer, σ represents the activation function, and FFN represents a feedforward neural network.

Next, we utilize the updated hyperedge representations to update the nodes. Similar to the hyperedge update method, for updating the node representations, for any node v_i,its representation $H_i^{v\ (l+1)}$ at layer $l+1$ is given by:

$$H_i^{v\ (l+1)} = \sigma\left(\sum_{e_j \in v_i} H_j^{e\ (l+1)} \odot \alpha_{i,j}^{(l)} \cdot W_n\right) + H_i^{v\ (l)} \quad (13)$$

$$\alpha_{i,j}^{(l)} = \text{Softmax}\left(\frac{\text{FFN}(H_i^{v\ (l)})\cdot\text{FFN}(H_j^{e\ (l+1)})^\top}{\sqrt{d}}\right) \quad (14)$$

where $\alpha_{i,j}^{(l)}$ denotes the association strength of hyperedge e_j with respect to node v_i at the l-th layer, and W_n is a learnable parameter. Specifically, when $l=0$, $\alpha_{i,j}^{(0)} = \bar{a}_{i,j}$.

To better optimize the node representations during training, we introduce a node distance loss function:

$$\mathcal{L}_d = \sum_{k=1}^{L}\sum_{e\in E}\sum_{v_i \hat{\in} e, v_j \hat{\in} e} d(H_i^{v\ (k)}, H_j^{v\ (k)}) \quad (15)$$

where $H_i^{v\ (k)}$ and $H_j^{v\ (k)}$ respectively represent in the k-th layer of hypergraph attention, the i-th and j-th nodes contained in any hyperedge e, d represents the Euclidean distance between nodes, L represents the number of layers for hypergraph updates, $\hat{\in}$ represents the node v_i affiliated to the hyperedge e.

Through iteratively refining the attention weights and synchronously updating the representations of nodes and hyperedges, our method effectively captures the high-order role correlations among cross-modal arguments, ultimately enabling the role hyperedges to acquire strong cross-modal argument relationship aggregation capabilities.

3.3 Multi-modal Event Argument Extraction

For a multi-modal event and its role set $R = \{r_i\}_{i=1}^{|R|}$, we represent each role r_i as a role query q_{r_i}. Then, we determine whether candidate argument arg_j^c is an argument for role r_i by calculating the matching score between each candidate argument and the role query. The specific calculation is as follows:

$$\varphi(i,j) = \text{Sigmoid}(q_{r_i}, H_j^v) \quad (16)$$

where q_{r_i} is the query representation of role r_i and H_j^v is the updated node representation of candidate argument arg_j^c.

During the training, we define the loss for extracting multi-modal event arguments as:

$$\mathcal{L}_{eae} = -\frac{1}{|V|}\sum_{i=1}^{|V|}\sum_{j=1}^{|R|} y_{i,j}\log\varphi(i,j) \quad (17)$$

where $y_{i,j}$ represents the ground truth label distribution, V is the set of candidate argument nodes. The loss function of our model is:

$$\mathcal{L} = \mathcal{L}_{eae} + \gamma \cdot \mathcal{L}_{d} \quad (18)$$

where γ is the loss weight hyperparameter.

In the inference, we set an argument extraction hyperparameter τ to determine whether candidate argument arg_j^c corresponds to role r_i. When $\varphi(i,j) \geq \tau$, we determine arg_j^c as an argument of r_i.

4 Experiments

4.1 Experimental Setup

Dataset: We adopt the multi-modal event extraction dataset M2E2 (Li et al., 2020) for our experiments, which contains 8 event types and 15 role types. The dataset comprises a total of 245 multimedia documents, specifically including 6,167 sentences and 1,014 images. There are 1,297 textual events and 391 visual events, among which 192 textual events and 203 visual events are combined into 309 multimedia events. During the training, we follow the practice of (Du et al., 2023), use ACE05 (Walker et al., 2006) and SWiG (Pratt et al., 2020) for training. ACE05 is a textual event extraction dataset containing 33 event types and 36 role types. SWiG is a situational recognition dataset comprising 126,102 images, annotated with 504 activity verbs and 1,788 semantic roles for visual events.

Baselines: To verify the effectiveness of our proposed method, we selected 8 multi-modal event argument extraction methods for comparison. WASE (Li et al., 2020), CLIP-EVENT (Li et al., 2022), UniCL (Liu et al., 2022), CAMEL (Du et al., 2023), MGIM (Liu et al., 2024), MMUTF (Seeberger et al., 2024), VEGSRF (Liu et al., 2025b), MGFSG-EE (Wang et al., 2025b). The specific introduction can be found in Appendix 1.

Implementation Details: We follow the experimental setup of (Seeberger et al., 2024). In the backbone model, we employ the T5 model (Raffel et al., 2019) as the text encoder and the CLIP model (Radford et al., 2021) as the visual encoder. For the visual entity detection task, we select YOLOv8 (Varghese and M., 2024) as the object detector and filter out detection results with confidence scores less than 0.8. For event type detection, we adopt the results from CAMEL (Du et al., 2023) as the event mention set. The number of hypergraph attention layers L in the model is set to 3. We train the model using the AdamW optimizer for 10 epochs on an A100 GPU, with a batch size of 32 and a learning rate of 1e-5.

4.2 Main Results

Table 1 shows the comparison between CVRH and other methods on the M2E2 dataset. The results show that our method achieves the best F1-score in textual event argument extraction, visual event argument extraction, and multi-modal event argument extraction. For textual event argument extraction and visual event argument extraction, compared with the optimal MMUTF and CAMEL methods, CVRH achieves improvements of 2% and 0.9% in F1-score, respectively. This can be attributed to CVRH high-quality modeling of the latent relationships among arguments, effectively capturing the high-order semantic correlations and underlying dependencies. Notably, the most significant performance improvement occurs in multi-modal event argument extraction. Compared to the best existing result, our method achieves an improvement of 6.9%, which fully demonstrates that CVRH effectively captures the potential role semantic associations existing among cross-modal event arguments. Furthermore, CVRH effectively captures the latent semantic associations among arguments even without relying on image-text paired data (such as CAMEL) for explicit cross-modal alignment. This robustly validates the inherent effectiveness of the "role-centric" paradigm for cross-modal arguments relation mining.

4.3 Ablation Study

To further demonstrate the effectiveness of each component in CVRH, we conduct corresponding ablation experiments, and the results are shown in Table 2.

(1)w/o RII indicates the removal of role interpretation information from the role hyperedges. As the results show, removing RII information will reduce the performance of event argument extraction across all three modalities. Notably, role interpretation information has the greatest impact on purely textual event argument extraction and the least im-

Model	Text			Vision			Multi		
	P	R	F1	P	R	F1	P	R	F1
WASEatt (Li et al., 2020)	27.5	33.2	30.1	9.7	11.1	10.3	18.6	21.6	19.9
WASEobj (Li et al., 2020)	23.5	30.3	26.4	14.5	10.1	11.9	19.5	18.9	19.2
CLIP-EVENT (Li et al., 2022)	-	-	-	21.1	13.1	17.1	-	-	-
UNICL (Liu et al., 2022)	27.8	34.3	30.7	16.9	13.8	15.2	24.3	22.6	23.4
CAMEL (Du et al., 2023)	24.8	41.8	31.1	21.4	**28.4**	24.4	31.4	35.1	33.2
MGIM (Liu et al., 2024)	28.2	34.7	31.2	24.1	14.1	17.8	25.2	21.7	24.6
MMUTF (Seeberger et al., 2024)	**33.6**	44.2	38.2	**23.6**	18.8	20.9	39.9	20.8	27.4
VEGSRF (Liu et al., 2025b)	-	-	-	-	-	-	**44.8**	18.6	25.3
MGFSG-EE (Wang et al., 2025b)	29.1	35.3	31.9	26.2	16.6	20.3	28.4	26.5	27.4
CVRH(Ours)	33.2	**51.0**	**40.2**	23.4	27.6	**25.3**	41.1	**39.2**	**40.1**

Table 1: Main Results of MEAE on three modes. We use **bold** text to indicate the best performance and underline to indicate the second best.

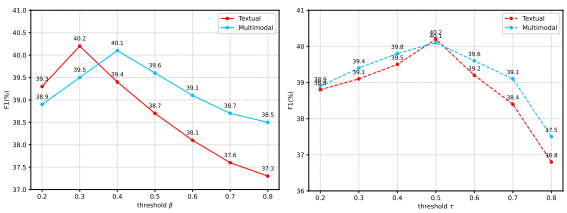
pact on visual event argument extraction. This is because textual RII and textual arguments reside within the same modal space; their interaction does not need to cross the modal gap, so the establishment of semantic association is more direct and efficient. In contrast, visual arguments require an additional cross-modal mapping step, where information loss leads to a relatively weakened association with the RII.

(2)w/o dv indicates the removal of the dynamic variational vector from the role hyperedges. The results show that all results have a certain degree of decrease. This is because, after removing the dynamic variational vector, the role hyperedges degenerate into a static representation, causing them to lose the ability to update and learn. Additionally, during hyperedge initialization, it becomes difficult to capture the representational differences between similar roles, which leads to a reduced discriminability in the representation space of the hyperedges, consequently resulting in more misjudgments during argument nodes clustering.

(3)w/o DiffTrans represents replacing the differential Transformer-based modality-shared encoder with a standard Transformer module. **w/o** $\mathcal{L}_d$ denotes removing the node distance loss function. The results show that after removing either method, the argument extraction performance for all three modalities declines to some extent. This indicates that both modules have a positive impact on the overall performance of the model.

4.4 Parameter Sensitivity Analysis

To evaluate the impact of different hyperedge aggregation thresholds β and argument discrimination thresholds τ on the model, we conducted experiments across the three modalities using varying values.



(a) hyperparameter β (b) hyperparameter τ

Figure 3: The impact of the two hyperparameters on textual MEAE and multimodal MEAE.

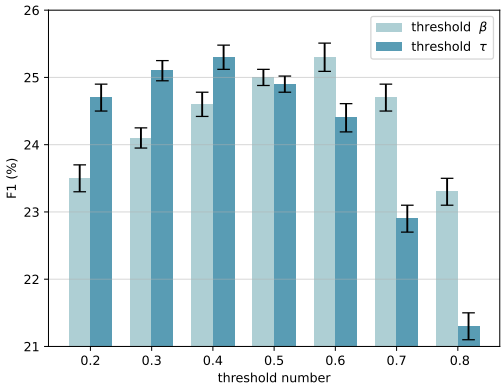


Figure 4: The impact of the two hyperparameters on visio MEAE.

Textual modality and multi-modal analysis: Figure 3 illustrates the impact of different thresholds on model performance for text and multimodal event argument extraction. The results show

Model	Text			Vision			Multi		
	P	R	F1	P	R	F1	P	R	F1
CVRH	**33.2**	**51.0**	**40.2**	**23.4**	**27.6**	**25.3**	**41.1**	**39.2**	**40.1**
O/w RII	29.7	46.1	36.1	22.4	26.3	24.2	38.4	35.9	37.1
O/w dv	31.2	48.7	38.0	21.2	25.6	23.2	38.7	36.1	37.4
O/w DiffTrans	32.3	49.8	39.2	22.5	26.8	24.5	40.2	37.9	39.0
O/w $\mathcal{L}_d$	32.4	50.2	39.4	22.6	27.1	24.6	40.8	38.9	39.8

Table 2: Research results on module ablation based on three modalities

that, regarding the hyperedge aggregation threshold β and the argument discrimination threshold τ, the text modality achieves the best performance at 0.3 and 0.5, respectively, while the multi-modal modality reaches its optimal performance at 0.4 and 0.5, respectively. From the trend, it can be see that the sensitivity of the two modalities to both thresholds is quite similar. Notably, when the threshold values deviate from the optimal value, multi-modal performance exhibits a smaller decline compared to the text modality, further demonstrating the effectiveness and robustness of our proposed method in establishing cross-modal argument relationships.

Visual modal analysis: Figure 4 illustrates the impact of the two threshold values on the model performance for visual event argument extraction. The results show that the model achieves the best performance at thresholds of $\beta = 0.6$ and $\tau = 0.4$, respectively. However, the performance of the visual modality is far inferior to the other two modalities. This is because, in our method, the hyperedges primarily consist of the text modality information, causing the representations of the hyperedges and the visual arguments to originate from different representational spaces. Additionally, unlike multimodal data that includes paired text, purely visual data lacks textual arguments that serve as crucial semantic supervision signals, which prevents the establishment of cross-modal semantic transfer and guidance mechanisms. Therefore, mapping visual features into text-dominated hyperedges inevitably incurs information loss, which directly results in a lower sensitivity of visual arguments to the hyperedge aggregation threshold.

4.5 MLLM for MEAE

Multimodal Large Language Models (MLLMs) leveraging their powerful reasoning and semantic understanding capabilities, have achieved impressive results across various multimodal tasks. We evaluate the performance of Qwen2.5-VL-7B and LLaVA1.5-7B on MEAE, with the results shown in Table 3, and the specific prompt information can be found in Appendix 2. The results indicate that, compared to existing methods, both MLLMs achieved superior performance, which fully demonstrates the powerful semantic understanding capabilities of MLLMs in multimodal tasks. However, the visual EAE performance of both models is significantly lower than their EAE performance in the other two modalities. This further supports our conclusion that the bottleneck in multimodal event argument extraction lies in the recognition of the visual modality arguments.

Modal	Method	P	R	F1
Text	**Qwen2.5-VL-7B**	24.2	47.1	31.9
	LLaVA1.5-7B	25.1	46.7	32.6
Vision	**Qwen2.5-VL-7B**	11.9	16.2	13.7
	LLaVA1.5-7B	12.4	17.5	14.5
Multi	**Qwen2.5-VL-7B**	29.7	36.3	32.6
	LLaVA1.5-7B	30.1	37.6	33.4

Table 3: The results of two MLLMs on MEAE

5 Conclusion

This paper proposes a Cross-modal Variational Role Hypergraph Network via Semantic Enhancement (CVRH), aiming to model the potential role semantic associations among cross-modal event arguments. First, CVRH designs a differential Transformer-based modality-shared encoder, which effectively enhancing the independence of each candidate argument representation. Next, CVRH constructs a cross-modal role hypergraph using event roles as hyperedges. This design enables the role hyperedge representations to possess both high-quality role information and the capability for adaptive dynamic updates. Experimental results demonstrate that our proposed CVRH achieves the best performance on M2E2 dataset.

Limitations

Although our proposed CVRH effectively models the implicit role correlations among cross-modal arguments, its performance in visual EAE remains unsatisfactory. This is because the role hyperedge information in CVRH is primarily text-based. How to design role hyperedges based on multimodal information will be a key focus of our future research.

On the other hand, due to the scarcity of the MEAE dataset, CVRH has only shown effectiveness on a few highly discriminative role types. Whether it still has effectiveness in handling easily confused role types will be our main research direction in the future. In addition, multimodal explainability (Dang et al., 2024) with diverse explanations (Zhao et al., 2024; Xu et al., 2025b) for EAE is also a focus point worth paying attention.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U24A20335, 62376143, 62473241, 62476162, 62272286), the Fundamental Research Program of Shanxi Province, China (202503021211239, 202303021211021).

References

Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 445–455, New York, NY, USA. Association for Computing Machinery.

Yiming Cui, Bin Sun, Tao Jiang, and Hongrui Cui. 2025. Multimedia event extraction based on multimodal low-dimensional feature representation space. *Signal, Image and Video Processing*, 19(5):1863–1711.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *Preprint*, arXiv:2412.02104.

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

Ujun Jeong, Kaize Ding, Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2022. Nothing stands alone: Relational fake news detection with hypergraph neural networks. *2022 IEEE International Conference on Big Data (Big Data)*, pages 596–605.

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

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

Qian Li, Cheng Ji, Shu Guo, Kun Peng, Qianren Mao, and Shangguang Wang. 2025a. Variational multimodal hypergraph attention network for multi-modal relation extraction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25.

Qian Li, Cheng Ji, Shu Guo, Kun Peng, Qianren Mao, and Shangguang Wang. 2025b. Variational multimodal hypergraph attention network for multi-modal relation extraction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25.

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

Junrui Liu, Tong Li, Di Wu, Zifang Tang, Yuan Fang, and Zhen Yang. 2025a. An aspect performance-aware hypergraph neural network for review-based recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, page 503–511, New York, NY, USA. Association for Computing Machinery.

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

Wanlong Liu, Shaohuan Cheng, Dingyi Zeng, and Qu Hong. 2023. Enhancing document-level event argument extraction with contextual clues and role relevance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12908–12922, Toronto, Canada. Association for Computational Linguistics.

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

Haoran Luo, E. Haihong, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Ze min Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. 2025. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

C Walker, S Strassel, J Medero, and K Maeda. 2006. Ace 2005 multilingual training corpus. *Progress of Theoretical Physics Supplement*, 110(110):261–276.

Qizhi Wan, Tao Liu, Changxuan Wan, Rong Hu, Keli Xiao, and Yuxin Shuai. 2025. Event pattern-instance graph: A multi-round role representation learning strategy for document-level event argument extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1865–1877, Vienna, Austria. Association for Computational Linguistics.

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

Guanghui Wang, Dexi Liu, Jian-Yun Nie, Qizhi Wan, Rong Hu, Xiping Liu, Wanlong Liu, and Jiaming Liu. 2025a. DEGAP: Dual event-guided adaptive prefixes for templated-based event argument extraction with slot querying. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7598–7613, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. 2022. Rethinking the two-stage framework for grounded situation recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2651–2658.

Bo Xu, Erchen Yu, Jiahui Zhou, Hongfei Lin, and Linlin Zong. 2025a. HyperHatePrompt: A hypergraph-based prompting fusion model for multimodal hate detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3825–3835, Abu Dhabi, UAE. Association for Computational Linguistics.

Hao Xu, Yunxiao Zhao, Jiayang Zhang, Zhiqiang Wang, and Ru Li. 2025b. LOG: A local-to-global optimization approach for retrieval-based explainable multi-hop question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9085–9095, Abu Dhabi, UAE. Association for Computational Linguistics.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2025. Differential transformer. In *International Conference on Representation Learning*, volume 2025, pages 144–164.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 3342–3352, Online. Association for Computational Linguistics.

Yawen Zeng, Qin Jin, Tengfei Bao, and Wenfeng Li. 2023. Multi-modal knowledge hypergraph for diverse image retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3376–3383.

Yunxiao Zhao, Zhiqiang Wang, Xiaoli Li, Jiye Liang, and Ru Li. 2024. AGR: Reinforced causal agent-guided self-explaining rationalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 510–518, Bangkok, Thailand. Association for Computational Linguistics.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5298–5306, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Baselines

We compare our proposed approach with a wide range of SOTA models.

- **WASE** (Li et al., 2020) utilizes unimodal datasets to separately train a text model and a visual model, and then performs modal alignment using the VOA Caption dataset.
- **CLIP-EVENT** (Li et al., 2022) integrates event-level structured knowledge into the CLIP pre-trained model to enhance CLIP model understanding of events and their corresponding arguments.
- **UNICL** (Liu et al., 2022) constructs a multimodal shared space based on contrastive learning techniques to enhance event arguments extraction performance in both text and vision.
- **CAMEL** (Du et al., 2023) leverages data augmentation methods by generating images and corresponding image descriptions to improve cross-modal event arguments extraction effectiveness.
- **MGIM** (Liu et al., 2024) aligns image and text information through two stages: graph structure representation and multi-round progressive reasoning, enhancing the correspondence of cross-modal information.
- **MMUTF** (Seeberger et al., 2024) designs an event template and uses text prompts along with the event template structure to model relationships among multimodal events.
- **VEGSRF** (Liu et al., 2025b) leverages textual event information to focus on image details and employs an image event template to mine potential relationships among image event arguments.
- **MGFSG-EE** (Wang et al., 2025b) enhances the capability of MEE by modeling the co-occurrence and background information between text and images.

A.2 Detailed prompt information for the LLM

The detailed prompt information for obtaining role interpretation information using the LLM in Section 3.2 is as follows:

> Given a pair $\langle e, r \rangle$, where e is an event type and r is a role under that event type, please describe the meaning of role b as fully as possible from three perspectives: its definition, usage scenarios, and synonyms, in one coherent sentence not exceeding 100 words.

The detailed prompt information used when testing the effect of MLLM on MEAE in Section 4.5 is as follows:

> Here is a set of image-text pairs $\langle m, t \rangle$, which describe event e in detail in a multimodal form. Please extract the argument for role r in event e, where the argument can be a continuous text span or a visual entity bounding box.